Deep Learning for Health Data: Attention, Activations and Beyond

DISSERTATION

for the attainment of the degree of
Doctor of Engineering (Doktor-Ingenieur)
at the
Faculty of Applied Computer Science
of the
University of Augsburg

by

Srividya Tirunellai Rajamani

2023

Referees:	Prof. Dr. Björn W. Schuller Prof. Dr. Frank Kramer
Examiners:	Prof. Dr. Björn W. Schuller Prof. Dr. Frank Kramer Prof. Dr. Matthias Schlesner
Date of the oral exam:	26th September 2024

Abstract

This thesis explores deep learning based methods for health data, specifically on novel enhancements to attention mechanisms in diverse tasks of image and signal analysis. We demonstrate the effectiveness of our proposed attention mechanism enhancements in performance improvement, model complexity reduction, outlier detection as well as dealing with sparse and irregularly sampled time series data.

In the context of medical image segmentation, effective handling of outliers is vital to ensure translation of research into clinical practise. Standard metrics used for reporting the performance of medical image segmentation algorithms report aggregate metrics across all patients. Due to this reporting, models that report superior performance could end up producing completely erroneous results, or even anatomically impossible results in a few challenging cases (corner-cases), albeit without being noticed. To counter this drawback, we propose a framework that helps to identify and report corner cases. Further, we propose a novel balanced checkpointing scheme capable of finding a solution that has superior performance even on these corner cases.

Deep neural networks with attention mechanism have shown promising results in many computer vision and medical image processing applications. One way to enhance attention is to build on the concept of deformability which was introduced in the context of convolutions. We propose a new Deformable Attention Network (DANet) that enables a more accurate contextual information computation in a similarly efficient way. Our novel technique is based on learning the deformation of the query, key and value attention feature maps in a continuous way. A deep segmentation network with this attention mechanism is able to capture attention from only the pertinent non-local locations.

Deformability indicates that the attention mechanism could be further regularised. Hence we explore ways to regularise attention. We introduce a simple and low-overhead approach of adding noise to the attention block which we discover to be very effective when using an attention mechanism. Our proposed methodology of introducing regularisation in the attention block by adding noise makes the network

more robust and resilient, especially in scenarios where there is limited training data. We incorporate this regularisation mechanism in the criss-cross attention block. This criss-cross attention block enhanced with regularisation is integrated in the bottleneck layer of a U-Net for the task of medical image segmentation.

In the context of attention mechanism utilization in time-series data, we demonstrate the efficacy of using sparsely and irregularly sampled data when used in tandem with state-of-the-art existing attention based networks that are capable of handling sparse data. With our proposed sub-sampling approach, we demonstrate that time-series data could be further coarsely acquired. This could be of immense help for various applications where data acquisition and labeling is a significant challenge.

By utilizing attention mechanisms in non-linear blocks in the context of GRU, we propose a novel Attention based GRU module. We demonstrate the effectiveness of this module to improve performance in the context of speech emotion recognition. Additionally, we also propose a novel metric for image quality assessment to compute the quality of a given image without a reference pristine quality image. Many of the image acquisition processes, especially in medical imaging, would immensely benefit from such a metric which can indicate if the quality of an image is improving or worsening based on adaptation of the acquisition parameters.

List of Publications

Author Profiles

- ORCID:
  `https://orcid.org/0000-0002-1571-7229`

- Google Scholar:
  `https://scholar.google.com/citations?hl=en&user=-bLcuQMAAAAJ`

Journal Articles

- **Srividya Tirunellai Rajamani**, Kumar Rajamani, Ashwin Venkateshvaran, Andreas Triantafyllopoulos, Alexander Kathan & Björn Schuller: *Toward Detecting and Addressing Corner Cases in Deep Learning Based Medical Image Segmentation*, IEEE Access, volume 11, pp.95334-95345, 2023, IEEE (IF:3.9).

- Andreas Triantafyllopoulos, Alexander Kathan, Alice Baird, Lukas Christ, Alexander Gebhard, Maurice Gerczuk, Vincent Karas, Tobias Hübner, Xin Jing, Shuo Liu, Adria Mallol-Ragolta, Manuel Milling, Sandra Ottl, Anastasia Semertzidou, **Srividya Tirunellai Rajamani**, Tianhao Yan, Zijiang Yang, Judith Dineley, Shahin Amiriparian, Katrin D. Bartl-Pokorny, Anton Batliner, Florian B. Pokorny & Björn Schuller: *HEAR4Health: a blueprint for making computer audition a staple of modern healthcare*, Frontiers in Digital Health, volume 5, pp.1-12, 2023.

- Alexander Kathan, Mathias Harrer, Ludwig Küster, Andreas Triantafyllopoulos, Xiangheng He, Manuel Milling, Maurice Gerczuk, Tianhao Yan, **Srividya Tirunellai Rajamani**, Elena Heber, Inga Grossmann, David D. Ebert & Björn Schuller: *Personalised depression forecasting using mobile sensor data and ecological momentary assessment*, Frontiers in Digital Health, volume 4, pp.1-15, 2022.

Publications in Conference Proceedings

- **Srividya Tirunellai Rajamani**,Kumar Rajamani & Björn Schuller: *A novel and simple approach to regularise attention frameworks and its efficacy in segmentation*, Proc. Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, pp.1-4, 2023, IEEE.

- **Srividya Tirunellai Rajamani**, Kumar Rajamani, Priya Rani, Rashmita Barick, Ramasubramanya M.S, Sridevi V Aithal, Rajkumar ElagiriRamalingam, Sahana D Gowda & Björn Schuller: *Novel No-Reference Multi-Dimensional Perceptual Similarity Metric*, Proc. Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, pp.2045-2048, 2022, IEEE.

- **Srividya Tirunellai Rajamani**, Kumar Rajamani, Alexander Kathan & Björn Schuller: *Novel Insights of Induced Sparsity on Multi-Time Attention Networks*, Proc. Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, pp.2615-2618, 2022, IEEE.

- Kumar Rajamani, Sahana D Gowda, Vishwa Tej N, **Srividya Tirunellai Rajamani**: *Deformable Attention (DANet) for Semantic Image Segmentation*, Proc. Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, pp.3781-3784, 2022, IEEE.

- Alexander Kathan, Andreas Triantafyllopoulos, Xiangheng He, Manuel Milling, Tianhao Yan, **Srividya Tirunellai Rajamani**, Ludwig Küster, Mathias Harrer, Elena Heber, Inga Grossmann, David Daniel Ebert & Björn Schuller: *Journaling Data for Daily PHQ-2 Depression Prediction and Forecasting*, Proc. Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, pp.2627-2630, 2022, IEEE.

- Xiangheng He, Andreas Triantafyllopoulos, Alexander Kathan, Manuel Milling, Tianhao Yan, **Srividya Tirunellai Rajamani**, Ludwig Küster, Mathias Harrer, Elena Heber, Inga Grossmann, David Daniel Ebert & Björn Schuller: *Depression Diagnosis and Forecast Based on Mobile Phone Sensor Data*, Proc. Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, pp.4679-4682, 2022, IEEE.

- **Srividya Tirunellai Rajamani**, Kumar Rajamani & Björn Schuller: *Towards an Efficient Deep Learning Model for Emotion and Theme Recognition in Music*, Proc. Annual IEEE International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, pp.1-5, 2021, IEEE.

- **Srividya Tirunellai Rajamani**, Kumar Rajamani & Björn Schuller: *Emotion and Theme Recognition in Music using Attention-based Methods*, Proc. MediaEval Multimedia Benchmark Workshop, Online, 2020.

- Maurice Gerczuk, Shahin Amiriparian, Sandra Ottl, **Srividya Tirunellai Rajamani** & Björn Schuller: *Emotion and Themes Recognition in Music with Convolutional and Recurrent Attention-Blocks*, Proc. MediaEval Multimedia Benchmark Workshop, Online, 2020.
- **Srividya Tirunellai Rajamani**, Kumar Rajamani, Adria Mallol-Ragolta, Shuo Liu & Björn Schuller: *A Novel Attention-Based Gated Recurrent Unit and its Efficacy in Speech Emotion Recognition*, Proc. Annual International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Ontario, Canada, pp.6294-6298, 2021, IEEE.

Contents

Part I

INTRODUCTION

1

Introduction

1.1 Motivation

Health and well-being are indispensable aspects of our life. Moreover, they are crucial for a fulfilling and productive life. Health encompasses physical, mental, emotional and social aspects, to name a few. Physical health refers to the state of our body. It includes factors such as nutrition, sleep, exercise and the absence of illness or disease. Mental health is as important as physical health. It refers to our psychological well-being. Anxiety and depression are among the major mental health issues that need to be addressed. Emotional well-being encompasses our ability to manage and express our emotions in a healthy way. It involves self-awareness, self-regulation, and interpersonal relationships.

In today's world, technology plays a significant and ever-growing role in health and well-being. It is not only transforming the healthcare industry but also empowering individuals to take control of their health. Some of the key ways in which technology influences is through health monitoring, health apps, telemedicine, robotics and artificial intelligence, big data and analytics, public health surveillence and health gamification.

The role of artificial intelligence in health and well-being is an area of active research. Machine learning and deep learning based approaches have shown promising results in diagnosing diseases, analyzing medical images, and predicting patient outcomes. They have also been shown to identify patterns and trends in healthcare data, which can aid in decision-making and personalised treatment plans.

One of the areas that deep learning based approaches are revolutionalising is medical images analysis, of which medical image segmentation is a key task. Another type of clinical data that is predominantly used is electronic health record (EHR) data. Deep learning is also assisting in the way EHR data can be analysed and interpreted. In the context of emotional health, deep learning is also making its impact in analysing emotions from different modalities like speech and audio signals.

1.2 Research objectives

Despite the huge advancements in all these areas, there are several open challenges that still exist. Towards addressing some of the challenges, we aim to address the following research questions in this thesis:

1. Are there scenarios where deep learning based segmentation model could yield erroneous segmentation results for some subjects. How do we spot these corner cases? Once identified, how do we handle these corner cases?

2. Can deep learning based image segmentation frameworks be made more robust and resilient even in scenarios with limited training data?

3. Do deep learning medical image segmentation models efficiently capture contextual information across pixels

4. During the process of medical image acquisition, can we determine whether or not the quality of an image is improving based on the adaptation of the acquisition parameters?

5. To what extent could time-series based health-care data be sparsely and efficiently acquired without impacting performance?

6. Can deep learning based models be made to capture long range interactions while analysing audio signals like speech emotion data, thereby improving their performance?

1.3 Contributions

To address these research questions, the main contributions of this thesis are as follows:

- A framework for detecting and handling corner-cases in deep learning based medical image segmentation methods.

- A mechanism to regularise attention networks to increase model robustness.

- A methodology to deform attention to efficiently capture relevant long range contextual information in medical images.

- A new multidimensional No-Reference Perceptual Similarity Metric (NR-PSIM) to determine whether or not the image quality is improving when acquisition parameters are adapted.

- An analysis of the effect of sparse data on the predictive performance of networks that can handle irregular time series data.

- A novel self-attention based VGG-like network (SA-VGG) to reduce complexity of audio analysis models in terms of number of parameters and FLOPS.

- A novel attention based Gated Recurrent Unit (AR-GRU) module to improve performance of audio signal analysis tasks like speech emotion recognition.

1.4 Thesis Structure

The following chapters of this thesis are organised as follows:

- **Chapter 2** discusses fundamental deep learning architectures like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Encoder-Decoder architecture, UNet and attention mechanisms. It also briefly covers activation functions, regularisation techniques as well as metrics in the context of medical image segmentation as well as image quality and similarity.

- **Chapter 3** analyses state of the art approaches for medical image segmentation, speech emotion recognition and time series analysis that are related to the research objectives of this thesis.

- **Chapter 4** proposes a framework for identifying and reporting corner cases in the context of medical image segmentation, for which the segmentation results are erroneous or even anatomically impossible. It also proposes a novel balanced model checkpointing scheme that enables finding a solution that performs well even on these corner cases.

- **Chapter 5** explores the effect of introducing regularisation in attention and demonstrates its effectiveness in making a network more robust and resilient.

- **Chapter 6** proposes a Deformable Attention Network by introducing deformation of the query, key and value feature maps such that the attention mechanism is able to capture attention from the pertinent non-local locations.

- **Chapter 7** describes a novel no-reference perceptual similarity metric that can compute the quality of an image without a reference pristine quality image. This metric would be of immense benefit in image acquisition processes, especially in medical imaging, to indicate if the quality of an image is improving or not based on adaptation of the acquisition parameters.

- **Chapter 8** investigates the effect of inducing varying degrees of sparsity on the predictive performance of state of the art networks like Multi-Time Attention Networks (mTAN) (Shukla *et al.*, 2021) that can handle sparse and irregular time series data.
- **Chapter 9** proposes a novel integration of stand-alone self-attention into a Visual Geometry Group (VGG)-like network to significantly reduce the number of model parameters and FLOPS while retaining or improving performance for the task of multi-label emotion and theme recognition in music.
- **Chapter 10** proposes a novel attention based Gated Recurrent Unit (GRU) module and demonstrates its efficacy for speech emotion recognition.
- **Chapter 11** concludes the thesis with a summary, a discussion of the limitations of the proposed methods and directions for future work.

Part II

BACKGROUND

2

Background

The ability to build more complex functions by composing shallow neural networks or developing networks with more than one hidden layer had been understood even before the modern era of rapid advancements in deep learning. Even though the term "deep learning" was first used by Dechter (1986), interest was limited due to practical concerns such as the lack of ability to train such networks well. However, startling improvements in image classification reported by Krizhevsky *et al.* (2012) resulted in a resurge of research in deep learning. The confluence of four factors contributed significantly to this tremendous progress, namely, larger training datasets, improved processing power for training, the use of the ReLU activation function, and the use of stochastic gradient descent. In this chapter, we provide a brief summary of the widely-used deep learning architectures, activation functions, regularisation methods as well as metrics that constitute the foundation for the deep learning based solutions to our research problems.

2.1 Deep learning architectures

2.1.1 Convolutional Neural Networks

Images have three properties due to which fully connected networks turn out to be not well suited for their processing.

1. High dimensionality: A typical image for a classification task comprises of 224×224 RGB values. Hence, the training data required as well as the memory and computation needs pose practical challenges.

2. Statistical relation between nearby pixels: Nearby pixels in images are statistically related. Fully connected networks however, are unable to leverage this since they treat the relationship between every input equally.

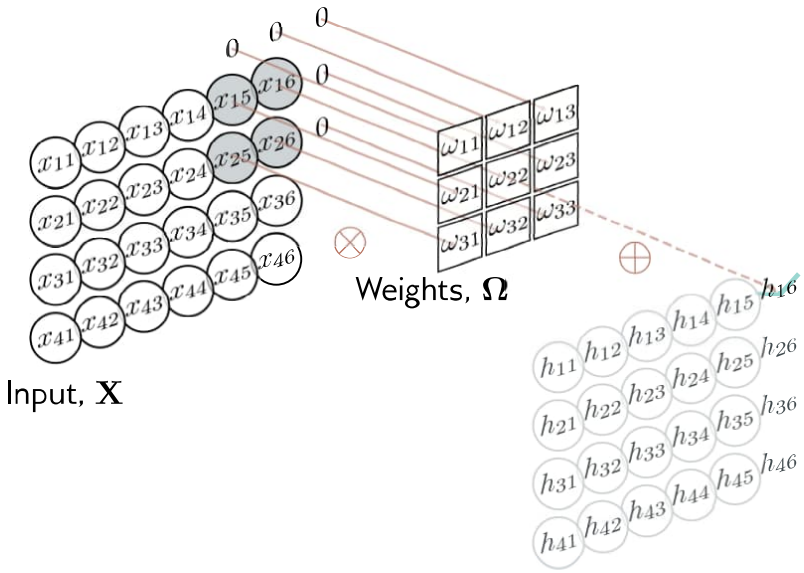

Figure 2.1: 2D convolutional layer. Each output $\mathbf{h_{ij}}$ computes a weighted sum of the 3×3 nearest inputs, adds a bias, and passes the result through an activation function. With zero padding, positions beyond the image's edge are considered to be zero. Illustration adapted from (Prince, 2023)

3. Invariance to geometric transformations: An image is generally invariant to geometric transformations. However, even minor geometric transformations changes every input to the network. A fully connected model would therefore have to learn the patterns of pixels at every position, making it extremely inefficient.

The above reasons indicate that a specialised model architecture is required for processing images. Convolutional neural network (CNN), a network that predominantly consists of convolutional layers, has been shown to be very effective at handling this. Each local image region is processed separately by the convolutional layers, using parameters shared across the whole image. They have been shown to be well-suited to handle images since they use fewer parameters than fully connected layers and can effectively leverage the spatial relationships between nearby pixels. Furthermore, they do not have to re-learn the interpretation of the pixels at every position. Figure 2.1 depicts a 2D convolution layer and how the output is computed.

2.1.1.1 VGG

Neural network architecture design has grown more abstract over the years. Researchers have moved from thinking in terms of individual neurons to whole layers, and then to blocks which comprises of repeating patterns of layers. The idea of using blocks first emerged in VGG network (Simonyan *et al.*, 2015) from the Visual

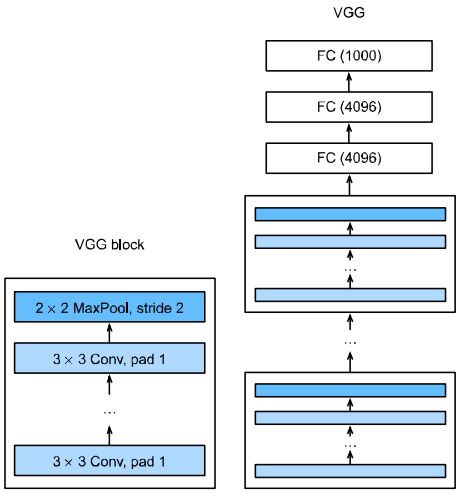

Figure 2.2: Figure on the left depicts a VGG block. It comprises of a sequence of convolutions followed by a max-pooling layer. Figure on right depicts a VGG Network. It comprises of 2 parts, i). a convolutional part which connects several VGG blocks in succession and ii). fully connected layers. Figure adapted from (Zhang, Aston *et al.*, 2023)

Geometry Group (VGG) at Oxford University. VGG is not just a specific manifestation but rather represents a family of networks. Fig 2.2 depicts a VGG block and a VGG network.

2.1.2 Recurrent Neural Networks

Recurrent neural networks (RNNs) are deep learning models that use recurrent connections to capture the dynamics of sequences. This might seem counter-intuitive at first since the feed-forward nature of neural networks is what makes the order of computation unambiguous. However, it is ensured that no such ambiguity can arise by defining the recurrent edges in a precise way. By applying the same underlying parameters at each time step, RNNs are unrolled across time or sequence steps. To propagate each layer's activations to the subsequent layer at the same time step, standard connections are applied synchronously. Information is passed across adjacent time steps through the use of the recurrent connections that are dynamic. In other words, RNNs are feed-forward neural networks where each layer's conventional and recurrent parameters are shared across time steps. This is depicted in the unfolded view in Figure 2.3. On the left side of the figure, recurrent connections are depicted via cyclic edges. On the right, the RNN is unfolded over time steps. Here, recurrent edges span adjacent time steps, while conventional connections are computed synchronously.

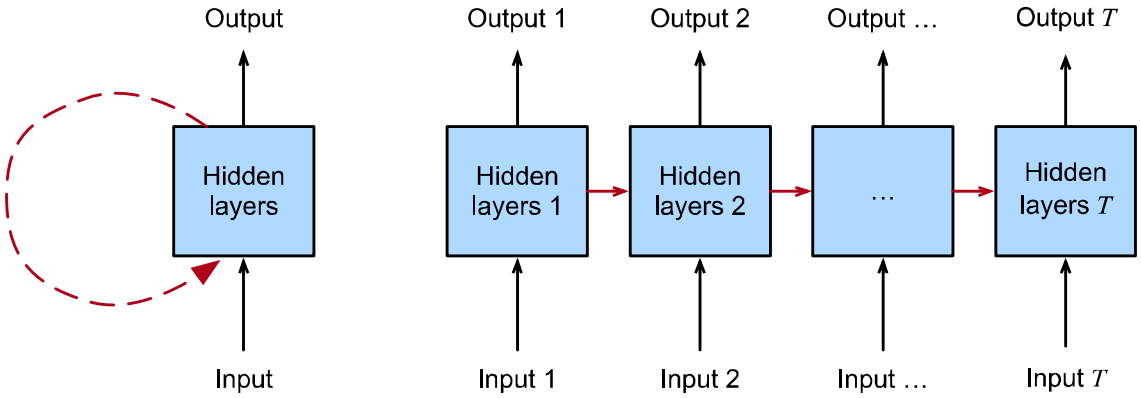

Figure 2.3: On the left, cyclic edges are used to depict recurrent connections. On the right, the RNN is unfolded over time steps. Here, conventional connections are computed synchronously whereas recurrent edges span adjacent time steps. Figure taken from (Zhang, Aston *et al.*, 2023)

2.1.2.1 LSTM

RNNs suffer from the problem of vanishing and exploding gradients and hence face challenges while learning long-term dependencies (Bengio *et al.*, 1994; Hochreiter, Bengio, *et al.*, 2001). Techniques like gradient clipping were able to address the problem of exploding gradients. But handling vanishing gradients turned out to be more challenging. Long short-term memory (LSTM) model by Hochreiter and Schmidhuber (1997) were one of the first and most successful techniques for addressing vanishing gradients. Though LSTMs are similar to standard recurrent neural networks, the main difference is that each ordinary recurrent node is replaced by a memory cell. Each memory cell has an internal state which ensures that the gradient can pass across many time steps without vanishing or exploding. Simple recurrent neural networks have long-term memory in the form of weights. This long-term memory encodes general knowledge about the data and changes gradually during training. Ephemeral activations that pass from each node to successive nodes constitutes its short-term memory. Using the memory cell, the LSTM model 2.4 introduces an intermediate type of storage. A memory cell is built from simpler nodes in a specific connectivity pattern. It is a composite unit with the novel inclusion of multiplicative nodes.

Each memory cell has an internal state and a number of multiplicative gates. The "input gate" decides if a given input should impact the internal state. The "forget gate" is responsible to decide if the internal state should be flushed to 0. The "output gate" determines if the internal state of a given neuron should be allowed to impact the cell's output.

LSTMs support gating of the hidden state. This signifies that there are established mechanisms which regulate when a hidden state should be updated and also

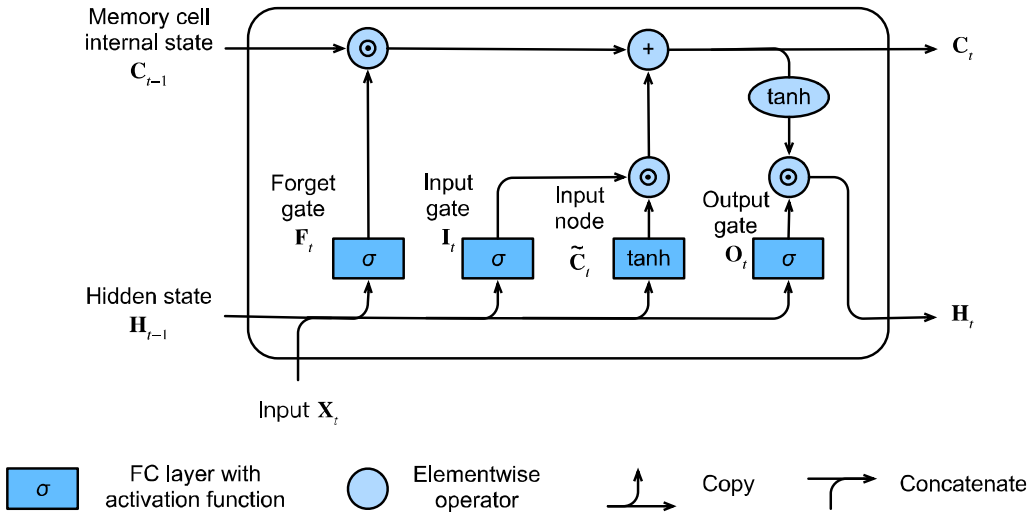

Figure 2.4: LSTM model. Figure taken from (Zhang, Aston *et al.*, 2023)

for when it should be reset. These mechanisms are learned. For instance, it learns not to update the hidden state after the first observation, if the first token is of great importance. Similarly, it learns to skip irrelevant temporary observations. It also learns to reset the latent state whenever needed. This is one of the main differentiation between vanilla RNNs and LSTMs.

Though LSTMs were intially published in 1997, victories in prediction competitions in the mid-2000s led to their rise in prominence. They played a dominant role in sequence learning until 2017 after which Transformers rose to prominence. It is important to note that some of the key ideas of Transformers are inspired from the architecture design innovations introduced by the LSTM.

2.1.2.2 GRU

With RNNs and specifically the LSTM architecture gaining popularity in the 2010s, the research community began to focus on experimenting with simplified architectures to speed up computation, though retaining the key idea of incorporating an internal state and multiplicative gating mechanisms. A modified version of the LSTM memory cell with comparable performance but yet faster to compute (Chung *et al.*, 2014) was achieved by the gated recurrent unit (GRU) (Cho *et al.*, 2014) .

In GRU 2.5, the three gates of LSTMs are replaced by two, namely the reset gate and the update gate. These gates are given sigmoid activations, similar to LSTMs, to force their values to lie in the interval (0, 1). Short-term dependencies in sequences are captured by reset gates while the long-term dependencies in sequences are captured by the update gates.

In conclusion, gated RNNS like LSTMs and GRUs can better capture dependencies for sequences with large time step distances as compared to simple RNNs.

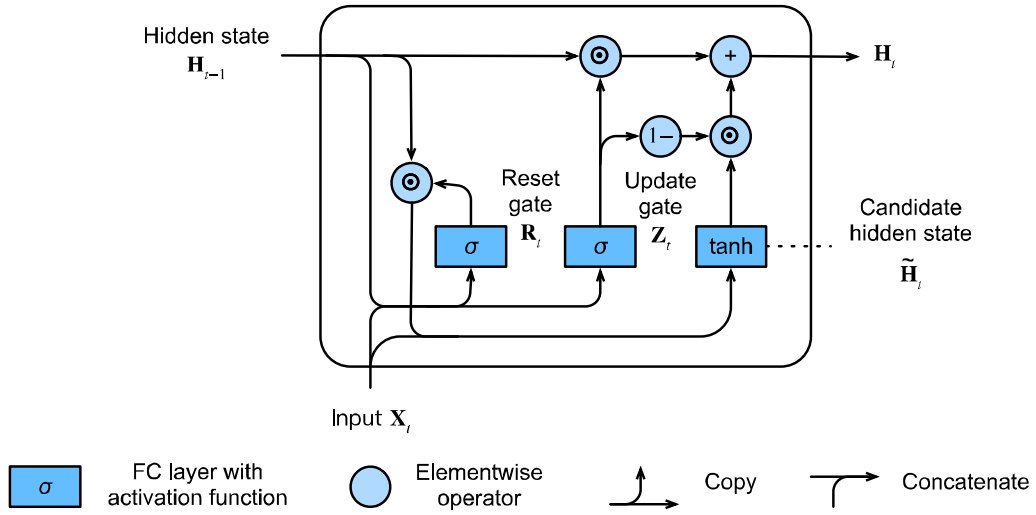

Figure 2.5: GRU model. Figure taken from (Zhang, Aston *et al.*, 2023)

GRUs achieve similar performance as LSTMs but tend to be computationally lighter. Basic RNNs are the extreme case of GRUs whenever the reset gate is switched on. By turning on the update gate, they can also skip sub-sequences.

