ANALYZING THE MEMORY OF BLSTM NEURAL NETWORKS FOR ENHANCED EMOTION CLASSIFICATION IN DYADIC SPOKEN INTERACTIONS

Martin Wöllmer[1], *Angeliki Metallinou*[2], *Nassos Katsamanis*[2], *Björn Schuller*[1], *Shrikanth Narayanan*[2]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA
woellmer@tum.de, metallin@usc.edu, nkatsam@sipi.usc.edu, schuller@tum.de, shri@sipi.usc.edu

ABSTRACT

Recent studies indicate that bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks are well-suited for automatic emotion recognition systems and may lead to better results than systems applying other widely used classifiers such as Support Vector Machines or feedforward Neural Networks. The good performance of BLSTM emotion recognition systems could be attributed to their ability to model and exploit contextual information self-learned via recurrently connected memory blocks which allows them to incorporate information about how emotion evolves over time. However, the actual amount of bidirectional context that a BLSTM classifier takes into account when classifying an observation has not been investigated so far. This paper presents a methodology to systematically investigate the number of past and future utterance-level observations that are considered to generate an emotion prediction for a given utterance, and to examine to what extent this temporal bidirectional context contributes to the overall BLSTM performance.

Index Terms— emotion recognition, Long Short-Term Memory, sequential Jacobian, context modeling

1. INTRODUCTION

Automatic emotion recognition (AER) has become an important research area and finds many applications in modern human-computer interaction scenarios, including call-center dialogue systems, conversational agents [1], and behavioral bioinformatics [2]. To cope with the challenge of extracting affective states from audio and video data captured during naturalistic spontaneous interactions, various techniques for feature extraction and classification have been proposed. Partly, these methods have been inspired by related pattern recognition fields such as automatic speech recognition or image processing, leading to a variety of emotion recognition systems based, e. g., on Hidden Markov Models (HMM), Support Vector Machines (SVM), or neural networks.

In contrast to static classification scenarios, modern AER is influenced by the growing awareness that context plays an important role in expressing and perceiving emotions [3]. Human emotions tend to evolve slowly over time and utterances observed in isolation might not be sufficient to recognize the expressed emotion. This motivates the introduction of some form of context-sensitivity in emotion classification frameworks. For example, it was shown that AER performance in dyadic interactions profits from taking into account speech cues from the past utterance of a speaker and his interlocutor [4].

Recently, bidirectional Long Short-Term Memory (BLSTM) neural networks were introduced in order to overcome the vanishing gradient problem of conventional Recurrent Neural Networks (RNNs) [5, 6]. BLSTM neural networks make use of an arbitrary, self-learned amount of past and future contextual information. Therefore, they seem well suited for emotion recognition applications where modeling the *emotional history* during a conversation is of interest. Application of BLSTM networks for speech-based [7] and audiovisual [8, 9] emotion recognition has led to performance gains in context-sensitive AER compared to systems that do not make use of context information, such as context-free HMM or SVM-based approaches.

Yet, the actual *amount* of contextual information that is exploited within a BLSTM network for emotion classification has not been investigated so far and networks are often seen as a 'black box' being less transparent than, e. g., HMM systems. This paper presents a methodology firstly, to systematically determine the amount of context that is used by BLSTM networks to classify utterances of a speaker during a conversation and, secondly, examine the extent that this available context contributes to the overall BLSTM performance. Our goal is to better understand the effect of BLSTM modeling of human emotions and to gain insights supporting future AER system design. For our analyses, we train and evaluate our recently proposed audiovisual BLSTM emotion recognition framework [8] on the IEMOCAP database [10], a large multimodal emotional database.

2. BIDIRECTIONAL LONG SHORT-TERM MEMORY

A popular technique for context-sensitive classification based on neural networks is the application of RNNs. RNNs are able to model a certain amount of context by using cyclic connections and can in principle map from the entire *history* of previous inputs to each output. However, the analysis of the error flow in conventional recurrent neural nets resulted in the finding that long-range context is inaccessible to standard RNNs since the backpropagated error either blows up or decays over time (vanishing gradient problem). An effective approach to overcome the vanishing gradient problem is the Long Short-Term Memory architecture [5], which is able to store information in linear memory cells over a longer period of time and can learn the optimal amount of contextual information relevant for the classification task. An LSTM hidden layer is composed of multiple recurrently connected subnets which will be referred to as *memory blocks* in the following. Every memory block consists of self-connected *memory cells* and three multiplicative *gate* units (input, output, and forget gates). Since these gates allow for write, read, and reset operations within a memory block, an LSTM block can be interpreted as (differentiable) memory chip in a digital computer. The overall effect of the gate units is that the LSTM memory cells can store and access information over long periods of time and thus avoid the vanishing gradient problem (for details see [6]).

