MASTAF: A Model-Agnostic Spatio-Temporal Attention Fusion Network for Few-shot Video Classification

Xin Liu Huanle Zhang Hamed Pirsiavash Xin Liu
University of California, Davis
{rexliu,dtzhang,hpirsiav,xinliu}@ucdavis.edu

Appendix

A. Positive cases Analysis

In the main paper, Table 3 shows that the attention fusion mechanism can take advantage of the self-attention module and the cross-attention module to improve the performance of our model. Figure 1 displays three positive cases in SSv2-all[1]. The top figure of each sub-figure shows the angle between representations of each class's prototype and the query video in the feature space. These representations are closer if the angle is smaller. The bottom figure of each sub-figure displays the extracted frames from the videos corresponding to the top figure. For example, the label of the query video in Figure 1a is 'pretending to put something underneath something'. The cross-attention module predicts S^5 is the most similar to the query video. However, the real label of S^5 is 'picking something up'. In contrast, the self-attention predicts S^1 is the most similar to the query video, which is correct. And after fusing these two attention mechanisms, our MASTAF predicts correctly. Figure 1b shows the opposite situation in which the cross-attention module predicts correctly and the self-attention module predicts wrong. Figure 1c shows a positive case of fusion's ability to predict correctly even when the self-attention module and the cross-attention module are both wrong. Figure 1 demonstrates the effectiveness of our MASTAF model.

Table 1: Comparison results between the MASTAF without multi-task training setting and MASTAF for 5-way 1-shot video classification

Method	UCF101	SSv2-all
MASTAF-No-Global	89.4	49.5
MASTAF	**90.6**	**50.3**

B. Multi-task learning setting

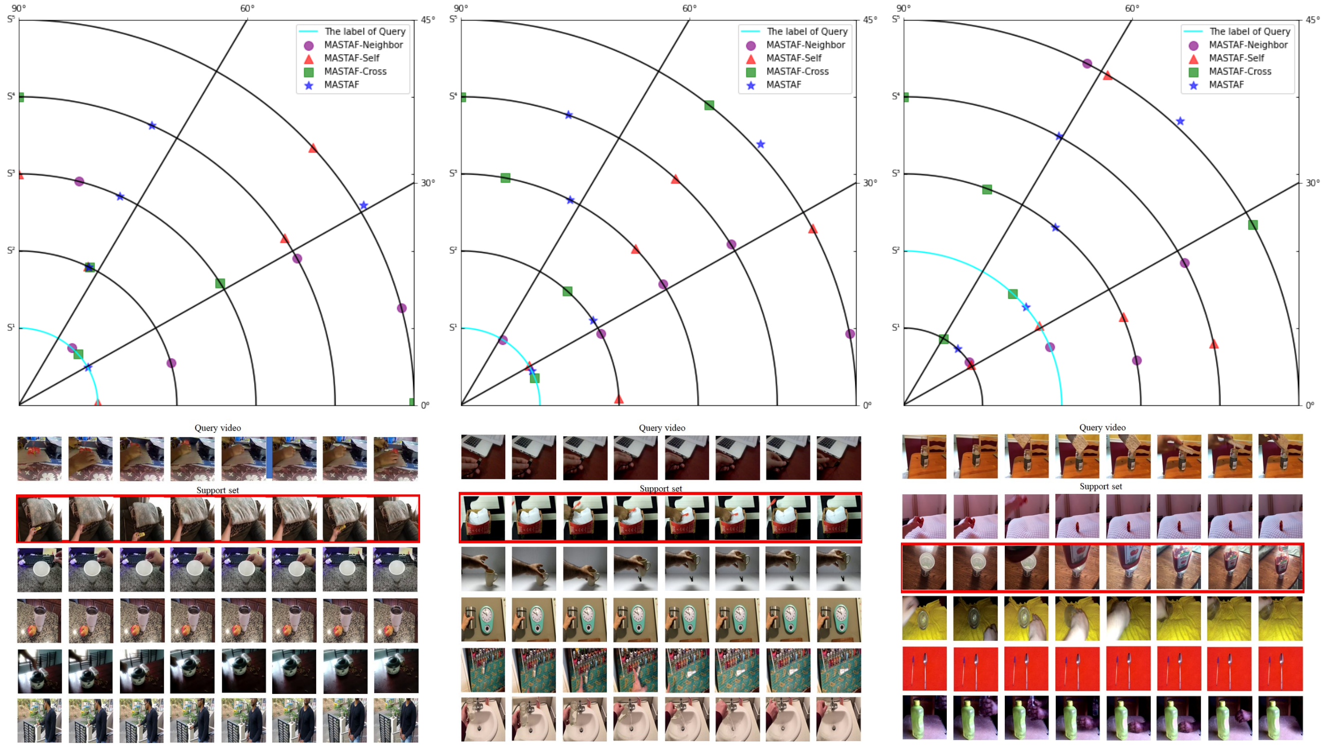
We add a global video classification task in the multi-task learning setting. Table 1 shows the comparison results in which we fixed other hyperparameters but without global video classification task in the baseline model(MASTAF-No-Global). From the results, we can see that the global classification task improves the performance, which demonstrates the benefits of the multi-task learning setting. We argue that the global classification task using the representations from the cross-attention module reduces the risk of overfitting for the nearest neighbor classification task in the training dataset and generates a general representation for unseen class in a few-shot scenario.

