INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Should we hard-code the recurrence concept or learn it instead ? Exploring the Transformer architecture for Audio-Visual Speech Recognition

George Sterpu[1], *Christian Saam*[2], *Naomi Harte*[1]

[1]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
[2]ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland
{sterpug, saamc, nharte}@tcd.ie

Abstract

The audio-visual speech fusion strategy AV Align has shown significant performance improvements in audio-visual speech recognition (AVSR) on the challenging LRS2 dataset. Performance improvements range between 7% and 30% depending on the noise level when leveraging the visual modality of speech in addition to the auditory one. This work presents a variant of AV Align where the recurrent Long Short-term Memory (LSTM) computation block is replaced by the more recently proposed Transformer block. We compare the two methods, discussing in greater detail their strengths and weaknesses. We find that Transformers also learn cross-modal monotonic alignments, but suffer from the same visual convergence problems as the LSTM model, calling for a deeper investigation into the dominant modality problem in machine learning.

Index Terms: Audio-Visual Speech Recognition, AV Align, Transformers

1. Introduction

Multimodal fusion [1] allows the exploitation of redundancies and complementarities in naturally occurring signals, boosting the overall robustness of data processing systems in the presence of noise. One notable application is Audio-Visual Speech Recognition (AVSR), where the structure of the same speech signal materialises under two coherent modalities conveying variable levels of information. Recent advancements in machine learning led to a renewal of interest in AVSR and its key concepts such as cross-modal alignment and fusion [2, 3, 4].

More recently, the AV Align strategy [3, 4], which explicitly models the alignment of the audio and video sequences based on dot-product attention, has shown significant performance improvements on the challenging conditions of the LRS2 dataset [5]. At the same time, AV Align also discovers block-wise monotonic alignments between the input modalities without alignment supervision, and outperforms a related strategy where the text modality is used as a proxy for fusion [2]. This concept has also been applied to emotion recognition [6, 7], or speech grounding from video [8].

In the neural network space, input sequences are traditionally processed by recurrent neural networks variants including Long Short-term Memory networks (LSTM [9]). A more recent model called Transformer [10] removes the recurrent connections and updates the sequence representations using instead self-attention connections. Without the sequential processing constraint of RNNs, the Transformer model can better leverage computational resources through parallelism and achieve performance comparable to LSTMs on speech recogntion tasks for a fraction of the training costs [11, 12, 13].

In this work, we explore a variant of the AV Align strategy where the LSTM cells previously used in [3, 4] are replaced with Transformer layers. Our contributions are as follows. We describe the equivalent Transformer block for the alignment operation of AV Align in Section 2. We train and evaluate Audio-only and Audio-Visual Transformer models on the LRS2 dataset under identical conditions with [4] in Section 3.3. In Section 3.4 we show that the Audio-Visual Transformer suffers from the same video convergence problem as the LSTM-based AV Align, and the auxiliary Action Unit loss helps recover the previously seen performance improvements.

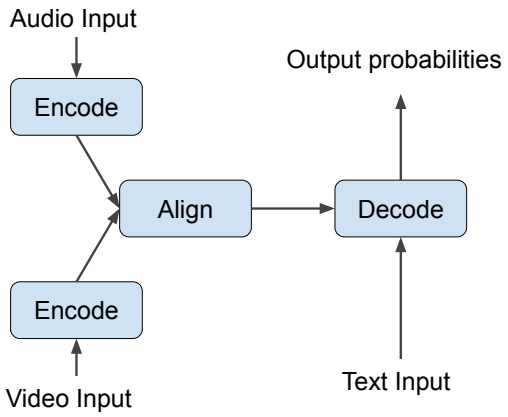


Figure 1: *Schematic representation of the Audio-Visual alignment and fusion strategy AV Align for sequence to sequence speech recognition.*

2. Audio-Visual Transformer

The AV Align architecture was first introduced in [3] and treated in more detail in [4]. The aim of this section is to show how the original LSTM block in AV Align can be replaced with the Transformer one.

In the most general case, given a variable length acoustic sentence $a = \{a_1, a_2, \dots, a_N\}$ and its corresponding visual track $v = \{v_1, v_2, \dots, v_M\}$, we transform the raw input signals into higher level latent representations (denoted by $o_A = \{o_{A_1}, o_{A_2}, \dots, o_{A_N}\}$ and $o_V = \{o_{V_1}, o_{V_2}, \dots, o_{V_M}\}$) using stacks of Transformer Encoders. An Align stack introduced in Section 2.1 soft-aligns the two modalities, and computes the fused sequence o_{AV}. Finally, an auto-regressive decoder is used to predict the output sequence of graphemes.

