LiRA: Learning Visual Speech Representations from Audio through Self-supervision

Pingchuan Ma\textsuperscript{1*}, Rodrigo Mira\textsuperscript{1*}, Stavros Petridis\textsuperscript{2}, Björn W. Schuller\textsuperscript{1,3} and Maja Pantic\textsuperscript{1,2}

\textsuperscript{1}iBUG Group, Imperial College London, UK
\textsuperscript{2}Facebook London, UK
\textsuperscript{3}Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

\texttt{pm4115@ic.ac.uk, rs2517@ic.ac.uk}

**Abstract**

The large amount of audiovisual content being shared online today has drawn substantial attention to the prospect of audiovisual self-supervised learning. Recent works have focused on each of these modalities separately, while others have attempted to model both simultaneously in a cross-modal fashion. However, comparatively little attention has been given to leveraging one modality as a training objective to learn from the other. In this work, we propose Learning visual speech Representations from Audio via self-supervision (LiRA). Specifically, we train a ResNet+Conformer model to predict acoustic features from unlabelled visual speech. We find that this pre-trained model can be leveraged towards word-level and sentence-level lip-reading through feature extraction and fine-tuning experiments. We show that our approach significantly outperforms other self-supervised methods on the Lip Reading in the Wild (LRW) dataset and achieves state-of-the-art performance on Lip Reading Sentences 2 (LRS2) using only a fraction of the total labelled data.

**Index Terms:** self-supervised learning, lip-reading, visual speech recognition, visual representations, conformer

1. Introduction

Self-supervised learning aims to leverage unlabelled data by extracting the training objective directly from the input itself, in an attempt to model meaningful representations of the proposed modality which capture its content and structure. In works adopting this methodology, this task is usually known as the “pretext task” and this initial training procedure is known as the “pre-training” stage. After pre-training, the network is trained on the “downstream task”, which generally involves a smaller set of manually labelled data. This methodology has received substantial attention in recent years within the computer vision community. Pretext tasks for visual self-supervision include image colourisation [42], jigsaw puzzle solving [21], as well as combinations of these and other tasks [12]. Self-supervised learning has also been explored in the speech community through works such as Contrastive Predicting Coding (CPC) [22] and wav2vec [34], which predict/discriminate future segments of audio samples; LIM (Local Info Max) [32], which maximises mutual information for the same speaker; and, more recently, PASE (Problem Agnostic Speech Encoder) [26, 33], which predicts established audio features such as STFT and MFCC.

Self-supervision has also been adopted in the audiovisual domain. Recent approaches include audiovisual fusion [27, 28], clustering [4], and distillation [31]; cross-modal discrimination [23]; cyclic translation between modalities [30]; and permutative predictive coding [38]. Shukla \textit{et al.} [35] focus on learning audio representations by facial reconstruction from waveform speech. Conversely, [24] predict frequency-based summaries of ambient sound from video, while other recent works apply audio-visual synchronisation [5, 7, 13] to learn visual embeddings. A task that can benefit from self-supervised learning is lip-reading. Current state-of-the-art lip-reading models rely on annotating hundreds of hours of visual speech data [18], which is costly. To solve this issue, Afouras \textit{et al.} [3] propose using a pre-trained Automatic Speech Recognition (ASR) model to produce machine-generated captions for unsupervised pre-training. This provides automatically labelled data but still relies on an ASR model trained on large amounts of labelled data.

In this work, we aim to leverage the vast amount of available audiovisual speech data to learn generic visual speech features and improve state-of-the-art lip-reading models by predicting audio features from visual speech. The targeted audio features are extracted from waveform audio without the need for additional labels using an established speech encoder (PASE+ [33]). Using the proposed approach, the learnt visual features are explicitly guided by audio which contains rich information about speech. This in turn can lead to learning visual features which are more suitable for speech recognition. After this training procedure, we apply our model (Fig. 2) for lip-reading on a transcribed visual speech dataset.

Our research contributions are as follows: 1) We present LiRA, which learns powerful visual speech representations by predicting acoustic features from raw video taken from large audio-visual datasets. 2) We demonstrate that LiRA provides a good initialisation for fine-tuning lip-reading models which consistently outperforms training from scratch, and that this method is particularly beneficial for smaller labelled datasets. 3) We show that LiRA outperforms previous self-supervised methods for word-level lip-reading, achieving an accuracy of 88.1\% on LRW by pre-training on unlabelled data. 4) Finally, we leverage our self-supervised approach towards sentence-level lip-reading, and find that our fine-tuned model achieves state-of-the-art performance for LRS2.

