On Attention Modules for Audio-Visual Synchronization

Naji Khosravan*
Center for Research in Computer Vision
University of Central Florida
Orlando, FL, USA
najikh@cs.ucf.edu

Shervin Ardeshir
Netflix, Inc
Los Gatos, CA, USA
sardeshirbehrostaghi@netflix.com

Rohit Puri
Netflix, Inc
Los Gatos, CA, USA
rpuri@netflix.com

1. Introduction

In this effort, we use a convolutional neural network (CNN) based architecture that is capable of identifying the important portions of a video, and using them to determine the synchronization between the audio and visual signals. We study whether introducing attention modules would help the network emphasize on corresponding parts of the input data in order to make a better decision. To conduct this study, we explored defining the problem of audio-video synchronization in two different ways.

Synchronization as a Regression Problem: Defining audio-visual synchronization as a regression problem, we directly estimate the amount of misalignment between the audio and visual signals. To bound the output of the regression function, we make the assumption that misalignment between the audio and visual domain is bounded and within 130 frames (approximately 4.3 seconds given the frame rate of 29.97 Hz).

Synchronization as a Binary Classification Problem Given a video, network is trained to be able to decide whether the audio and visual modalities of the video are synchronized with each other or not. In order to train the network for this task, we expose the network to synchronized and non-synchronized audio-video streams alongside their binary labels during training time. Defining the problem as classification, alleviates the limitation of being able to decide about unbounded misalignments. However, it would not estimate the amount of misalignment.

In this paper we propose an attention based framework, trained in a self-supervised manner, for the audio-visual synchronization problem. The proposed attention modules learn to determine what to attend to in order to decide about the audio-visual synchrony of the video in the wild. We evaluate the performance of each of the two approaches on publicly available data in terms of regression error, and classification accuracy. We observe that taking into account temporal and spatio-temporal attention leads to improvement in both metrics. We also evaluate the performance of the attention modules qualitatively, verifying that the attention modules are correctly selecting discriminative parts of the video.

*Work done during an internship at Netflix.

2. Framework

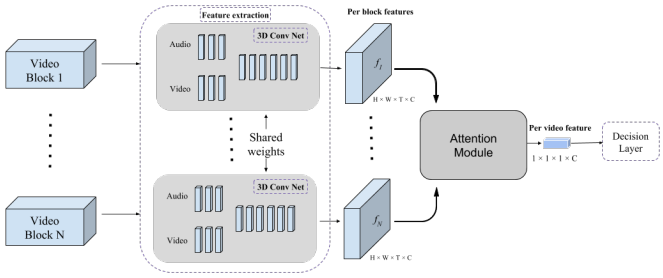

Figure 1. Architecture of the proposed approach. The video is split into several 25 frame temporal blocks. Each temporal block is passed through a 3D convolutional neural network, extracting spatio-temporal features from each block. The features are the input to the attention module, where they are evaluated in terms of discriminative power and combined into one video-level global feature. The decision is finally made based on the global feature.

Our proposed neural network architecture involves three main steps. The first step is feature extraction, where we split the input video into several blocks, in time, and extract joint audio-visual features from each block. In the second step, we apply attention modules, evaluating the importance of different (temporal or spatio-temporal) parts of the video. Finally, we combine features extracted from different parts of the video into a per video global feature based on the weights obtained from the attention modules. In the following, we provide a summary on the data representation, the two architectures used for temporal and spatio-temporal attention, and the training and testing procedures. For a more detailed report on this effort please refer to [2].

2.1. Joint Representation

The backbone of our architecture, is that of [3]. As shown in Figure 1, we divide the input video into N non-overlapping temporal blocks of length 25 frames (approximately 0.8 seconds given the frame rate of 29.97 for our videos), namely $B_1, B_2, ..., B_N$. We extract a joint audio-visual feature from each temporal block resulting in a $H \times W \times T \times C$ tensor f_i for block B_i. f_i is obtained by applying the convolutional network introduced in [3], where visual and audio features are extracted separately in the initial layers of the network and later concatenated across channels. The visual features result in a $H \times W \times T \times C_v$ feature and the audio feature results in a $T \times C_a$ feature. The audio feature is replicated $H \times W$ times and concatenated with the visual feature across channels, resulting in a $H \times W \times T \times (C_v + C_a)$ dimensional tensor where $C = C_v + C_a$. The network is followed by 5 convolution layers applied to the concatenated features, combining the two modalities and resulting in a joint representation. The joint representation is the input to the attention modules. We describe the details of applying temporal and spatiotemporal attention modules in the following sections.