2.1.2.3 Bi-directional Recurrent Neural Networks

The aim of sequence learning tasks like language modeling is to predict the next token given all previous tokens in a sequence. In this scenario, the uni-directional chaining of a standard RNN seems appropriate since it is required to only condition upon the leftward context. However, conditioning the prediction at every time step on both the leftward and the rightward context is perfectly fine in many other sequence learning tasks. Speech detection is one such task where assessing the part of speech associated with a given word requires the context in both directions to be taken into account. Another common task is to mask out random tokens in a text document and then train a sequence model to predict the values of the missing tokens. This is frequently used as a pre-training exercise before fine-tuning a model on an actual task of interest.

Any uni-directional RNN can be transformed by a simple technique into a bi-directional RNN (Schuster *et al.*, 1997). The technique involves implementing two uni-directional RNN layers chained together in opposite directions and acting on the same input 2.6. For the first RNN layer, $\mathbf{x_1}$ is the first input and $\mathbf{x_t}$ is the last input. However, for the second RNN layer, $\mathbf{x_t}$ is the first input and $\mathbf{x_1}$ is the last input. The corresponding outputs of the two underlying uni-directional RNN layers are concatenated to produce the output of this bi-directional RNN layer.

The data prior to and after the current time step is used simultaneously to compute the hidden state for each time step in bi-directional RNNs. Sequence

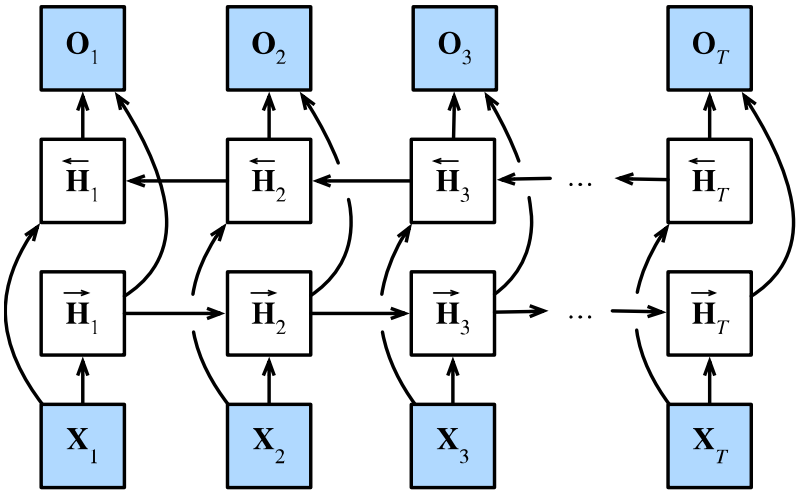

Figure 2.6: Architecture of a bi-directional RNN. Figure taken from (Zhang, Aston *et al.*, 2023)

Figure 2.7: Encoder-Decoder architecture. Figure adapted from (Zhang, Aston *et al.*, 2023)

encoding and the estimation of observations given bi-directional context are some tasks where bi-directional RNNs are most useful. However, the long gradient chains makes training of bi-directional RNNs very costly.

2.1.3 Encoder-Decoder Architecture

One of the characteristics of sequence-to-sequence problems like machine translation is that the inputs and outputs are of varying lengths that are unaligned. Designing an encoder–decoder architecture 2.7 is one of the standard approaches to handle this sort of data. The two major components of an encoder-decoder model are:

1. Encoder: The input for this is a variable-length sequence.

2. Decoder: It acts as a conditional language model. It predicts the next token in the target sequence by using the encoded input and the previous context of the target sequence.

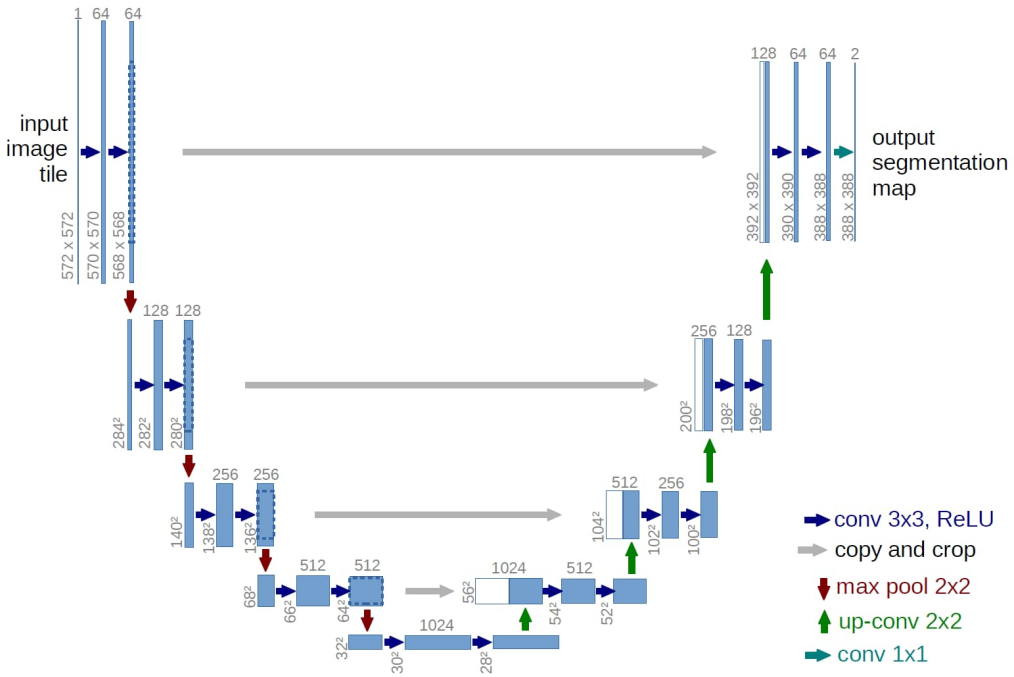

Figure 2.8: U-Net architecture. Figure taken from Ronneberger *et al.* (2015)

An encoder-decoder architecture is also used in the context of semantic image segmentation. The goal of semantic image segmentation is to assign a label to each pixel according to the object that it belongs to. In case a pixel does not correspond to anything in the training database, then no label is assigned. This is achieved using a series of convolution layers and max pooling operations for down-sampling, also referred to as encoder. The output of the encoder is transformed into latent space representation using a fully connected layer, that contains information about the entire image. This latent state representation is then up-sampled and de-convolved (transposed convolutions without up-sampling) by a series of max un-pooling layers and de-convolution layers, also referred to as decoder. A heuristic method is used to generate the final segmentation. This heuristic method greedily searches for the class that is most represented and infers its region by taking into account the probabilities and also by encouraging connectedness. The next most-represented class is then added, where it dominates at the remaining unlabeled pixels. This process is continued until it is no longer possible to add more due to insufficient evidence.

2.1.4 UNet

In a semantic segmentation network that utilises encoder-decoder architecture, the image is repeatedly down-sampled by the encoder until the receptive fields are large

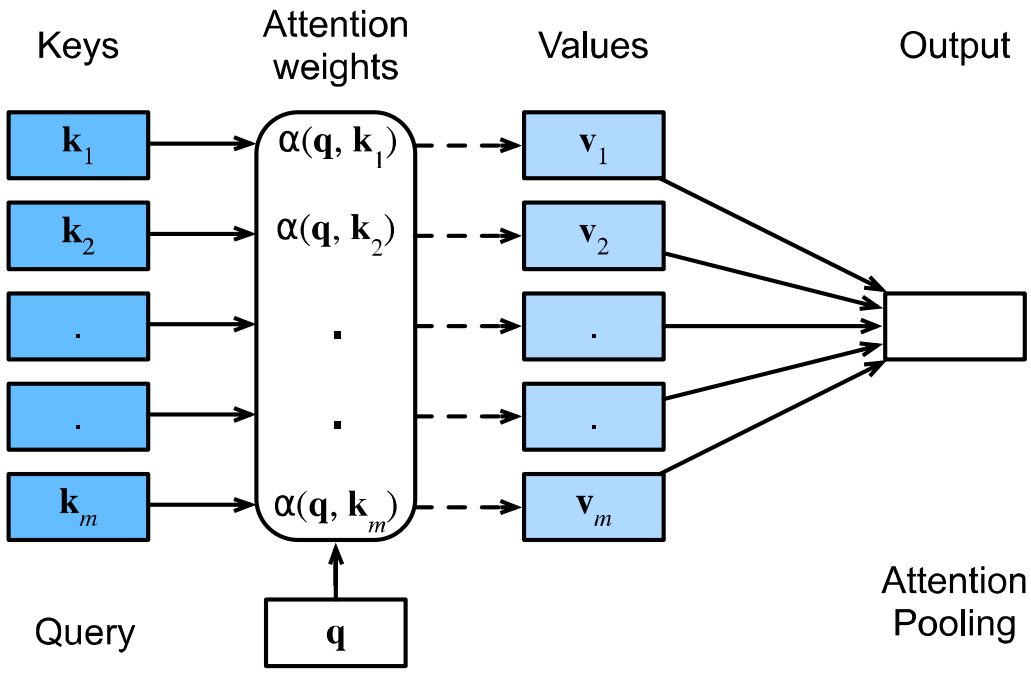

Figure 2.9: Attention mechanism uses attention pooling to compute a linear combination over values $\mathbf{v_i}$. The weights are computed using the query $\mathbf{q}$ and keys $\mathbf{k_i}$. Figure taken from (Zhang, Aston *et al.*, 2023)

and information is integrated from across the image. It is then up-sampled back to the size of the original image by the decoder. A probability over possible object classes at each pixel is the final output. One drawback of this architecture is that the high-resolution details should be "remembered" by the low-resolution representation in the middle of the network in order to make the final result accurate. However, if residual connections can be used to transfer the representations from the encoder to their partner in the decoder, this becomes unnecessary.

One such encoder-decoder architecture where the earlier representations are concatenated to the later ones is the U-Net architecture 2.8. Since U-Net is completely convolutional, after training, it can be run on an image of any size. U-Nets have also found many other uses in computer graphics and vision even though it was initially intended for segmenting medical images.

2.1.5 Attention mechanisms

One of the important aspects of human perception is the attention mechanism. Our ability to exploit partial glimpses and selectively focus on salient parts enables us to capture visual structure better. Based on the same principle, attention mechanisms in deep neural networks are one of the recent advances. Attention blocks help to

capture long range interactions between the elements. Attention mechanism was introduced as an enhancement for encoder–decoder RNNs in order to selectively focus on particular parts of the input sequence in sequence-to-sequence applications, such as machine translations (Bahdanau *et al.*, 2015).

Prior to this, the entire input was compressed by the encoder into a single fixed-length vector and fed into the decoder in such sequence-to-sequence machine translation models (Sutskever *et al.*, 2014). Attention mechanism, however, is based on the intuition that, rather than compressing the input, it might be better for the decoder to revisit the input sequence at every step. Further, it would be ideal if the decoder could selectively focus on particular parts of the input sequence at particular decoding steps, rather than always seeing the same representation of the input. The attention mechanism of Bahdanau *et al.* (2015) provided a simple means through which, at each decoding step, the decoder could dynamically attend to different parts of the input. The high-level idea behind this is that, a representation of length equal to the original input sequence would be produced by the encoder. Then, at decoding time, a context vector consisting of a weighted sum of the representations on the input at each time step would be sent as input to the decoder (via some control mechanism). Intuitively, the extent to which each step's context "focuses" on each input token is determined by the weights. Making this process of assigning the weights differentiable is extremely important to enable it to be learnt along with all of the other neural network parameters. A differentiable means of control through which a neural network can select (query) elements from a set (of keys) to construct an associated weighted sum over representations is provided by attention mechanism 2.9.

Attention mechanism successfully enhanced RNNs that already dominated machine translation applications. The original encoder–decoder sequence-to-sequence architectures were out-performed by such attention based models. However, soon, their usefulness beyond being an enhancement for encoder–decoder recurrent neural networks became evident. Moreover, their reputed usefulness for picking out salient inputs emerged. Dispensing with recurrent connections altogether, Vaswani *et al.* (2017) proposed the Transformer architecture for machine translation which relied on cleverly arranged attention mechanisms to capture all relationships among input and output tokens. Due to its remarkable performance, transformers began showing up in majority of state-of-the-art natural language processing systems by 2018. Another significant trend around the same time was the dominant practice in natural language processing of pre-training large-scale models on enormous generic background corpora. Such models were pre-trained to optimise some self-supervised pre-training objective and then were fine-tuned using the available downstream data. When applied in the paradigm of such pre-training, the gap between transformers and traditional architectures grew significantly wide. Thus the dominance of transformers coincided with the dominance of such large-scale pre-trained models that are also known as foundation models (Bommasani *et al.*, 2021).

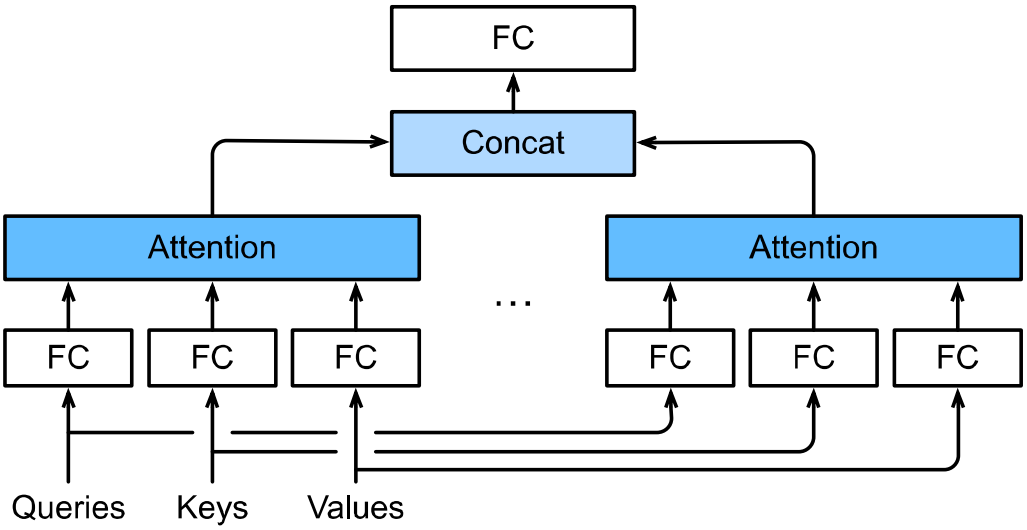

Figure 2.10: Multi-head attention, where multiple heads are first concatenated and then linearly transformed. Figure taken from (Zhang, Aston *et al.*, 2023)

2.1.5.1 Multi-head attention

Combining knowledge from different behaviors of the same attention mechanism, given the same set of queries, keys, and values would, in practice, be desirable for the model. This enables capturing dependencies of various ranges within a sequence. Such an attention mechanism jointly uses different representation subspaces of queries, keys, and values. Transforming the queries, keys, and values with h independently learned linear projections would enable going beyond performing only a single attention pooling. The next step would be to parallely feed these h projected queries, keys, and values into attention pooling. The final output is produced by concatenating the h attention pooling outputs and transforming them with another learned linear projection. This design where each of the h attention pooling outputs is a head, is called multi-head attention(Vaswani *et al.*, 2017). Fig. 2.10 depicts multi-head attention using fully connected layers to perform learnable linear transformations.

2.1.5.2 Stand-alone self-attention

Improving the performance of the network has been the focus for most of the proposed attention blocks like criss-cross attention (Huang, Zilong *et al.*, 2019), CBAM (Woo *et al.*, 2018), or attention augmented convolution (Bello *et al.*, 2019). However, in most cases, they also result in increasing the number of trainable parameters and FLOPS significantly. One of the attention mechanism that is demonstrated to significantly reduce the number of trainable parameters and FLOPS is the stand-alone self-attention (Ramachandran *et al.*, 2019).

When attention is applied to a single context instead of across multiple contexts (i. e., the query, keys, and values are extracted from the same context), it is known as self-attention. Rather than using attention as an augmentation on top of convolutions, in stand-alone self-attention, spatial convolutions are replaced with a form of self-attention. It uses local, spatial-relative attention using 2D relative position embeddings, instead of embeddings based on the absolute position, resulting in better accuracies. The spatial relative attention is defined as:

$$y_{ij} = \sum_{a,b \in N_k(i,j)} softmax_{ab}(q_{ij}^T k_{ab} + q_{ij}^T r_{a-i,b-j}) v_{ab}, \tag{2.1}$$

where *queries* q_{ij}, *keys* k_{ab}, and *values* v_{ab} are linear transformations of a pixel in position ij and the neighbourhood pixels, $a - i$ is the row offset, $b - j$ is the column offset, and $r_{a-i,b-j}$ is the concatenated row and column offset embedding.

Self-attention is shown by Ramachandran *et al.* (2019) to be translation equivariant (similar to convolutions) through the use of relative position information. Furthermore, the parameter count of attention is independent of the size of spatial extent as opposed to convolution whose parameter count grows quadratically with spatial extent. Also, the increase in computational cost of attention with spatial extent is slower as compared to convolution. For image classification tasks, the use of stand-alone self-attention in later layers of a network have been shown to outperform the baseline with far fewer FLOPS and parameters.

2.1.5.3 Spatial and channel attention

CNNs extract hierarchal information from images using convolutional filters. Information from the spatial and channel information of an image are fused to achieve this. Spatial features in each input channel are first identified by the different filters. Later, the spatial features across all available output channels is added. Typically, when creating the output feature maps, all the channels are equally weighted by the network. Convolutional Block Attention Module (CBAM) (Woo *et al.*, 2018) infers attention maps along two separate dimensions, channel and spatial sequentially. The input feature map is multiplied with these attention maps for adaptive feature refinement. Meaningful features along those two principal dimensions: channel and spatial axes are thereby emphasised. Each of the branches are hence able to learn 'what' and 'where' to attend in the channel and spatial axes respectively. Learning which information to emphasise or suppress makes the information flow within the network efficient.

Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{\mathbf{C} \times \mathbf{H} \times \mathbf{W}}$ as input, a 1D channel attention map $\mathbf{M_c} \in \mathbb{R}^{\mathbf{C} \times \mathbf{1} \times \mathbf{1}}$ and a 2D spatial attention map $\mathbf{M_s} \in \mathbb{R}^{\mathbf{1} \times \mathbf{H} \times \mathbf{W}}$ is inferred sequentially by CBAM as illustrated in Fig. 2.11. The overall attention process can be summarised as follows:

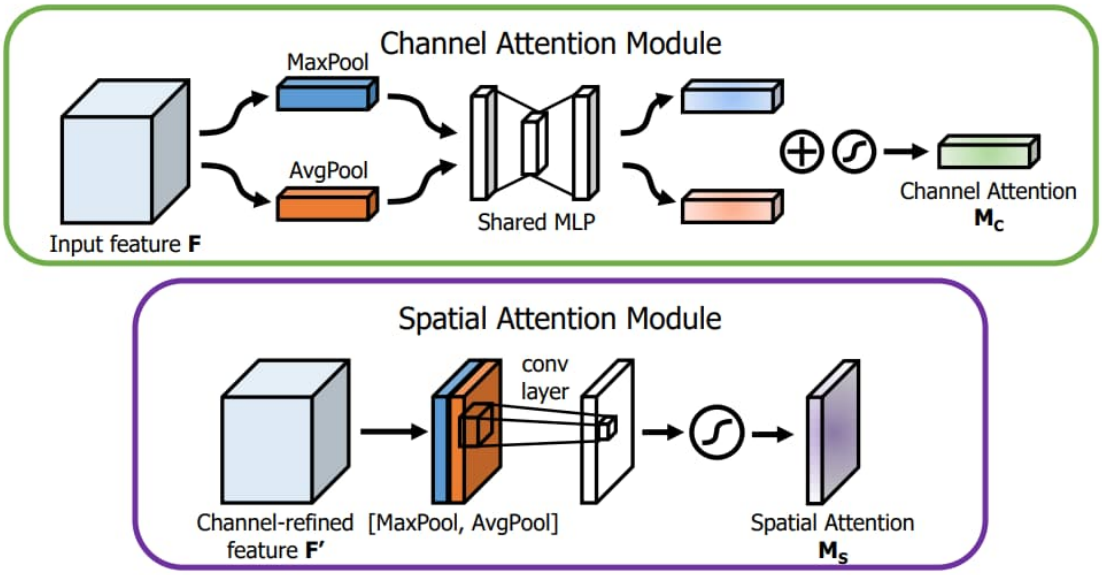

Figure 2.11: Computation process for channel and spatial attention map. Figure taken from Woo *et al.* (2018)

$$\begin{aligned} \mathbf{F}' &= \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F}, \\ \mathbf{F}'' &= \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}' \end{aligned} \tag{2.2}$$

where $\otimes$ denotes element-wise multiplication. During multiplication, the spatial attention values are broadcasted along the channel dimension, and vice versa. $\mathbf{F}''$ is the final refined output.

2.2 Activation functions

Activation functions play the role of deciding whether a neuron should be activated or not. This is done by calculating the weighted sum and then adding bias to it. Input signals are transformed to the output through these differentiable operators. Most of them also add nonlinearity.

2.2.1 Sigmoid function

Inputs whose values lie in the domain R are transformed to outputs that lie on the interval (0, 1) by the sigmoid function. Since it squashes any input in the range (-inf, inf) to some value in the range (0, 1), sigmoid is often called a squashing function.

$$sigmoid(x) = \frac{1}{1 + exp(-x)} \tag{2.3}$$

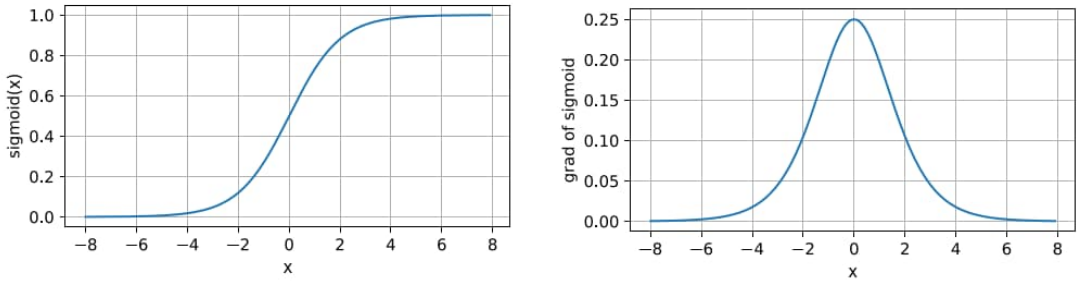

Figure 2.12: Sigmoid activation function and its derivative

Sigmoid function is a smooth, differentiable approximation to a thresholding unit. When the output has to be interpreted as probabilities for binary classification problems, sigmoids are used as activation functions on the output units. Sigmoid can be considered as a special case of the softmax. However, the simpler and more easily trainable ReLU has largely replaced the sigmoid for most use in hidden layers. One of the main reasons for this is that its gradient vanishes for large positive and negative arguments and therefore, sigmoid poses challenges for optimization (LeCun *et al.*, 1998). This could lead to plateaus that are not easy to escape from. Nonetheless, the importance of sigmoids continue and there are network architectures that leverage sigmoid units to control the flow of information across time.

Figure 2.12 shows the sigmoid activation function and its derivative. The graph shows that the sigmoid function approaches a linear transformation when the input is close to 0. Also, the derivative of the sigmoid function goes to a maximum of 0.25. The derivative approaches 0 as the input diverges from 0 in either direction.

2.2.2 Tanh function

The tanh (hyperbolic tangent) function also squashes its inputs, similar to the sigmoid function. The input is transformed into elements on the interval between -1 and 1.

$$sigmoid(x) = \frac{1 - exp(-2x)}{1 + exp(-2x)} \tag{2.4}$$

Figure 2.13 shows the *tanh* activation function and its derivative. The graph shows that the tanh function approaches a linear transformation when the input is close to 0. The shape of the tanh and sigmoid function are similar. However, about the origin of the coordinate system, the tanh function exhibits point symmetry (Kalman *et al.*, 1992).

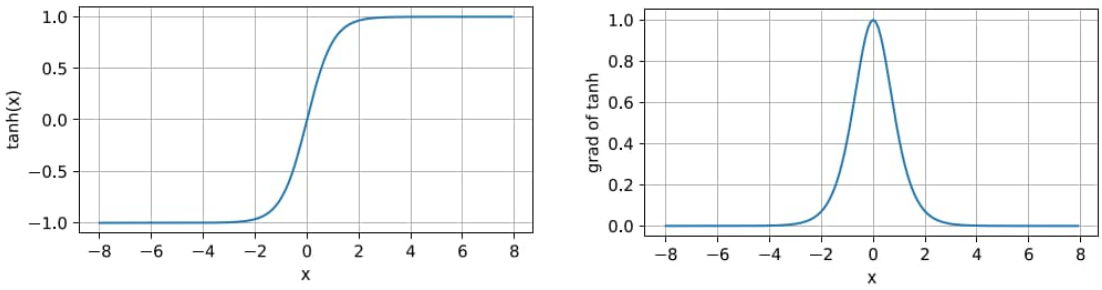

Figure 2.13: Tanh activation function and its derivative

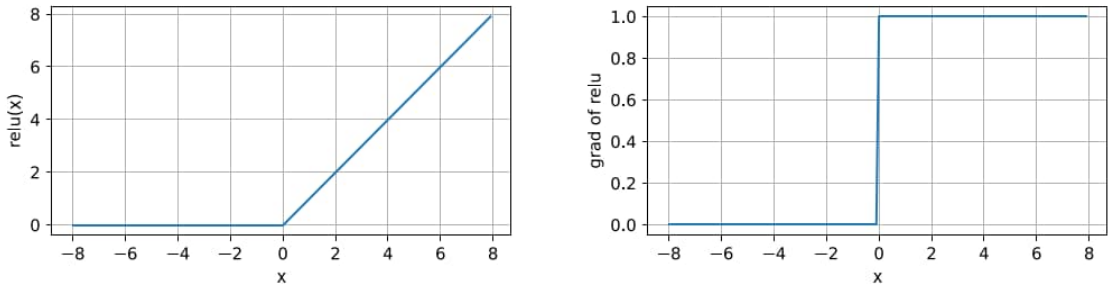

Figure 2.14: ReLU activation function and its derivative

2.2.3 Rectified Linear Unit (ReLU

Rectified linear unit (ReLU) (Nair *et al.*, 2010) is one of the most commonly used activation function due to the simplicity of its implementation as well as its good performance on a variety of predictive tasks. Furthermore, ReLU provides a very simple nonlinear transformation. Given an element x, the function is defined as the maximum of that element and 0

$$ReLU(x) = max(x, 0) \tag{2.5}$$

Only positive elements are retained by the ReLU function. All negative elements are discarded by setting the corresponding activations to 0. ReLU's derivatives either vanish or they just let the argument through and hence they are well behaved. Therefore, the optimization is well behaved and hence the problem of vanishing gradients is mitigated. Hence, ReLU is one of the most prominently used activation function.

