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

George Sterpu¹, Christian Saam², Naomi Harte¹

¹Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
²ADAPT 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 modulus 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 co-existent 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 recognition 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.

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 \(a = \{a_1, a_2, \ldots, a_N\}\) and its corresponding visual track \(v = \{v_1, v_2, \ldots, v_M\}\), we transform the raw input signals into higher level latent representations (denoted by \(o_A = \{o_{A1}, o_{A2}, \ldots, o_{AN}\}\) and \(o_V = \{o_{V1}, o_{V2}, \ldots, o_{VM}\}\)) 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
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:

$$c_V = \text{attention}(\text{query} = o_A, \text{source} = o_V)$$  \hspace{1cm} (1)

$$o_{AV} = c_V + o_A$$  \hspace{1cm} (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 representations, 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/visual keys. A linear combination style fusion was explored internally for the Transformer model [10], also being used to fuse the decoder’s input with the visual convolution front-end, reusing the existing implementations of attention and feed-forward layers. Our code is publicly available\(^1\).

\(^1\)https://github.com/georgesterpa/Taris

\[\text{Align stack}\]

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.

### 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.

<table>
<thead>
<tr>
<th>layer</th>
<th>operation</th>
<th>output shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Rescale [-1 ... +1]</td>
<td>36x36x3</td>
</tr>
<tr>
<td>1</td>
<td>Conv</td>
<td>36x36x8</td>
</tr>
<tr>
<td>2-3</td>
<td>Res block</td>
<td>36x36x8</td>
</tr>
<tr>
<td>4-5</td>
<td>Res block</td>
<td>18x18x16</td>
</tr>
<tr>
<td>6-7</td>
<td>Res block</td>
<td>9x9x32</td>
</tr>
<tr>
<td>8-9</td>
<td>Res block</td>
<td>5x5x64</td>
</tr>
<tr>
<td>10</td>
<td>Conv 5x5</td>
<td>1x1x256</td>
</tr>
</tbody>
</table>

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 uses a model size $d_{\text{model}} = 256$, a filter size $d_{\text{ff}} = \ldots$
partly due to the longer propagation path of the error signal for convergence problem of the visual module in AV Align was monotonic alignments. We have previously speculated that the AV Transformer presents the same learning difficulties as video representations. Without the auxiliary Action Unit loss, tonic based on the dot-product correlations between audio and in [4], the cross-modal alignments emerge as locally monotonic at the macro-block level with the AU loss (Figure 3a), thought to be caused by the improper learning of visual representations. The alignments emerge as responses (Figure 3a), thought to be caused by the improper learning of it.

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>
<thead>
<tr>
<th>System</th>
<th>clean</th>
<th>10db</th>
<th>0db</th>
<th>-5db</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Transformer</td>
<td>13.65</td>
<td>18.84</td>
<td>31.21</td>
<td>43.71</td>
</tr>
<tr>
<td>AV Transformer</td>
<td>13.06</td>
<td>17.77</td>
<td>30.60</td>
<td>43.28</td>
</tr>
<tr>
<td>AV Transformer + AU</td>
<td>12.08</td>
<td>14.82</td>
<td>23.52</td>
<td>31.74</td>
</tr>
</tbody>
</table>

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).

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