Table 2: Comparison results between MASTAF-NoML-Mean and MASTAF for 5-way 1-shot video classification

Method	UCF101	SSv2-all
MASTAF-NoML-Mean	89.9	49.3
MASTAF	**90.6**	**50.3**

C. Meta-learner

We evaluate the influence of meat-learner in the MASTAF by developing a model without the meta-learner,i.e., MASTAF-NoML-Mean. In MASTAF-NoML-Mean, we use the average pooling on each relation map(M^{self} in Eq 2 in the main paper) and correlation map($M^{cross}_{S_q \leftarrow S^c}$ in Eq 6 in the main paper and $M^{cross}_{S^c \leftarrow S_q}$ in Eq 7 in the main paper) as the kernel to compute the attention map in each self-attention module and cross-attention module. As we can see from Table 2, our MASTAF with meta-learner outperform MASTAF-NoML-Mean, which means the meta-learner dynamically generates the kernel to summarize the local features in each relation and correlation map.



(a) SSv2-all: Pretending to put something underneath something.Only MASTAF-Self and MASTAF predict correctly.

(b) SSv2-all: Putting something on a surface. Only MASTAF-Cross and MASTAF predict correctly.

(c) SSv2-all: Failing to put something into something because something does not fit.Only MASTAF predicts correctly.

Figure 1: Examples for MASTAF-Neighbor, MASTAF-Self, MASTAF-Cross, and MASTAF. The top figure in each subfigure shows the angle between each class prototype of each model and query video. A horizontal line indicates that the angle between the representation of the support class's prototype and the representation of query video is 0°, which means that the similarity score is 1. The vertical line indicates that the angle between the representation of the support class's prototype and the representation of query video is 90°, which means that the similarity score is 0. For comparison purposes, we specify that the representation of each class is on the same arc. Different symbols represent prototype representations extracted from different models and dimensionality reduction by t-Distributed Stochastic Neighbor Embedding.The bottom figure in each sub-figure shows the frames extracted from the query video and support set corresponding to the top figure in the same sub-figure. In the bottom figure, videos of classes from S^1 to S^5 are shown from top to bottom. The highlighted video(Red) in the support set has the same label as the query video. This figure shows that the fusion mechanism can take advantage of the self-attention module and cross-attention module to extract more discriminative spatiotemporal representations.

Table 3: Comparison results between MASTAF-NoRes and MASTAF for 5-way 1-shot video classification

Method	UCF101	SSv2-all
MASTAF-NoRes	88.9	49.2
MASTAF	**90.6**	**50.3**

D. Residual structure in the attention network

To verify the effectiveness of residual structure in the attention network, we create a baseline model, i.e., MASTAF-NoRes, in which we remove the residual design in both the self-attention module and cross-attention module. The result in Table 3 shows that our MASTAF outperforms the MASTAF-NoRes, which demonstrates the residual structure is beneficial for few-shot video classification because it helps to remain the similar representation for the videos from the same classes and call attention to the minor differences for videos from the different classes.

References

[1] K. Cao, J. Ji, Z. Cao, C. Chang, and J. Niebles. Few-shot video classification via temporal alignment. pages 10615–10624, 2020.