TOWARDS DIVERSE LIP READING REPRESENTATIONS

Xichen Pan[1]*, Zekai Li*[1]*, Yichen Gong*[2]*, Xinbing Wang*[1]*, Zhouhan Lin*[1*]
[1]Shanghai Jiao Tong University, [2]Horizon Robotics

ABSTRACT

Attention-based models like Transformer have been well-explored and proved effective on a wide range of tasks. However, the attention mechanism still cannot successfully model temporal correlation for word-level lip reading, a typical contextually informed video classification task aiming at recognizing isolated words. In this work, we find Transformer suffers from extremely over-smoothed representations, its representation capacity is limited by the incorrect using manner of contextual information. To tackle this problem, we propose a simple attention mask to maintain and incorporate contextual information, allowing Transformer to learn diverse representations and become powerful in dealing with word-level lip reading. Our method demonstrates state-of-the-art audio-visual speech recognition performances. The proposed lip reading model largely outperforms current attention-based models and achieves a top-1 accuracy of 89.0% on Lip Reading in the Wild (LRW). [1]

1. INTRODUCTION

Lip reading, i.e., Visual Speech recognition (VSR) shares a common goal of recognizing speech contents with audio-based automatic speech recognition (ASR), but it has to deal with less informative silent video inputs. Lip reading is a challenging task considering its inherent ambiguities that many phonemes can be mapped to one viseme. However, it is a task worth exploring because the involving of visual modality can significantly improve ASR performance in noisy environments.

Word-level lip reading is typically a video classification task with contextual information. The input is a fixed-length silent video clip containing a target word sampled within a sentence, and the model aims to classify the target word within target span, i.e., word boundary, referring to the provided contextual information. Modern word-level lip reading models typically consist of a front-end that extracts features from the region of interest and a back-end for temporal modeling. [1] first introduced 3-D CNN as a front-end, consisting of a 3-D Conv layer followed by a deep 2-D CNN, and has been widely used for its excellent temporal feature extraction capability [2, 3, 4]. For the back-end, RNN-based models, especially LSTM [5] and GRU [6] have been widely used by [7, 8, 9, 10, 1]. CNN-based back-ends are also popular in this field [2, 11, 3]. Attention-based back-ends are used in [12, 13], while its performance lags behind RNNs or CNNs by a large margin.

Attention-based models like Transformer [14] have been well-explored and proved effective in many fields, including text, image, speech, etc. Currently, the state-of-the-art sentence-level lip reading method [15] relies on an attention-based temporal modeling module. In contrast, Transformer performs poorly on word-level lip reading. We find that the main reason is the over-smoothed representations, and the cosine similarity across outputs can even reach 0.99. This phenomenon has been discussed in text [16] and image [17], and it is undesired as it degrades overall representation power and reduces the learning capacity of Transformer. [18] shows simply increasing the depth of vision transformers cannot boost the performance. [17] proposed that this can be attributed to the significant similarity across the extracted patch representations. Moreover, [19, 20, 16] also observed that shallow representations are better than deep representations owing to its better diversity.

In this work, we find the over-smoothed representations are mainly attributed to the incorrect contextual information using manner. We can tackle the over-smoothing problem in word-level lip reading by properly incorporating contextual information using a simple designed attention mask. Experimental results show that our proposed method gives diverse representations and significantly outperforms current attention-based baselines by a large margin. The proposed model achieves new state-of-the-art audio-visual speech recognition (AVSR) performances on LRW.

2. ATTENTION MASK

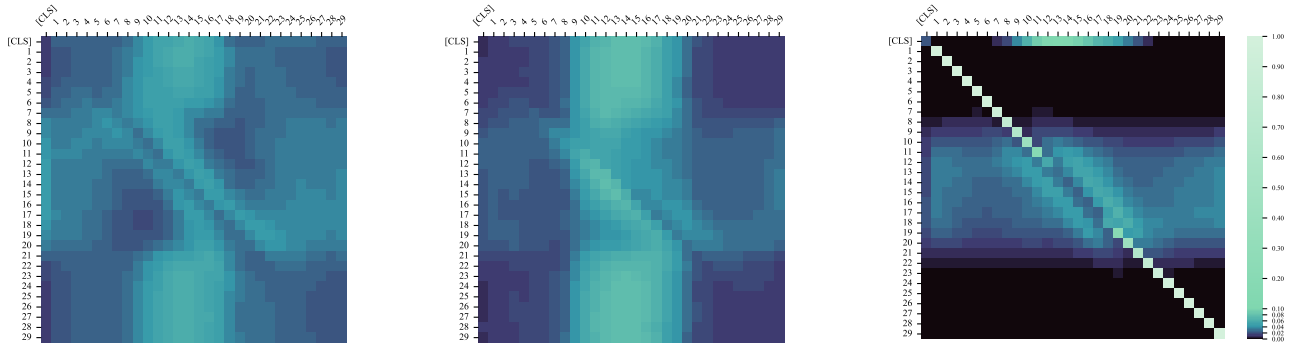
We explore the existence of over-smoothing for word-level lip reading task, by measuring the similarity between tokens in finial Transformer encoder layer. Specifically, we use the token-wise absolute cosine similarity [17] as our metric:

