TOWARD VISUAL VOICE ACTIVITY DETECTION FOR UNCONSTRAINED VIDEOS

Rahul Sharma *Krishna Somandepalli* *Shrikanth Narayanan*

University of Southern California, USA

ABSTRACT

The prevalent audio-based Voice Activity Detection (VAD) systems are challenged by the presence of ambient noise and are sensitive to variations in the type of the noise. The use of information from the visual modality, when available, can help overcome some of the problems of audio-based VAD. Existing visual-VAD systems however do not operate directly on the whole image but require intermediate face detection, face landmark detection and subsequent facial feature extraction from the lip region. In this work we present an end-to-end trainable Hierarchical Context Aware (HiCA) architecture for visual-VAD for videos obtained in unconstrained environments which can be trained with videos as input and audio speech labels as output. The network is designed to account for local and global temporal information in a video sequence. In contrast to existing visual-VAD systems our proposed approach does not rely on face detection and subsequent facial feature extraction. It can obtain a VAD accuracy of 66% on a dataset of Hollywood movie videos just with visual information. Further analysis of the representations learned from our visual-VAD system shows that the network learns to localize on human faces, and sometimes speaking human faces specifically. Our quantitative analysis of the effectiveness of face localization shows that our system performs better than sound-localization networks designed for unconstrained videos.

Index Terms— Cross-modal learning, visualization, localization, Visual-VAD

1. INTRODUCTION

A key processing step in most speech technology systems, whether the target application is automatic speech recognition, speech enhancement or emotion recognition, is Voice Activity Detection (VAD)[1, 2, 3, 4]. VAD is an audio segmentation problem, targeted to segregate the speech from the non-speech regions, which often may have noise and other interfering sources. While a simple classification task, VAD is severely challenged by the variety and variability in the noise seen in real world conditions[5]. Since VAD is at the initial stages of a speech processing pipeline, its performance degradation or failure will severely affect the downstream processing blocks such as speech recognition[1].

As an alternative, researchers have proposed complementing the audio-based system with information from the visual modality i.e., information from the talking face. The recent work by [5] reports the fusion of the two modalities using a bimodal RNN, modeling each modality using LSTMs. To handle the video modality they proposed 2D convolutional representations of the lip region. Previous methods such as [6] incorporated the visual information using handcrafted features describing the lip region and Navarathna et al.[7] used DCT features around the mouth region.

Furthermore, most of these methods [6, 7, 8, 9] have focused on constrained domains such as news broadcasts or meetings where the video is recorded in constrained settings such as controlled lighting, fixed camera, fixed background, etc [10]. This limits their applications to videos in which the face region is clearly visible and localized. Critically, explicit information about the mouth/lip landmark regions would be needed for feature extraction [11]. In contrast, videos in domains such as movies, or street cameras, are not constrained and methods optimized for domains such as broadcast news do not generalize.

Our work addresses the problem of VAD using visual information for unconstrained videos. Here, we propose an end-to-end trainable Hierarchical Context-Aware (HiCA) deep neural network to predict coarse VAD labels using just the visual information. In order to enable the network to learn from a longer context, which is a necessity in case of videos, we decentralize the temporal context in form of local 3D convolutions and a global LSTM. We do not explicitly detect the face of a speaker or extract facial features, neither for training nor for inference. We evaluate the proposed architecture with videos from Hollywood movies, which is a challenging domain due to its relatively uncontrolled settings in form of frequent shot changes and varying camera dynamics, and the variety and variability in the depiction of speaking characters.

In addition to evaluating the framework for VAD performance, we perform a formal analysis of the learned representations. Recently, with the proliferation of DNN architectures, there is an increased interest in developing tools to probe what a network learns [12]. Zhou et al [13] have proposed an approach called class activation maps (CAMs), which uses the global average pooling of convolutional output to visualize the activations learned by a CNN corresponding to a particular class. Their technique provided a CNN the ability to localize objects in an image pertaining to a particular class from a classification network trained with image-level labels. Selvaraju et al[14] proposed an efficient generalization to CAMs where they linearly approximated any network employed ahead of the CNNs, so that the idea of CAMs can be extended to any non-linear network on top of CNNs. More recently, [15] introduced a more accurate version of Grad-CAM where they further generalized the Grad-CAM by computing the importance of each pixel in the feature map towards the decision of interest. We use Grad-CAMs to visualize the representations learned by the network, as further described in sec. 2.3. Our performance analysis shows that the proposed network can robustly localize the human beings in the videos.