A shortcoming of standard RNNs is that they have access to past but not to future context. This can be overcome by using *bidirectional* RNNs [11], where two separate recurrent hidden layers scan the input sequences in opposite directions. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions. In this study, we use a combination of the principle of bidirectional networks and the LSTM technique (i. e., bidirectional LSTM) to exploit context between successive spoken utterances for context-sensitive emotion recognition.

3. DATABASE AND ANNOTATION

Our experiments are based on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [10] which contains approximately 12 hours of audio-visual data from five mixed gender pairs of actors. IEMOCAP includes detailed face information obtained from motion capture as well as video and audio of each session. Two techniques of actor training were used; scripts and improvisation of hypothetical scenarios. The goal was to elicit emotional displays that resemble natural emotional expression and are generated through a suitable context. As a result, context is an important factor in recognizing these emotional expressions, as is the case in most real-life interactions.

Dyadic sessions of approximately five minute length were recorded and were later manually segmented into utterances. Each utterance was annotated into nine categorical (such as anger, happiness, or neutrality) as well as dimensional tags (valence, activation, dominance) by multiple human annotators. Dimensional tags take integer values that range from one to five. The dimensional tag of an utterance is the average of the tags given by two or three annotators. We focus on the classification of valence and activation, which enables us to make use of all the available data, even utterances for which there was no categorical inter-annotator agreement, and thus no categorical label exists. We perform classification of three levels of valence and activation: level 1 contains ratings in the range [1,2], level 2 contains ratings in the range (2,4) and level 3 contains ratings in the range [4,5]. These levels intuitively correspond to low, medium and high activation respectively, and to negative, neutral and positive valence respectively.

In addition, we also examine the joint classification of the emotional dimensions by building three, four, and five clusters in the valence-activation space, as in [8]. The cluster midpoints in the emotional space are determined by applying the K-means algorithm on the annotations of the respective training sets. The ground truth of every utterance is assigned to one of the clusters using the minimum Euclidean distance between its annotation and the cluster midpoints. The intuition for clustering the valence-activation space is to build classifiers that provide richer and more complete emotional information, compared to classifying only valence or only activation.

4. AUDIO-VISUAL FEATURE EXTRACTION

Visual feature extraction is based on the normalized (x,y,z) coordinates from 46 Motion Capture (MoCap) facial markers, located as shown in [10]. In order to obtain a low-dimensional representation of the facial marker information, we use Principal Feature Analysis (PFA, see [12]). This method performs Principal Component Analysis (PCA) as a first step and selects features (here marker coordinates) so as to minimize the correlations between them. We select 30 features (covering approximately 95% of the total variability) and append the first derivatives, which results in a 60-dimensional representation of visual information. The MoCap framerate is 60 fps. The visual feature selection and normalization framework is described in detail in our previous work [13].

As low-level speech features, we extract mean and variance normalized 12 MFCC coefficients, 27 Mel Frequency Band coefficients (MFB), pitch, and energy, together with their first derivatives, using the Praat Toolbox. Both, audio and visual features are extracted at a framerate of 25 ms, with a window size of 50 ms. To obtain one static feature vector per utterance, we use a set of statistical functionals that are computed from the low-level acoustic and visual features. These functionals include means, standard deviations, linear and quadratic regression parameters (slope, offset, linear/quadratic approximation error), maximum and minimum positions, skewness, kurtosis, quartiles, inter-quartile ranges, and percentiles. All functionals are calculated using our openSMILE toolkit [14]. In order to reduce the size of the resulting feature space, we conduct a cyclic Correlation based Feature Subset Selection (CFS) on the training set. This results in an automatic selection of between 66 and 224 features, depending on the classification task.