The Transformer architecture defined in [10] is made of an Encoder and a Decoder stack. The Encoder stack contains repeated blocks of self-attention and feed forward layers. The decoder stack contains repeated blocks of self-attention, decoder-encoder attention, and feed-forward layers. The inputs

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-2480

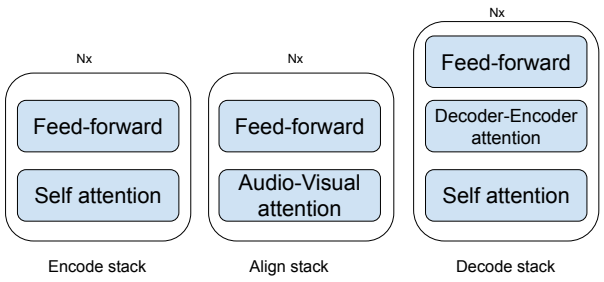


Figure 2: *The three main blocks of the Audio-Visual Transformer variant. The Encode and Decode stacks are the same as in the original model [10]. We introduce the Align stack which is based on generic attention and feed-forward layers. We keep the Nx notation from the original article to imply stacking together multiple blocks of the same structure.*

to these stacks are summed with positional encodings to embed information about the absolute position of timesteps within sequences. The Encoder and the Decoder stacks are schematically illustrated in Figure 2.

2.1. The Align stack

In order to adapt AV Align to the Transformer architecture, we define an additional Align stack as the repeated application of cross-modal attention and feed-forward layers. The Align stack is displayed in Figure 2 between the Encode and the Decode stacks. The cross-modal attention layer is a generic attention layer applied between the outputs of the two stream encoders. The align block takes video outputs o_V and audio outputs o_A as keys and queries respectively, whereas the regular encoder-decoder attention layer receives audio representations and graphemes. Consequently, the Audio-Visual Transformer model implements a single generic attention operation, as originally defined in [10], maintaining simplicity.

Formally, the audio-visual alignment and fusion steps of the attention layer in the Align stack can be described as:

$$c_V = \mathsf{attention}(\mathrm{query} = o_A, \mathrm{source} = o_V) \quad (1)$$

$$o_{AV} = c_V + o_A \quad (2)$$

where c_V are the visual context vectors computed as linear combinations of the video source o_V. Both the LSTM and Transformer variants of AV Align use the concept of dot-product attention to align the higher level audio and video representation, as in (1). However, whereas the LSTM model [3] fuses the visual context vector with audio representation by concatenating them and projecting to a shared space using linear combination, the fusion operation in the Transformer is a residual connection, seen in (2). Adding one layer's inputs to the attention output is the default fusion mechanism of the original Transformer model [10], also being used to fuse the decoder's input queries with the audio/audiovisual keys. A linear combination style fusion was explored internally for the Transformer model, without significant findings. No statistically significant differences were also reported in [4] when exploring multiple feature fusion strategies.

Our implementation forks the Transformer model officially supported in TensorFlow 2 [14] and only adds the high level Align stack, together with the visual convolution front-end, reusing the existing implementations of attention and feed-forward layers. Our code is publicly available[1].

Table 1: *CNN Architecture. All convolutions use 3x3 kernels, except the final one. The Residual Block is taken from [19] in its* full preactivation *variant.*

layer	operation	output shape
0	Rescale [-1 ... +1]	36x36x**3**
1	Conv	36x36x**8**
2-3	Res block	36x36x**8**
4-5	Res block	18x18x**16**
6-7	Res block	9x9x**32**
8-9	Res block	5x5x**64**
10	Conv 5x5	1x1x**256**

Compared to the multi-modal Transformer model proposed in [6], we do not make use of cross-modal attention at every layer in the alignment stack. As we argued in [3], there may be limited correspondences between audio and video at the lower levels of representations, and aligning only the higher level concepts is likely to speed up the training convergence.