Figure 2.14 shows the ReLU activation function and its derivative. The ReLU activation function is piecewise linear. The derivative of the ReLU function is 0 when the input is negative and 1 when the input is positive.

The ReLU function has many variants. One of them is the parametrized ReLU (pReLU) function (He *et al.*, 2015). This variant enables some information to get through even when the argument is negative by adding a linear term to ReLU.

$$pReLU(x) = max(0, x) + \alpha\, min(0, x) \tag{2.6}$$

2.2.4 Attention-based Rectified Linear Unit

One more variant of the ReLU that exploits an element-wise attention mechanism is the *Attention-based Rectified Linear Unit* (AReLU) (Chen *et al.*, 2020). It is also a learnable activation function. Similar to the ReLU activation function, the AReLU amplifies positive elements and suppresses negative ones, but with learnt, data-adaptive parameters. The attention module within AReLU learns element-wise residues of the activated part of the input. Hence, the network training is more resistant to gradient vanishing. The learnt attentive activation of AReLU results in well-focused activations of relevant regions of the feature map. It facilitates fast network training under small learning rates with only two extra learnable parameters (α and β) per layer.

AReLU (Chen *et al.*, 2020) represented as ($\mathcal{F}(x_i, \alpha, \beta)$) is defined using a combination of an element-wise sign-based attention mechanism $\mathcal{L}(x_i, \alpha, \beta)$ and a standard Rectified Linear Unit $\mathcal{R}(x_i)$, as described in equation 2.7.

$$\begin{aligned} \mathcal{F}(x_i, \alpha, \beta) &= \mathcal{R}(x_i) + \mathcal{L}(x_i, \alpha, \beta) \\ &= \begin{cases} C(\alpha)x_i & , x_i < 0 \\ (1 + \sigma(\beta))x_i & , x_i >= 0, \end{cases} \end{aligned} \tag{2.7}$$

where $X = \{x_i\}$ is the activation layer's input, $\{\alpha, \beta\} \in R^2$ are learnable parameters, C($\cdot$) clamps an input variable into [0.01, 0.99] to prevent α from becoming zero, and σ is the sigmoid activation.

2.3 Regularisation

Reducing the generalisation gap between training and test performance is an important aspect of machine learning. Regularisation techniques are a family of methods that aim to reduce this gap. Adding explicit terms to the loss function which favour certain parameter choices is what strictly constitutes regularisation. However, in machine learning, any strategy that improves generalisation is commonly referred to as regularisation. The 4 mechanisms through which generalisation can be improved are summarised in Figure 2.15.

Another way to group regularisation techniques is as follows:

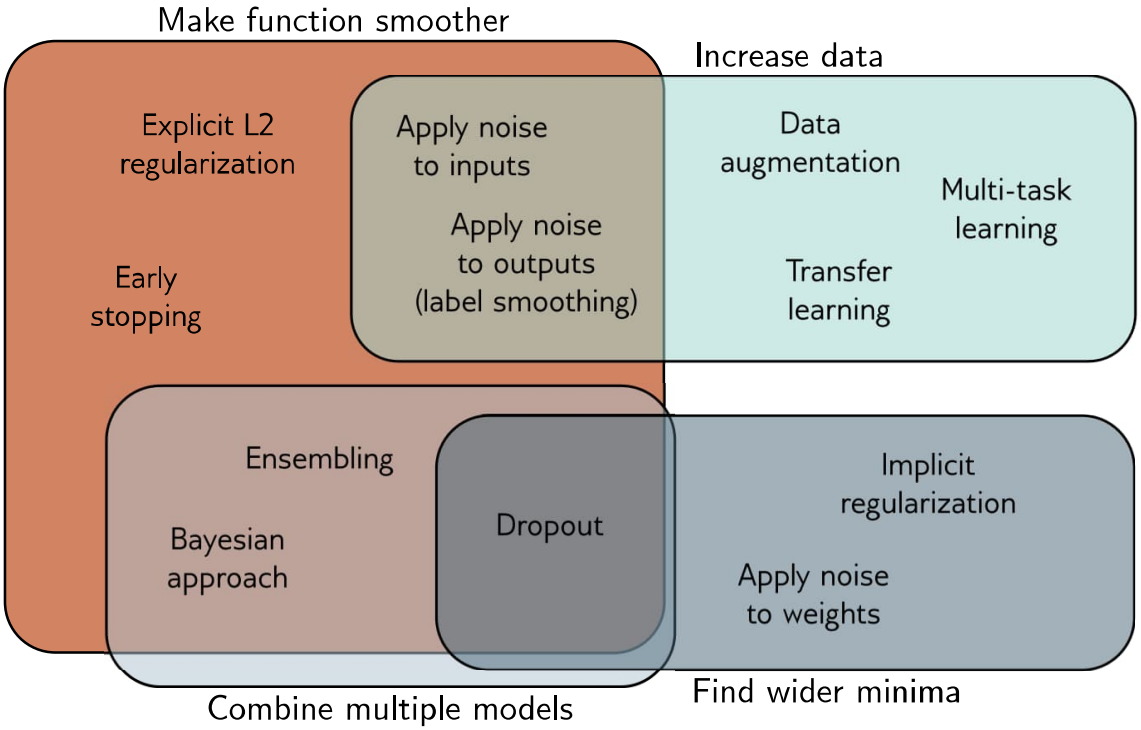

Figure 2.15: Four mechanisms for improving generalisation through regularisation methods. 1. Methods that make the modeled function smoother. 2. Methods that increase the effective amount of data. 3. Methods that combine multiple models and hence mitigate against uncertainty in the fitting process. 4. Methods that encourage the training process to converge to a wide minimum where small errors in the estimated parameters are less important. Figure taken from Prince (2023)

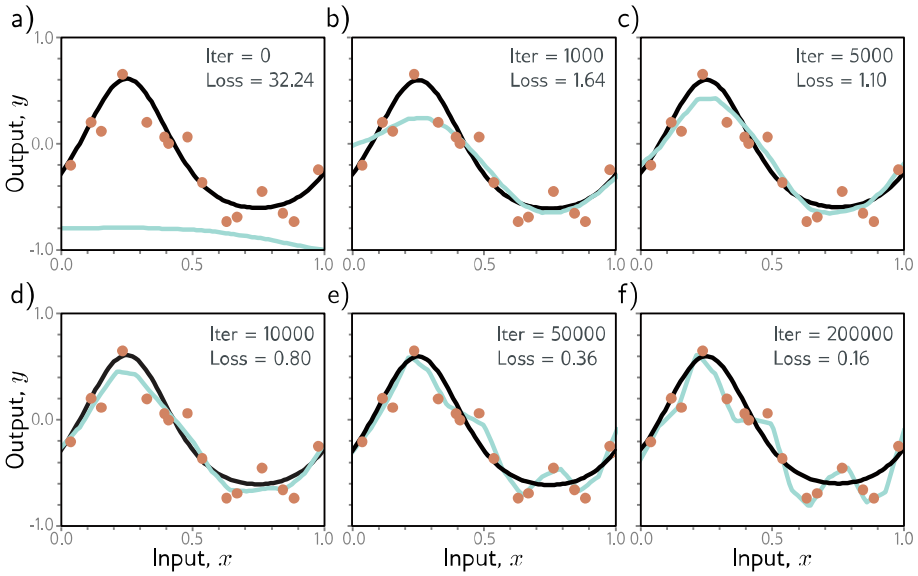

Figure 2.16: Early stopping. a) Simplified shallow network model with 14 linear regions is initialised randomly (cyan curve) and trained with SGD using a batch size of five and a learning rate of 0.05. b–d) As training proceeds, the function first captures the coarse structure of the true function (black curve) before e–f) overfitting to the noisy training data (orange points). Although the training loss continues to decrease throughout this process, the learned models in panels (c) and (d) are closest to the true underlying function. They will generalise better on average to test data than those in panels (e) or (f). Figure taken from Prince (2023)

2.3.1 Explicit regularisation

Explicit regularisation includes methods like probabilistic interpretation and L2 regularisation. In explicit regularisation, the training algorithm is encouraged to find a good solution by adding extra terms to the loss function.

2.3.2 Implicit regularisation

Implicit regularisation includes gradient descent and stochastic gradient descent since both of these exhibit an implicit yet intriguing behavior of giving preference for some solutions over others and do not move neutrally to the minimum of the loss function.

2.3.3 Heuristics based methods

2.3.3.1 Early stopping

The process of stopping the training procedure before it has fully converged is referred to as early stopping. By ensuring that the model captures the coarse shape of the underlying function but does not start to overfit to the noise 2.16, early stopping can reduce overfitting. Early stopping has an effect similar to that of explicit L2 regularisation. Since the weights are initialised to small values, with early stopping, they simply don't have time to become large.

2.3.3.2 Ensembling

Building several models and averaging their predictions is another approach to reduce the generalisation gap between training and test data. An ensemble is a group of such models. Test performance is reliably improved by this technique. However, this is at the cost of training and storing multiple models and performing inference multiple times.

Using different random initialisations is one of the ways to train different models. This may help in regions of input space far from the training data. Different models may produce different predictions and the fitted function is relatively unconstrained. Therefore, the average of several models may generalise better than any single model.

Re-sampling the training data with replacement and generating several different datasets that are then used to train different models is the other approach. This approach is known as bagging or bootstrap aggregating. The model will interpolate from nearby points if a data point is not present in one training set. The fitted function will be more moderate in this region if that point was an outlier. This approach thus has the effect of smoothing out the data. Training models with different hyperparameters or training completely different families of models are some of the other approaches.

2.3.3.3 Dropout

Dropout is a regularisation technique in which a subset (typically 50%) of hidden units are randomly clamped to zero at each iteration of SGD 2.17. The dependence of the network on any given hidden unit is thereby lessened. Furthermore, in order to ensure that the change in the function due to the presence or absence of any hidden unit is reduced, the weights are encouraged to have smaller magnitudes.

The positive benefit of this technique is that undesirable "kinks" in the function that are far from the training data and don't affect the loss can be eliminated.

At test time, the network is run as usual with all the hidden units active. However, a weight scaling inference rule is applied whereby the weights are multiplied by one minus the dropout probability. This is done in order to compensate for the fact

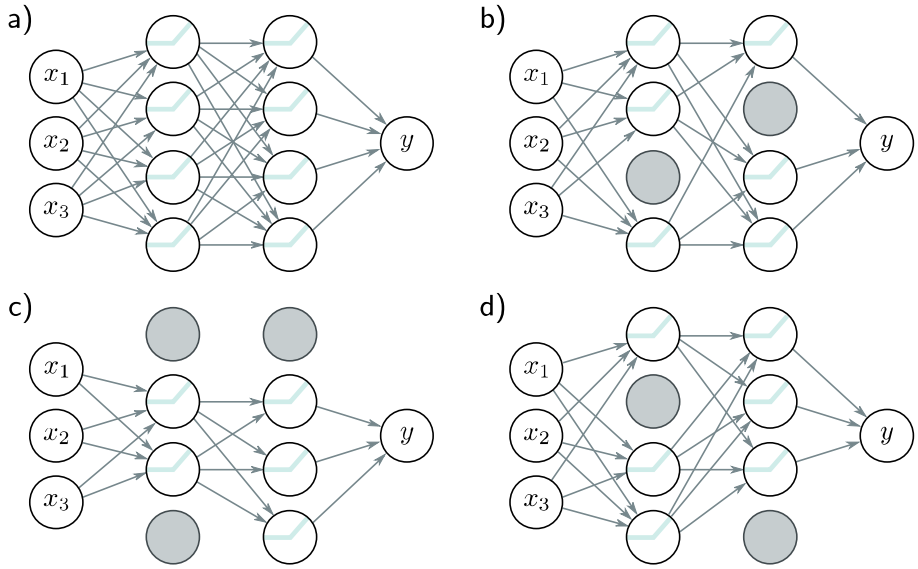

Figure 2.17: Dropout. a) Original network. b–d) A random subset of hidden units is clamped to zero (gray nodes) at each training iteration. Training therefore happens with a slightly different network each time, since the incoming and outgoing weights from these units have no effect. Figure taken from Prince (2023)

that the network now has more hidden units than it was trained with at any given iteration. Monte Carlo dropout is an alternate approach to inference. Here, as in training, the network is run multiple times with different random subsets of units clamped to zero and the results are combined. Since every random version of the network is a different model, this approach closely relates to ensembling. However, multiple networks need not be trained or stored in this case.

2.3.3.4 Applying noise

Applying noise to parts of the network during training makes the final model more robust. The idea of applying noise arises from dropout which can be regarded as the application of multiplicative Bernoulli noise to the network activations. The following are some of the options on how noise could be applied to different parts of the network:

- Adding noise to the input data: This has the effect of smoothing out the learnt function 2.18. For regression problems, this is equivalent to penalising the derivatives of the network's output with respect to its input by adding a regularising term. Adversarial training is an extreme variant. Here, small perturbations of the input that cause large changes to the output are actively searched by the optimization algorithm. These can be regarded as worst-case additive noise vectors.

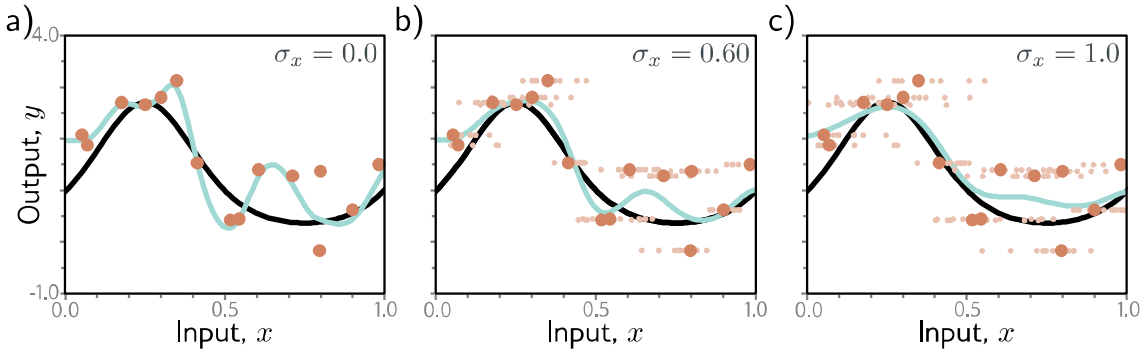

Figure 2.18: Adding noise to inputs. At each step of SGD, random noise with variance $\sigma_{\mathbf{x}}^{\mathbf{2}}$ is added to the batch data. a–c) Fitted model with different noise levels (small dots represent ten samples). Adding more noise smooths out the fitted function (cyan line). Figure taken from Prince (2023)

- Adding noise to the weights: Here, the network is encouraged to make sensible predictions inspite of small perturbations in the weights. The training thereby converges to the local minima in the middle of wide, flat regions, where changing of the individual weights do not have much effect.

- Perturbing the labels: Here, methods like label smoothing are used to perturb the labels. The cross-entropy between the predicted distribution and a distribution where the true label has probability $\mathbf{1} - \rho$ is minimised by changing the loss function, and the other classes have equal probability. Thereby, the network is made more capable to generalise to various scenarios.

2.3.3.5 Transfer learning and multi-task learning

Transfer learning is mainly useful in scenarios where there is lack of sufficient data for network training. In such scenarios, the network is pre-trained on a related task for which sufficient training data is available. Then, the last layer of this network is removed and one or more layers that produce a suitable output are added. Finally, this new network is trained for the original task by either keeping the main model fixed and training the new layers only, or by fine-tuning the entire model.

The intuition behind transfer learning is that the network will be able to transfer the internal representation of the data that it built from the related task to the original task. An alternate way to view this is that, the final network is more likely to produce a good solution when most of its parameters are initialised in a sensible part of the solution space.

Another technique that is related to transfer learning is multi-task learning. Here, the network is trained to solve multiple related tasks concurrently. One example for such multi-task learning is to simultaneously learn to segment the scene,

estimate the pixel-wise depth, and predict a descriptive caption for a given input image. Gaining some understanding of the input image is essential for all these tasks. The intuition behind multi-task learning is that solving multiple related tasks simultaneously would aid the model to perform better on each of those individual tasks.

2.3.3.6 Self supervised learning

Self-supervised learning methods enable creation of large amounts of "free" labeled data that can be used for transfer learning. Generative and contrastive learning are the 2 well-known families of self-supervised learning.

- Generative self-supervised learning: Here, a part of every data example is masked. Predicting the missing part is the secondary task in this case.
- Contrastive self-supervised learning: Here, pairs of examples with commonalities are compared to unrelated pairs.

2.3.3.7 Augmentation

Transfer learning leverages different dataset to improve performance while multi-task learning aims to achieve this by leveraging additional labels. Augmenting or expanding the dataset is yet another option that is often used to improve the network performance. One of the data augmentation approaches that is commonly used is the generation of additional training data by transforming the input data in such a way that the label stays the same. The model is thereby taught to be in-different to irrelevant data transformations.

2.4 Metrics

2.4.1 Medical image segmentation metrics

Taha *et al.* (2015) group medical image segmentation metrices into six categories depending on the relations between the metrics, their nature and definition. These 6 categories are overlap based, volume based, pair-counting based, information theoretic based, probabilistic based, and spatial distance based.

2.4.1.1 Dice coefficient

The Dice coefficient (Dice, 1945) (DICE) is one of the often used metric in validating medical volume segmentation. It is also known as Dice score or Dice Similarity Coefficient. It is a similarity measure between two sets of data. It is therefore also called as the overlap index. In the context of image segmentation, in order

to evaluate the similarity between a predicted segmentation mask and the ground truth segmentation mask, Dice score is used. The Dice score ranges from 0 to 1, where 0 indicates no overlap and 1 indicates perfect overlap.

The Dice score is calculated as follows:

$$DICE = \frac{2|(X \cap Y)|}{|X| + |Y|} \tag{2.8}$$

The Dice score is equal to twice the size of the intersection divided by the sum of the sizes of the two sets. The Dice score therefore measures the proportion of overlap between the two sets, normalised by the size of the sets.

Apart from the direct comparison between automatic and ground truth segmentations, Dice score is also commonly used to measure repeatability or reproducibility. Zou *et al.* (2004) used the Dice score as a statistical validation metric to evaluate the performance of the reproducibility of manual segmentations.

2.4.1.2 Jaccard Index

The Jaccard index (JAC) (Jaccard, 1912) between two sets is defined as the intersection between them divided by their union, that is

$$JAC = \frac{|(X \cap Y)|}{|X \cup Y|} \tag{2.9}$$

At the extrema 0, 1, JAC and DICE are equal. At all other values, JAC is always larger than DICE. Furthermore the two metrics are related according to

$$JAC = \frac{DICE}{2 - DICE} \tag{2.10}$$

JAC and DICE therefore measure the same aspects and hence result in the same system ranking.

2.4.1.3 Hausdorff Distance

Hausdorff Distance (HD) is a spatial distance-based metric which is widely used in the evaluation of image segmentation as a dissimilarity measure. It is recommended when the overall accuracy, e.g the boundary delineation (contour), of the segmentation is of importance (Fenster *et al.*, 2005). Distance-based measures take the spatial position of voxels into consideration. Since the distances are calculated in voxel, the voxel size is not taken into account.

The Hausdorff Distance (HD) between 2 finite point sets (A, B is defined by

$$HD(A, B) = max(h(A, B), h(B, A)) \tag{2.11}$$

where $\mathbf{h}(\mathbf{A}, \mathbf{B})$ is called the directed Hausdorff distance and given by

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \tag{2.12}$$

where $\|\mathbf{a} - \mathbf{b}\|$ is some norm, e.g. Euclidean distance.

The HD is generally sensitive to outliers. Hence, it is not recommended to use the HD directly, since noise and outliers are common in medical segmentations (Gerig *et al.*, 2001; Zhang, Dengsheng *et al.*, 2004). The Average Distance, or the Average Hausdorff Distance (AVD)- is known to be stable and less sensitive to outliers than the HD. The AVD is the HD averaged over all points. It is defined by

$$AVD(A, B) = max(d(A, B), d(B, A)) \tag{2.13}$$

where $\mathbf{d}(\mathbf{A}, \mathbf{B})$ is called the directed Average Hausdorff distance given by

$$d(A, B) = \frac{1}{N} \sum_{a \in A} \min_{b \in B} \|a - b\| \tag{2.14}$$

2.4.2 Image quality and similarity metrics

2.4.2.1 Perception-based Image Quality Evaluator (PIQUE)

Perception-based Image Quality Evaluator (PIQUE) (Venkatanath *et al.*, 2015) extracts local features from perceptually significant spatial regions to predict quality. Block-level analysis is done to determine distortions in local blocks. A score is assigned to each distortion block based on distortion type. The block-level scores are pooled in order to arrive at the overall image quality score.

2.4.2.2 Structural Similarity (SSIM) Index

Structural Similarity (SSIM) Index (Wang, Zhou, Bovik, *et al.*, 2004) uses luminance, contrast, and structure to compute similarity measurements. The mean intensity of the reference and the target image is compared to estimate the luminance. In order to estimate the contrast, the standard deviation of the two image signals is compared. The normalised signals are transformed to have unit standard deviation in order to compare the structure. SSIM is computed using the formula:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{2.15}$$

2.4.2.3 Learnt Perceptual Image Patch Similarity (LPIPS)

LPIPS (Zhang, R *et al.*, 2018) computes deep feature spaces from pre-trained CNN architectures like SqueezeNet, AlexNet, and VGG. Feature stacks are first extracted

from L layers of a pre-trained CNN. The features are then normalised in the channel dimension. The activations are scaled and the l_2 distance is computed between these activations. The final score is computed by spatial averaging and channel-wise summation. In summary, cosine distance is computed in the channel dimension. This is then averaged across spatial dimensions. The distance between the reference and distorted patches x, x_0 is computed using the formula:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2. \tag{2.16}$$

Part III

RELATED WORK

3

Related Work

3.1 Medical Image Segmentation

In the healthcare domain, one of the widely researched machine learning area is medical image segmentation (Long *et al.*, 2015; Lin *et al.*, 2017; Yu, Changqian *et al.*, 2018; Zhou, Zongwei *et al.*, 2018; Isensee *et al.*, 2019; Rajamani, Kumar, Gowda, *et al.*, 2022; Rajamani, Kumar, Rani, *et al.*, 2022). Some of the commonly used modalities for medical image acquisition include X-ray, ultrasound, computed tomography (CT), magnetic resonance imaging (MRI) and mammography. Medical image segmentation involves segmenting of the organs, tissues or specific pathologies in these medical images. It mainly involves the correct identification of a *region of interest* (ROI) in medical images. Segmenting the different anatomical parts of the heart from cardiac magnetic resonance (MR) image is one such example. Automatic and accurate segmentation holds the potential of reducing the time to diagnose and reduce the manual segmentation efforts by clinicians.

One of the important factor to enable any research to be utilized in real world applications in clinical domain is the holistic measurement of performance of the research methods or algorithms. The topic of identifying appropriate evaluation protocols is therefore of great prominence in the research community, more so in the deep learning research community.

In this context, the lack of reliability in medical image segmentation performance assessments was explored by Müller, Soto-Rey, *et al.* (2022). Renard *et al.* (2020) and Parikh *et al.* (2019) noted that metrics typically used for reporting are often overoptimistic of model performance. Furthermore, their potential weaknesses are often not revealed. Therefore, translating research to clinical settings often is encountered by problems (El-Naqa *et al.*, 2021; Parikh *et al.*, 2019). Müller, Soto-Rey, *et al.* (2022) provided an overview of commonly-used evaluation scores, such as the Dice similarity coefficient, Jaccard coefficient or Cohen's Kappa and described which metric is best suited for which usecases and scenarios. They also established a set of guidelines for interpretation and a standardised evaluation. Müller, Hartmann,

et al. (2022) proposed a library named MISeval, for metric evaluation, to further advance standardisation and reproducibility.

A set of boundary overlap metrics was explored by Yeghiazaryan *et al.* (2018). Their aim was to capture the segmentation errors that occur in the context of utilising size-based, overlap-based, and boundary distance based segmentation metrices. The existence of large differences between existing evaluation scores was highlighted by them. High dependencies of the utilised metrics on the clinical use case was also demonstrated by them. All of these underscore the fact that the applicability of the methods to real-world data still faces challenges even when these methods achieve high values of well-known metrics, such as the Dice score.

Several contemporary research demonstrate the prevelance of these issues across the general medical image segmentation field (Hesamian *et al.*, 2019; Jungo *et al.*, 2019; Fu *et al.*, 2021; Reinke, Eisenmann, *et al.*, 2021; Reinke, Tizabi, *et al.*, 2021). In this thesis, we focus on the specific facets of this problem appear for the task of Cardiac MR Image segmentation. A comprehensive summary of the performance of state of the art deep learning methods for Cardiac MR Image segmentation is presented by Bernard *et al.* (2018). The challenges that are still prevalent in this area are also highlighted by them. Some of the significant challenges that they identify are:

- Segmenting the Right Ventricle (RV) and calculating the RV ejection fraction.
- Segmenting the Myocardium at the End Systole (ES) phase due to the difficulty in precisely delineating the Left Ventricle (LV) and Right Ventricle (RV) walls.
- Segmenting slices near the apex and base: Small structures in the apex and difficulty in differentiating between multiple structures at the basal slices pose significant challenges.
- Manual segmentation of the apex and basal slices by experts results in differing outcomes (inter-observer variability).
- Deep learning based segmentation methods often generate anatomically impossible results in some of the slices.

These aspects highlight that cardiac MR segmentation is a field that is technically challenging. Therefore, it becomes more relevant to ensure that the boundary conditions and limitations of each deep learning based solution for this is precisely identified before they can be used in a clinical context.