5. EXPERIMENTS

5.1. Emotion Recognition using BLSTM Networks

To assess speaker independent emotion recognition performance of the applied BLSTM networks we carry out a cyclic leave-one-speaker-out cross validation. The mean and standard deviation of the number of test and training utterances across the ten folds is 498 ± 60 and 4475 ± 61, respectively. All BLSTM networks consist of 128 memory blocks per input direction, with one memory cell per block. The number of input nodes corresponds to the number of different features per utterance whereas the number of output nodes corresponds to the number of target classes. For comparison reasons, we also train SVMs using our utterance-level features.

In Table 1, we present the average unweighted F1-measure over the 10 speakers (folds) that is obtained for SVMs and the proposed audio-visual BLSTM classifier. The BLSTM approach outperforms context-free SVMs for all classification tasks. To investigate the importance of having meaningful available context information during BLSTM network training and decoding, we repeated all BLSTM classification experiments using randomly shuffled data. Specifically, we processed the utterances of a given conversation in arbitrary order so that the network is not able to make use of meaningful context information. As can be seen in Table 1, this downgrades recognition performance (average F1-measure) for all classification tasks. To test the statistical significance of this result, we performed paired t-tests to compare the average F1-measures and we found that BLSTM performs significantly worse (p=0.05) when we shuffle the input utterances. The normality assumption of the paired t-tests regarding the F1 distribution are satisfied according to the Shapiro-Wilk test. The performance gap suggests that the good performance of the BLSTM classifiers is to a large extent due to their ability to effectively learn an adequate amount of relevant emotional context from past and future observations. It can also be interpreted as evidence that learning to incorporate temporal context information is relevant for human emotion modeling.

5.2. Sequential Jacobian Analysis

An impression of the amount of contextual information that is used by the BLSTM network can be gained by measuring the sensitivity of the network outputs to the network inputs. When using feedforward neural networks, this can be done by calculating the Jacobian matrix J whose elements J_{ki} correspond to the derivatives of the network outputs y_k with respect to the network inputs x_i. To extend the Jacobian to recurrent neural networks, we have to specify the timesteps (representing utterances) at which the input and output

classification task	classifier		
	SVM	BLSTM	BLSTM (shuffled)
valence	61.61 ± 4.75	65.12 ± 5.13	59.71 ± 4.51
activation	51.29 ± 3.84	54.90 ± 5.02	52.10 ± 6.86
three clusters	67.86 ± 5.36	72.35 ± 5.10	67.86 ± 5.08
four clusters	57.03 ± 6.05	62.80 ± 6.69	59.27 ± 6.40
five clusters	48.34 ± 7.58	54.60 ± 5.85	51.68 ± 6.48

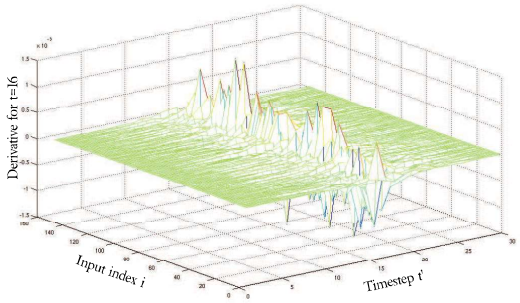
Table 1. Recognition performances [%] of SVMs and of BLSTM networks for the five classification tasks. For BLSTMs we train on the original sequence of utterances and on utterances that are randomly shuffled: mean and standard deviation of F1-measure across the 10 folds.

variables are measured. Thus, we calculate a four-dimensional matrix called the *sequential Jacobian* [6] to determine the sensitivity of the network outputs at time t to the inputs at time t':

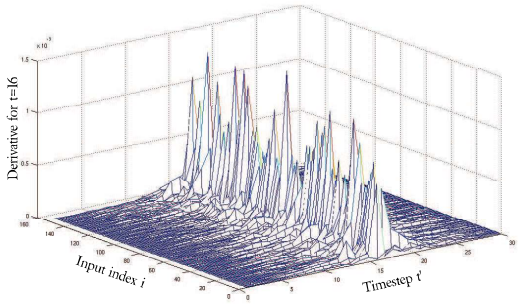
$$J_{ki}^{tt'} = \frac{\partial y_k^t}{\partial x_i^{t'}}$$