3. Experiments and Results

3.1. Setup

LRS2 [5] contains 45,839 spoken sentences from BBC television recorded in uncontrolled illumination conditions, challenging head poses, and a low image resolution of 160x160 pixels. LRS2 is the largest AVSR dataset publicly available for research, and allows the comparison of results with more recent work [15, 16, 17].

Our system takes auditory and visual input concurrently. The **audio** input is the raw waveform signal of an entire sentence. The **visual** stream consists of video frame sequences, centred on the speaker's face, which correspond to the audio track. We use the OpenFace toolkit [18] to first detect and align the faces. We then crop the lip area to a static window determined heuristically, covering the bottom 40% of the image height and the middle 80% of the image width.

Audio input. The audio waveforms sampled at 16,000 Hz. Following the procedure of [4], we add cafeteria acoustic noise to the clean signal at three different Signal to Noise Ratios (SNR) of 10db, 0db, and -5db, in order to study how the audio information loss affects learning. We compute the log magnitude spectrogram of the input, choosing a frame length of 25ms with 10ms stride and 1024 frequency bins for the Short-time Fourier Transform (STFT), and a frequency range from 80Hz to 11,025Hz with 30 bins for the mel scale warp. We stack the features of 8 consecutive STFT frames into a larger window, leading to an audio feature vector a_i of size 240, and we shift this window right by 3 frames, thus attaining an overlap of 5 frames between windows.

Visual input. We down-sample the 3-channel RGB images of the lip regions to 36x36 pixels. A ResNet CNN [19] processes the images to produce a feature vector v_j of **256 units** per frame. The network architecture is the same as in [4], and is detailed in Table 1.

3.2. Neural network details

The transformer model uses 6 layers in the Encoder and Decoder stacks, a model size $d_{model} = 256$, a filter size $d_{ff} =$

[1] https://github.com/georgesterpu/Taris

256, one attention head, and 0.1 dropout on all attention weights and feedforward activations. The Align stack is made of a single block of cross-modal attention and feed-forward layers, with one attention head. We performed an ablation study, noting that an increase in width and depth was not worth the additional computation time with respect to accuracy.

3.3. Audio-Visual Speech Recognition Performance

We train both audio and audio-visual Transformer models on the LRS2 dataset, with the audio modality corrupted in four stages of cafeteria noise. As in [4], we train an additional audio-visual model with the Action Unit loss enabled. The results are shown in Table 2.

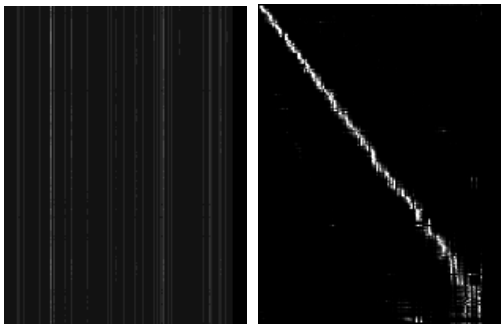
Table 2: *Character Error Rate [%] on LRS2*

System	clean	10db	0db	-5db
Audio LSTM [4]	16.38	21.85	36.27	49.08
AV Align LSTM + AU [4]	15.57	18.28	26.57	33.98
Audio Transformer	13.65	18.84	31.21	43.71
AV Transformer	13.06	17.77	30.60	43.28
AV Transformer + AU	12.08	14.82	23.52	31.74

We notice that the AV Transformer achieves a similar performance to the Audio Transformer, suggesting that the video modality was uninformative. We start seeing performance improvements only when the AU loss is used, reproducing the finding in [4]. The relative performance improvements of the AV Transformer + AU over the Audio Transformer start at 7.5% in clean speech, and go up to 26.6% in noised speech. Thus, the visual modality brings similar levels of relative improvements over the audio-only modality to both the Transformer and the LSTM trained in [4]. The absolute error differences between the LSTM and the Transformer models are partly owed to the larger model size of the Transformer used in this work (25 MB Audio, 36 MB AV) over the LSTM one in [4] (9.3 MB Audio).

3.4. Audio-Visual Alignments

As in [4], we inspect the alignment weights between the audio and visual representations, which are displayed in Figure 3. Without the AU Loss, the AV Transformer has the same difficulty as the Audio-Visual LSTM of [4] to learn cross-modal correspondences (Figure 3a), thought to be caused by the improper learning of visual representations. The alignments emerge as monotonic at the macro-block level with the AU loss (Figure 3b).