One of the significant efforts towards this is the formation of a consortium of multiple academia and industry researchers as well as practitioners to analyse the flaws in machine learning algorithm validation. In their seminal research (Maier-Hein, Reinke, *et al.*, 2022), this consortium has identified various pitfalls in the choice of validation metrics which they group into the following categories:

- Inappropriate problem phrasing
- Poor selection of metrics
- Poor application of metrics

With their "Metrics Reloaded" framework, they discuss ways to address these challenges and propose a problem fingerprinting framework as well as a metrics selection methodology.

In the context of large-scale international challenges that benchmark different models, Maier-Hein, Eisenmann, *et al.* (2018) emphasise that the outcomes have to be interpreted with care. The influence of choice of metrics as well as criteria for aggregated ranking across metrics in deciding the winning method is highlighted by them. One of the significant design choice that is shown by them to change the winners based on the aggregation method chosen is the metric-based vs case-based ranking scheme.

Evaluating disagreggated model performance is a research area being actively focused on even in areas other than medical image segmentation. Generally, the focus here is on evaluating the fairness of models with respect to different groups (e. g., age and gender groups) or sub-populations. However, evaluating how models perform across different individuals is also being researched (Doddington *et al.*, 1998; Kathan *et al.*, 2022). 'Individual fairness' is a related notion in which it is contested that "similar individuals should receive equal treatment" (Sharifi-Malvajerdi *et al.*, 2019).

Corner cases in classification tasks was explored by Ouyang *et al.* (2021) in their work. They propose a metric which targets the characteristics of corner cases and this is calculated on the basis of modified 'surprise' adequacy. Additionally, in order to achieve fairer classication performance for all subjects in a dataset, they generated artificial corner cases and used them to improve a model. A "Deep Validation" framework was proposed by Wu *et al.* (2019) for classification tasks. When the system is perceived to be working incorrectly, this framework aims to identify the error-inducing inputs and flags them for human intervention. In the context of medical image segmentation, the goal to strive for is that the model needs to generalise well to different patients, irrespective of the anatomical or pathological differences.

3.2 Speech emotion recognition

(Gunes *et al.*, 2011) demonstrate that the emotional state of a speaker is revealed in the paralinguistic information embedded in the human voice. Humans use emotions to adjust the content of their message or the tone of their voice. Through this, they aim to smooth the interaction and empathise with their interactant. Hence,

this information plays a significant role in *Human-Human Interaction*. Similarly, in order to boost the *Human-Computer Interaction* (HCI) experience and to better mimic HHI, machines need to be able to leverage *Speech Emotion Recognition* (SER) capabilities. This capability is also important is many use-cases other than HCI.

Tradititional methods for Speech Emotion Recognition (SER) tried to capture the salient information from the human voice by extracting hand-crafted features like pitch, energy etc. from acoustic signals (El Ayadi *et al.*, 2011). Conventional machine learning techniques, such as *Hidden Markov Models* (HMMs) or *Support Vector Machines* (SVMs) (Schuller *et al.*, 2003; Yi-Lin Lin *et al.*, 2005) were used to process these hand-crafted acoustic features. Either these hand-crafted acoustic features or the raw audio itself have been used as input for deep learning techniques by more recent approaches. These include *Convolutional Neural Networks* (CNNs) (Alif Bin Abdul Qayyum *et al.*, 2019), *Recurrent Neural Networks* (RNNs) (Mirsamadi *et al.*, 2017; Zhao, Ziping *et al.*, 2019; Yu, Yeonguk *et al.*, 2020), or the combinations of CNNs and RNNs (Zhao, Jianfeng *et al.*, 2019).

The temporal dynamics of sequential data are captured well by RNNs, such as *Long Short-Term Memory* (LSTM) (Hochreiter and Schmidhuber, 1997), and *Gated Recurrent Units (GRU)* (Cho *et al.*, 2014). Their ablility to capture the temporal dependencies of the acoustic features makes such techniques well-suited for SER tasks. Recent research (Mirsamadi *et al.*, 2017; Zhao, Ziping *et al.*, 2019; Yu, Yeonguk *et al.*, 2020) has also demonstrated that attention mechanisms can be used to assist RNNs to focus on the most emotionally salient information. (Ramet *et al.*, 2018; Yeh *et al.*, 2019) also used contextual information to improve the performance of SER systems. Through their *Interaction-Aware Attention Network* (IAAN), Yeh *et al.* (2019) successfully leveraged contextual information. In a two-speaker dialog scenario, IAAN detects the emotional state of one speaker's utterance by using the previous speaker turns and learning its attention scores.

3.3 Time series analysis

One of the major challenges in time series analysis is to be able to deal with sparse and irregularly sampled time-series data. Rubanova *et al.* (2019) proposed a generalised recurrent neural network (RNN) with continuous-time hidden dynamics defined by ordinary differential equations (ODEs) to overcome the challenge of non-uniform time intervals. They showed that the RNN-based equivalents were outperformed by these ODE-based models on irregularly-sampled data. Furthermore, the ability to handle arbitrary time intervals between observations and explicitly model the likelihood of observation times using Poisson processes is handled naturally by ODE-RNNs (Rubanova *et al.*, 2019).

These models however face difficulties when the input data has long-term dependencies. An algorithm which is based on the long short-term memory (LSTM)

approach is therefore proposed by Lechner *et al.* (2020). This method separates the memory from its time-continuous state and a continuous-time dynamical flow is thereby encoded within the RNN. New inputs arriving at arbitrary time-lags can be handled with this. At the same time, a constant error propagation through the memory path is ensured. Lechner *et al.* (2020) call these RNN models ODE-LSTMs. They also demonstrated that, on non-uniformly sampled data with long-term dependencies, these models can outperform their advanced RNN-based counterparts .

Considering irregular sampling from the perspective of missing data, Li, Steven Cheng-Xian *et al.* (2020) propose an encoder-decoder framework to learn from generic indexed sequences. Additionally, they also propose the following:

- Learning methods for this framework based on variational autoencoders and generative adversarial networks
- Continuous convolutional layers that can be efficiently combined with existing neural network architectures

.

Tan *et al.* (2020) propose a novel end-to-end dual-attention time-aware gated recurrent unit (DATA-GRU) to handle irregular multivariate time series data. They demonstrate that the DATA-GRU is especially able to address the following two aspects:

1. Preservation of the informative varying intervals: This is achieved through the introduction of a time-aware structure which can directly adjust the influence of the previous status in coordination with the elapsed time.
2. Tackling missing values: This is accomplished through their proposed dual-attention structure which jointly considers data-quality and medical-knowledge.

Moreover, using a novel unreliability-aware attention mechanism, they handle the diversity in the reliability of different data. A symptom-aware attention mechanism that extracts medical reasons from original clinical records is proposed by Tan *et al.* (2020). Multi-Time Attention Networks (mTAN) is proposed by Shukla *et al.* (2021) where an attention mechanism is used to produce a fixed-length representation from a time series containing a variable number of observations and an embedding of continuous time values is learnt.

Part IV

CONTRIBUTIONS

4

Detecting and handling corner cases in medical image segmentation

This chapter discusses the research work "Toward Detecting and Addressing Corner Cases in Deep Learning Based Medical Image Segmentation" published in IEEE Access (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Venkateshvaran, *et al.*, 2023).

4.1 Motivation

There are several challenges in translating machine learning research into clinical practice. In this research, we identify some of these challenges in the area of medical image segmentation. Furthermore, we propose strategies to deal with these challenges systematically. Corner cases or cases where deep learning model yields wrong segmentation is our main focus. The performance of medical image segmentation algorithms is generally reported using standard metrics like the average Dice score across all patients. We uncover an important aspect that reporting based on aggregation across all patients has the drawback that corner cases go unnoticed. Models with superior performance could potentially provide erroneous or even anatomically impossible segmentation results on some challenging cases without being noticed due to such reporting. Using the Automated Cardiac Diagnosis Challenge (ACDC) challenge's Magnetic Resonance (MR) cardiac image segmentation task, we demonstrate how corner cases go unnoticed. Additionally, to address this challenge and to help identify and report such corner cases, we propose a framework. Furthermore, we propose a novel balanced checkpointing scheme. This scheme enables determining a solution that performs well even on the corner cases. On our identified corner case in the ACDC segmentation challenge, by using our balanced checkpointing

scheme, the Dice score on the corner case improves by 44.6 % for LV, 46.1 % for RV and 38.1 % for the Myocardium. Our proposed framework is also generalisable and can be applied in contexts other than MR cardiac segmentation. We demonstrate this generalisability by using it for chest X-ray lung segmentation. Even for deep learning tasks beyond medical image segmentation, this framework has significant applicability.

4.2 Dataset and Baseline Network Architecture

In this section, we provide details about the dataset as well as the baseline network architecture used in our experiments.

4.2.1 The ACDC segmentation dataset

Our experiments are conducted on the Automated Cardiac Diagnosis Challenge (ACDC)'s segmentation dataset (Bernard *et al.*, 2018). Evaluating the performance of deep learning methods in segmenting the myocardium (MYO), left ventricle (LV) and right ventricle (RV) and classifying the pathologies from cardiac MRI is the objective of the challenge. 3D cine-Magnetic Resonance (MR) cardiac scans of 100 unique patients from the University Hospital of Dijon comprises the training dataset of this challenge. These 100 patients comprise of 20 patients belonging to following five classes each:

1. Normal case
2. Heart failure with Infarction
3. Dilated Cardiomyopathy
4. Hypertrophic Cardiomyopathy
5. Abnormal Right Ventricle

The End Systole (ES) and End Diastole (ED) frames, identified based on the motion of the mitral valve from the long axis orientation by a single expert, is provided for each patient. Thus, for these 100 patients, there are 200 volumes in total. Also, the ground truth segmentation masks for the Left Ventricle (LV), Right Ventricle (RV) and Myocardium (MYO) is made available. The challenge test set constitutes another 50 patients, with 10 patients per class.

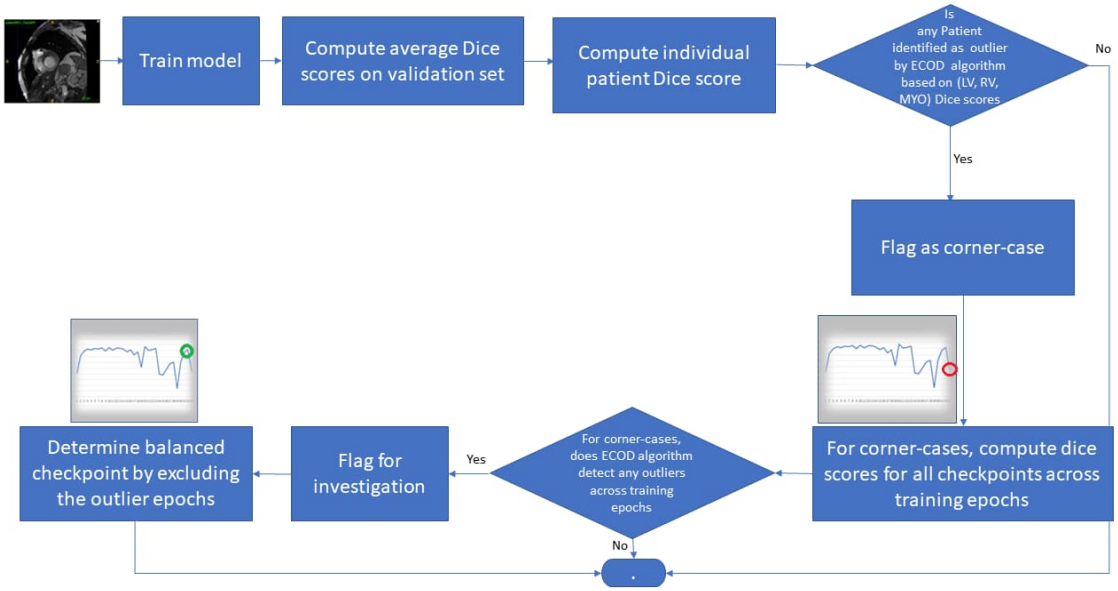

Figure 4.1: Schematic of our proposed framework for detecting and addressing corner cases in deep learning based medical image segmentation. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Venkateshvaran, *et al.*, 2023)

4.2.2 SAUNet architecture

Sun *et al.* (2020) recently proposed a U-Net based network for medical image segmentation named SAUNet – Shape Attentive U-Net for Interpretable Medical Image Segmentation. On the ACDC Cardiac MR segmentation challenge dataset, SAUNet not only achieves high average Dice scores but also provides good interpretability. SAUNet comprises of a texture stream and a gated shape stream. The texture stream has a U-Net like structure (Ronneberger *et al.*, 2015). However, the encoder is replaced with dense blocks from DenseNet-121 (Huang, Gao *et al.*, 2017). Further, each decoder block is a dual attention block. Additionally, shape features are learnt through a secondary stream which processes shape features of the image. Furthermore, in the decoder, at every resolution of the U-Net, the interpretability of features is enabled using spatial and channel-wise attention paths. In our experiments, we utilise SAUNet as the baseline architecture. Also, we use the same hyperparameters and training-validation split as used by Sun *et al.* (2020).

4.3 Novel framework for detecting and handling corner cases

Figure 4.1 depicts the schematic of our proposed methodology for identifying and addressing corner-cases. The following sections explain this methodology.

4.3.1 Methodology for detecting and reporting of corner-cases

currently, aggregate metrics is reported by Deep learning based medical image segmentation methods. In order to determine potential outliers, we propose that the characteristics of patient-wise metrics should be analysed.

Empirical-Cumulative-distribution-based Outlier Detection (ECOD) (Li, Zheng *et al.*, 2022) is one of the recent unsupervised approaches for outlier detection in large, high-dimensional datasets. ECOD is a method for multivariate statistical anomaly detection. It leverages the fact that outliers are often the "rare events" that appear in the tails of a distribution (right-tail and left-tail). In this method, an empirical cumulative distribution is first computed along each data dimension. In the next step, this empirical distribution is utilised to estimate the left and right tail probabilities ($\hat{\mathbf{F}}^{(j)}_{left}$ and $\hat{\mathbf{F}}^{(j)}_{right}$). Finally, by aggregating the estimated tail probabilities across all dimensions, the outlier score is computed in a non-parametric way (Li, Zheng *et al.*, 2022).

Given input data $\mathbf{X} = \{X_i\}_{i=1}^{n} \in \mathbb{R}^{n \times d}$ with n samples and d features where $X_i^{(j)}$ refers to the value of j-th feature of the i-th sample,

$$\hat{\mathbf{F}}^{(\mathbf{j})}_{\mathbf{left}}(\mathbf{z}) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbb{1}\{\mathbf{X}^{(\mathbf{j})}_{\mathbf{i}} \leq \mathbf{z}\} \text{ for } z \in \mathbb{R} \tag{4.1}$$

$$\hat{\mathbf{F}}^{(\mathbf{j})}_{\mathbf{right}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbb{1}\{\mathbf{X}^{(\mathbf{j})}_{\mathbf{i}} \geq \mathbf{z}\} \text{ for } z \in \mathbb{R} \tag{4.2}$$

where $\mathbb{1}\{.\}$ is the indicator function that is 1 when its argument is true and is 0 otherwise (Li, Zheng *et al.*, 2022).

For cardiac image segmentation, we propose to analyse the Dice scores of LV, RV and MYO jointly by representing them as a 3D vector. For every patient, this 3D vector is computed. Then, the ECOD algorithm is used to determine the corner cases by using these 3D vector values for all patients. Those outliers that are detected with this approach are flagged so that they can be analysed in detail. Additionally, for these cases, the segmentation outcomes should also be reported so that clinicians can analyse and gain understanding as to why the model fails in these cases.

4.3.2 Strategy for getting further insights into the corner cases

Generally, the training process is monitored by plotting the average Dice scores across the training epochs. But, insights into the model performance on the corner-cases cannot be obtained from this. We therefore propose that the Dice score curves of the corner-cases should be analysed across the training epochs to gain insights. We again utilise the ECOD algorithm (Li, Zheng *et al.*, 2022) to detect if there are any outliers across the training epochs. The analysis is done using the 3D (LV, RV, MYO) Dice scores across the training epochs for the corner-cases. This is different from the previous step where the analysis is done across patients.

4.3.3 Approach for identifying a balanced checkpoint

The usual approach used for model checkpointing during training is based on least-loss or highest average-IoU (Intersection Over Union). However, in scenarios where the Dice score varies significantly across different epochs for corner cases, this traditional checkpointing scheme could compromise the performance on corner-cases. Such a model could also produce segmentation outcomes that are anatomically impossible in clinical practice. Therefore, in order to enable deep learning research to be usable in clinical context, it is important to use a balanced checkpointing approach.

For identifying a more balanced checkpoint, as a first step, we propose to exclude all those epochs which are detected as outlier epochs for the corner-case in the previous step. Out of the remaining epochs, we propose to use the final epoch as the balanced checkpoint.

4.4 Experimental setup and results for medical image segmentation

4.4.1 Corner case detection and reporting

In Table 4.1, we report the results on the ACDC segmentation challenge dataset. In column 2, the average Dice scores obtained using our model trained with a SAUNet network architecture (Sun *et al.*, 2020) is reported. Furthermore, we utilise the methodology proposed above to identify corner-cases. For this, we first compute patient-wise Dice scores for LV, RV and MYO and provide these 3-dimensional scores to the ECOD algorithm (Li, Zheng *et al.*, 2022) to detect outliers. We use the default contamination rate of 0.1 of ECOD algorithm from the PyOD toolbox (Zhao, Yue *et al.*, 2019). With this approach, we detect Patient057_ES as the only outlier. We report the Dice-scores of this corner case patient in column 3 of the table.

Organ	(a). Avg Dice score	(b). Dice score for corner case Patient057_ES	Dice score difference (a-b)
LV	0.912	0.351	0.561
RV	0.833	0.201	0.632
MYO	0.848	0.441	0.407

Table 4.1: Average Dice score and Dice scores for the corner-case identified for LV, RV, and MYO on ACDC validation set

The difference between the average Dice scores and the Dice scores of Patient057_ES is report in column 4 of the table. From the table, it is evident that the difference between average Dice scores and that of the corner-case Patient057_ES is 56.1 % for LV, 63.2 % for RV and 40.7 % for MYO.

The visual segmentation results for all the 8 slices at End Systole for the corner-case Patient057_ES is presented in Figure 4.2. For the first 4 slices, we can see that the predicted segmentation is not only completely incorrect but also anatomically impossible. The left ventricle region is identified as the myocardium and the myocardium region is identified as the right ventricle in these 4 slices.

4.4.2 Insights into the corner cases

By utilising our strategy for getting further insights into the corner-cases, for the Patient057 which is identified as corner-case, we analyse its Dice scores across the training epochs. Using the 3D (LV, RV, MYO) Dice scores across the training epochs with the ECOD algorithm (Li, Zheng *et al.*, 2022), we observe that this patient has outliers across the training epochs, unlike the other patients. Therefore, we flag Patient057 for careful analysis by clinicians and researchers.

In Figure 4.3, the plot of the Dice scores for the entire validation set as well as for Patient057 is presented. The average Dice score plot for the entire validation set is depicted in the top row. The individualised Dice score plot for the corner-case, Patient057_ES is depicted in the bottom row. The plots for LV, RV, MYO and a consolidated view for all these 3 anatomies is depicted in the 4 columns, respectively. The plot contains Dice scores for those epochs where the model was checkpointed. Least average-loss is the criteria that we use to create these checkpoints.

All the curves in the first row pf this figure seem to indicate that the model is training effectively. Model performance and metrics are normally reported in this manner. However, a different scenario is evident from the bottom row where for the corner-case, Patient057_ES, we observe that the Dice scores varies considerably

across the training epochs for LV, RV and MYO. The Dice score between the 24th and 25th checkpoint for instance has a very large variation. Such abnormal variations could indicate erroneous model training or model performance, but these get masked when only average Dice scores are considered.

4.4.3 Balanced checkpoint determination

By using our proposed balanced checkpoint determination approach, the outlier epochs determined by ECOD algorithm are excluded and of the remaining ones, we choose the final epoch. As observed in row 2 of Figure 4.3, the balanced checkpoint identified with this approach is the 32nd (penultimate) checkpoint.

Table 4.2 reports the results of using this identified balanced checkpoint. As noted in column (a) of the table, the average Dice scores for the entire validation dataset based on least-loss checkpoint are 0.912 for LV, 0.833 for RV, and 0.848 for the Myocardium. This seems like a reasonably well performing solution at face value. However, this same checkpoint results in extremely low Dice scores for Patient057 of 0.352 for LV, 0.201 for RV, and 0.441 for the Myocardium as observed in column (c) of the table. Thus, the performance on the corner case is considerably compromised by following such a classical approach of saving the model based on least-loss. At our proposed checkpoint, the corner case Patient057 has significantly higher Dice scores as observed in column (d) of the table. At the proposed checkpoint, Patient057 has a Dice score of 0.798 for LV, 0.662 for RV, and 0.822 for the Myocardium which is an improvement of 44.6 % for LV, 46.1 % for RV, and 38.1 % for the Myocardium as compared to the previously identified checkpoint. Moreover, as observed in column (b), at this new identified checkpoint, the average Dice scores on the entire validation set also increases by about 1 to 2 % for each of LV, RV, and MYO.

4.5 Benchmarking with various segmentation metrices

The average Dice score has been used as the evaluation metric in our analysis thus far since it is a well-established as well as commonly used metric for evaluating segmentation models. Dice score is defined as twice the area of overlap between the predicted segmentation and the actual labels, divided by the sum of the areas of the predicted segmentation and the ground truth labels, leading to a range between 0 (worst) and 1 (best)(Bertels *et al.*, 2019).

We now analyse if the failure to detect the low performance in corner cases is because of averaging across all patients or is a characteristic of Dice score. For this analysis, we evaluate other metrics for benchmarking segmentation results.

A metric that is closely related to the Dice score is the Jaccard Coefficient. It is also known as the intersection over union and is often used to determine the perfor-

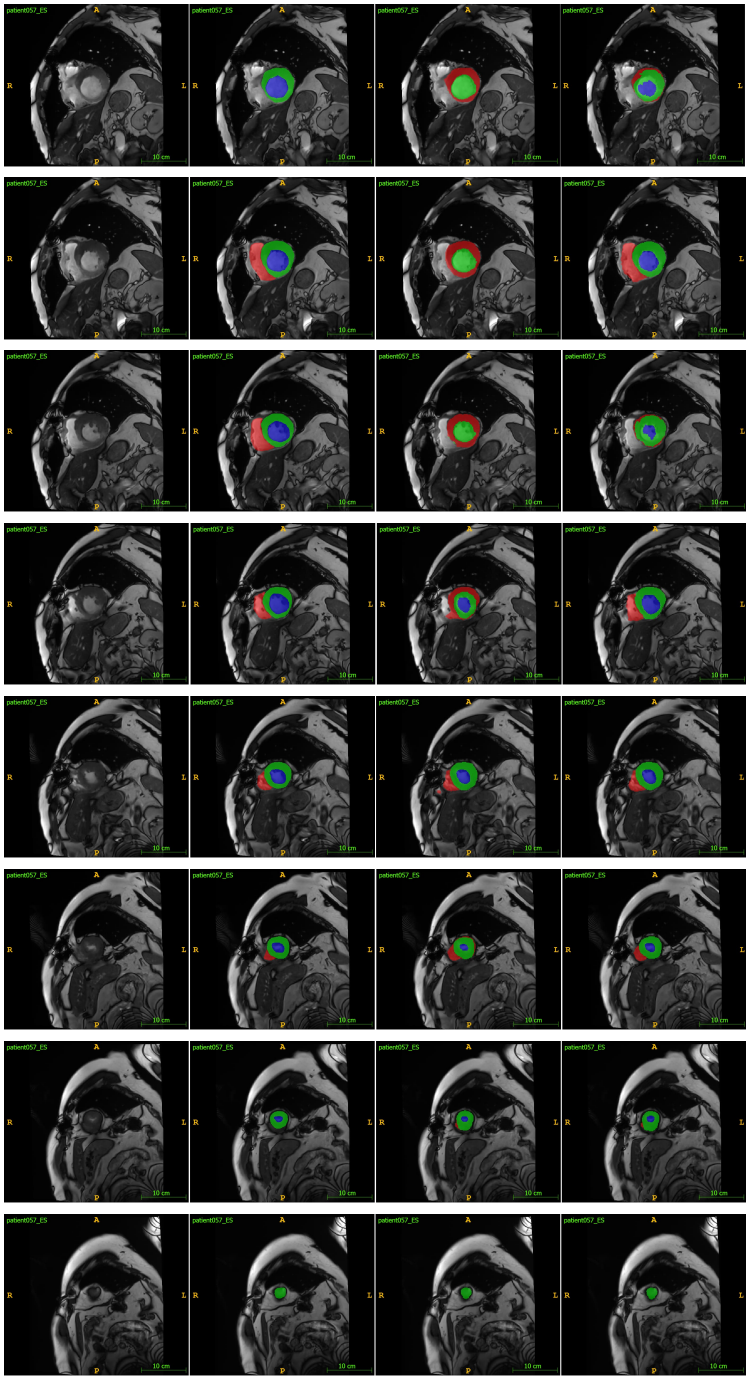

Figure 4.2: The rows contain, from top to bottom, slices 1 to 8 of End Systole frames for Patient057 from ACDC dataset. The columns from left to right are: (a). Original image, (b). Ground truth, (c). Predicted segmentation with the least-loss checkpoint, and (d). Predicted segmentation with the proposed balanced checkpoint, respectively. The colour coding used is blue for LV, green for MYO, and red for RV. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Venkateshvaran, *et al.*, 2023)

(A) Results for entire validation set

Organ	(a). Dice Scores based on Least-loss checkpoint (checkpoint 33)	(b). Dice Scores based on proposed balanced checkpoint (checkpoint 32)	Percentage gain (b - a)
LV	0.912	**0.925**	**1.3**
RV	0.833	**0.856**	**2.3**
MYO	0.848	**0.863**	**1.5**

(B) Results for Patient057

Organ	(c). Dice Scores based on Least-loss checkpoint (checkpoint 33)	(d). Dice Scores based on proposed balanced checkpoint (checkpoint 32)	Percentage gain (d - c)
LV	0.352	**0.798**	**44.6**
RV	0.201	**0.662**	**46.1**
MYO	0.441	**0.822**	**38.1**

Table 4.2: Effect of choosing a balanced checkpoint.
Table (A). Average Dice scores for entire validation set with least-loss checkpoint and proposed balanced checkpoint. Table (B): Dice scores for the corner case, Patient057. Proposed balanced checkpoint significantly improves performance on corner-case (d-c). Furthermore, average Dice scores also improves (b-a)

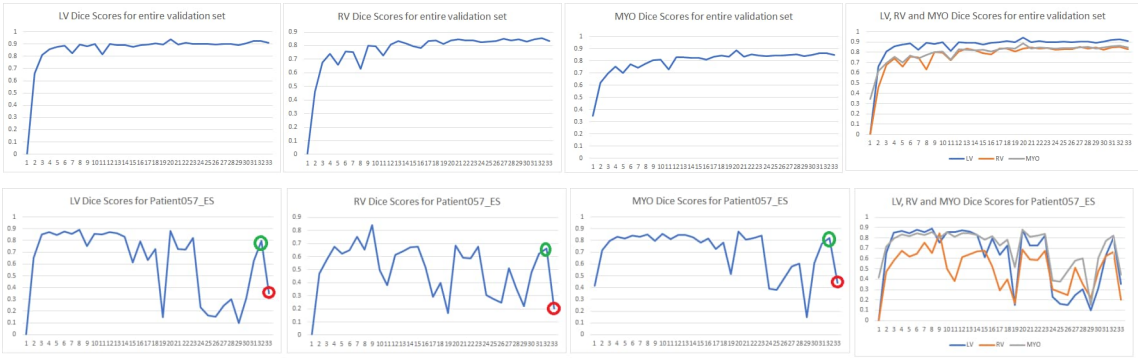

Figure 4.3: Plot of Dice scores at the least-loss based checkpoints over the training epochs. The 4 columns, from left to right, contain Dice scores for (a). LV, (b). RV, (c). MYO, and (d). consolidated-view, respectively. The top row contains the plots of average Dice score for the validation set. The bottom row contains the plots for the corner-case, Patient057_ES (the Dice score at least-loss based checkpoint is marked in red and at the proposed balanced checkpoint marked in green, respectively). Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Venkateshvaran, *et al.*, 2023)

Organ	(a). Avg Jaccard Coefficient	(b). Jaccard Coefficient for corner case Patient057_ES	Jaccard Coefficient difference (a-b)
LV	0.849	0.213	0.636
RV	0.731	0.112	0.619
MYO	0.745	0.283	0.462

Table 4.3: Average Jaccard Coefficient and Jaccard Coefficient for the corner-case identified for LV, RV, and MYO on ACDC validation set

Organ	(a). Avg bAHD	(b). bAHD for corner case Patient057_ES	bAHD difference (b-a)
LV	0.152	1.938	1.786
RV	0.851	24.177	23.326
MYO	0.218	1.689	1.471

Table 4.4: Average bAHD and bAHD for the corner-case identified for LV, RV, and MYO on ACDC validation set

mance of image segmentation algorithms (Long *et al.*, 2015). Similar to average Dice score, it also calculates the ratio of the overlapping regions. However, the Jaccard Coefficient is more sensitive to false positives in contrast to the average Dice score which focuses on balancing precision and recall.