The contributions of this work are as follows: i) We investigate the problem of visual-VAD for unconstrained videos. We propose a cross-modal learning approach where the input is visual modality and the output is audio-speech labels. ii) This work propose HiCA deep architecture to learn from longer contexts. The architecture incorporates temporal context at two hierarchical levels. iii) Furthermore, taking advantage of the interpretability of the architecture, we present a detailed and systematic analysis of the learned representations. Our analysis shows that the visual-VAD can localize to humans in videos.

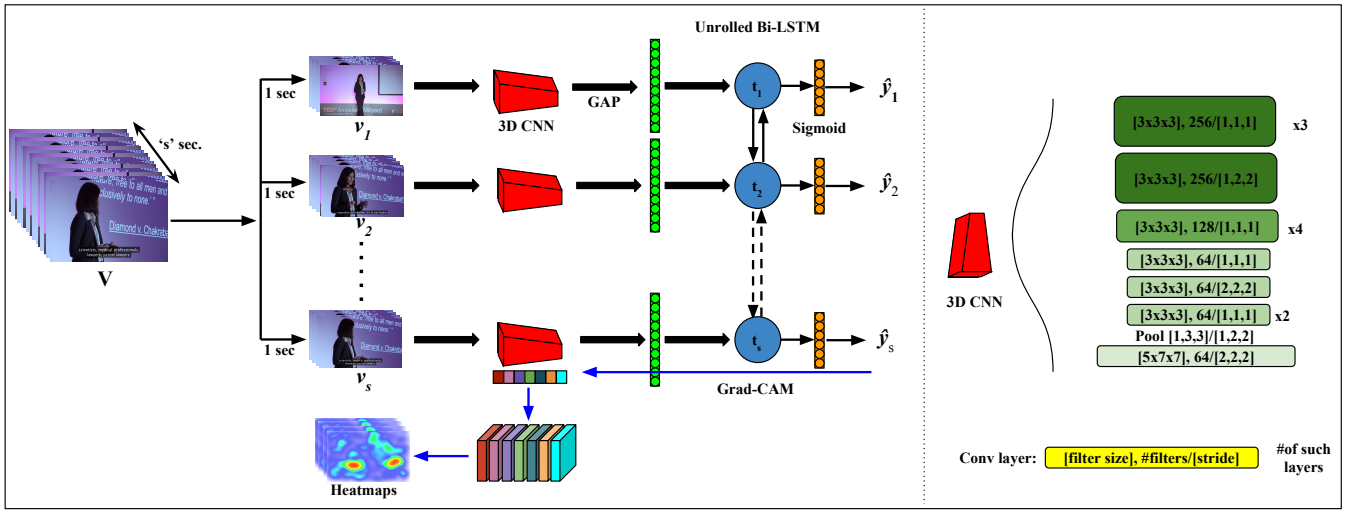

Fig. 1: Complete architecture of the proposed hierarchically context aware (HiCA) framework.

2. PROPOSED APPROACH

In this section, we formalize the cross-modal learning problem. Next, we introduce the HiCA deep neural network architecture specifically for the cross-modal VAD objective. Subsequently, we will detail the procedure for visualizing the optimized 3D CNNs using a modified version of Grad-CAMs.