$$\mathrm{CosSim}\left(\boldsymbol{h}\right) = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{\left|h_i^{\mathsf{T}} h_j\right|}{\|h_i\|_2 \|h_j\|_2} \tag{1}$$

where $\boldsymbol{h} = [h_{\texttt{[CLS]}}, h_1, \ldots, h_T]$ is a sequence of representation, and T denotes the temporal length of input video sequence. Larger values of $CosSim$ indicate a higher correlation among tokens' representations and vice versa.

*Corresponding author.

[1]Our code will be publicly available after the anonymity period.



(a) Not used (b) Binary indicator (c) Attention mask

Fig. 1: Transformer attention score heatmaps of different word boundary using manners sharing the same color scale. Each row is the attention weight of one token towards the others. The attention weight is obtained by averaging all self-attention layers across 8 attention heads. High weight values are represented by a lighter color while low values by a deeper color.

We test the $CosSim$ of a popular CNN-based model MS-TCN [2] at the last layer and the result is **0.4156**. This result indicates diverse representations because CNN-based models only perform message passing locally. While for Transformer back-end, the $CosSim$ reaches **0.9971**. As shown in Fig. 1a, Transformer back-end is actually able to infer the word boundary from the extracted features, and the `[CLS]` token mainly focuses on the target word frames[2]. But Transformer cannot use the inferred word boundary information properly. In particular, target word frames and context frames attend mutually to each other, and the representations become more and more similar as the Transformer goes deeper.

We also try to use binary indicators proposed by [21] to indicate word boundaries for Transformer. In detail, we use an augmented feature $\mathbf{x}^+$ as input, which has a form of $x_t^+ = [x_t, b_t]$. x_t is the feature extracted from the t-th frames, and $b_t = \mathbb{1}_{t \in B}$, B is the set of all target word frames. The tested $CosSim$ is **0.9983**, indicating Transformer still cannot use contextual information correctly and provides diverse representations. As shown in Fig. 1b, all frames mainly focus on target word frames and rarely attend to contextual information. This is because the context frames are less informative, the bi-directional message passing through target word frames and context frames make contextual information easily dropped by Transformer. The final representations become the uniform integrated features of target word frames.

We find the contextual information is very useful for learning diverse representations. The critical point is how to preserve and properly incorporate it throughout multiple self-attention layers. In particular, we propose to use a simple attention mask $\boldsymbol{M} \in \mathbb{R}^{(T+1)\times(T+1)}$ shown in Fig. 3 in the self-attention calculation of Transformer encoder:

$$\boldsymbol{M} = [M_{\texttt{[CLS]}}, M_1, \ldots, M_t, \ldots, M_T]^{\mathsf{T}} \tag{2}$$

[2]Target words were placed right in the middle of the video clips

where $M_t \in \mathbb{R}^{T+1}$ is the attention mask of the t-th frame:

$$M_t = \begin{cases} [m_{\texttt{[CLS]}}, m_1, \ldots, m_T], & t \text{ is } \texttt{[CLS]} \\ [0, 0, \ldots, 0], & t \in B \\ [-\inf, -\inf, \ldots, -\inf], & t \notin B \end{cases} \tag{3}$$

the i-th element m_i in M_t equals to 0 when $i = t$, or it will be $-\inf$ if $i \neq t$.

The proposed attention mask only allows the context frame to attend to itself, which means it is not involved in the attention calculation, but only performs feed forward. This design aims at maintaining the contextual information during multiple self-attention calculations. The target word frames can attend to the whole sequence, thus performing temporal modeling and having a capacity to integrate contextual information, including speaker's lip movement, head pose, lighting condition, backgrounds, etc. The `[CLS]` token can attend to the target word frames so that it is able to aggregate features from these frames and output an overall representations for the whole video clip. We can finally obtain much more diversified representations, and our $CosSim$ drops to **0.3584**. If we only consider $CosSim$ for target word frames $\boldsymbol{h}' = \{h_t\}, t \in B$, the result is **0.5673**. This demonstrates that the proposed attention mask $\boldsymbol{M}$ significantly promotes token diversification for Transformer.