Aydin *et al.* (2021)'s balanced Average Hausdorff Distance (bAHD) is another recent yet popular metric. It is derived from the Hausdorff distance, which calculates the closeness of each point in a segmentation set to the nearest point in the ground truth label set and vice-versa. The balanced Average Hausdorff Distance (bAHD), however, averages these distances. This therefore results in a more robust way to account for outlier points in segmentation tasks. Lower bAHD scores indicate higher segmentation quality.

The average Dice score and Dice scores for the corner-cases are presented in Table 4.1. Similarly, the results evaluated using the Jaccard Coefficient and the balanced Average Hausdorff Distance (bAHD) are presented in Table 4.3 and 4.4, respectively. EvaluateSegmentation tool (Taha *et al.*, 2015) is used to compute these metrics. Patient057_ES is detected as a corner-case when the ECOD algorithm is run on the patient-wise metrices. These results validate that, even with other well established and state-of-the-art metrices, averaging across patients is indeed the major reason for failure to detect the corner cases.

4.6 Generalizability of the proposed framework

The generalisability of our proposed framework is validated in this section by using the chest X-ray lung segmentation task. The NIH chest X-ray dataset (Wang, Xiaosong *et al.*, 2017) contains both posterior-anterior and anterior-posterior views. Tang *et al.* (2019) used 100 abnormal chest X-ray images from this dataset with

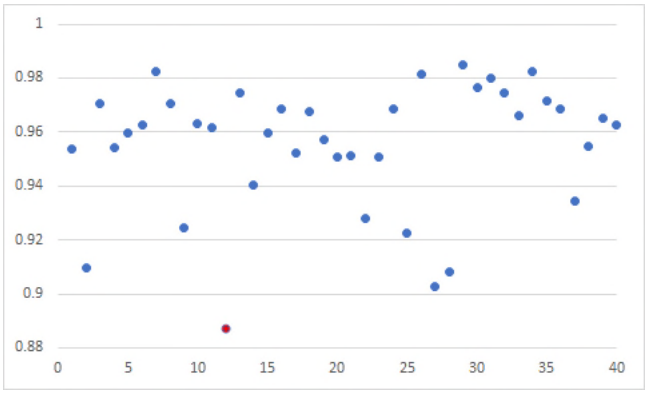

Figure 4.4: Scatter plot of patient-wise Dice scores for the NIH validation set. The outlier Dice score detected with ECOD (which corresponds to patient NIH_0072) is highlighted in red. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Venkateshvaran, *et al.*, 2023)

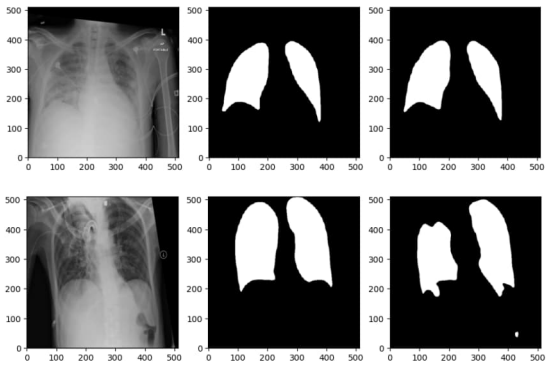

Figure 4.5: Lung segmentation results for couple of images from the NIH dataset. The first row contains results for patient NIH_0090 which is an exemplar patient. The second row contains results for the identified outlier patient NIH_0072. The columns from left to right are (a). Original image, (b). Ground truth and (c). Predicted segmentation. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Venkateshvaran, *et al.*, 2023)

various severity of lung diseases and manually annotated the lung masks[1]. We conduct our experiments on this abnormal chest X-ray dataset.

To conduct our experiments, we utilise a U-Net structure from Oktay *et al.* (2018) and Schlemper *et al.* (2019) as our baseline architecture. It contains four blocks in the downsampling path and similarly four blocks in the upsampling path. In every block, there is batch normalization, 2D convolution and ReLU, repeated twice. The last block is a 1×1 2D convolution block. Max-pooling is used after every block in the down-sampling path so as to reduce the spatial dimension of the feature map by 2 each time. In the up-sampling path, the opposite is done where, using ConvTranspose2d, the spatial dimension of the concatenated feature maps is doubled. The feature channels in the down-sampling path is increased as as $(1 - 64 - 128 - 256 - 512)$ and then decreased similarly in the up-sampling path. The number of label classes for semantic segmentation would determine the number of feature channels in the last layer of the U-Net.

The U-Net's output from the last down-sampling block (which is the reduced dimension of the feature maps $\mathbf{H}$) is given as input to the criss-cross attention module (CCA) (Huang, Zilong *et al.*, 2019). The attention maps need to be small in order to keep its time and space complexity under control. Therefore, the attention module is inserted at the bottleneck of the U-Net. The contextual information in the criss-cross path of each pixel is gathered by criss-cross attention module leading to feature maps $\mathbf{H}'$. After 2 iterations of criss-cross attention, the resulting feature maps are passed through the U-Net's up-sampling path.

Using this model on the validation set of 40 patients of the NIH dataset, an average Dice score of 0.955. Figure 4.4 visualises the scatter plot of the patient-wise Dice scores. By using the ECOD algorithm with the default contamination factor of 0.1, patient NIH_0072 is detected as an outlier. Hence, it is flagged for detailed analysis (marked in red in the scatter plot). In figure 4.5, the visual segmentation result for an exemplar patient, NIH_0090 and for the detected outlier patient NIH_0072, is presented. It is evident from these visual results that the outlier detected by our framework does have sub-optimal segmentation outcomes.

With this, we demonstrate the generalisability of our proposed framework for detecting corner cases across other modalities, anatomies and network architectures.

4.7 Discussion

This section details the clinical insights gained from the corner case that our proposed approach identified on the ACDC cardiac image segmentation dataset. Furthermore, other potential solutions for addressing corner-cases are also outlined. Potential other alternatives for optimal checkpoint determination are also elaborated.

[1]Data: https://nihcc.app.box.com/s/r8kf5xcthjvvvf6r7l1an99e1nj4080m

Organ	(a). Dice Scores based on highest average-IoU checkpoint	(b). Dice Scores based on least-loss checkpoint	Percentage difference (b-a)
LV	0.899	0.912	**1.3**
RV	0.848	0.833	**-1.5**
MYO	0.844	0.848	**0.4**

Table 4.5: Checkpoint based on least-loss vs highest average-IoU on validation set

4.7.1 Clinical insights into identified corner-case

"To understand the observed aberration in the predicted segmentation of Patient057, we obtained clinical insights from an experienced cardiac imaging specialist. Careful inspection of the short axis images from the apex to the base of the LV in addition to the corresponding long axis images revealed prominent anterolateral and posteromedial papillary muscles that are generally underrepresented in the dataset. Further, the segmentation prediction based on least-loss checkpoint inaccurately identified this region of pronounced musculature as myocardium. Current international recommendations advise that papillary muscles are included in the LV cavity, as seen in the ground truth analysis where experts carefully cut through this region during cavity delineation. A plausible explanation for this aberration is the underrepresentation of such variants in the current dataset. This hypothesis, however, requires further investigation in larger databases" (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Venkateshvaran, *et al.*, 2023).

4.7.2 Checkpoint determination using Least-loss vs highest average-IoU

The standard approach to model checkpointing during training is based on least-loss or highest average-IoU. Table 4.5 presents the Dice scores computed based on both of these approaches on the validation set. Using either of these checkpointing approaches yields comparable performance as observed in the 2nd and 3rd column of this table. We utilise the least-loss based checkpoint in this current work.

4.7.3 Other potential approaches for corner-case handling

Subjects/patients could potentially be identified as being corner-cases due to several factors. This could either be due to the characteristics of the data or due to flaws in annotation, or model/network's deficiencies, based on our current insights. However,

an active collaboration between researchers and clinical experts is required for the precise reasons to be identified and for potential mitigation approaches to be derived.

Similarly, there could be various approaches to address such corner-cases. If, for instance, a corner case is due to under-representation of the unique data in the training dataset, it could be addressed through the following ways:

- Using a data approach: We address this in our proposed approach by handling the corner-case separately. However, adding more real or synthetic data with similar characteristics to the dataset could be another approach for addressing this. Excluding such corner cases from the training and validation data and including a disclaimer that the solution cannot be utilised in such outlier scenarios could also potentially be an approach that could be considered. This would provide clinicians with a better understanding of the capabilities and potential pitfalls of the model and complement standardised model reporting (Mitchell *et al.*, 2019).

- Through the model: During the model training, further attributes of the data could be provided as context. In the ACDC challenge dataset, for instance, there are 5 different classes and this class information could be utilised as additional input during the model training.

- Through ground-truth refining: A separate class can be used to mark those regions that confuse the model. The papillary muscles, when prominently visible, could, for instance, be labelled as a separate class.

- Through anomaly classification as a step prior to segmentation: Corner and regular cases could be separated using a standalone classifier before segmentation. However, since the number of corner-cases could be very few, this could be challenging to realise.

4.7.4 Other potential approaches for optimal checkpoint determination

In order to determine a balanced checkpoint such that corner-cases also obtain reasonable results, in our proposed balanced checkpointing approach, we suggest to exclude the outlier epochs and choose the final epoch out of the remaining epochs. However, rather than the global optimum, this approach could result in a local-optimum. In order to find the global optima, several factors play a role such as

- the number of corner-cases.

- the behaviour of the solution over the various training epochs on the corner-cases.

- the behaviour of the solution over the various training epochs on the non-corner cases .

This is therefore a complex multi-factor optimisation problem and is an area of active research (Bertels *et al.*, 2019; Eelbode *et al.*, 2020; Renard *et al.*, 2020).

4.8 Conclusion and future work

In this research work, we have revealed a fundamental but thus far overlooked aspect of deep-learning based segmentation models. Average metrics indicate the model performance on the majority of the cases. However, such approaches tend to overlook the method's performance on the corner-cases. Identifying these corner-cases and handling them is crucial when deploying such solutions in a clinical setup.

We have proposed strategies that enable systematic addressing of these challenges. Firstly, our framework helps to easily identify any corner-cases. Secondly, we have detailed the approaches to get deeper insights into the specific corner cases. Finally, we have outlined an approach to get a balanced model which not only significantly improves the performance on the identified corner-case but also improves the overall average Dice scores.

Utilising our proposed framework for biomedical image analysis tasks other than medical image segmentation, such as medical image classification and object detection is a interesting area to explore. Another potential research area is to determine the balanced checkpoint based on global optima automatically.

5

Attention regularisation

This chapter discusses the research work "A novel and simple approach to regularise attention frameworks and its efficacy in segmentation" published in Proceedings of Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (Rajamani, Srividya Tirunellai, Rajamani, Kumar, and Schuller, 2023)

5.1 Motivation

Attention mechanism is one of the recent research advancement that has demonstrated significant potential in deep learning networks for medical image processing related tasks. Attention mechanism effectively captures long range interactions. Furthermore, in order to make the computation of attention blocks efficient, recent research work like criss-cross attention have been proposed. We propose a simple yet effective enhancement to attention mechanism that involves regularisation of the attention block. A low-overhead way to perform this attention block regularisation that improves the network robustness and resilience is by addition of noise. We incorporate this regularisation in the criss-cross attention block which is integrated in the bottleneck layer of a U-Net. This enhances the performance in medical image segmentation, more so when there is limited training data.

5.2 Novel attention regularisation framework

Our approach involves the simple but novel way to increase a network's robustness through the introduction of regularisation in attention module. In our current work involving semantic image segmentation, we utilise this concept within the criss-cross attention module (Huang, Zilong *et al.*, 2019). We introduce regularisation within this criss-cross attention module. Figure 5.1 depicts a block diagram of our proposed Recurrent Criss-Cross Attention (RCCA) module with regularisation. After each

pass of the recurrent criss-cross attention, regularisation is performed, as shown in the block diagram. Through empirical experiments, we discover this manner of interleaving regularisation and criss-cross attention as the most promising approach.

This regularised recurrent criss cross module is integrated within the U-Net architecture (Schlemper *et al.*, 2019) to evalute its effectiveness for medical image segmentation task. In order to capture non-local contextual information in a robust way, this is included in the U-Net's bottleneck layer.

We now discuss the details of

- The criss-cross attention module (Huang, Zilong *et al.*, 2019)
- The baseline U-Net + CCA architecture used in our experiments
- Our proposed regularised attention module

5.2.1 Criss-Cross Attention Module

Huang, Zilong *et al.* (2019) proposed the criss-cross attention module (CCA) which aggregates contextual information in horizontal and vertical directions for each pixel. Feature maps $\mathbf{H}$ of reduced dimension are computed from the input image by using convolutional neural network (CNN). The CCA module consists of 3 convolutional layers applied on $\mathbf{H} \in \mathbb{R}^{C \times H \times W}$ with 1×1 as kernel size.

The contextual information is aggregated by

$$\mathbf{H'_u} = \sum_{\mathbf{i} \in |\mathbf{\Phi_u}|} \mathbf{A_{i,u}} \mathbf{\Phi_{i,u}} + \mathbf{H_u}, \tag{5.1}$$

with $\mathbf{H'_u}$ being a feature vector in the module's output feature maps $\mathbf{H'} \in \mathbb{R}^{C \times H \times W}$ at position u and $\mathbf{A_{i,u}}$ being a scalar value at channel i and position u in the attention map $\mathbf{A}$. The set $\mathbf{\Phi_u}$ is a collection of feature vectors in the feature map V obtained for feature adaption by applying another convolutional layer with 1×1 filters on $\mathbf{H}$.

5.2.2 Baseline Network Architecture: U-Net + Criss-Cross Attention Module

To conduct our experiments, we utilise a U-Net structure from Oktay *et al.* (2018) and Schlemper *et al.* (2019) as our baseline architecture. It contains four blocks in the downsampling path and similarly four blocks in the upsampling path. In every block, there is batch normalization, 2D convolution and ReLU, repeated twice. Max-pooling is used after every block in the down-sampling path so as to reduce the spatial dimension of the feature map by 2 each time. In the up-sampling path, the opposite is done where, using ConvTranspose2d, the spatial dimension of the

concatenated feature maps is doubled. The feature channels in the down-sampling path is increased as $(1 - 64 - 128 - 256 - 512)$ and then decreased similarly in the up-sampling path. The number of label classes for semantic segmentation would determine the number of feature channels in the last layer of the U-Net.

The local representation feature maps $\mathbf{H}$ is used as the reduced dimension input to the criss-cross module. This is basically the output of last block of the U-Net in the downsampling path. Since these feature maps are of reduced dimension, the attention module is inserted in the bottleneck. This also ensures that the attention maps are smaller, and have lesser complexity in terms of time and space. Huang, Zilong *et al.* (2019)'s CCNet attention module computed feature maps $\mathbf{H}'$ by utilising contextual information in the criss-cross path of each pixel. When using recurrent criss cross attention, $R = 2$ loops through the attention module is done to obtain the contextual features $\mathbf{H}''$. This is then concatenated with the feature maps $\mathbf{X}$ and merged with a convolution layer. The U-Net's upsampling path uses this resulting feature maps as its input.

5.2.3 Our proposed regularised attention sampling

Non-local information on a feature map of height H and width W is computed by the criss-cross attention module. We compute a noise mask which has similar dimension as that of the attention feature map. The noise is randomly sampled from a Gaussian distribution. A grid search methodology was used to empirically determine the mean and variance values for the Gaussian noise that would result in the best outcomes. As seen in Figure 5.1, such a noise mask is added to the attention feature map after each criss-cross attention module.

5.3 Experimental setup and results for medical image segmentation

Method	**Mean Dice Score**	**% gain**
U-Net + RCCA	0.931	-
U-Net + our proposed regularised RCCA	**0.955**	**2.5**

Table 5.1: Mean dice score values averaged over 5 runs

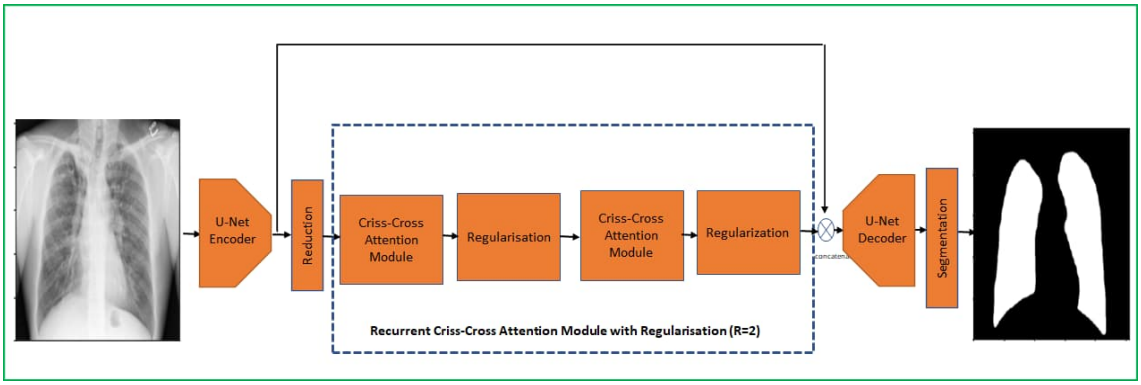

Figure 5.1: The block diagram of our proposed Recurrent Criss-Cross Attention (RCCA) module with regularisation is enclosed in dotted blue line. This proposed module is utilised in the bottleneck layer of the U-Net (enclosed in solid green line) for our experiments on medical image segmentation task. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, and Schuller, 2023)

5.3.1 Dataset

Our experiments are conducted on the NIH chest X-ray dataset (Wang, Xiaosong *et al.*, 2017). This dataset contains posterior-anterior as well as anterior-posterior views. From this dataset, Tang *et al.* (2019) utilised in their work 100 abnormal chest X-ray images with various severity of lung diseases for which the lung masks were manually annotated[1]. We utilise these 100 abnormal chest X-ray images, each of size (512 x 512) in our experiments.

5.3.2 Experiments and Result

Our experiments were done using 60 images for training and the rest for validation and testing and training was done for 60 epochs. This choice for using only 60 % for training was to evaluate the scenario of training segmentation networks with limited training data. In Table 5.1, the mean dice score obtained by averaging over 5 runs are reported. Mean dice score results obtained when using U-Net + vanilla Recurrent Criss-Cross Attention (RCCA) is reported in the first row. In the second row, the mean dice score obtained when using U-Net with our proposed regularised RCCA is reported. The mean and variance for the Gaussian-noise based regularisation is empirically determined in our experiments. Our proposed approach results in an improvement of about 2.5 % over the baseline U-NET + vanilla RCCA with a mean dice score of 0.955.

Figure 5.2 demonstrates that our proposed regularised attention block estimates the lung masks much closer to the ground truth.

[1]Data: https://nihcc.app.box.com/s/r8kf5xcthjvvvf6r7l1an99e1nj4080m

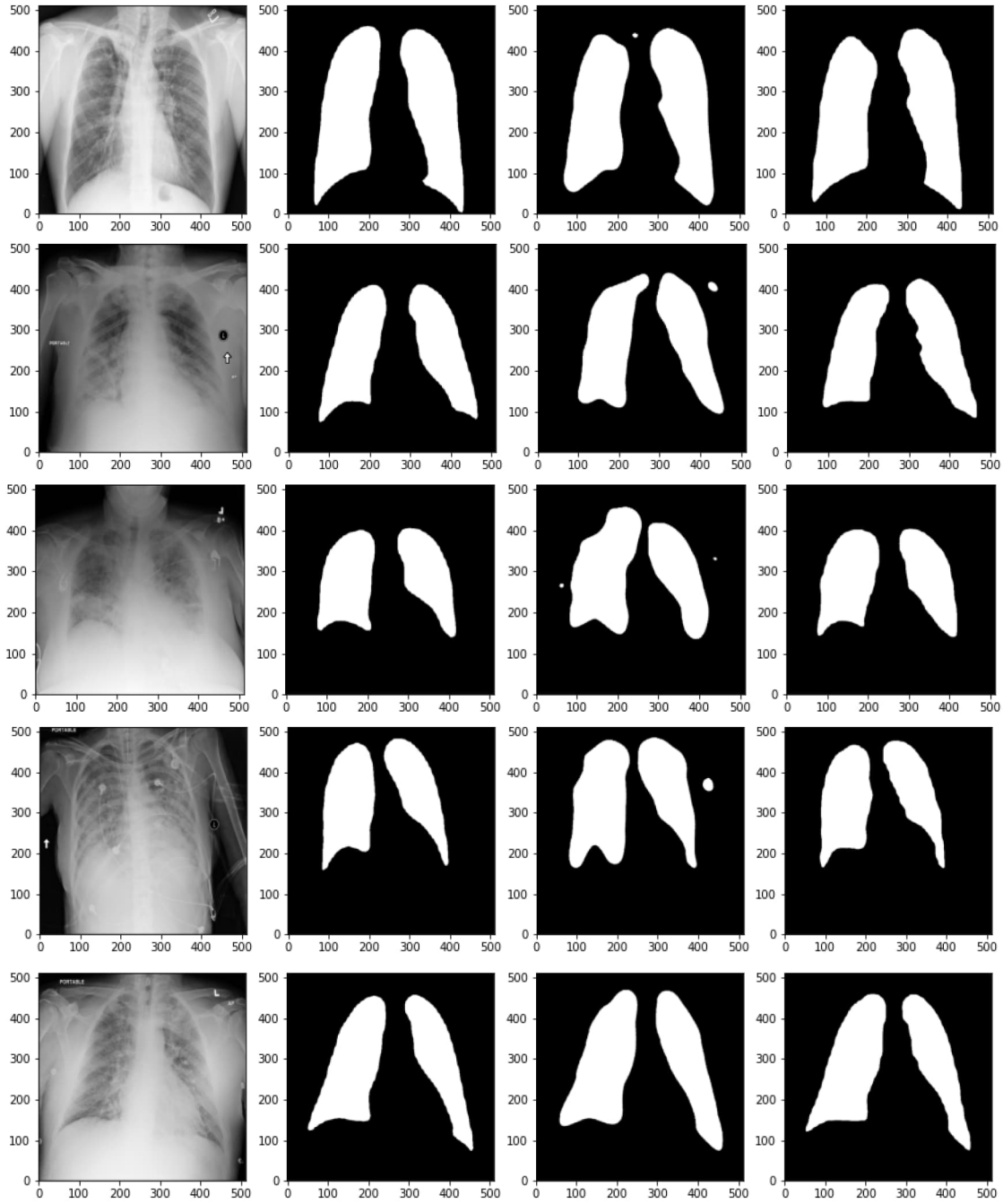

Figure 5.2: Visual comparison of lung segmentation results. In each of the 5 rows, results for different test images are shown. The 4 columns contain, from left to right, (a). the input image, (b). the ground-truth segmentation mask, (c). the segmentation mask predicted with U-Net + the vanilla RCCA, and (d). the segmentation mask predicted with U-Net + the regularised RCCA, respectively. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, and Schuller, 2023)

5.4 Conclusion

We proposed a novel regularisation of the attention mechanism using additive Gaussian noise. Incorporating this in the chosen U-Net + RCCA framework improves the segmentation of lung lobes. Through our experiments, we demonstrate that segmentation network become robust and also generate better segmentation results as compared to the baseline when regularisation is added in a CCNet. Our regularised attention enables the network to segment the objects of interest with an improved dice score.

One of the areas of future work is to utilise this regularised attention mechanism in attention-based classification tasks. Furthermore, other approaches to attention regularisation as well as changes to noise distribution is an interesting area for futher exploration.