2. Methodology

2.1. Pretext task

LiRA predicts PASE+ features from raw video and is composed of three distinct components. The first is the spatial encoder, which is a traditional 2D ResNet-18 preceded by a 3D front-end layer. The second component is the temporal encoder – the conformer – which receives as input the frame-wise features produced by the spatial encoder and returns a set of features of the same size. The conformer encoder contains traditional attention-based transformer blocks, which excel at capturing global temporal dependencies, with convolutional layers, which
To evaluate the visual speech representations, we run three variations of end-to-end lip-reading procedures, while the sub-figures in the bottom row ((d),(e),(f)) refer to sentence-level lip-reading. From left to right, (a) and (d) denote training from scratch (the whole model is initialised randomly); (b) and (e) are feature extraction experiments based on visual features extracted from the pre-trained model; and (c) and (f) are fine-tuning experiments. Blue coloured blocks are trained from scratch on the downstream task; yellow coloured blocks are loaded from the pre-trained model and kept frozen during the downstream task; and green coloured blocks are loaded from the pre-trained model and are then fine-tuned for the downstream task. We abbreviate the following model layers: TM: Transformer; FC: Fully-Connected layer; MS-TCN: Multi-Scale Temporal Convolutional Network.

A greater challenge since it features a largely unconstrained vocabulary of more than 40,000 words. Both datasets are collected from BBC programs.

LRS3 [2] similarly contains approximately 150,000 utterances of varying lengths (438.9 hours in total) taken from TED talks. However, these utterances are substantially longer than the ones featured in LRS2, resulting in effectively double the total amount of hours of video and a larger vocabulary. This dataset guarantees no overlap between the speakers featured in the train and test sets, meaning that the test set is entirely comprised of speakers that were not seen in other sets.

3.2. Pre-processing
To crop the mouth Regions of Interest (ROIs), we start by detecting the 68-point facial landmarks using dlib [11]. We then normalise each frame using a neural reference frame to remove rotation and size differences. Given the transformed facial landmarks, a fixed bounding box is used to crop mouth ROIs with a size of $96 \times 96$.

3.3. Data augmentation
Following [15], we produce augmented visual streams by applying the techniques of horizontal flipping with a probability of 0.5 and random cropping to a size of $88 \times 88$. During the testing phase, instead of randomly cropping, we crop a patch of size $88 \times 88$ from the centre of the image. For the word-level classification, mixup with a weight of 0.4 is employed.
which varies according to the dataset used for the downstream (as proposed in [9]) – followed by a set of conformer [10] blocks (ReLU – Rectified Linear Unit) and relative positional encoding (Fig. 2a) and sentence-level lip-reading models (Fig. 2f). The 3D front-end module preceding our ResNet consists of Table 1: A comparison of the performance between the baseline methods and ours (pre-trained on LRS3) on the LRW dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Strategy</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet + BLSTM [37]</td>
<td>Supervised</td>
<td>83.0</td>
</tr>
<tr>
<td>Two-stream 3D CNN [40]</td>
<td>Supervised</td>
<td>84.1</td>
</tr>
<tr>
<td>ResNet + BLSTM [36]</td>
<td>Supervised</td>
<td>84.3</td>
</tr>
<tr>
<td>ResNet + DenseTCN [17]</td>
<td>Supervised</td>
<td>88.4</td>
</tr>
<tr>
<td>PerfectMatch [7]</td>
<td>Self-supervised</td>
<td>71.6</td>
</tr>
<tr>
<td>PT-CDDL [8]</td>
<td>Self-supervised</td>
<td>75.9</td>
</tr>
<tr>
<td>AV-PPC [38]</td>
<td>Self-supervised</td>
<td>84.8</td>
</tr>
<tr>
<td>LiRA-Supervised [15]</td>
<td>Supervised</td>
<td>87.4</td>
</tr>
<tr>
<td>LiRA-Frozen</td>
<td>Self-supervised</td>
<td>83.1</td>
</tr>
<tr>
<td>LiRA-FineTuned</td>
<td>Self-supervised</td>
<td>88.1</td>
</tr>
</tbody>
</table>

3.4. Training settings in the pretext task

The 3D front-end module preceding our ResNet consists of a convolutional layer with kernel size (5, 7, 7) followed by a max pooling layer. The conformer, on the other hand, is comprised of an initial embedding module – feed forward layer combined with layer normalisation, dropout (0.1), activation (ReLU – Rectified Linear Unit) and relative positional encoding (as proposed in [9]) – followed by a set of conformer [10] blocks which varies according to the dataset used for the downstream task (6 blocks for LRW, 12 blocks for LRS2). The conformer blocks feature the following parameters: $d^q = 2048$, $n_{head} = 4$, $d^k = 256$, $d^v = 256$, $d^d = 256$; where $d^d$ is the hidden dimension of the feed-forward modules, $n_{head}$ is the number of self-attention heads, and $d^q$, $d^k$, $d^v$ are the dimensions of the key (K), query (Q), and value (V) in the self-attention layers respectively. The MLP consists of a linear layer with a hidden dimension of 256 units, ReLU activation, dropout, and a linear layer to project the representation to 256-dimensional latent space. For prediction, we average the PASE+ features, which are computed at 100 frames per second (fps), over time to match the frame rate of the input visual features (25 fps). We optimise our model using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$) combined with the Noam scheduler [39] (25000 warm-up steps). The model is trained on LRS3 with a batch size of 32. For simplicity, we randomly sample 1 second from each clip and use it as the input to our network, discarding any utterances with less than 1 second in length.