2.1. Problem Formulation

Let V be a video segment of duration t seconds, and $v_i, i \in 1, ..., N$ be N partitions of the video segment, of equal duration s seconds such that $N \cdot s = t$. Formally,

$$V = \{v_1; v_2; \dots v_N\} \tag{1}$$

For each v_i, we are given a binary VAD label y_i. $y_n = 1$ indicates presence of speech, and $y_n = 0$ otherwise. Thus

$$Y = \{y_1; y_2; \dots y_N\} \quad y_i \in \{0, 1\} \tag{2}$$

We frame this task as the supervised learning problem to map from $V \longrightarrow Y$. In all our experiments, we set $s = 1s$ and $N = 10$

2.2. Network Architecture

Inspired by the recent success of CNN-LSTM architectures to model contextual information [16, 17, 18], we propose a combination of 3D convolutional network and bi-directional LSTM. 3D CNNs and LSTMs enable us to model the local (v_i in eq.1) and a longer (global) temporal context ($[v_1 \dots v_n]$ in eq. 1). Due to nature of this multi-scale context modeling, we refer to our proposed architecture as *Hierarchical Context Aware (HiCA)* architecture. The schematic of HiCA architecture is shown in Fig.1. The network can be visualized as a 2 stage pipeline, i) local spatiotemporal convolutions and ii) global bi-directional LSTM, stitched by a global average pooling (GAP) layer.

*i) Local spatiotemporal convolutions (3D conv):*The smaller video segments v_i from V is input to a ResNet-like [19] convolutional network with 3-dimensional convolutions i.e., convolution operations are performed along the height, width and time of the video frames (see Fig 1). The weights for the convolution layers are shared among v_n for $n \in \{1, 2, \dots s\}$. These 3D convolutions account for the local temporal context within the smaller segments. The output of the 3D-convs is passed through a 3D-GAP, which is capable of learning class discriminative localizations[13]. The average pooling is performed spatially as well as temporally.

ii) Global bi-directional LSTM (B-LSTM): The output obtained from the GAP layer is input to B-LSTM. The N outputs corresponding to each of $v_1, v_2 \dots v_N$ are given as input to N nodes of bi-directional LSTM. This BLSTM accounts for the temporal context for the complete N second long video segment V. The output at each node of the BLSTM is passed through a sigmoid layer to obtain final logits $\hat{P} = \{\hat{p_1}, \hat{p_2} \dots \hat{p_s}\}$. The weights for the sigmoid layer are shared among all the output nodes of BLSTM. The network is optimized to minimize the cross-entropy loss between $\hat{P}$ and Y.

2.3. Visualizations

In order to understand the visual constructs captured by the 3D-conv, we modified the Grad-CAMs[14] to accommodate 3D convolutions. Because we apply sigmoid activation to the final layer in our network, we get only one class score as output which represents the confidence for the output class 'speech' for the video segment v_n. Thus, in order to obtain the class discriminative localization map G, we first compute the gradients of the posterior score $\hat{p_i}$, of the prediction $\hat{y_i}$, with respect to each of the feature map F^m of the last convolutional layer which has m number of filters. The gradients computed are average along spatial as well as temporal dimensions to obtain the weight of the feature map m towards the output class 'speech'.

$$\alpha_m = \frac{1}{Z} \sum_i \sum_j \sum_k \frac{\partial \hat{p}}{\partial F^m_{ijk}} \tag{3}$$

where Z is the dimension of vectorized F^m. Then we perform a weighted sum of all the feature maps m of the last convolution layer and apply ReLU to obtain the required localization map G. The magnitude at each pixel, in map G, is proportional to the "attention" of the network, which can be visualized as a heatmap.

$$G = ReLU(\sum_m \alpha_m F^m) \tag{4}$$

3. PERFORMANCE EVALUATION

3.1. Implementation Details

In order to train the proposed network, we compiled Hollywood movies with labels derived from the timestamps of associated subtitles. It is important to note the scale at which we obtain VAD labels is coarse. Typically in audio experiments, VAD labels are acquired

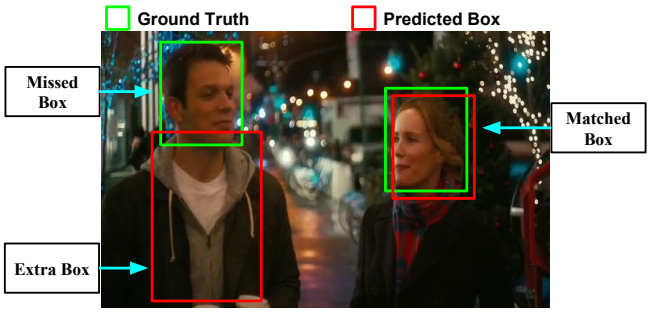

Fig. 2: Examples of missed, matched and extra boxes in a frame.