Figure 1(a) shows the derivatives of the network outputs at time $t = 16$ with respect to the different network inputs (i. e., features) at different timesteps t' for a randomly selected session consisting of 30 utterances when using a BLSTM network for the discrimination of five emotional clusters. Since we use BLSTM networks for utterance-level prediction, each timestep corresponds to one utterance. Note that the absolute magnitude of the derivatives is not important. We are rather interested in the relative magnitudes of the derivatives to each other, since this determines the sensitivity of outputs with respect to inputs at different timesteps. Of course the highest sensitivity can be detected at timestep $t' = 16$, which means that the current input has the most significant influence on the current output. However, also for timesteps smaller or greater than 16, derivatives different from zero can be found. This indicates that also past and future utterances affect the current prediction. As positive and negative derivatives are of equal importance, Figure 1(b) shows the absolute values of the derivatives in Figure 1(a). Finally, Figure 1(c) displays the corresponding derivatives summed up over all inputs and normalized to the magnitude of the derivative at $t' = 16$.

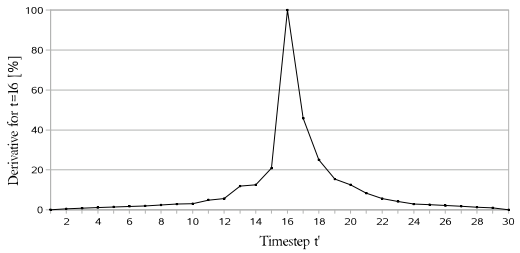
In order to systematically evaluate how many past and future inputs are relevant for the current prediction, we determined how many utterances before and after the current utterance (e. g., utterance 16 in the example given in Figure 1) have a sensitivity greater or equal to 3 % of the maximum sensitivity. To this end, we calculated projections of the sequential Jacobian as in Figure 1(c) for each timestep t in each session and each fold. Figure 2(a) shows the number of relevant past and future utterances dependent on the position in the sequence (i. e., dependent on the utterance number within a session) when using a BLSTM network for the discrimination of five clusters in the emotional space (the corresponding figures for the other classification tasks are very similar and are omitted). The number of past utterances for which the sensitivity lies above the 3 % threshold increases approximately until the eighth utterance in a session. As more and more past utterances become available, the graph converges to a value of between seven and eight, meaning that roughly seven to eight utterances of past context are used for a prediction. For the first few emotion predictions the network uses about eight utterances of future context. The slight decrease of the number of used future utterances for higher utterance numbers (i. e., for utterances occurring later in a session) is simply due to the fact that some sessions consist of less than 30 utterances, which means that towards the end of a session, less future utterances are available on average. Figure 2(b) shows the number of relevant preceding and



(a) Derivatives at time $t = 16$.



(b) Absolute values of the derivatives in Figure 1(a).



(c) Derivatives summed up over all inputs and normalized.

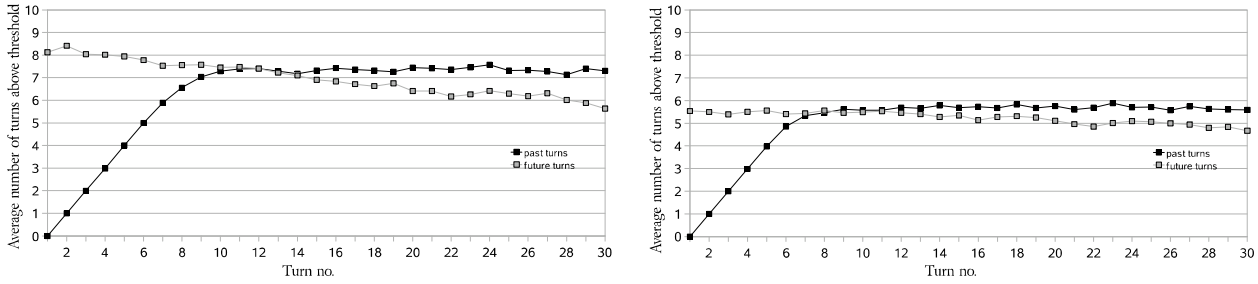
Fig. 1. Derivatives of the network outputs at time $t = 16$ with respect to the different network inputs at different timesteps t'; randomly selected session consisting of 30 utterances (BLSTM network for the discrimination of five emotional clusters).

successive utterances for the BLSTM network trained on randomly shuffled data. As can be seen, the amount of used context is less than for the BLSTM trained on correctly aligned utterances. Even though no reasonable emotional context can be learned when training on arbitrarily shuffled data, the network still uses context. One reason for this could be that BLSTM attempts to learn other session-specific charactieristics, such as speaker characteristics.