In contrast, if we drop all contextual information in attention mask setting and only use $\boldsymbol{h}'$ as input, the $CosSim$ reaches **0.9981**. The main reason is that attention mask allows extra context frames to serve as a knowledge base. These context frames are in different environments, and able to augment the representation of the original frame. To further illustrate the importance of contextual information, We also test several intermediate cases by dropping some context frames. As we involve more contextual information, $CosSim$ drops, indicating the model can learn more diversified representations, thus promoting final performance (see Fig. 2).

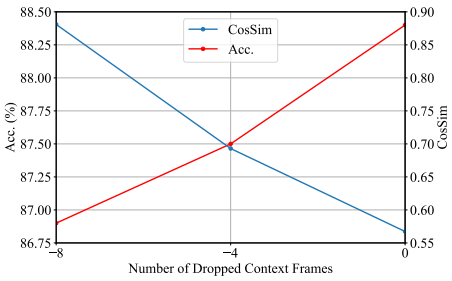


Fig. 2: $CosSim$ of target word frames $\boldsymbol{h}'$ and accuracy when we drop some context frames in attention mask setting.

3. ARCHITECTURE

As shown in Fig. 3, We employ the proposed attention mask in a lip reading model. The model is comprised of front-end, back-end, and classifier. For the AVSR model, we simply use an extra 1-D ResNet-18 audio front-end to extract audio features at 25Hz so that they can exactly match with visual features. Then we concatenate the two modalities' features channel-wise and use a 1-D Conv module to fuse.

Front-ends We use a self-supervised pre-trained MoCo v2 [22] visual front-end proposed by [23], it has been proved to be extremely effective in lip reading task. At the end of front-end, we use a 1-D Conv module to reduce the channels to $d_{model} = 512$. While for audio modality, the 3-D Conv and MoCo v2 modules in front-end are replaced by the abovementioned 1-D ResNet-18 module.

Back-end We use a 6-layers Transformer encoder [14] as our back-end. A special randomly initialized and learnable `[CLS]` token is added at the beginning of the feature sequence to obtain the overall representation [24]. We also use absolute positional encoding [14] to incorporate position information. Note we use the attention mask introduced in Sec. 2 in attention calculation.

Classifier The classifier is a module for classifying the overall representation. We use a same 1-D Conv module as the one in front-end, only adding an additional 1-D Conv layer with output channel N to project features into word classes. We use cross-entropy loss as our loss function.

4. EXPERIMENT

4.1. Experimental Settings

We use large-scale publicly available word-level lip reading dataset Lip Reading in the Wild (LRW) [25] as our testbed. We trained our models for 100 epochs[3] using AdamW optimizer [27] with an initial learning rate of 3×10^{-4} and a weight decay of 10^{-2}. We use dropouts with probabilities of 0.2 and label smoothing with weights set to 0.1 for regularization. Cosine scheduler and 10 epochs linear learning rate warm-up are used during training.

[3]Around 2 days using 6 NVIDIA GeForce RTX 3090 GPUs

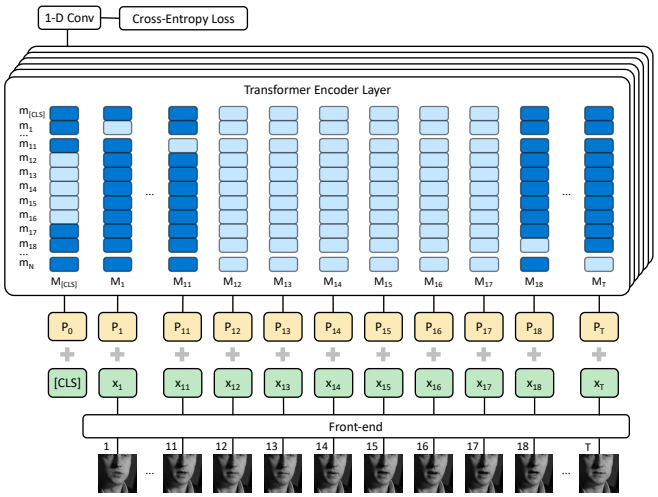


Fig. 3: Overall architecture of our lip reading model. Green blocks denotes extracted feature of each frame and the `[CLS]` token embedding, yellow blocks denote positional encoding. Blue blocks mean attention mask, the deeper ones are masked during attention calculation. In this example illustration, #12 to #17 are target word frames.