(a) *Base model* (b) *Base model + AU Loss*

Figure 3: *The Audio-Visual alignments learnt by the Transformer models*

4. Transformer or LSTM for AVSR?

The results show that the self-attention connections of the Transformer model can successfully substitute the recurrent ones originally used in the LSTM-based AV Align [3]. As in [4], the cross-modal alignments emerge as locally monotonic based on the dot-product correlations between audio and video representations. Without the auxiliary Action Unit loss, the AV Transformer presents the same learning difficulties as the LSTM variant of AV Align, and does not manage to learn monotonic alignments. We have previously speculated that the convergence problem of the visual module in AV Align was partly due to the longer propagation path of the error signal for the visual CNN and RNN in the sequence to sequence structure. Despite the great reduction of this path length in a Transformer network, our AV Transformer still required the AU loss. This demands a deeper investigation into the dominant modality problem in multi-modal machine learning, where patterns need to be discovered in the weaker visual signal. Also seen in [3, 4], sequence to sequence models are known to be susceptible to encoder-decoder disconnect when the information distribution in the target signal can be exploited for localised optimisation of the decoder, and the audio-visual disconnect is another ramification of the same problem.

Our study does not reflect an analysis of the parameter efficiency of the Transformer network compared to the LSTM for this particular dataset. We opted for commonly used hyper-parameters for datasets of this size, noting that the Transformer model is larger, partly explaining the improvements in error rates. This is because the advantages and disadvantages of both strategies go beyond parameter efficiency, being reflected in hardware throughput and engineering effort, and are discussed in greater detail in [11].

It has been suggested before that Transformers do learn the concept of recurrence from self attention connections. However, despite their highly parallel design conveying significant performance advantages over LSTMs, there is still a sense of wastefulness, particularly in speech, where distant inputs are unlikely to require connectivity. Additionally, the information from one speech frame to another does not change so much as to demand a full update of every representation in a layer.

Despite these inefficiencies, the Transformer architecture achieves faster computation speeds than LSTM on modern hardware for the majority of today's benchmarks. The LSTM blocks are facing more technical and engineering challenges in modern machine learning frameworks, which additionally leads to higher maintenance and development costs. In [20] it is argued that general purpose algorithms that best leverage computation scaling appear to be the most successful ones in the long run. The quintessential question becomes: is recurrence a concept that we want to embed into neural networks by hand, or is it preferable to opt for simpler architectures that allow the automatic learning of it ?

5. Acknowledgements

Our work is supported by a GPU grant from NVIDIA. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

6. References

[1] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb 2019.

[2] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] G. Sterpu, C. Saam, and N. Harte, "Attention-based Audio-Visual Fusion for Robust Automatic Speech Recognition," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. New York, NY, USA: ACM, 2018, pp. 111–115.

[4] G. Sterpu, C. Saam, and N. Harte, "How to teach dnns to pay attention to the visual modality in speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1052–1064, 2020.

[5] BBC and O. University, "The Oxford-BBC Lip Reading Sentences 2 (LRS2) Dataset," http://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html, 2017, online, Accessed: 2 May 2020.

[6] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6558–6569. [Online]. Available: https://www.aclweb.org/anthology/P19-1656

[7] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3507–3511.

[8] G. Paraskevopoulos, S. Parthasarathy, A. Khare, and S. Sundaram, "Multiresolution and multimodal speech recognition with transformers," 2020.

[9] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *IET Conference Proceedings*, pp. 850–855, January 1999.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[11] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and lstm encoder decoder models for asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 8–15.

[12] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 449–456.

[13] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[14] The TensorFlow Model Garden, "Transformer Translation Model," https://github.com/tensorflow/models/tree/r2.1.0/official/nlp/transformer, 2020, online, Accessed: 14 May 2020.

[15] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 513–520.

[16] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[17] J. Yu, S. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, and D. Yu, "Audio-visual recognition of overlapped speech for the lrs2 dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6984–6988.

[18] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *13th IEEE International Conference on Automatic Face Gesture Recognition*, May 2018, pp. 59–66.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV 2016*. Springer International, 2016, pp. 630–645.

[20] R. Sutton, "The bitter lesson," *Incomplete Ideas (blog), 13 March*, 2019. [Online]. Available: http://www.incompleteideas.net/IncIdeas/BitterLesson.html