2.2. Attention Modules

Our attention modules consist of two layers of $1 \times 1 \times 1$ convolutions applied to the joint audio-visual features, resulting in a scalar confidence value per block (temporal or spatio-temporal). The confidences are then passed through a softmax function to obtain a weight for each of these blocks. The weights are used to obtain a weighted mean of all the features of the video. The weighted mean is passed to the decision layer (as depicted in Figure 1). In other words, the attention modules evaluate each portion of the video (a temporal or spatio-temporal block) in terms of its importance and therefore, its contribution to the final decision. In the following, we will go over a more detailed description of the two attention modules studied in this work.

2.2.1 Temporal Attention

As explained in Section 2.1, a video results in a set of features $f_1, f_2, ..., f_N$. For the temporal attention module, we apply global average pooling to each $H \times W \times T \times C$ dimensional feature f_i across spatial and temporal dimensions, resulting in a $1 \times 1 \times 1 \times C$ dimensional feature f_i^{gap}. Therefore, representing each block of the video using a single global feature vector f_i^{gap}. We apply $1 \times 1 \times 1$ convolution layers on the global average pooled features, resulting in a single scalar confidence value c_i for each temporal block B_i. The confidence value c_i is intuitively capturing the absolute importance of that specific temporal block. Applying a softmax normalization function over all the confidence values of different time-blocks of the same video, we obtain a weight w_i for each feature f_i^{gap}. The normalization is performed to enforce the notion of probability and keep the norms of the output fearures in the same range as each individual global feature. The weighted mean of the features $\Sigma_i w_i f_i^{gap}$ is passed to the decision layer (see figure 2).

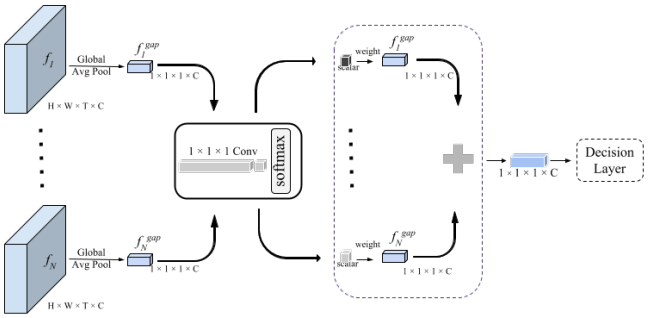

Figure 2. The temporal attention module: $1 \times 1 \times 1$ convolutions are used to obtain a confidence score for each single temporal block. All the confidences are passed through a softmax function and the resulting weights are applied to the temporal features ($1 \times 1 \times 1 \times C$). The weighted average of all features (a C dimensional vector) is then passed through the decision layer.

2.2.2 Spatio-temporal Attention

For the spatio-temporal attention module, we apply the $1 \times 1 \times 1$ convolution layers directly on the $H \times W \times T \times C$ dimensional features, resulting in a set of confidence values $c_{H \times W \times T}$ for each block. We then enforce the notion of probability across all the confidence values of all the blocks ($H \times W \times T$ scalar values). The decision is made based on the weighted average on the spatio-temporal features $\Sigma_{n=1}^{N} \Sigma_{i=1}^{T} \Sigma_{j=1}^{H} \Sigma_{k=1}^{W} w_{nijk} f_n^{ijk}$, where f_n^{ijk} is a feature vector extracted from a single spatial block i, j, k at temporal block n (see figure 3).

2.3. Baseline

In order to evaluate the effect of our temporal and spatio-temporal attention modules, we compare their performance with the performance of a uniform weighting baseline. As the attention modules simply calculate weights for the features, and the decision is made based on the weighted average of those features, as a baseline, we simply feed the average of the input features directly into the decision layer. In other words, we evaluate the effect of bypassing the weighting step.

3. Experiments

In this section, we go over the dataset used for training and evaluating the performance of the proposed approach in Section 3.1. We report the quantitative results in Section 3.2 and go over some qualitative examples in Section 3.3.