Preprocessing we use gray-scale aligned faces as visual input following [2]. Raw audio waveforms are normalized to zero mean and unit variance following [23].

Data Augmentation Following [4], random cropping and horizontal flipping are performed consistently across all frames of a video clip. For each audio waveform, babble noise is added with an SNR level uniformly selected from [-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, clean] following [7]. We synthesized 10K different babble noise samples by mixing 20 randomly selected audio samples from LRW.

4.2. Results

We present experimental results of our AVSR model on LRW in Table 1, reporting top-1 accuracy on visual-only (VO), audio-only (AO), and audio-visual (AV) models. We provide a model size comparison of our model with that of the other two popular CNN- and RNN-based models in Table 2.

Our method demonstrates state-of-the-art AVSR performaces. Specifically, the VO model achieves a top-1 accuracy of 89.0%, largely surpassing other attention-based models [13, 12, 23]. Moreover, our models do not need to be trained in 2 phases or initialized by pre-trained AO and VO models. The final AVSR model can be directly trained from scratch using an end-to-end manner. As shown in Fig. 4, our AV model has much better noise robustness over AO model.

Ablations We show the impact of each proposed module in the final accuracy by comparing with multiple baselines. We summarize the results of this study in Table 3. All the proposed modules result in clear performance improvements. Our final result utilizes MoCo v2 to leverage visual self-

Method	Front-ends	Back-ends	Acc. (%)
Visual-Only			
LRW [25]	VGG-M	-	61.1
ResNet-34 + Bi-GRU [7]	3-D Conv + ResNet-34	Bi-GRU	82.0
SpotFast + Transformer [12]	SpotFast Network	Lateral Transformers	84.4
MoCo v2 + Transformer [23]	3-D Conv + MoCo v2	Transformer	85.0
MS-TCN [2]	3-D Conv + ResNet-18	MS-TCN	85.3
ResNet-18 + Bi-LSTM + Binary Indicators [21]	3-D Conv + ResNet-18	Bi-LSTM	88.1
Densely Connected TCN [3]	3-D Conv + ResNet-18	DC-TCN	88.4
SE-ResNet-18 + Bi-GRU + Binary Indicators [10]	3-D Conv + SE-ResNet-18	Bi-GRU	88.4
Knowledge Distillation (ensemble) [11]	3-D Conv + ResNet-18	MS-TCN	88.5
Multi-Head Visual-Audio Memory [26]	3-D Conv + ResNet-18	MS-TCN + MH Visual-Audio Memory	88.5
MoCo v2 + Transformer + Attention Mask (Ours)	3-D Conv + MoCo v2	Transformer	**89.0**
Audio-Only			
MFCC + Bi-GRU [7]	-	Bi-GRU	97.7
1-D ResNet-18 + Bi-GRU [7]	1-D ResNet-18	Bi-GRU	97.7
1-D ResNet-18 + Bi-LSTM + Binary Indicators [21]	1-D ResNet-18	Bi-LSTM	98.6
1-D ResNet-18 + Transformer + Attention Mask (Ours)	1-D ResNet-18	Transformer	**98.9**
Audio-Visual			
1-D ResNet-18 (Audio) + 3-D Conv + ResNet-34 (Visual) [7]		Bi-GRU	98.0
1-D ResNet-18 (Audio) + 3-D Conv + MoCo v2 (Visual) (Ours)		Transformer	**99.0**

Table 1: Audio-only, visual-only and audio-visual results of top-1 accuracy tested on LRW.

Modules	Ours	MS-TCN	Bi-GRU
Front-end	24.6M	11.2M	11.2M
Back-end	18.9M	24.8M	47.2M
Classifier	0.5M	0.4M	1.0M

Table 2: The parameters comparison of ours, MS-TCN [2] and Bi-GRU [10] models.

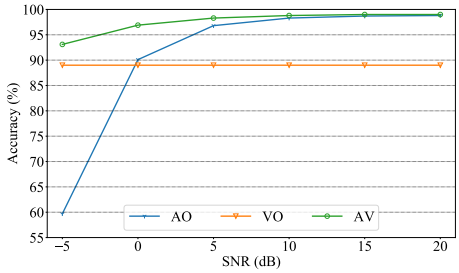


Fig. 4: Accuracy under different SNR levels. The noises are babble speech generated from LRW.

supervised learning, the top-1 accuracy reaches 89.0% with a further improvement of 0.6% over ResNet-18 front-end.

