Supplementary Material for “CAD - Contextual Multi-modal Alignment for Dynamic AVQA”

Asmar Nadeem[1], Adrian Hilton[1], Robert Dawes[2], Graham Thomas[2], Armin Mustafa[1]
[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom.
[2]BBC Research and Development, United Kingdom.

Datasets and Implementation Details

Dataset. For AV fine temporal alignment-based pre-training, we employ ACAV100M [3] dataset. The ACAV100M dataset is a very large-scale collection of 100 million audio-visual clips designed for audio-visual representation learning. It covers diverse topics including human sounds, music, animal, and nature sounds, etc and facilitates the exploration of synchronisation, alignment, and semantic associations between audio and visual streams. We leverage this dataset to advance the state-of-the-art in dynamic audio-visual question answering (AVQA), one of the toughest applications in audio-visual learning, by pre-training our novel network on the music category of this dataset.

Implementation. The music category in ACAV100M dataset contains around 26.3 million videos and we sample 6 million videos out of these. Each video has a length of 10 seconds and for our pre-training task, we stitch 6 videos together. Then, we create 60 cues of one second each and in the next step, we extract features of audio and visual streams using PANNs [2] and ViT [1] respectively.

This dataset does not contain any annotations and for question queries, first, we use GIT_L [7], with standard settings, to generate a caption for each 10-second video out of the total 6 videos which are stitched into one. In the next step, we generate a question query as did in BEIR [6] from each caption. In this manner, we end up with 6 question queries for a combined 60-second video where each query belongs to each 10-second clip. We use question queries in a way that when we send a positive pair (audio and visual of the same cue) as input, a question query of that 10-second clip that contains the cue is employed. In the case of the negative pair (audio and visual of the different cues), we alternatively use the question query of either the audio stream clip or the visual stream clip if they belong to different clips. Here, the question query is encoded using [5].

We use selection probability of 60% and 40% (60-40) for the positive and negative pairs respectively. In Figure 1, we demonstrate this by evaluating different combinations of positive-negative pairs' selection probability in %. We start with a 90-10 split and pre-train the model till 10-90. For each split configuration, we evaluate it by training the model and then, testing it to find the overall accuracy. The model achieved the highest accuracy for the 60-40 split as shown in Figure 1.

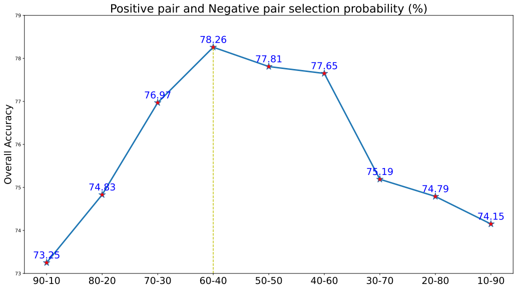

Figure 1. Effect of positive and negative pairs selection probability on the overall performance.

Another contribution to our work is the Contextual Block where we select 80% of the visual features to be passed through it. This is based on the experiments as well as shown in Figure 2. We iterate from 10% to 100% where initially there is no effect on the performance but it picks up from 40% to 80% and beyond that it decreases slightly.

The last of these experiments is the selection of % of either contextual or non-contextual features for masking within the Contextual Block. In this work, we zero out or mask 90% based on the experimentation shown in Figure 3.

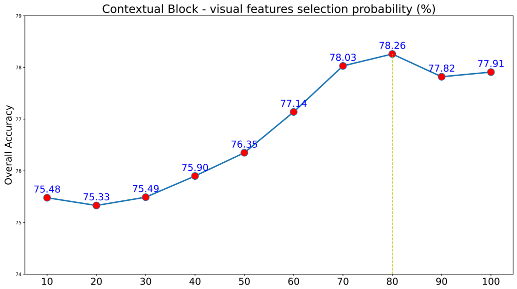

Figure 2. Effect of % of visual features to go through Contextual Block on the overall performance.

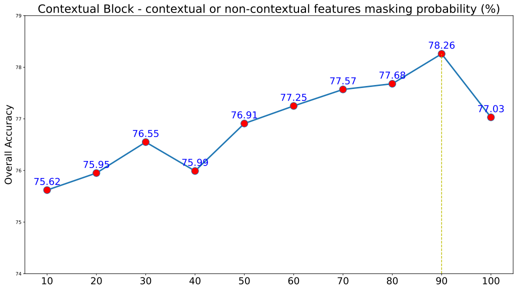

Figure 3. Effect of % of contextual or non-contextual features masked in the Contextual Block on the overall performance.

We also calculate the GPU hours of pre-training and training stages for our method (CAD) as well as the state-of-the-art (SOTA) [4] to analyse the effect of the Contextual Block. As shown in Table 1, GPU hours decrease for both the pre-training and training stages for our method as well as the SOTA when implemented with the Contextual Block. In this case, we implement the SOTA with the pre-training and also, the Contextual Block.

Task	Pre-training (GPU hours)	Training (GPU hours)
w/o Contextual Block - ST-AVQA (SOTA) [4]	205.7	2.1
w Contextual Block - ST-AVQA (SOTA) [4]	131.2	1.6
w/o Contextual Block - CAD (Ours)	257.1	2.4
w Contextual Block - CAD (Ours)	149.4	1.7

Table 1. w/o Contextual Block and w Contextual Block describe the effect of without and with the Contextual Block on the employed compute for the SOTA [4] as well as our method CAD.

By comparing the first and last row of Table 1, our method demonstrates efficiency over the SOTA. When coupled with the Contextual Block, the SOTA also demonstrates more efficiency as shown in row 2 but the performance is still lower than our method as demonstrated in Table 2 of the main paper.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 1

[3] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 1

[4] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 2

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1

[6] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021. 1

[7] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1