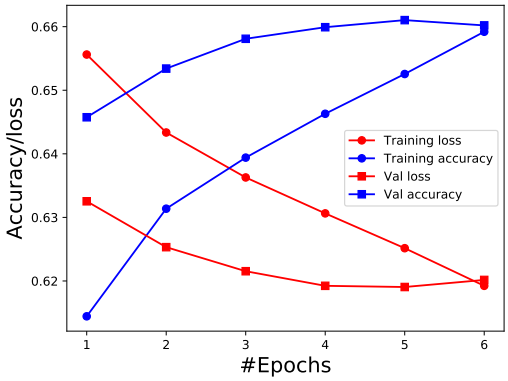

Fig. 3: Training stats for HiCA network, averaged over each epoch.

at 25-100ms precision. But here, our precision is limited by the subtitle timestamps. As Hollywood movies do not have a fixed structure with respect to the appearance of the characters, they are a good representation of the videos "in the wild", albeit with a higher quality compared with surveillance videos for example. Additionally, the subtitles are available as a separate stream in the videos and are time synchronized so we were able to obtain VAD labels automatically.

Our dataset consists of video clips from 96 movies released during the period 2014 to 2016. It comprises about 60k video segments, 10 seconds each. We obtain the coarse VAD labels for each one second video clips, v_n using the subtitles timestamps. For the duration of each dialogue in the subtitles, the corresponding video segment is labeled as speaking. For edge cases, we use the labels associated with majority of the 1 sec segment. The network is trained using a $80:20$, $train:validation$ split. The network is optimized to minimize the cross-entropy loss using the Adam[20] optimizer with an learning rate of 10^{-5}. The BLSTM consists of 2 single-layered LSTM cells with state size 512 each. Because of the size of the model and computational limitations we trained the network on single 12GB GPU with a mini-batch of size 2. The network is trained for 180000 iterations and each epoch took nearly 25 GPU hours.

3.2. VAD Performance

In order to evaluate the performance of the proposed network for VAD, we use video segments from 19 Hollywood movies, not included in the training or validation set. The segments are distributed as 51:49 for speech:non-speech labels. The network attains an accuracy of 66.10%. The evolution of VAD accuracy and loss over training epochs is shown in Fig.3. As mentioned before, our approach to visual-VAD is novel, in the sense that we can learn the model end-to-end with the entire video frame as input. We do not need additional face-detection or specific feature extraction methods which could lead to additional errors. To the best of our knowledge all methods in the literature involve explicitly extracting facial features (specifically around lip region) thus requiring the presence of frontal face, that too with good resolution. Additionally, existing visual-VAD approaches do not handle cases having multiple faces in a frame. Hence it is not feasible to compare our approach to visual-VAD models.

3.3. Visualization Analysis of Learned Representations

In order to have a detailed understanding of the representations learned by the proposed deep network in a cross-modal scenario, we present a formal analysis of the localization capabilities of the network. We evaluated the learned representations over a set of 113 video clips chosen from six Hollywood movies: About Last Night, How to be Single, Keanu, Krampus, Max, and Tomorrowland, none included in the training set. Each clip is nearly 30sec long (32.5 ± 7.5, a total of about 62 minutes). The clips were chosen arbitrarily from the movies, ensuring that each clip had sufficient speech/non-speech parts. Using the visualization method described in sec 2.3, we obtained the heatmaps corresponding to discriminative image regions for all the videos in the test set. For five videos, we show the qualitative localization performance in Fig.4, using 5 different frames in chronological order from the videos. These 5 videos along with more test videos can be found here[1].

As shown in the Fig.4, it is evident that the proposed network can robustly localize human faces in videos. $Seq1(i)$ and $seq2(v)$ highlight the capability of the network to locate multiple faces present in a frame irrespective of the view of faces (frontal or profile). $Seq2(iii)$, $seq4(v)$, and $seq4(iii)$ shows that the network can localize not just the human faces but the human body too. To further scrutinize the above observations, we present a quantitative study to evaluate the network's capability to localize human faces and bodies.