Figure 3 shows the number of relevant past utterances when considering different classification tasks and sensitivity-thresholds from 1 to 10 %. Again, we can see that networks trained on randomly shuffled data use less context (see dashed lines in Figure 3) while the amount of context exploited for the different classification tasks is relatively similar.

6. CONCLUSION AND OUTLOOK

In the light of recent studies which showed that context modeling via Long Short-Term Memory networks is well-suited for emotion



(a) BLSTM network trained on utterances in the correct order.

(b) BLSTM network trained on randomly shuffled data.

Fig. 2. Average number of relevant past and future utterances dependent on the position in the sequence when using a BLSTM network for the discrimination of five emotional clusters (3 % sensitivity-threshold).

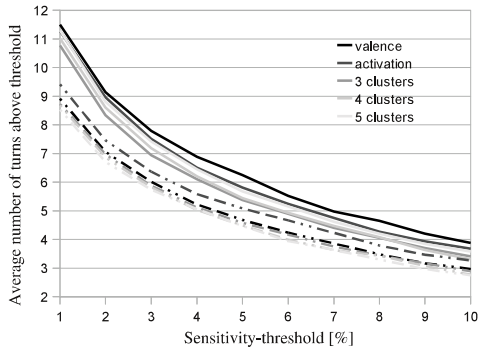


Fig. 3. Average number of relevant past utterances dependent on the sensitivity-threshold; straight lines: utterances in correct order; dashed lines: randomly shuffled data.

recognition applications [8] [9], we propose a methodology to analyze the amount of past and future context that is used by a BLSTM network to predict the emotional expression of a spoken utterance. In addition, we investigated the contribution of contextual information to the overall BLSTM performance, by randomly shuffling the order of utterances within a conversation so that the network fails to learn and exploit meaningful context. Systematic evaluations of the sequential Jacobian of trained BLSTM networks revealed that approximately eight past (and if available, also future) utterances are considered by the network as contextual information, when using a 3 % sensitivity-threshold as defined in Section 5. When the input utterances are randomly shuffled, the BLSTM network uses fewer past and future utterances (around six). Emotion recognition results showed that performance significantly decreases when networks are trained on randomly shuffled data. This suggests that good BLSTM performance is due to the network's ability to learn an adequate amount of relevant emotional context around the current observation. When such meaningful context is not present, performance degrades. Furthermore, this result illustrates that modeling typical emotional evolution during a conversation could provide useful information for emotion recognition systems.

These findings are specific to the emotion recognition database that we examine, since the dynamics of emotional states may generally vary across different types of interaction. Yet, our experiments on the IEMOCAP corpus present a first attempt to quantify the amount of context that is automatically learned during training of an emotion recognition system based on BLSTM. Future studies could apply the proposed context analysis method for other databases and scenarios, such as human-computer interactions, human-robot dialogues, and call-center data. This could help us gain insights regarding the flexibility and adaptiveness of LSTM context modeling, as well as the characteristics of different emotion recognition use-cases.

7. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] M. Black, A. Katsamanis, C.-C. Lee, A. Lammert, B. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *Proc. of Interspeech*, 2010, pp. 2030–2033.

[3] L. F. Barrett and E. A. Kensinger, "Context is routinely encoded during emotion perception," *Psychological Science*, vol. 21, pp. 595–599, 2010.

[4] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. of Interspeech*, 2009, pp. 1983–1986.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. Graves, *Supervised sequence labelling with recurrent neural networks*, Ph.D. thesis, Technische Universität München, 2008.

[7] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

[8] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. of Interspeech*, 2010, pp. 2362–2365.

[9] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, pp. 92–105, 2011.

[10] C. Busso, M. Bulut, C-C Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S.Lee, and S.Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, 1997.

[12] I. Cohen, Q. T. Xiang, S. Zhou, X. Sean, Z. Thomas, and T. S. Huang, "Feature selection using principal feature analysis," 2002.

[13] A. Metallinou, C.Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *Proc. of ICASSP*, 2010, pp. 2474–2477.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, 2010, pp. 1459–1462.