Variable Lengths We evaluate our model's temporal robustness against variable lengths input [2]. Specifically, models are trained without variable length augmentation. We test the models' performances by randomly removing n context frames from the input video clip, results are shown in Table 4. The performance of End-to-End AVSR[4] [7] drops significantly because of the sequentiality inductive bias of RNNs. CNNs' locality inductive bias also do harm to their performance. While with a weak prior, our attention-based models have strong temporal robustness against frames drop scenario.

[4]3-D-Conv + ResNet-18 as front-end and Bi-GRU as back-end.

Front-end	Word Boundary	Acc. (%)	Δ (%)
ResNet-18	Not Used	83.4	-
ResNet-18	Frame Removal	84.6	+1.2
ResNet-18	Binary Indicators	87.0	+3.6
ResNet-18	Attention Mask	88.4	+5.0
MoCo v2	Attention Mask	89.0	+5.6

Table 3: A comparison of the proposed method and baselines.

Drop n Frames →	0	1	2	3	4	5
End-to-End AVSR [7]	84.6	80.2	71.3	59.5	45.9	32.9
MS-TCN [2]	86.3	83.8	80.2	77.1	72.2	65.0
Ours	88.4	87.1	84.2	80.5	75.6	70.0

Table 4: The top-1 accuracy of different methods on LRW where n frames are randomly removed.

5. CONCLUSION

In this work, we explore the over-smoothing phenomenon in Transformer for word-level lip reading task, we propose to use a simple attention mask to tackle this problem. Our method results in highly diverse representations and better contextual information incorporation capacity. We demonstrate state-of-the-art AVSR performance.

6. REFERENCES

[1] Themos Stafylakis and Georgios Tzimiropoulos, "Combining residual networks with lstms for lipreading," in *Proc. of INTERSPEECH*, 2017.

[2] Brais Martínez, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Lipreading using temporal convolutional networks," in *Proc. of ICASSP*, 2020.

[3] Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proc. of WACV*, 2021.

[4] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. of ICASSP*, 2021.

[5] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. of EMNLP*, 2014.

[7] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic, "End-to-end audiovisual speech recognition," in *Proc. of ICASSP*, 2018.

[8] Xing Zhao, Shuang Yang, Shiguang Shan, and Xilin Chen, "Mutual information maximization for effective lip reading," in *Proc. of FG*, 2020.

[9] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition," in *Proc. of FG*, 2020.

[10] Dalu Feng, Shuang Yang, Shiguang Shan, and Xilin Chen, "Learn an effective lip reading model without pains," *ArXiv preprint*, 2020.

[11] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic, "Towards practical lipreading with distilled and efficient models," in *Proc. of ICASSP*, 2021.

[12] Peratham Wiriyathammabhum, "Spotfast networks with memory augmented lateral transformers for lipreading," in *Proc. of ICONIP*, 2020.

[13] Mingshuang Luo, Shuang Yang, Xilin Chen, Zitao Liu, and Shiguang Shan, "Synchronous bidirectional learning for multilingual lip reading," in *Proc. of BMVC*, 2020.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. of NeurIPS*, 2017.

[15] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman, "Sub-word level lip reading with visual attention," in *Proc. of CVPR*, 2022.

[16] Han Shi, JIAHUI GAO, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen MS Lee, and James Kwok, "Revisiting over-smoothing in bert from the perspective of graph," in *Proc. of ICLR*, 2021.

[17] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu, "Vision transformers with patch diversification," *ArXiv preprint*, 2021.

[18] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou, "Going deeper with image transformers," in *Proc. of ICCV*, 2021.

[19] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei, "Bert loses patience: Fast and robust inference with early exit," *Proc. of NeurIPS*, 2020.

[20] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras, "Shallow-deep networks: Understanding and mitigating network overthinking," in *Proc. of ICML*, 2019.

[21] Themos Stafylakis, Muhammad Haris Khan, and Georgios Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and lstms," *Computer Vision and Image Understanding*, 2018.

[22] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He, "Improved baselines with momentum contrastive learning," *ArXiv preprint*, 2020.

[23] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin, "Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition," in *Proc. of ACL*, 2022.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL*, 2019.

[25] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Proc. of ACCV*, 2016.

[26] Minsu Kim, Jeong Hun Yeo, and Yong Man Ro, "Distinguishing homophenes using multi-head visual-audio memory for lip reading," in *Proc. of AAAI*, 2022.

[27] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Proc. of ICLR*, 2019.