3.5. Training settings in downstream tasks

LiRA-Supervised In LiRA-Supervised, we train word-level (Fig. 2a) and sentence-level lip-reading models (Fig. 2d) from scratch. In particular, for the task of word-level lip-reading, we add a MS-TCN followed by a linear classifier with an output dimension of 500 on top of the encoder like [17]. A cross-entropy loss is employed to optimise the whole model using Adam with decoupled Weight decay (AdamW) [14] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and a weight decay of 0.01 for 80 epochs with a batch size of 32. The initial learning rate is set to 0.0003. For the task of sentence-level lip-reading, we use 12 multi-head attention blocks ($d^q = 2048$, $n_{head} = 4$, $d^k = 256$, $d^v = 256$, $d^d = 256$) together with a linear layer on the top of conformer blocks like [16]. Following [20], we use a combination of CTC and cross-entropy loss to train a hybrid CTC/Attention architecture for 50 epochs with a batch size of 8. In this case, we use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-8}$ with the first 25000 steps for warm-up. The initial learning rate is set to 0.0004. At the decoding phase, we use a beam size of 20 for beam search. During decoding, we also apply a transformer-based language model trained on LRS2, LRS3, and Librispeech 960h [25] (16.2 million words in total). Due to graphic memory limitations, we exclude utterances with more than 600 frames during training.

LiRA-Frozen At the end of self-supervised training, the features extracted from the pre-trained frozen encoder are fed to a classifier for evaluation. For word-level lip-reading, we use a MS-TCN, followed by a linear layer with an output size of 500 for classification (Fig. 2b). For the sentence-level lip-reading, the LiRA features are first fed to 12 conformer blocks, and then the encoded representations are used for CTC/attention joint training (Fig. 2c).

LiRA-FineTuned We follow the same hyperparameter setting as LiRA-Supervised, but instead of training from scratch, we initialise the encoder with the pre-trained weights from the pretext task and then fine-tune the entire model for word-level lip-reading (Fig. 2c) and sentence-level lip-reading (Fig. 2f).

4. Results

4.1. Word-level lip-reading

We first evaluate the performance of LiRA-Supervised by training the model from scratch. This leads to an accuracy of 87.4% on LRW which is very close to the state-of-art performance. For LiRA-Frozen, which is pre-trained on LRS3, the learnt visual speech representations are evaluated on word-level lip-reading by training a MS-TCN classifier on top of the frozen representations, as illustrated in Fig. 2b. Feature extraction performance (LiRA-Frozen) for different layers is portrayed in Fig. 3. We observe that the representations extracted from the last layer of the ResNet-18 achieve a maximum accuracy of 83.1 % as seen in Table 1. It is clear that the performance generally decreases as the layer becomes deeper, which may indicate that the features extracted in deeper layers are further tuned towards the pretext task and therefore fail to generalise as well for other tasks.

The performance of the 3 downstream scenarios while varying the amount of training data on LRW is shown in Fig. 4a. We use LRS3 for self-supervised pre-training. We observe that the feature extraction approach leads to superior performance compared to LiRA-Supervised when using smaller fractions of the labelled training set (1-2%). This indicates that the pre-trained model learns useful visual features which also work well on LRW. By adopting this methodology, we can simply train the classification layers while the encoder remains frozen, and hence
To investigate the performance of visual speech representations on a more challenging task, we run training from scratch (Fig. 2d) and fine-tuning (Fig. 2f) experiments on LRS2 after pre-training in a more challenging task, we run training from scratch (Fig. 2d) and fine-tuning our models for this new task, we achieve an accuracy of 88.1 % on LRW and report a WER of 38.8 % on LRS2. Given the extent of modern audiovisual corpora, we believe it would be promising to leverage this method towards other visual tasks such as emotion recognition and speaker recognition in the future.

5. Conclusion

We present LiRA, which learns visual speech representations by cross-modal self-supervised learning. We train a visual model by predicting acoustic features from visual speech, and observe that it can be adapted for lip-reading with remarkable success. By fine-tuning our models for this new task, we achieve an accuracy of 88.1 % on LRW and report a WER of 38.8 % on LRS2. Given the extent of modern audiovisual corpora, we believe it would be promising to leverage this method towards other visual tasks such as emotion recognition and speaker recognition in the future.

6. Acknowledgements

The work of Pingchuan Ma has been partially supported by Honda. The work of Rodrigo Mira has been funded by Samsung. All datasets used in the experiments and all training, testing and ablation studies have been conducted at Imperial College.
7. References