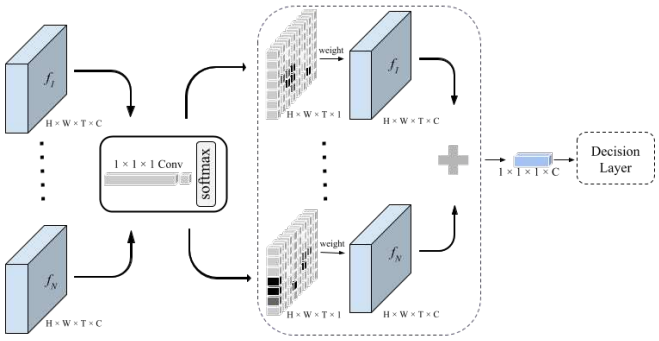

Figure 3. The spatio-temporal attention module: $1 \times 1 \times 1$ convolutions are used to obtain a confidence score for each single spatio-temporal block (each spatial block within each temporal block). All the confidences are passed through a softmax function and the resulting weights are applied to the spatio-temporal features ($1 \times 1 \times 1 \times C$). The weighted average of all features (a C dimensional vector) is then passed through the decision layer.

3.1. Dataset

We evaluate the proposed approach on the publicly AudioSet [1] dataset, which contains an ontology of 632 audio event categories. We train the temporal and spatio-temporal modules on 3000 examples of the speech subset of the dataset, and test the proposed approach on 7000 examples randomly selected from all 632 categories of the dataset. Furthermore, to show explicitly how our attention modules perform, we evaluated our method on 800 examples from each of the two selected categories: speech, and generic impact sound. These two classes are good examples for evaluation of our method as they include speech or sound classes such as breaking, hitting, bouncing etc. in which attention plays an important role. For the classification task, we used each video as a positive example, and a misaligned version of the video as a negative example. For the regression task, we use randomly misaligned videos alongside with the amount of their misalignment for training the network.

In our experiments the input videos are resized to 224×224. The length of each input sample is selected to be 125 frames which is broken into blocks of 25 frames.

3.2. Quantitative Evaluation

We evaluate the performance of the proposed approaches in terms of binary classification accuracy and regression Mean Absolute Error (MAE).

The classification accuracies are reported in Table 1. The first row shows the performance of the baseline method, where no attention module is used. Comparing the first two rows of the table, we can observe the effect of using temporal attention in the classification accuracy. We can see that in the speech category, using temporal attention leads to 4.9% improvement in classification accuracy. In the generic sound class, temporal attention yields a higher accuracy boost of 9.3%. We attribute the lower margin in the speech class to the fact that in speech videos, most of the temporal blocks of the video do contain discriminative features (lip movement) and therefore, the weights are generally more uniform (see Figure 6). The last row shows the performance of our network with the spatio-temporal attention module. In the speech class, incorporating spatio-temporal attention leads to 3.8% compared to using temporal attention, and 8.7% compared to not incorporating attention at all. In the generic sound class a 9.8% improvement is achieved using the spatio-temporal attention (compared to not using attention). Spatio-temporal attention has a lower margin of improvement over temporal attention in generic sound class compared to speech. This lower margin could be associated to the fact that speech videos tend to be more spatially localized (towards the face of the speaker).

Method	Random subset	Speech	Generic sound
Baseline network [3]	0.611	0.716	0.658
Temporal attention	0.733	0.765	0.751
Spatio-temporal attention	**0.765**	**0.803**	**0.756**

Table 1. Classification accuracy: Left column contains the methods being evaluated in terms of binary classification accuracy. The rest of the columns show the performance of methods on the Random subset, Speech and Generic sound category of the AudioSet [1] dataset, respectively.

The regression MAEs are reported in Table 2. As it can be seen, trends similar to that of the classification formulation, can be observed. The temporal attention module contributes more to generic sounds, while spatio-temporal attention module boosts the performance in the speech category.

Method	Random subset	Speech	Generic sound
Baseline network [3]	65.15	63.61	65.28
Temporal attention	28.67	26.67	**28.67**
Spatio-temporal attention	**28.67**	**28.54**	28.80

Table 2. Regression error (number of frames): Left column contains the methods being evaluated in terms shift errors in the scale of number of frames, on the same categories of the AudioSet [1] dataset.

To further illustrate the effect of attention modules, we plot and compare the distributions of output scores from our classification network in Figure 4. As it can be observed, attention modules help in better separating of the two classes of sync and un-sync data.

We also plot the distribution of misalignment error (MAEs) of the regression models is Figure 4. The desired distribution of errors is the one with a peak closer to 0. It can be observed that the base line error (without any attention) does not yield to errors less than 40 frames, while our attention modules significantly improve the performance and have lower values of error.