6

Deformable attention

This chapter discusses the research work "Deformable Attention (DANet) for Semantic Image Segmentation" published in Proceedings of Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (Rajamani, Kumar, Gowda, *et al.*, 2022)

6.1 Motivation

One of the widely researched topic in deep learning is medical image segmentation. Deep learning methods using attention mechanism have been demonstrated to improve the performance of semantic segmentation tasks. Criss-cross-attention module (Huang, Zilong *et al.*, 2019) is one such recent attention based methods that captures global self-attention and is memory and time efficient as well. However, the accuracy of semantic segmentation networks could be further improved if attention is captured from only the relevant non-local locations. Our novel Deformable Attention Network (DANet) computes contextual information in a more accurate yet efficient way. In DANet, the query, key and value attention feature maps are deformed and the deformation is learnt in a continous manner. Using such a deformable attention mechanism, a deep segmentation model can capture attention from important non-local locations. As demonstrated by our experiments, by recursively applying deformable attention blocks within a U-Net, the network is able to perform better owing to its ability to capture dynamic and precise attention context.

6.2 Novel Deformable Attention Network (DANet)

Self attention mechanisms have shown great success in recent times. Furthermore, with their sparse deformable convolutions, Heinrich *et al.* (2019) demonstrated that

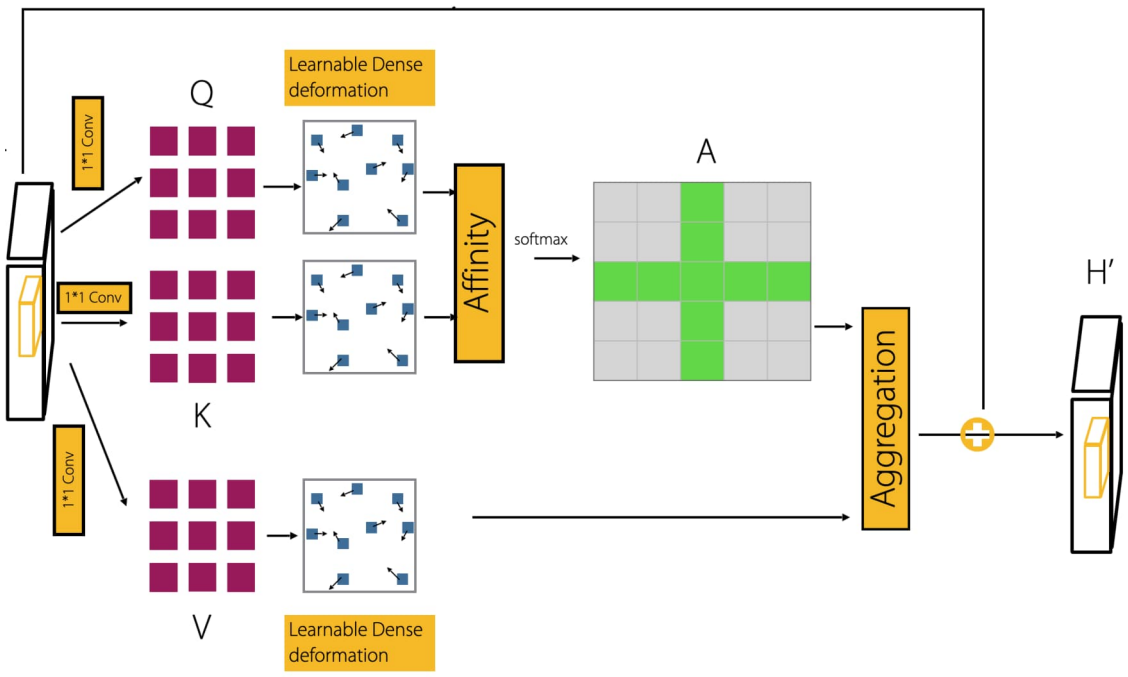

Figure 6.1: A block diagram of the proposed deformable criss-cross attention module. In our deformable attention, we have deformed the query, key and value attention feature maps. Differentiable bi-linear interpolation is used for deformation. Figure taken from (Rajamani, Kumar, Gowda, *et al.*, 2022)

attention computing blocks viz. query, key and value feature maps do not have to be regularly structured. Motivated by these research and improvising over criss-cross attention (Huang, Zilong *et al.*, 2019), we explore spatially-adaptive query, key and value feature maps. We propose a novel Deformable Attention Network (DANet) for medical image segmentation which deforms the query, key and value feature maps in a continuous space. Figure 6.1 contains a block diagram of our proposed deformable criss-cross attention module. We demonstrate that using this deformable criss-cross attention module within our custom baseline U-Net architecture improves the segmentation performance. Our proposed deformable criss-cross attention module can be easily utilized within any state-of-art segmentation network.

We perform experiments by utilising our proposed deformable criss-cross attention module within a custom U-Net consisting of three blocks each in the down-sampling and up-sampling path. In every block, batch normalization followed by 2D convolution and ReLU was done twice. In the down-sampling path, the spatial dimension of the feature maps is reduced by 2 after every block using max-pooling. The reverse is done in the up-sampling path where the spatial dimension of the concatenated feature maps is doubled using ConvTranspose2d. The feature channels

in the down-sampling path is increased from 1 to 512 and then decreased similarly in the up-sampling path. The number of feature channels in the last layer of the U-Net is set to match the number of label classes for semantic segmentation.

The U-Net's output from the last down-sampling block (which is the reduced dimension of the feature maps **H**) is given as input to the criss-cross attention module (CCA) (Huang, Zilong *et al.*, 2019). The attention maps need to be small in order to keep its time and space complexity under control. Therefore, the attention module is inserted at the bottleneck of the U-Net. The contextual information in the criss-cross path of each pixel is gathered by the attention module in the original CCNet (Huang, Zilong *et al.*, 2019). In our proposed deformable attention network, first, a learnable deformation is used to deform the query, key and value feature maps. Then, using these deformed attention feature maps, a regular criss-cross attention is performed. The U-Net's up-sampling path takes these resulting feature maps as input.

6.3 Experimental setup and results for medical image segmentation

We evaluate the performance of our proposed DANet on Chest CT images for COVID-19 lesion segmentation. In cross-sectional images of CT scans, the usual signs indicating primitive stages of COVID-19 is ground glass opacities (GGO). Clinicians are able to effectively treat COVID-19 if these regions can be detected in cross-sectional images of CT scans. However, automating this detection is important since detecting is manually is highly time consuming as well as error-prone. Figure 6.2 shows few slices to provide a visual example of how ground-glass opacity lesions and consolidation lesions can be distinguished in the images.

We conduct our experiments using 2 public COVID-19 CT segmentation datasets. The first one consists of 100 axial CT images from different COVID-19 patients (MedicalSegmentation.com, n.d.). This data collection is from the Italian Society of Medical and Interventional Radiology. The second dataset comprises of axial volumetric CTs of 9 patients from Radiopaedia. This dataset comprises of whole volumes having both positive (373 positive) and negative slices (455 negative slices).

On this combined dataset consisting of 471 two-dimensional axial lung CT images and their segmentations for ground glass opacities (GGO) and consolidation lesions, we conduct our experiments with 3-fold cross validation. Data from three different patients plus one third of images from the 100 slice CT stack taken from more than 40 different patients is contained in each fold. The CT images are cropped and rescaled to 256×256 size. Random affine deformations are used to augment the data during training.

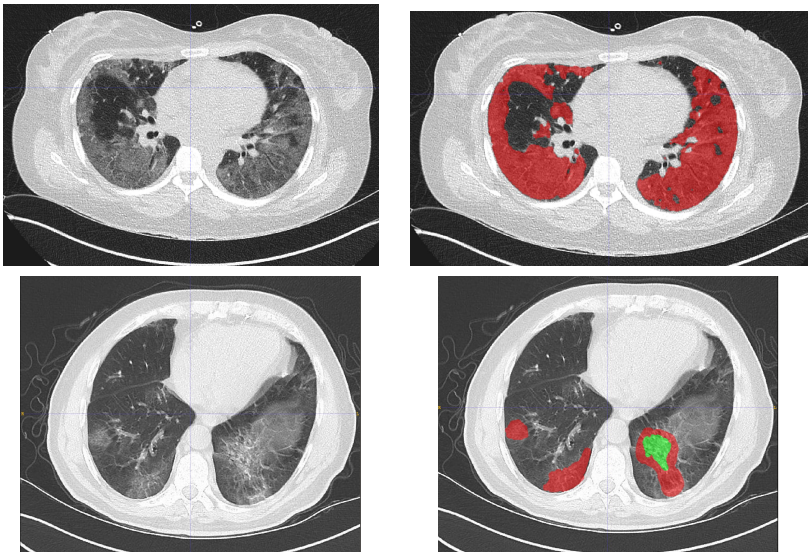

Figure 6.2: The first row contains a sample slice from one of the dataset and its corresponding ground-glass opacity lesion (GGO) marking. The second row contains another sample slice and its corresponding GGO and consolidation lesion marking. Dataset from website (MedicalSegmentation.com, n.d.). Figure taken from (Rajamani, Kumar, Gowda, *et al.*, 2022)

Training is performed for 500 epochs. Adam optimizer and an initial learning rate of 0.002 is used. Furthermore, a cyclic learning rate with an upper boundary of 0.005 is used. Class-weighted cross-entropy loss is used to address the problem of training from imbalanced data .

We compared our model with two cutting-edge models, namely U-Net and criss-cross attention (Huang, Zilong *et al.*, 2019) for the infection region experiments and multi-class labeling. The number of trainable parameters for the U-Net is 611K. For the U-Net enhanced with criss-cross attention, the parameter count is 847K. Our proposed variant using deformable CCNet has slightly more parameters of 849K. The dice score of ground-glass opacities and consolidation as well as the number of parameters is presented in Table 6.1.

Our proposed DANet method achieves the best competitive Dice score of **0.66** for GGO lesion and **0.55** for consolidation lesion averaged across all the patients. It outperforms the baseline U-Net model's Dice score by **4.4**% on multi-label segmentation.

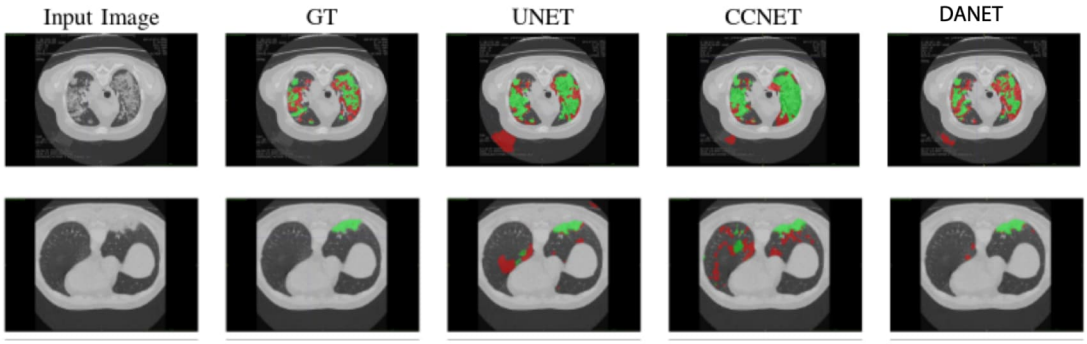

Figure 6.3: Visual comparison of multi-class lung segmentation results. The red labels indicate GGO and green labels indicate consolidation. Figure taken from (Rajamani, Kumar, Gowda, *et al.*, 2022)

Table 6.1: Quantitative results (Dice score) of Ground-Glass opacities and consolidation. The results are averaged across multiple folds and multiple runs. The best results are shown in Blue font.

Model	GGO	Consol.	Avg	%Gain	#Params
UNet	0.63	0.52	0.58	-	611.7 K
UNET+CCA	0.65	0.52	0.59	2.5	847.3K
DANet	**0.66**	**0.55**	**0.60**	**4.4**	849.7K

6.4 Conclusion

We have proposed a novel enhancement to the criss-cross attention module using deformable attention maps (DANet). When incorporated into the U-Net framework, our proposed novel DANet considerably improves COVID-19 CT lesion segmentation.

In this work, the query, key as well as value feature maps have been used to learn the deformation of the attention maps. An ablation study to analyse these deformations independently and understand how much they contribute individually to the performance improvement would be an interesting future work to pursue to gain further insights.

7

Novel metric for image quality assessment

This chapter discusses the research work "Novel No-Reference Multi-Dimensional Perceptual Similarity Metric" published in Proceedings of Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Rani, *et al.*, 2022)

7.1 Motivation

In healthcare applications, one of the important aspects is the acquisition of high quality images. The process of acquisition of medical data using Computed Tomography (CT), Magnetic Resonance (MR), Ultrasound has seen great strides in automation. Integration of automated image quality measures into the acquisition workflow yields significant benefits. However, in the image acquisition workflow, it is difficult to determine if changes to the acquisition parameters are aiding in the improvement of image quality or not. For instance, in medical image acquisition using MR, there are several hundred acquisition parameters that could be potentially fine-tuned to get the ideal or best quality image and hence the task of automatic measurement of image quality is challenging.

Some of the commonly used approaches for image quality assessment are Structural Similarity (SSIM) index (Wang, Zhou, Bovik, *et al.*, 2004), Multiscale Structural Similarity (MSSIM) index (Wang, Z., Simoncelli, *et al.*, 2003), Functional Similarity FSIM index (Zhang, L. *et al.*, 2011), or HDR-VDP (Mantiuk *et al.*, 2011). Perceptual metric based on deep learning based approach (Zhang, R *et al.*, 2018) have also been recently proposed. However, in the context of live acquisition of images where the quality would be continuously changing, these approaches are insufficient to determine the quality of images. For instance, in MR image acquisition, in order to to achieve optimal image quality, several acquisition parameters

like image width, depth, image contrast and artifact suppression needs to be optimised. Therefore, the need is to determine if the quality of the image is improving or not when each of these settings are modified in a automatic and quantitative manner. However, in such live acquisition scenarios, there is no reference pristine image quality image that is available.

Having a continuous measure of quality when the acquisition parameters are modified on each of the axis like noise, blur, contrast would be extremely useful for the medical imaging community. The design and deployment of artificial intelligence based algorithms could also leverage such a measure towards the aim of automating the acquisition process. But it is challenging to arrive at a multi-dimensional quality metric which is continuous and also representative.

We propose a novel multi-dimensional no-reference perceptual similarity metric. By combining no-reference image quality metric (PIQUE) and perceptual similarity, this metric can compute the quality of a given image without a reference pristine quality image. The axis of noise, blur and contrast are the dimensions of quality that we have currently taken into consideration. The correlation of our proposed metric with quality of an image in a multi-dimensional sense is demonstrated by our experiments.

7.2 Novel No-Reference Perceptual Similarity Metric

Our proposed method is based on a combination of full-reference image quality assessment (FR-IQA) and no-reference image quality assessment (NR-IQA). During the acquisition or pre-processing process, the quality of a given image at that instance is determined through the no-reference image quality assessment. When a particular acquisition parameter is modified, if there is an improvement in the no-reference image quality metric, then this is considered to indicate that the original image was degraded due to that particular aspect, in the first place. A distortion correction technique is applied when a degradation in a particular dimension/axis is discovered to obtain a best-possible pristine image. For the full reference image quality assessment, this pristine image is used as the good quality reference.

The no-reference perception-based image quality evaluator (PIQUE) is used as our NR-IQA in this work. Learnt Perceptual Image Patch Similarity (LPIPS) (Zhang, R *et al.*, 2018) is used for perceptual similarity. However, our proposed concept could also be used with any other NR-IQA or FR-IQA. Furthermore, it can be easily extended to diverse types of distortions other than noise, blur, or contrast as introduced in TID2013 (Ponomarenko *et al.*, 2015) or the BAPPS dataset (Zhang, R *et al.*, 2018).

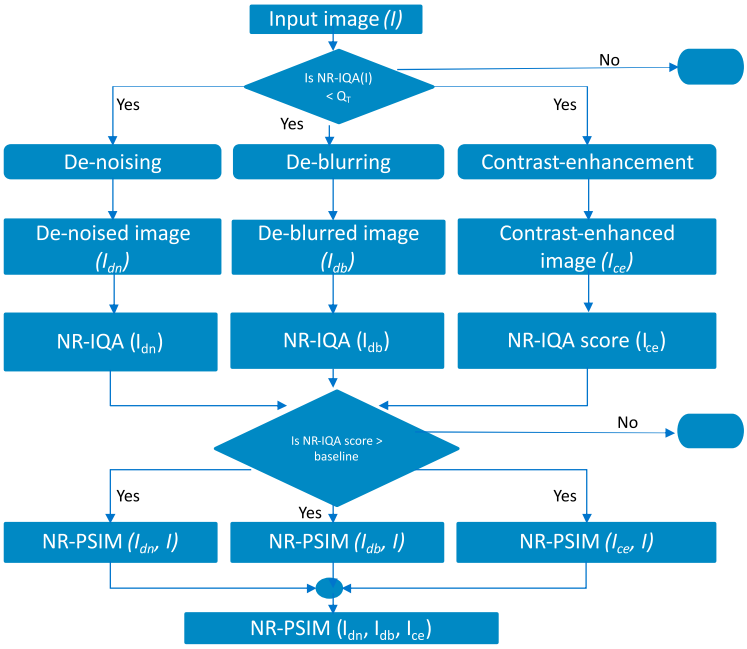

Figure 7.1: Schematic for computation of proposed novel NR-PSIM score. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Rani, *et al.*, 2022)

Figure 7.1 contains the schema of our proposed NR-PSIM. First, a no-reference image quality metric is computed for a given input image. Based on a pre-defined threshold Q_t, if the score indicates poor image quality the image is restored through techniques like de-noising, de-blurring, contrast enhancement, etc. The threshold itself is empirically determined based on the problem domain. Next, we check if this results in an improvement compared to the original image's baseline score, by computing the NR-IQA of the restored image. The restored image is used as the pristine image quality for our NR-PSIM computation, if the score indicates that the restored image's quality has improved. By using the restored image as reference and the baseline image as the distorted image, our proposed NR-PSIM is computed as full-reference image quality metric (FR-IQA). In this work, we use PIQUE as the NR-IQA and LPIPS as the FR-IQA. For each of the potential degradation scenarios like noise, blur, contrast, the metric of the restored image is computed and then aggregated to arrive at the final multi-dimensional NR-PSIM score.

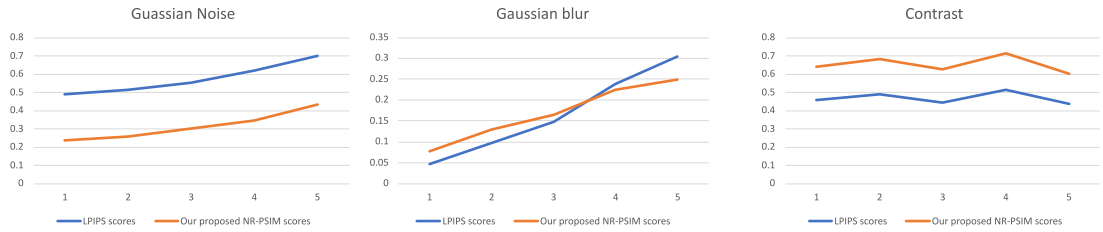

Figure 7.2: Graph of the LPIPS scores (pristine, distorted) vs our proposed NR-PISM scores. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Rani, *et al.*, 2022)

7.3 Experimental setup and results on generic and medical images

7.3.1 Datasets

- **Tampere Image Database 2013 (TID2013)** The Tampere Image Database 2013 (TID2013) is a reference database for full-reference image visual quality assessment metrics evaluation. It has 25 input images with 24 distortions types, each sampled at 5 levels, resulting in 3000 distortions in all. It has 500k judgments on these 3000 distortions. In this work, for our experiments, we utilise all the 5 levels of distortions for 3 distortion types, namely, additive Gaussian noise, Gaussian blur and contrast change for one of the reference image, *i21.bmp*.

- **MR dataset** MR Brain images[1] is also used for conducting our experiments. We perform similar experiments on this dataset by inducing different levels of distortion using Gaussian noise and Gaussian blur. This is mainly because there are no reference datasets similar to TID2013 that is available for medical images to benchmark image quality.

7.3.2 Results and discussion

We conduct experiments with additive Gaussian noise, Gaussian blur as well as contrast change. To recover Gaussian noise, we use median filter. Blind de-convolution is used to recover Gaussian blur. Flat-field correction method is used to recover contrast. Table 7.1 presents our proposed NR-PSIM scores in the context of additive Gaussian noise. For the given input image, distortion level 1 having a NR-PSIM

[1]https://www.kaggle.com/sartajbhuvaji/brain-tumor-classification-mri

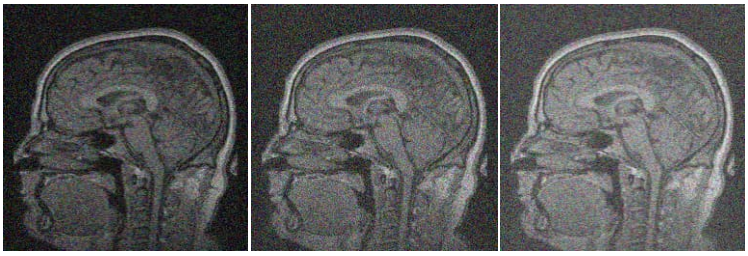

Figure 7.3: High-quality, medium-quality and low-quality MR-Brain images distorted with Gaussian noise and the corresponding NR-PSIM scores. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Rani, *et al.*, 2022)

Table 7.1: PIQUE and NR-PSIM score for TID2013 image: 5 levels of distortion with additive Gaussian noise; de-noising done with median filter

Distortion level	**PIQUE score: distorted image**	**PIQUE score: de-noised image**	**Proposed NR-PSIM score**
1	33.5712	25.2716	0.238
2	40.7428	21.8926	0.259
3	48.3527	16.2969	0.299
4	55.6393	12.9928	0.346
5	61.1238	10.0932	0.432

score of 0.238 and level 5 having a score of 0.432 demonstrating the good correlation between the computed NR-PSIM scores and the level of distortion in the image.

Results of our Gaussian blur experiments is presented in Table 7.2. Results of our contrast correction is detailed in Table 7.3. As can be seen from the PIQUE scores of the distorted images, the 5 levels of distortion in contrast are not in progressive levels of distortion. Our NR-PSIM scores also reflect the same. Using the pristine image as reference, we also compute LPIPS score to demonstrate the similarity in trend of our NR-PSIM scores. The joint plot of LPIPS and NR-PSIM scores for each of noise, blur and contrast depicted in Figure 7.2 demonstrates this.

The results of our Gaussian blur experiments on MR brain image are presented in table 7.4. The correlation between the distortion (blur) levels in the input image and the NR-PSIM scores can be observed here as well. Figure 7.3 depicts the computed NR-PSIM scores for high-quality, medium-quality and low-quality MR brain image distorted with Gaussian noise.

Table 7.2: PIQUE and NR-PSIM score for TID2013 image: 5 levels of distortion with Gaussian blur; de-blurring done with blind de-convolution

Distortion level	**PIQUE score: distorted image**	**PIQUE score: de-blurred image**	**Proposed NR-PSIM score**
1	24.4965	21.4093	0.040
2	35.0990	26.9279	0.028
3	55.1862	51.5675	0.009
4	83.6344	83.1691	0.002
5	92.2498	83.8991	0.002

Table 7.3: PIQUE and NR-PSIM score for TID2013 image: 5 levels of distortion with contrast change; contrast correction done with flat-field correction

Distortion level	**PIQUE score: distorted image**	**PIQUE score: contrast-corrected image**	**Proposed NR-PSIM score**
1	19.8784	19.1912	0.644
2	21.0720	20.9462	0.683
3	20.4942	20.3195	0.626
4	27.1940	25.8898	0.714
5	19.3961	19.3659	0.605

7.4 Conclusion

In this work, we propose a novel no-reference multi-dimensional similarity metric, NR-PSIM. The effectiveness of our NR-PSIM scores against FR-IQA (LPIPS) is demonstrated. Though currently demonstrated in a limited context, our results are very promising. Diverse types of distortions as well as combinations of several distortions could be further experiments to explore.

Currently there are no reference datasets for benchmarking image quality in the context of medical imaging. It would be immensely beneficial to create a large-scale reference dataset like TID2013 or BAPPS in the context of medical imaging to advance image quality related research, especially to improve automated image acquisition.

Table 7.4: PIQUE and NR-PSIM score for MR brain image: 5 levels of distortion with Gaussian blur; de-blurring done with blind de-convolution

Distortion level	**PIQUE score: distorted image**	**PIQUE score: de-blurred image**	**Proposed NR-PSIM score**
1	78.7308	33.1687	0.079
2	80.9443	39.5337	0.129
3	84.6468	42.0443	0.166
4	90.0815	54.1547	0.226
5	91.3957	67.8806	0.248

8

Sparse data and attention networks

This chapter discusses the research work "Novel Insights of Induced Sparsity on Multi-Time Attention Networks" published in Proceedings of Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Kathan, *et al.*, 2022)

8.1 Motivation

Dealing with sparse irregularly sampled time-series data is an area of active research. Current deep learning approaches that deal with such data have not explored the extent to which the input data can be sparse. Our work is based on physiological time series data in electronic health records that by the its very nature is sparse and irregularly sampled.

Multi-Time Attention Networks (mTAN) is a recent attention work proposed by Shukla *et al.* (2021) that is capable of handling sparse irregularly sampled time-series data. We induce varying degrees of sparsity and study its effect on the predictive performance of Multi-Time Attention Networks (mTAN). This is done by sub-sampling the time-series before it is input to the mTAN network. The sub-sampling is done by using a range of 10 to 90 % while performing our empirical experiments.

The 2 datasets that we use for our experiments are the Human Activity dataset and Physionet 2012 mortality prediction dataset. On the Human Activity dataset, with our proposed time-point sub-sampling coupled with mTAN, even with 80 % lesser time-points for training, the performance is still improved by 2 %. On the Physionet dataset, even with 30 % lesser time-points, our approach achieves comparable performance as the baseline. Thus, when used in tandem with state-of-the-art networks capable of handling sparse data like mTAN, we demonstrate that time-series data could be further coarsely acquired. various applications where data

acquisition and labeling is a significant challenge would significantly benefit from these insights.

8.2 Methodology

In this section, the recently proposed mTAN network (Shukla *et al.*, 2021) and our proposed methodology are presented. Sub-sampling a sparse and irregularly sampled time-series before it is provided as input to the mTAN module is our proposed technique. In order to obtain maximum performance from the subsequent deep learning inference stages, the input must be sub-sampled to the right extent. We study and discover this through our empirical experiments.

8.2.1 Multi-Time Attention Network

Shukla *et al.* (2021) recently proposed a Multi-Time Attention Module (mTAN) which is capable of transforming sparse and irregularly sampled time series into a fixed dimensional space. It produces a fixed dimensional representation at the query time points by taking irregularly sampled time points and corresponding values as keys and values. Multiple continuous-time embeddings and attention-based interpolation are used to realise this.