For the quantitative evaluation, for each test clip, we first obtain the bounding boxes (further referred as $pbox$) corresponding to heatmaps generated using the procedure described in sec.2.3. We then compare the obtained $pbox$ against state-of-the-art face detector and human body detector using the following measures.

i) Face Detection: We first obtain the ground truth bounding boxes, $gbox$, for the detected faces in each frame of the video segment using Google's API for face detection [2]. Next, we classify each $pbox$ into one of 3 categories based on the overlap of the predicted boxes with the ground-truth boxes, as described below. The schematic in the Fig. 2 shows the examples of all the three cases.

Matched boxes, ϕ: Predicted boxes that match the ground-truth boxes at a given IoU [21] threshold, ϵ. The i^{th} $pbox$ in frame f, $pbox_i^f$, is said to be matched if: $IOU(pbox_i^f, gbox_j^f) \geq \epsilon$ for some j^{th} $gbox$ in frame f and a particular threshold ϵ.

Missed boxes, θ: Ground truth boxes that were missed by the predictions, either due to failing matching criterion of the IOU-threshold or fewer predicted boxes. Formally, the $j^{th} gbox$ in frame f is said to be missed if: $IOU(pbox_i^f, gbox_j^f) < \epsilon$ for all $i^{th} pbox$ in frame f and a particular threshold ϵ.

Extra boxes, γ: Predicted boxes that do not match any of the ground truth boxes at a given IOU threshold, ϵ. The i^{th} $pbox$ in frame f, $pbox_i^f$, is said to be extra if: $IOU(pbox_i^f, gbox_j^f) < \epsilon$ for all j^{th} $gbox$ in frame f and a particular threshold ϵ.

We use the $F-score$ to quantify the efficiency of localization, with precision and recall computed as follows:

$$recall = \frac{|\phi|}{|\phi| + |\theta|} \quad \text{and} \quad precision = \frac{|\phi|}{|\phi| + |\gamma|} \tag{5}$$

[1] http://bit.ly/vvad_icip

[2] Neven Vision fR^{TM} API

Fig. 4: Qualitative localization performance of the proposed HiCA network for various test videos.

where $|.|$ represents the cardinality of a set. Here, ϕ,θ &, γ can be seen as true positives, false negatives and false positives respectively.

We observed that the proposed network can detect human faces with average $F-score$ of 0.79 for a liberal IOU threshold choice of 0.1. Although there are no existing cross-modal DNN models proposed for visual-VAD tasks, a recent work [22] has developed methods for sound localization using a cross modal approach. Hence we use [22] as baseline for quantifying localization ability. The models proposed in [22] detects synchrony between audio and visual data and showed that their model can localize sound sources in a video, which settles to localizing human faces in the current setup. They localize the faces with an average $F-score$ of 0.62 for the same IOU threshold of 0.1. More importantly, we note that their model requires both modalities, visual as well as audio. Fig.5 shows the variation of the average $F-score$ for various IOU thresholds.

ii) Human body Detection: This setup aims to validate the extent to which the predicted localizations can conform to a human body detector. Here we use the state-of-the-art Faster-RCNN [23] models for obtaining bounding boxes containing human body. Once we get the bounding boxes, we follow the same procedure as described in the previous setup to obtain the $F-score$. We observed that the proposed network can localize human bodies with an average $F-score$ of 0.59 for the IOU threshold of 0.1. For the same IOU threshold, [22] attains an $F-score$ of 0.63.

iii) Person Detection: In this experiment we focus on understanding the significance of the extra boxes γ, in case of face detection experiment setup. We computed the percentage overlap of all the boxes in γ with the available human body boxes in the corresponding frames, with a conservative matching threshold of 90% overlap. The Fig.5 shows the trend of the matching percentage for various IoU thresholds in face detection experiment. This suggests that the majority of the prediction boxes those do not attend to faces, attend to human bodies.