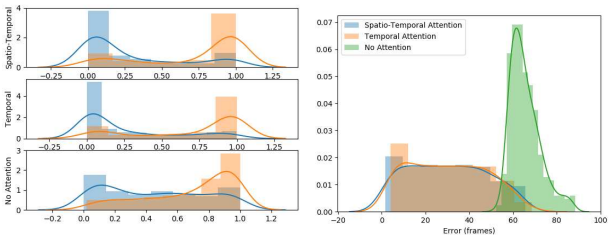

Figure 4. The distribution of scores predicted by the proposed approach is compared to that of the baseline. The distribution of the scores obtained from negative examples (non-synced videos) are shown in blue, and the distribution of the scores for the aligned videos are shown in orange. It can be observed that using attention modules causes more successful separation of positive and negative examples by the network. Distributions of Mean Absolute Errors (MAE) of the regression models in terms of number of frames. Attention modules significantly shift the peak of the error distribution closer to 0.

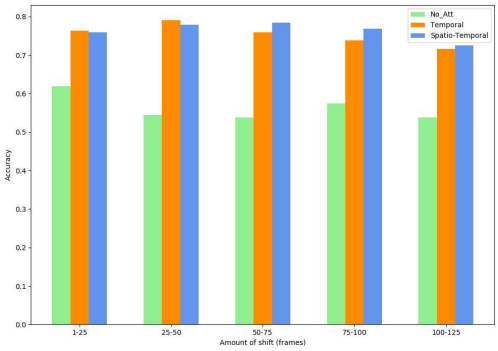

Figure 5. Classification accuracy vs misalignment values.

Furthermore, in Figure 5 we plot the classification accuracy vs misalignment values of the models. This plot shows that our proposed attention models are robust to the amount of misalignment, and maintains similar performance.

3.3. Qualitative Evaluation

Here we visualize some examples of the weights estimated by the network. We expect the informative parts of the video to lead to higher values. Two examples of the temporal attention weights are shown in Figure 6 (one from each class of dataset). In each example, each row contains one of the temporal blocks. We also show the score for each temporal block. As it can be observed, in the example on the left, a high weight has been assigned to the informative moment of the shoe tapping the ground. In the example on the right, the moments when the words are uttered by the actor are selected as the most informative parts.

In Figure 6, we show the weights obtained from the spatio-temporal module on the same examples. It can be obsersved that the network correctly assigns higher values to more discriminative regions of the video (e.g. shoe tapping the floor, and the speakers face).

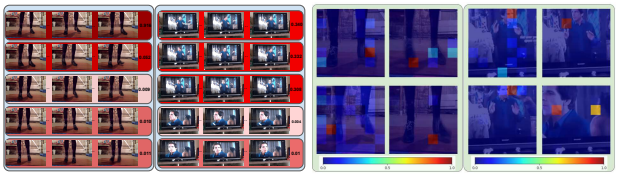

Figure 6. Left:Qualitative examples from the temporal attention module: Each row shows a temporal block of a video, highlighted with its corresponding attention weight (color-coded). Right:Qualitative examples from the spatio-temporal attention module. We picked the same examples as the temporal attention scores. It can be seen that the location of the shoe tapping on the floor and the face of the speaker are localized by the network.

4. Conclusion

In this work we studied the effect of incorporating temporal and spatio-temporal attention modules in the problem of audio-visual synchronization. We modeled the audio-visual synchronization both as a regression problem and a binary classification problem. While the regression is defined as a bounded problem and predicts the misalignment error directly from the input, the classification handles the case with unbounded shifts and makes a decision about inputs synchronization.

Our experiments suggest that a simple temporal attention module could lead to substantial performance gains in both regression and classification problems. Also, a more general spatio-temporal attention module could even achieve better performance as it is additionally capable of focusing on more discriminative spatial blocks of the video. Visualizing the weights generated by the temporal and spatio-temporal attention modules, we observe that the discriminative parts of the video are correctly given higher weights. To conclude, our experiments suggest that incorporating attention models in the audio-visual synchronization problem could lead to higher accuracy. Other variations of this approach, such as using different backbones for feature extraction, adopting different architectures such as recurrent models, could be potentially explored in the future.

References

[1] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 776–780. IEEE, 2017. 3

[2] N. Khosravan, S. Ardeshir, and R. Puri. On attention modules for audio-visual synchronization. *arXiv preprint arXiv:1812.06071*, 2018. 1

[3] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. *arXiv preprint arXiv:1804.03641*, 2018. 2, 3