A query time point t and a set of keys and values in the form of a D-dimensional multivariate sparse and irregularly sampled time series s is provided as input to the multi-time attention embedding module $mTAN(t, s)$. By leveraging a continuous-time attention mechanism applied to the H time embeddings, it returns a J-dimensional embedding at time t. This is formulated as follows:

$$mTAN(t, s)[j] = \sum_{h=1}^{H} \sum_{d=1}^{D} \hat{x}_{hd}(t, s) \cdot U_{hdj}, \tag{8.1}$$

where univariate continous-time functions is represented by $\hat{x}_{hd}$ and the learnable weights are represented by U_{hdj}. Further details can be found in Shukla *et al.* (2021).

Since this module defines a continuous function, in neural network architectures that expect fixed dimensional vectors as input, that this module cannot be directly used. Shukla *et al.* (2021) therefore propose the discretised mTAN module (mTAND) to addresses this aspect. In any deep neural network layer that has convolution and recurrent layers, the discretised mTAN module (mTAND) can be used to input sparse and irregularly sampled multivariate time series data. Our proposed technique of sub-sampling is incorporated into a temporal encoder-decoder architecture that leverages the mTAND module.

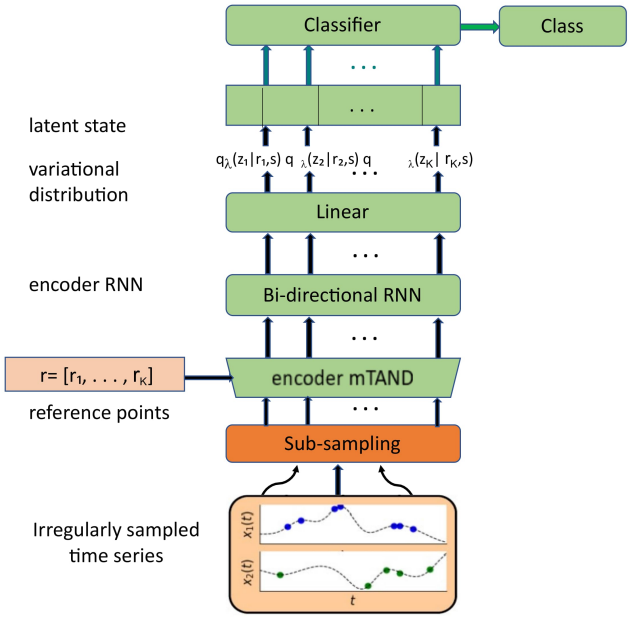

Figure 8.1: Our proposed inference network with time-series sub-sampling of input passed to the mTAND-Enc network. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Kathan, *et al.*, 2022)

8.2.2 Novel insights into attention networks and data sparsity

For the task of time-series interpolation and classification, Shukla *et al.* (2021) utilised their mTAND module within a temporal encoder (mTAND-Enc) to effectively represent sparse and irregular sampled time series. What has not yet been explored is the effect of the inherent redundancy in the input data. the predictive performance of mTAND-Enc network when the input is having varying degrees of data sparsity is what we explore in this work. As depicted in Figure 8.1, we perform this analysis by introducing a pre-processing step of sub-sampling the time-series training data in order to induce sparsity. Random locations are chosen from among those time-points that have measurements for this sub-sampling. The data dimensions provided to the network needs to be retained. Hence, the remaining time-points are retained but set to 0 instead of being excluded. Varying levels of such induced sparsity ranging from 10 to 90 percent is used in our empirical experiments.

8.3 Experimental setup and results on sparse time-series data

The results of the classification experiments on two real-world data sets, Human Activity and the Physionet Challenge 2012 datasets, by using our proposed approach is presented in this section.

8.3.1 Datasets

- Human Activity dataset

  5 activity sequences each from 5 individuals is contained in this dataset [1]. While they performed various activities like walking, sitting, lying, standing, etc, the 3D positions of the waist, chest and ankles were collected. The same data preprocessing steps as mentioned by Rubanova *et al.* (2019) is followed in our experiments. 6554 sequences with 12 channels and 50 time points is used to construct the dataset. We use seven activity types "falling", "lying", "on all fours", "sitting", "sitting on the ground","standing up" and "walking". This is done by combining classes out of the 11 original classes that correspond to very similar activities and hence hard to distinguish. Each time point in the sequence needs to be classified into one of these seven classes.

- Physionet Challenge 2012 dataset

  Multivariate time series data extracted from intensive care unit (ICU) records is contained in this dataset (Silva *et al.*, 2012). Sparse and irregularly spaced measurements from the first 48 hours after admission to ICU is contained in each record. To pre-process the data, we follow the same procedure as that of Rubanova *et al.* (2019). There are 2880 possible measurement times per time series by rounding the observation times to the nearest minute. In order to predict the in-hospital mortality, we use the 4000 labelled instances of this dataset for classification experiments. The dataset is randomly divided into a training set containing 80 % of the time series and a test set containing the remaining 20 % of instances. 20 % of the training set is used for validation. Using different random seeds to initialise the model, each experiment is repeated five times and the mean and standard deviation over these 5 runs is reported.

[1] https://archive.ics.uci.edu/ml/datasets/Localization+Data+for+Person+Activity

Table 8.1: Classification Performance on Human Activity dataset and % Reduction in time-points for training.

Model	**% Reduction**	**Accuracy**
mTAND-Enc (Shukla *et al.*, 2021)	-	0.907 ± 0.002
10 % time-point sampled data	90	0.915 ± 0.004
20 % time-point sampled data	**80**	**0.927 ± 0.002**
30 % time-point sampled data	70	0.926 ± 0.003
40 % time-point sampled data	60	0.924 ± 0.002
50 % time-point sampled data	50	0.922 ± 0.003
60 % time-point sampled data	40	0.921± 0.001
70 % time-point sampled data	30	0.918 ± 0.004
80 % time-point sampled data	20	0.914 ± 0.006
90 % time-point sampled data	10	0.908 ± 0.006

8.3.2 Results and discussion

The classification performance on the Human Activity dataset is summarised in Table 8.1. The classification performance on the Physionet Challenge 2012 dataset is presented in Table 8.2. In each table, the results of the baseline mTAND-Enc network is reported in the first row. The results of our proposed approach of sub-sampling the time-series before feeding it to the mTAND-Enc network is presented in the remaining rows using various sub-sampling rates ranging from 10 to 90 %. For the Human Activity dataset, we report the classification accuracy in accordance with previous works. Due to the inherent class imbalance in the dataset, for the Physionet dataset, we report the Receiver Operating Curve (ROC) – Area Under Curve (AUC).

The best performance for the Human Activity dataset is observed with 80 % lesser time-points, where the accuracy is boosted by 2 %. Performance comparable to the baseline is obtained with 30 % lesser time-points for the Physionet dataset using our proposed sub-sampling approach.

Every data point is vital in healthcare outcome prediction. Our novel discovery is that even in such mission critical applications, one could obtain comparable performance even with 30 % lesser time-points in the data. This opens up the possibility to acquire data in a more efficient and sparse manner which could potentially be a game changer for future work in this space. To get the best performance from time-series based deep learning models, already acquired data could also be aptly sub-sampled.

Table 8.2: Classification Performance on Physionet dataset and % Reduction in time-points for training.

Model	**% Reduction**	**AUC**
mTAND-Enc (Shukla *et al.*, 2021)	-	0.854 ± 0.001
10 % time-point sampled data	90	0.823 ± 0.005
20 % time-point sampled data	80	0.843 ± 0.006
30 % time-point sampled data	70	0.845 ± 0.002
40 % time-point sampled data	60	0.841 ± 0.007
50 % time-point sampled data	50	0.848 ± 0.006
60 % time-point sampled data	40	0.847 ± 0.004
70 % time-point sampled data	**30**	**0.854 ± 0.001**
80 % time-point sampled data	20	0.852 ± 0.003
90 % time-point sampled data	10	0.854 ± 0.001

8.4 Conclusion

The effectiveness of inducing sparsity in time-series data is demonstrated for the task of classification on two different datasets when used in combination with the recent mTAN network. Time-series data can therefore by coarsely acquired. In scenarios where data acquisition and labelling is a major challenge, this would be of great benefit. As demonstrated in our experimental results on the Human Activity and Physionet datasets, our approach could be utilised to further sub-sample a sparse and irregular time-series data before interpolating and classifying using an mTAND-like module in scenarios where the data is already acquired.

The level of sub-sampling for each classification task is currently determined empirically. One of the directions for future research is to explore more sophisticated sub-sampling techniques such as wavelet-based schemes. Learning the sparsity level in a dynamic and task specific way is yet another interesting area to explore (Huijben *et al.*, 2019; Van Gorp *et al.*, 2021).

9

Model complexity reduction using attention

This chapter discusses the research work "Towards an Efficient Deep Learning Model for Emotion and Theme Recognition in Music" published in Proceedings of the Annual IEEE International Workshop on Multimedia Signal Processing (MMSP) (Rajamani, Srividya Tirunellai, Rajamani, Kumar, and Schuller, 2021).

9.1 Motivation

Developing deep learning models that are efficient and can be deployed to resource constrained hardware is an area of active research. Towards the aim of reducing the number of floating point operations per second (FLOPS) and model parameters through optimum network configurations, we propose a novel integration of stand-alone self-attention into a Visual Geometry Group (VGG)-like network. Detecting emotion and theme in music is one of the important aspects in music information retrieval and recommendation systems and deep learning based techniques have demonstrated great potential in this regard. We demonstrate the effectiveness of our proposed self-attention based VGG-like network (SA-VGG) for multi-label emotion and theme recognition in music.

9.2 Baseline architecture

Bogdanov, Porter, *et al.* (2020) demonstrated that a VGG-like architecture is well-suited for the task of music tagging. This network had five 2D convolutional layers with a kernel size of 3×3. Each convolution layer was followed by max-pooling layer. A fully connected layer was used as the final layer. A log-amplitude mel-spectrogram is used as to this network. In every convolutional layer except the output layer, Exponential Linear Unit (ELU) is used as the activation function. In

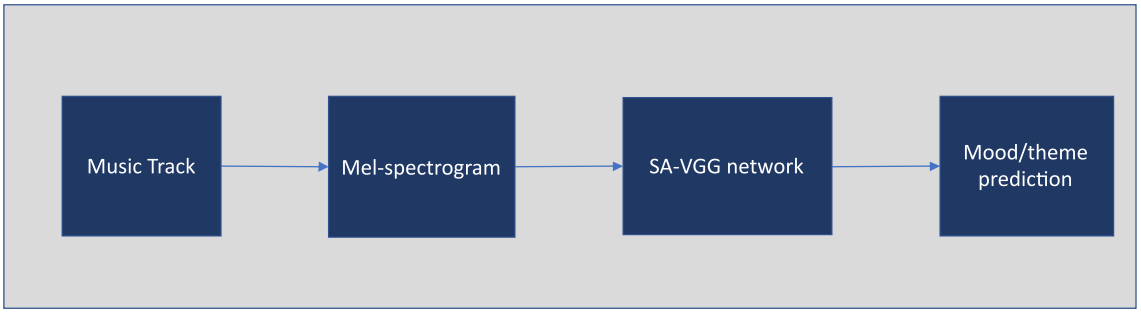

Figure 9.1: Overview of our approach for Emotion and Theme Recognition in Music. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, and Schuller, 2021)

the output layer, sigmoid activation is used to squeeze the output within [0, 1]. After every convolution and before activation, batch normalisation is added. The loss function used is binary cross entropy function. We utilize this proven VGG-like architecture as the baseline architecture. We contribute towards optimising this network by reducing the number of trainable parameters and FLOPS through a novel integration of stand-alone self-attention network elements.

9.3 Novel self-attention based VGG-like network (SA-VGG)

Figure 9.1 provides an schematic of our approach for emotion and theme recognition task in music. Figure 9.2 describes our novel SA-VGG network. In order to achieve maximum reduction in number of FLOPS and model parameters but still retaining the network performance, we optimally integrate a series of convolution layers and self-attention layers. At the end, dense connections are used. Prior research has recommended that the input should be sufficiently down-sampled before applying self-attention since the input size and the amount of memory required to hold the activations are proportional. However, though the input would be down-sampled the most at the final layers, applying self-attention here would miss out on modeling the long-range dependencies sufficiently. In order to obtain the optimum trade-off between performance and resource efficiency, we extensively experiment with various configurations for integration. We demonstrate that our SA-VGG that uses self-attention in the middle layers yields the most optimal integration. Not only does this configuration result in best gains in terms of model optimisation but also improves the performance. Our experiments results performed on the 5-layer VGG-like baseline network are described in Section 9.4.

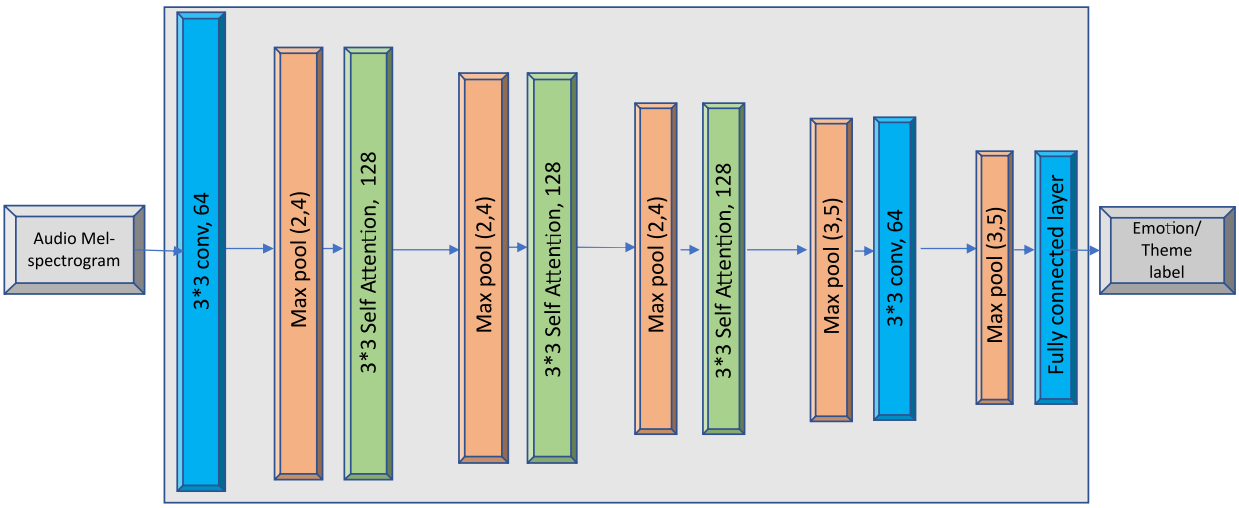

Figure 9.2: Our proposed novel self-attention based VGG-like Network (SA-VGG) for automatic tagging. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, and Schuller, 2021)

9.4 Experimental setup and results on music emotion recognition

9.4.1 Data

We conduct our experiments on the *autotagging-moodtheme* subset of the MTG-Jamendo dataset (Bogdanov, Won, *et al.*, 2019). This dataset is used in the *Emotions and Themes in Music Task* of MediaEval challenge (Bogdanov, Porter, *et al.*, 2020). Audio data from Jamendo[1], an open community of independent artists and music lovers, and made available under Creative Commons licenses is used as the basis to build the MTG-Jamendo dataset. The audio quality level is ensured to be consistent with commercial music streaming services. Basic technical quality assessment is also ensured. The dataset contains curated music which is is closer to commercial music collections with over 55 000 high quality full-length audio tracks. It is labeled with 195 tags from the categories of genre, instrument, and mood/theme. With a median track duration of 224 seconds, the audio tracks are encoded as 320 kbps MP3. In our experiments, we use the *split-0* subset. This comprises of 18 486 audio tracks with mood and theme annotations.

Mood annotations are used as a proxy in this dataset to understand the emotions conveyed by the music. The appropriate context for listening to the music or the concept or meaning that is sought to be conveyed by the artist with the mu-

[1]https://jamendo.com

sic is described through themes. The task is one of multi-label classification where each track is labeled with atleast one or more of the 56 distinct mood/themes tags. A train-validation-test split of 60-20-20 % is considered. Mel-spectrogram representation of each audio track is used as input. This transformation is done using the ESSENTIA library with 12 kHz sampling rate, 256 FFT bins, 256 samples hop-size and 96 mel-bands(Bogdanov, Wack, *et al.*, 2013). 1 400 time bins of the Mel-spectrogram of each track is considered as input. For tracks shorter than 1 400 time bins, looping/repetition is done to ensure that the track length used as input to the model is of the same size.

9.4.2 Experimental setup and results

We empirically determine the best location for the optimal usage of self attention in the baseline architecture comprising of 5 convolution layer VGG-like network. Our experiments included evaluating the use of self-attention in different layers of this network, comprising of:

- Self-attention in individual layers
- Self-attention in multiple layers with different combinations
- Self-attention in all layers

We focus our experiments on combinations involving the layers other than the first layer since using self-attention in the first layer requires significant amount of memory to hold the activations but only results in minor reduction in number of parameters or FLOPS.

A mixed optimisation approach introduced by (Won *et al.*, 2019) is used to ensure generalisation. For the first 25 epochs, ADAM with a learning rate of $1e-4$ is used. For the next 3 epochs, stochastic gradient descent (SGD) with a learning rate of $1e-3$ is used. After that, SGD with a learning rate of $1e-4$ is used. The model that is saved with the best Area Under the Receiver Operating Characteristic curve (ROC-AUC) is loaded at every switch point. Training is done for a maximum of 100 epochs. We use early stopping if the validation ROC-AUC does not increase for over 35 epochs. In all our experiments, the best model was learnt within 30 epochs. We report ROC-AUC as well as PR-AUC. This is because, when the data is unbalanced as in our case, ROC-AUC can result in over-optimistic scores (Davis *et al.*, 2006). The average PR-AUC is low since sparse tags report extremely poor PR-AUC.

The results of our experiments is summarised in Table 9.1. A minor reduction in number of parameters (.06 %) and number of FLOPS (3 %) but also a slight drop in performance is observed when using self-attention in Layer 1 alone. An improvement in the ROC-AUC and PR-AUC with 22 % fewer parameters and 11 % fewer FLOPS

Model	# Parameters	% Reduction	Giga FLOPS	% Reduction	ROC-AUC	PR-AUC
VGG-like baseline (Bogdanov, Porter, *et al.*, 2020)	448 122	-	3.32	-	.725	.107
Self-attention in Layer 1	447 866	.06	3.22	3	.724	.103
Self-attention in Layer 2	399 226	11	1.74	48	.723	.105
Self-attention in Layer 3	350 074	22	2.94	11	.730	.118
Self-attention in Layer 4	350 074	22	3.28	1	.724	.110
Self-attention in Layer 5	399 098	11	3.32	0	.716	.101
Self-attention in Layers 2 and 3	301 178	33	1.36	59	.731	.113
Self-attention in Layers 3 and 4	252 026	44	2.9	13	.725	.114
Self-attention in Layers 4 and 5	301 050	33	3.28	1	.717	.098
Self-attention in Layers 2 and 4	301 178	33	1.7	49	.731	.114
Self-attention in Layers 3 and 5	301 050	33	2.94	11	.725	.108
Self-attention in Layers 2, 3 and 4 **(Proposed SA-VGG network)**	**203 130**	**55**	**1.32**	**60**	**.726**	**.110**
Self-attention in Layers 3, 4 and 5	203 002	55	2.9	13	.714	.099
Self-attention in Layers 1, 2, 3, 4 and 5	153 850	66	1.2	64	.694	.079

Table 9.1: Results on the MediaEval *Emotions and themes in Music* dataset (subset of the MTG-Jamendo dataset)

is observed when using self-attention in Layer 3 alone. An improvement in the ROC-AUC and PR-AUC with 33 % fewer parameters and 59 % fewer FLOPS is observed when using self-attention in Layers 2 and 3. A drop in ROC-AUC by 4 % and PR-AUC by 26 % is observed when self-attention in used in all layers even though this is the most efficient in terms of number of trainable parameters (fewer by 66 %) and number of FLOPS (fewer by 64 %). Our proposed SA-VGG network uses self-attention in Layers 2, 3 and 4. This results in improvement of PR-AUC and ROC-AUC with **55 %** fewer parameters and **60 %** fewer FLOPS.

The results demonstrate that though replacing every convolution layer with self-attention layer in a VGG-like network is the most efficient in terms of number of parameters and FLOPS, it is not optimal due to the significant drop in performance. Also, no significant reduction in number of parameters or FLOPS is observed when using self-attention in the first layer or in the last layer.

9.5 Conclusion

We proposed a novel self-attention based SA-VGG network. Further, its effectiveness for multi-label emotion and theme recognition in music is demonstrated. Especially when executing the model inference on mobile device or other resource constrained computing hardware, the computational efficiency of this network becomes particularly relevant. It can also be applied to other music information retrieval tasks like genre classification or rhythm classification, since the proposed architecture is not task specific.

One of the directions for future work is the dynamic determination whether each layer of any network should use convolution or self-attention to achieve optimum balance between model complexity and performance.

Currently, one single headed self attention is used in the proposed SA-VGG network. A future area to explore is analysing the effect of using multiple attention heads to learn multiple distinct representations of the input. Furthermore, the effectiveness of other attention-based techniques like Convolutional Block Attention Modules (CBAMs) (Woo *et al.*, 2018) and attention augmented convolution (Bello *et al.*, 2019) for this task is to be evaluated.

10
Attention-based Gated Recurrent Unit

This chapter discusses the research work "A Novel Attention-Based Gated Recurrent Unit and its Efficacy in Speech Emotion Recognition" published in Proceedings of Annual International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Mallol-Ragolta, *et al.*, 2021).

10.1 Motivation

The basic long short-term memory (LSTM) or Gated Recurrent Unit (GRU) units have predominantly remained unchanged despite significant advancements in deep learning. By rightly adapting and enhancing the various elements of these units, it is possible to advance the state of the art. One such key element is activation functions. The use of diverse activation functions within GRU and bi-directional GRU (BiGRU) cells in the context of speech emotion recognition (SER) is explored. We also propose a novel Attention ReLU GRU (AR-GRU). Here, an attention-based Rectified Linear Unit (AReLU) activation(Chen *et al.*, 2020) is used within GRU and BiGRU cells. Using the recently proposed network for SER namely Interaction-Aware Attention Network (IAAN) (Yeh *et al.*, 2019), we demonstrate the effectiveness of AR-GRU on one exemplary application.

10.2 Novel Attention based Gated Recurrent Unit (AR-GRU)

The classical activation function used in conventional GRUs is *Hyperbolic Tangent* (tanh). Using the tanh activation function has inherent advantages but it is susceptible to the vanishing gradient problem.

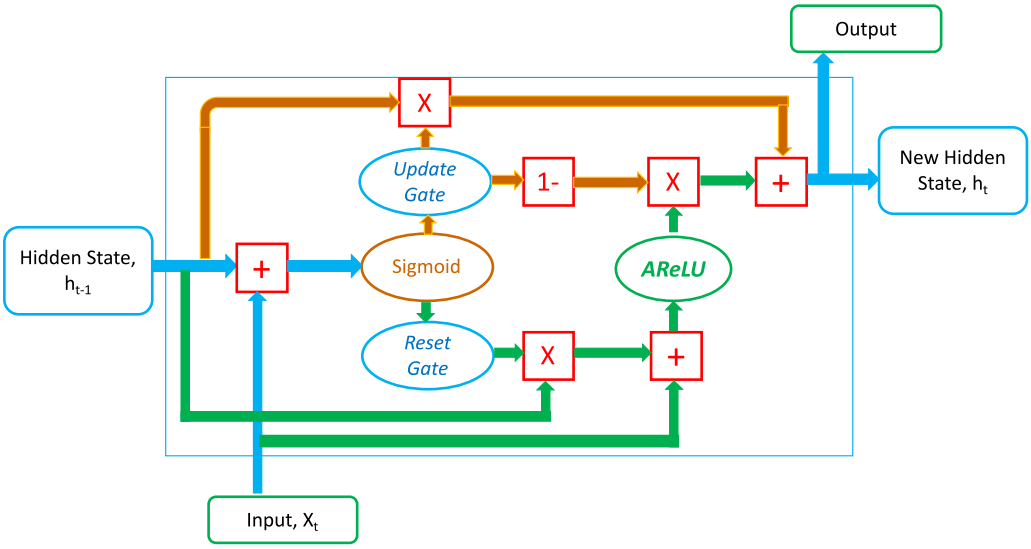

Figure 10.1: Our novel AR-GRU architecture: The classical tanh activation in a GRU is replaced by an Attention-based Rectified Linear Unit. Figure taken from (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Mallol-Ragolta, *et al.*, 2021)

Attention mechanism is one of the recent techniques in deep learning that has demonstrated significant improvements. One such realisation of a learnable attention mechanism in activation functions is Attention-based ReLU (Chen *et al.*, 2020). We propose an Attention ReLU activation based GRU unit described in Figure 10.1.

When attention-based ReLU is integrated within GRUs, it helps to capture long range interactions among the features. In speech recognition and more so in speech emotion recognision, capturing long range interactions plays an important role. This is mainly because of the supra-segmental nature of the phenomenon. Therefore, the performance of SER systems is improved through the use of AReLU-GRU which helps to capture these dependencies. Furthermore, the problem of vanishing gradient is also addressed.

10.3 Experimental setup and results on speech emotion recognition

10.3.1 Dataset Description

Using the IEMOCAP dataset (Busso *et al.*, 2008), we conduct experiments to examine the effectiveness of the different activation functions in GRU and BiGRU

in the context of Speech emotion recognition (SER). In the field of SER research, IEMOCAP dataset is a benchmark dataset that is widely used. This dataset contains conversations of 10 speakers. It comprises of five sessions where each session involves two speakers engaging in different conversational scenarios during their dialogue. A four emotion class classification, i. e., anger, happiness, sadness, and neutral, is performed using 5 531 utterances, where happiness and excitement are considered together as happiness, in order to compare with previous baseline performances. These four emotion classes have the following distribution in the 5 531 utterances: anger: 19.9 %, happiness: 29.5 %, neutral: 30.8 %, and sadness: 19.5 %.

10.3.2 Experimental Setup

The baseline model on which our empirical experiments of using different activation functions within GRU and BiGRU is done is the interaction-aware attention network (IAAN) (Yeh *et al.*, 2019). IAAN models the emotion of the current utterance by utilising the contextual information and affective influences from previous utterances. The current utterance of the speaker is handled through a BiGRU. The preceding utterances of the speaker and the interlocutor is handled through two GRUs. The openSMILE toolkit (Eyben *et al.*, 2010) with the Emobase 2010 Config is used to extract the acoustic low-level descriptors (LLDs) as well as features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and their statistics in each short frame of an utterance.