4. DISCUSSION AND CONCLUSION

Why localizing to faces?: It can be speculated that the network is learning to recognize salient motions or actions rather than just localizing faces in each frame independently. A face with moving lips is one such salient action, which is present in the majority cases for the current scenario of movies. The localization to human body parts

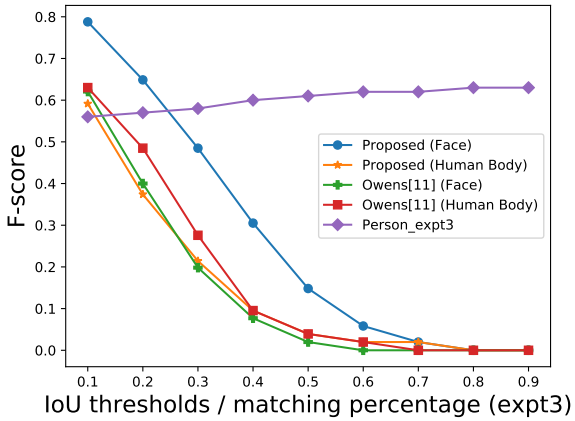

Fig. 5: Trend of $F-score$, for different experiments and matching percentage for expt 3

can be explained as the part of body gestures or actions considered salient by the network. We hypothesize that in ideal scenarios, in order to decide for speech regions in a video, the network should consider the talking faces as the most salient action. Since the VAD performance attained by the network is just 66%, it is still not close to the ideal case. One reason is that visual information relevant to human speech-activity (such as talking faces) may not be available in the video.

Our work is an initial effort toward incorporating visual information from videos about potential speech activity, with no explicit assumptions or steps about talking faces, for the VAD task. We proposed a network which decentralizes the temporal context and thus is able to learn from longer contexts. The learned network maps can be easily visualized for a better understanding of the representations. We analyzed the learned representations and showed that the network indeed automatically captures human faces as salient to decide for the presence of speech activity. This work opens up a new approach to the challenging problem of speaking face detection in videos. One immediate future next step to this work is to include audio to enhance the VAD performance.

5. REFERENCES

[1] Juan Manuel Górriz, Javier Ramírez, Elmar Wolfgang Lang, Carlos García Puntonet, and Ignacio Turias, "Improved likelihood ratio test based voice activity detector applied to speech recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 664–677, 2010.

[2] Javier Ramírez, José C Segura, Juan M Górriz, and Luz García, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2177–2189, 2007.

[3] E Verteletskaya and Kirill Sakhnov, "Voice activity detection for speech enhancement applications," *Acta Polytechnica*, vol. 50, no. 4, 2010.

[4] Pavol Harár, Radim Burget, and Malay Kishore Dutta, "Speech emotion recognition with deep learning," in *Signal Processing and Integrated Networks (SPIN), 2017 4th International Conference on*. IEEE, 2017, pp. 137–140.

[5] Fei Tao and Carlos Busso, "Bimodal recurrent neural network for audiovisual voice activity detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1938–1942.

[6] Peng Liu and Zuoying Wang, "Voice activity detection using visual information," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–609.

[7] Rajitha Navarathna, David Dean, Sridha Sridharan, Clinton Fookes, and Patrick Lucey, "Visual voice activity detection using frontal versus profile views," in *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*. IEEE, 2011, pp. 134–139.

[8] Andrew Aubrey, Bertrand Rivet, Yulia Hicks, Laurent Girin, Jonathon Chambers, and Christian Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filtering," in *Signal Processing Conference, 2007 15th European*. IEEE, 2007, pp. 2409–2413.

[9] Bart Joosten, Eric Postma, and Emiel Krahmer, "Visual voice activity detection at different speeds," in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[10] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (CASSP*. IEEE, 2002, pp. II–2017.

[11] F. Patrona, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Visual voice activity detection in the wild," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 967–977, June 2016.

[12] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International Conference on Machine Learning*, 2018, pp. 2673–2682.

[13] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 618–626.

[15] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2018, pp. 839–847.

[16] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[17] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.

[18] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] C Lawrence Zitnick and Piotr Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*. Springer, 2014, pp. 391–405.

[22] Andrew Owens and Alexei A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.