We experiment using non-learnable and learnable activation functions within the GRU and BiGRU cells of the IAAN. Using both unweighted accuracy (UA) and weighted accuracy (WA), we evaluate the performance. Our experiments were done using 5-fold leave-one-session-out (LOSO) cross validation and early stopping based on the performance on validation set in every 100 training epochs.

10.3.3 Experimental Results

The performance of our proposed method is compared with the following previous baseline networks:

BiLSTM+ATT(Mirsamadi *et al.*, 2017): A BiLSTM network that uses an attention-based pooling layer on frame-level features.

MDNN(Zhou, Suping *et al.*, 2018): A multi-path deep neural network comprising of several local classifiers and a global classifier.

IAAN(Yeh *et al.*, 2019): A GRU based network which incorporates the influence of contextual information between interlocutors within a transactional frame using interaction-aware attention.

As detailed in the Methodolody section, discovering the ideal integration of diverse activation functions within the GRU cell such that its performance is enhanced is the main novelty of our work.

Model		AReLU parameters alpha	beta	% UA	% WA
BiLSTM +ATT	Mirsamadi *et al.* (2017)	-	-	58.8	63.5
MDNN	Zhou, Suping *et al.* (2018)	-	-	62.7	61.8
IAAN	Yeh *et al.* (2019)	-	-	66.3	64.7
R-GRU based network (I)	Experiment 1	-	-	67.7	65.8
AR-GRU based network (I)	Experiment 2	0.9	2.0	35.7	38.7
AR-GRU based network (II)	Experiment 3	0	2.0	66.3	64.7
AR-GRU based network (III)	Experiment 4	0.01	-4.0	66.9	65.4
AR-GRU based network (IV)	Experiment 5	0.01	2.0	67.9	66.6
AR-GRU based network (V)	Experiment 6 : Proposed method	0.01	1.0	**68.3**	**66.9**

Table 10.1: The performance of the proposed models in comparison to state-of-the-art (upper part) and different network variants (lower part) on the IEMOCAP corpus for 4-way SER. UAR chance level resembles 25 %.

Table 10.1 presents the results on the baseline state-of-the-art networks and from all our diverse experiments for comparison. The results of our novel integration of ReLU activation units within the GRU cells (R-GRU) is first presented. Then we present the results of using AReLU as a learnable activation within GRU, detailing five different variants of integrating AReLU within GRU (AR-GRU).

Detailed description of the experiments and results are presented in (Rajamani, Srividya Tirunellai, Rajamani, Kumar, Mallol-Ragolta, *et al.*, 2021).

10.4 Conclusion

Our novel AR-GRU based network with an alpha of 0.01 and beta of 1 improves the performance of GRU for the considered task of speech emotion recognition. This proposed approach is also generalisable to other applications, such as other speech-related tasks or *Natural Language Processing* (NLP) tasks.

One of the directions for future research is to experiment with other activation functions within GRU. Several non-learnable activations like EELU (Kim *et al.*, 2020), Mish (Misra, 2020), and learnable activations such as Comb (Manessi *et al.*, 2018) and PAU (Molina *et al.*, 2020) have been recently proposed in the area of activation function. An interesting area to explore is to conduct a comparative study on the usage of such learnable and non-learnable activations within GRU. Based on our experimental findings that a small contribution of the negative values aids in getting improved results, the usage of non-learnable activation functions like Leaky ReLU as well as other learnable activations that handle negative values similar to AReLU could also be evaluated to analyse the impact on accuracy.

Part V

DISCUSSION

11

Concluding Remarks

11.1 Summary

This thesis explores deep learning methods for health data. The use of deep learning in healthcare and well-being is discussed, highlighting its impact on disease diagnosis, medical image analysis, and emotion analysis. It contributes several enhancements to attention mechanisms in image and signal analysis tasks. It proposes a novel integration of self-attention into a VGG-like network, a gated recurrent unit (GRU) module with attention for speech emotion recognition, and methodologies for image quality assessment and sparse time-series data acquisition. The thesis also addresses challenges in medical image segmentation such as handling of corner cases where the model outcome could be significantly erroneous for few subjects. Further, it aims to improve the performance, robustness, and efficiency of deep learning models in health data analysis.

Deep learning based medical image segmentation is one of the key areas for which explore avenues for improving its performance as well as other factors that are important for its utilisation in clinical practice. One of our key contribution is the identification of corner-cases in deep-learning based medical image segmentation methods. Additionally, a framework to address them is proposed. Potential reasons for the segmentation model to under-perform on corner cases are discussed by also taking into account clinical insights gained on few of these cases. Furthermore, two different approaches for checkpoint determination based on least-loss as well as highest IoU during model training are compared. Potential other approaches for handling corner-cases, such as adding more data with similar characteristics or refining the ground truth are also discussed.

Towards improving the robustness of segmentation methods, a novel regularisation technique using additive Gaussian noise in the attention mechanism of the U-Net + RCCA framework for lung lobe segmentation is introduced. This regularisation is shown to not only improve the robustness but also improve performance of segmentation models. This is the first time such a regularisation technique has

been proposed in the criss Cross attention network (CCNet) for medical image segmentation.

For improving the accuracy of semantic segmentation in medical images, we propose Deformable Attention Network (DANet). By integrating the DANet into the U-Net architecture, the model captures attention from relevant non-local locations, resulting in enhanced segmentation performance compared to the criss-cross attention mechanism. The DANet achieves this by learning the deformation of the query, key, and value attention feature maps in a continuous space. This allows the network to dynamically and precisely determine the locations from which to obtain non-local attention, leading to better segmentation results.

For determining image quality, a multi-dimensional no-reference perceptual similarity metric is proposed, which can be particularly useful in medical imaging where a good quality reference image may not always be available. The proposed metric combines no-reference image quality metric (PIQUE) and perceptual similarity and explores the dimensions of quality in the axis of noise, blur, and contrast. The experiments show that the proposed metric correlates very well with the quality of an image in a multi-dimensional sense. The unique challenge in quality determination in image acquisition workflows is also highlighted, where it is difficult to ascertain if a certain acquisition parameter is aiding in improving or worsening the final image quality. The proposed metric can help determine if a particular acquisition process or image pre-processing step is positively or negatively impacting the quality of the image.

In the context of time-series data, we demonstrate the effectiveness of inducing sparsity in time-series data for the task of classification when used in tandem with the recent multi-time attention (mTAN) network that is capable of learning from sparse and irregular data. The proposed approach of coarsely acquiring time-series data could be of immense help for various applications where data acquisition and labeling is a significant challenge.

Further, we explore using diverse activation functions within GRU and bi-directional GRU cells and propose a novel Attention ReLU GRU (AR-GRU) that employs attention-based Rectified Linear Unit activation. The effectiveness of AR-GRU on speech emotion recognition using the Interaction-Aware Attention Network is demonstrated.

11.2 Future work

One of the future directions for research is to leverage our proposed corner case detection framework for medical image analysis tasks other than medical image segmentation. Instance segmentation is one such task that is closely linked to semantic segmentation and impact of corner cases on instance segmentation tasks is still to be explored. Instance segmentation mainly relies on being able to segregate objects of

the same kind as separate instances. Corner-cases or outliers that are not detected or handled could significantly impair the performance of instance segmentation algorithms.

Another closely related downstream application of semantic segmentation is object tracking. Object tracking relies on the effectiveness of semantic segmentation at every given frame. Presence of corner-cases even in single frame could impact the tracking process. Multiple outliers across different frames could only exacerbate the scenario. Yet another important downstream task of semantic segmentation is content-based image retrieval. This is generally used to retrieve similar patients from the existing repository for comparative analysis by radiologists. Presence of corner-cases could lead to erroneous content-based retrieved images which may further impact the diagnostic outcomes. Hence further research is required to analyse the impact of corner-cases in these applications. Another future area to explore is the automatic determination of a balanced and optimal checkpoint for medical image segmentation models based on global optima.

The concept of attention regularisation could also be leveraged in attention-based classification tasks. Additionally, different noise distributions for regularisation of attention could be explored. Several new enhancements for attention are being proposed by researchers in recent times. One of the recent works has been to explore linear complexity for the attention blocks. These linear complexity attention blocks could be further regularised using our proposed approach. In the context of deformable attention, further exploration into the contribution of individual deformations of the query, key, and value feature maps towards performance improvement is an interesting area to analyse. Conducting an ablation study in this regard could provide additional insights into the effectiveness of each deformation.

For time-series data, more sophisticated sub-sampling techniques such as wavelet-based schemes could be explored. Furthermore, the sparsity level for irregular time series data could be learnt in a dynamic and task-specific way. Yet another exciting area to pursue is to compare the performance of the proposed attention based GRU (AR-GRU) with other types of recurrent neural networks, such as LSTM as well as exploring the interpretability of the proposed model and investigating how it can be used to gain insights into the underlying mechanisms of speech emotion recognition.

Acronyms

ACDC ..... Automated Cardiac Diagnosis Challenge
ADAM ..... Adaptive Moment estimation
AReLU ..... Attention based Rectified Linear Unit
AR-GRU ..... Attention ReLU Gated Recurrent Unit
AUC ..... Area under ROC curve
AVD ..... Average Hausdorff Distance
bAHD ..... Balanced Average Hausdorff Distance
BAPPS ..... Berkeley-Adobe Perceptual Patch Similarity
BiGRU ..... Bidirectional Gated Recurrent Unit
BiLSTM ..... Bidirectional Long Short-term Memory
CBAM ..... Convolutional Block Attention Module
CCA ..... Criss Cross Attention
CNN ..... Convolutional Neural Network
CT ..... Computed Tomography
COVID-19 ..... Coronavirus Disease of 2019
DANet ..... Deformable Attention Network
DICE ..... Dice coefficient
ECOD ..... Empirical-Cumulative-distribution- based Outlier Detection
ED ..... End Diastole
EELU ..... Elastic Exponential Linear Units
EHR ..... Electronic Health Record

ELU ..... Exponential Linear Unit
ES ..... End Systole
FC ..... Fully Connected
FFT ..... Fast Fourier Transform
FLOPS ..... Floating-point operations per second
FR-IQA ..... Full-Reference Image Quality Assessment
FSIM ..... Feature Similarity
GGO ..... Ground Glass Opacity
GRU ..... Gated Recurrent Unit
GT ..... Ground Truth
HD ..... Hausdorff Distance
HDR-VDP ..... High Dynamic Range Visible Difference Predictor
IAAN ..... Interaction Aware Attention Network
IEMOCAP ..... Interactive Emotional Dyadic Motion Capture
ICU ..... Intensive Care Unit
IoU ..... Intersection Over Union
IQA ..... Image Quality Assessment
JAC ..... Jaccard Index
kHz ..... kiloHertz
LLD ..... Low Level Descriptors
LOSO ..... Leave One Speaker Out
LPIPS ..... Learnt Perceptual Image Patch Similarity
LSTM ..... Long Short-Term Memory
LV ..... Left Ventricle
MDNN ..... Multi-path Deep Neural Network
MFCC ..... Mel-Frequency Cepstral Coefficients
MP3 ..... MPEG-1 Audio Layer 3
MR ..... Magnetic Resonance
MSSIM ..... Mean Structural Similarity
mTAN ..... Multi-Time Attention Network

MYO ..... Myocardium
NIH ..... National Institutes of Health
NLP ..... Natural Language Processing
NR-IQA ..... No-Reference Image Quality Assessment
NR-PSIM ..... No-Reference Perceptual Similarity
PAU ..... Padé Activation Units
PIQUE ..... Perception-based Image Quality Evaluator
PR-AUC ..... Area Under the Precision Recall (PR) Curve
RCCA ..... Recurrent Criss Cross Attention
ReLU ..... Rectified Linear Unit
RGB ..... Red Green Blue
RNN ..... Recurrent Neural Network
ROC ..... Receiver Operating Curve
ROC-AUC ..... Area Under the Receiver Operating Characteristic Curve
RT-PCR ..... Reverse Transcription-Polymerase Chain Reaction Test
RV ..... Right Ventricle
SAUNet ..... Shape Attentive U-Net
SA-VGG ..... Self-Attention based VGG
SER ..... Speech Emotion Recognition
SELU ..... Scaled Exponential Linear Unit
SGD ..... Stochastic Gradient Descent
SSIM ..... Structural Similarity
tanh ..... Hyperbolic Tangent
TID ..... Tampere Image Database
UA ..... Unweighted Average
VGG ..... Visual Geometry Group
VLSI ..... Very Large Scale Integration
WU ..... Weighted Average

List of Symbols

f . . . . . Frequency

$\mathbb{R}$ . . . . . Real number set

$\cup$ . . . . . Union

$\cap$ . . . . . Intersection

$\sum$ . . . . . Sum of all samples

σ . . . . . Standard deviation

$\otimes$ . . . . . Element-wise multiplication

μ . . . . . Mean of all samples

α . . . . . Alpha

β . . . . . Beta

Bibliography

[1] S. N. Shukla and B. M. Marlin, "Multi-Time Attention Networks for Irregularly Sampled Time Series," *International Conference on Learning Representations*, 2021.

[2] R. Dechter, "Learning While Searching in Constraint-Satisfaction-Problems," in *AAAI Conference on Artificial Intelligence*, Jan. 1986, pp. 178–183.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, pp. 1097–1105, Jan. 2012.

[4] S. J. Prince, *Understanding Deep Learning*. MIT Press, 2023. [Online]. Available: `http://udlbook.com`.

[5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015, pp. 1–14.

[6] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. Cambridge University Press, 2023, `https://D2L.ai`.

[7] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[8] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, IEEE Press, 2001, pp. 1–15.

[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," in *NeurIPS 2014 Deep Learning and Representation Learning Workshop*, Montréal, Canada, 2014, pp. 1–9.

[11] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1–15.

[12] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[13] O. Ronneberger, P.Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351, 2015, pp. 234–241.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.

[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, 2014, pp. 3104–3112.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 30, California, United States, 2017, pp. 5998–6008.

[17] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, *et al.*, "On the Opportunities and Risks of Foundation Models," *ArXiv*, 2021. [Online]. Available: `https://crfm.stanford.edu/assets/report.pdf`.

[18] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-Cross Attention for Semantic Segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 2019.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.

[20] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. Le, "Attention Augmented Convolutional Networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, Oct. 2019, pp. 3285–3294.

[21] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-Alone Self-Attention in Vision Models," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, vol. 32, Vancouver, Canada, 2019, pp. 68–80.

[22] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*. 1998.

[23] B. Kalman and S. Kwasny, "Why tanh: choosing a sigmoidal function," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, vol. 4, 1992, pp. 578–581.

[24] V. Nair and G. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of 27th International Conference on Machine Learning (ICML)*, vol. 27, Haifa, Israel, Jun. 2010, pp. 807–814.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[26] D. Chen, J. Li, and K. Xu, "AReLU: Attention-based Rectified Linear Unit," *arXiv*, arXiv:2006.13858, 2020.

[27] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, Sep. 2015.

[28] L. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[29] H. K. Zou, K. S. Warfield, A. Bharatha, M. C. C. Tempany, R. M. Kaus, J. S. Haker, M. W. Wells, A. F. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index: Scientific reports," *Academic Radiology*, vol. 11, no. 2, pp. 178–189, 2004.

[30] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[31] A. Fenster and B. Chiu, "Evaluation of segmentation algorithms for medical imaging," in *Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Shanghai, China, 2005, pp. 7186–7189.

[32] G. Gerig, M. Jomier, and M. Chakos, "Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2001, pp. 516–523.

[33] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, no. 1, pp. 1–19, 2004.

[34] N. Venkatanath, D. Praneeth, M. B. Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Twenty First National Conference on Communications (NCC)*, 2015, pp. 1–6.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[36] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[38] G. Lin, A. Milan, C. Shen, and I. D. Reid, "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, Honolulu, Hawaii, USA, 2017, pp. 5168–5177.

[39] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a Discriminative Feature Network for Semantic Segmentation," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[40] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pp. 3–11.

[41] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "Abstract: nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation," in *Bildverarbeitung für die Medizin 2019*, 2019, pp. 22–22.

[42] K. Rajamani, S. D. Gowda, V. N. Tej, and S. T. Rajamani, "Deformable attention (DANet) for semantic image segmentation," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2022, pp. 3781–3784.

[43] K. Rajamani, P. Rani, H. Siebert, R. ElagiriRamalingam, and M. Heinrich, "Attention-augmented U-Net (AA-U-Net) for semantic segmentation," in *Signal, Image and Video Processing*, 2022.

[44] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, pp. 1–8, 2022.

[45] F. Renard, S. Guedria, N. De Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, Aug. 2020.

[46] R. B. Parikh, S. Teeple, and A. S. Navathe, "Addressing Bias in Artificial Intelligence in Health Care," *JAMA*, vol. 322, no. 24, pp. 2377–2378, 2019.

[47] I. El-Naqa, H. Li, J. Fuhrman, Q. Hu, N. Gorre, W. Chen, and M. L. Giger, "Lessons learned in transitioning to AI in the medical imaging of COVID-19," *Journal of Medical Imaging*, vol. 8, no. S1, pp. 010902-1–010902-15, 2021.

[48] D. Müller, D. Hartmann, P. Meyer, F. Auer, I. Soto-Rey, and F. Kramer, "MISeval: A Metric Library for Medical Image Segmentation Evaluation," *Studies in health technology and informatics*, 2022.

[49] V. Yeghiazaryan and I. Voiculescu, "Family of boundary overlap metrics for the evaluation of medical image segmentation," *Journal of Medical Imaging*, vol. 5, no. 1, pp. 015006-1–015006-19, 2018.

[50] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," *Journal of Digital Imaging*, vol. 32, pp. 582–596, 2019.

[51] A. Jungo and M. Reyes, "Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019, pp. 48–56.

[52] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "A review of deep learning based methods for medical image multi-organ segmentation," *Physica Medica*, vol. 85, pp. 107–122, 2021.

[53] A. Reinke, M. Eisenmann, M. D. Tizabi, C. H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. J. Cardoso, V. Cheplygina, *et al.*, "Common limitations of performance metrics in biomedical image analysis," in *Medical Imaging with Deep Learning*, 2021.

[54] A. Reinke, M. D. Tizabi, C. H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, *et al.*, "Common Limitations of Image Processing Metrics: A Picture Story," in *arXiv preprint arXiv:2104.05642*, 2021.

[55] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, *et al.*, "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[56] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Büttner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, *et al.*, "Metrics reloaded: Pitfalls and recommendations for image analysis validation," in *arXiv preprint arXiv:2206.01653*, 2022.

[57] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, *et al.*, "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature communications*, vol. 9, pp. 1–13, 2018.

[58] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, ISCA, 1998, pp. 1–4.

[59] A. Kathan, M. Harrer, L. Küster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, I. Grossmann, D. D. Ebert, and B. Schuller, "Personalised depression forecasting using mobile sensor data and ecological momentary assessment," *Frontiers in Digital Health*, vol. 4, pp. 1–15, 2022.

[60] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth, "Average individual fairness: Algorithms, generalization and experiments," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[61] T. Ouyang, V. S. Marco, Y. Isobe, H. Asoh, Y. Oiwa, and Y. Seo, "Corner Case Data Description and Detection," in *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, IEEE, 2021, pp. 19–26.

[62] W. Wu, H. Xu, S. Zhong, M. R. Lyu, and I. King, "Deep Validation: Toward Detecting Real-World Corner Cases for Deep Neural Networks," in *49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, IEEE, 2019, pp. 125–137.

[63] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, CA, 2011.

[64] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[65] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, 2003, pp. II-1 –II-4.

[66] Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on HMM and SVM," in *Proceedings of the International Conference on Machine Learning and Cybernetics*, vol. 8, 2005, pp. 4898–4901.

[67] Alif Bin Abdul Qayyum, Asiful Arefeen, and Celia Shahnaz, "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition," in *Proceedings of the IEEE International Conference on Signal Processing, Information, Communication Systems*, 2019, pp. 122–125.

[68] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, 2017.

[69] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition," *IEEE Access*, 2019.

[70] Y. Yu and Y.-J. Kim, "Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database," *Electronics*, 2020.

[71] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[72] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *Proceedings of the IEEE Spoken Language Technology Workshop*, Athens, Greece, 2018, pp. 126–131.

[73] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An Interaction-aware Attention Network for Speech Emotion Recognition in Spoken Dialogs," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, 2019.

[74] Y. Rubanova, R. T. Chen, and D. Duvenaud, "Latent ODEs for Irregularly-Sampled Time Series," *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, vol. 32, pp. 5320–5330, 2019.

[75] M. Lechner and R. Hasani, "Learning long-term dependencies in irregularly-sampled time series," *arXiv preprint arXiv:2006.04418*, 2020.

[76] S. C.-X. Li and B. Marlin, "Learning from irregularly-sampled time series: A missing data perspective," in *International Conference on Machine Learning*, PMLR, 2020, pp. 5937–5946.

[77] Q. Tan, M. Ye, B. Yang, S. Liu, A. J. Ma, T. C.-F. Yip, G. L.-H. Wong, and P. Yuen, "DATA-GRU: Dual-Attention Time-Aware Gated Recurrent Unit for Irregular Multivariate Time Series," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 930–937.

[78] S. T. Rajamani, K. Rajamani, A. Venkateshvaran, A. Triantafyllopoulos, A. Kathan, and B. Schuller, "Toward detecting and addressing corner cases in deep learning based medical image segmentation," *IEEE Access*, vol. 11, pp. 95 334–95 345, 2023.

[79] J. Sun, F. Darbehani, M. Zaidi, and B. Wang, "SAUNet: Shape Attentive U-Net for Interpretable Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020, pp. 797–806.

[80] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[81] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. Chen, "ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–13, 2022.

[82] Y. Zhao, Z. Nasrullah, and Z. Li, "PyOD: A Python Toolbox for Scalable Outlier Detection," *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019.

[83] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019, pp. 92–100.

[84] O. U. Aydin, A. A. Taha, A. Hilbert, A. A. Khalil, I. Galinovic, J. B. Fiebach, D. Frey, and V. I. Madai, "On the usage of average hausdorff distance for segmentation performance assessment: Hidden error when used for ranking," *European radiology experimental*, vol. 5, pp. 1–7, 2021.

[85] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[86] Y. Tang, Y. Tang, J. Xiao, and R. M. Summers, "XLSor: A Robust and Accurate Lung Segmentor on Chest X-Rays Using Criss-Cross Attention and Customized Radiorealistic Abnormalities Generation," in *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2019.

[87] O. Oktay, J. Schlemper, L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," in *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[88] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, 2019.

[89] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

[90] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for Medical Image Segmentation: Theory and Practice When Evaluating With Dice Score or Jaccard Index," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[91] S. T. Rajamani, K. Rajamani, and B. Schuller, "A novel and simple approach to regularise attention frameworks and its efficacy in segmentation," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Sydney, Australia, 2023, pp. 1–4.

[92] M. P. Heinrich, O. Oktay, and N. Bouteldja, "OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions," *Medical Image Analysis*, 2019.

[93] MedicalSegmentation.com, *COVID-19 CT segmentation dataset*. [Online]. Available: `http://medicalsegmentation.com/covid19/`.

[94] S. T. Rajamani, K. Rajamani, P. Rani, R. Barick, R. M.S, S. V. Aithal, R. ElagiriRamalingam, S. D. Gowda, and B. Schuller, "Novel No-Reference Multi-Dimensional Perceptual Similarity Metric," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Glasgow, Scotland, United Kingdom, 2022, pp. 2045–2048.

[95] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proceedings of the Thirty-Seventh IEEE Asilomar Conference on Signals, Systems & Computers*, vol. 2, Pacific Grove, CA, USA, 2003, pp. 1398–1402.

[96] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[97] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph*, vol. 30, no. 4, pp. 1–14, 2011.

[98] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. C. Jay Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[99] S. T. Rajamani, K. Rajamani, A. Kathan, and B. Schuller, "Novel Insights of Induced Sparsity on Multi-Time Attention Networks," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Glasgow, Scotland, United Kingdom, 2022, pp. 2615–2618.

[100] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012," in *2012 Computing in Cardiology*, IEEE, 2012, pp. 245–248.

[101] I. A. Huijben, B. S. Veeling, and R. J. van Sloun, "Deep probabilistic subsampling for task-adaptive compressed sensing," in *International Conference on Learning Representations*, 2019.

[102] H. Van Gorp, I. Huijben, B. S. Veeling, N. Pezzotti, and R. J. Van Sloun, "Active Deep probabilistic Subsampling," in *International Conference on Machine Learning*, PMLR, 2021, pp. 10 509–10 518.

[103] S. T. Rajamani, K. Rajamani, and B. Schuller, "Towards an Efficient Deep Learning Model for Emotion and Theme Recognition in Music," in *IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, Tampere, Finland, 2021, pp. 1–5.

[104] D. Bogdanov, A. Porter, P. Tovstogan, and M. Won, "MediaEval 2020: Emotion and Theme Recognition in Music Using Jamendo," in *Proceedings of the MediaEval 2020 Workshop*, Online, 2020.

[105] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo Dataset for Automatic Music Tagging," in *Proceedings of the Machine Learning for Music Discovery Workshop, 36th International Conference on Machine Learning (ICML)*, California, United States, 2019.

[106] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: an Audio Analysis Library for Music Information Retrieval," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, 2013, pp. 493–498.

[107] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," *ArXiv*, vol. abs/1906.04972, 2019.

[108] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning, ACM*, vol. 06, Pittsburgh, Pennsylvania, United States, 2006, pp. 233–240.

[109] S. T. Rajamani, K. Rajamani, A. Mallol-Ragolta, S. Liu, and B. Schuller, "A Novel Attention-Based Gated Recurrent Unit and its Efficacy in Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, 2021, pp. 6294–6298.

[110] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[111] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the ACM Multimedia International Conference*, Florence, Italy, 2010.

[112] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, and K. Lei, "Inferring Emotion from Conversational Voice Data: A Semi-Supervised Multi-Path Generative Neural Network Approach," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, Louisiana, USA, 2018.

[113] D. Kim, J. Kim, and J. Kim, "Elastic Exponential Linear Units for Convolutional Neural Networks," *Neurocomputing*, vol. 406, pp. 253–266, 2020.

[114] D. Misra, "Mish: A Self Regularized Non-Monotonic Activation Function," in *Proceedings of the 31st British Machine Vision Conference*, 2020.

[115] F. Manessi and A. Rozza, "Learning Combinations of Activation Functions," in *Proceedings of the 24th International Conference on Pattern Recognition*, Beijing, China, 2018, pp. 61–66.

[116] A. Molina, P. Schramowski, and K. Kersting, "Padé Activation Units: End-to-end Learning of Flexible Activation Functions in Deep Networks," in *Proceedings of the International Conference on Learning Representations*, 2020.