Effectiveness of Max-Pooling for Fine-Tuning CLIP on Videos

Supplementary Material

1. Implementation Details

The bulk of our experiments are conducted using a ViT-B/16 model architecture, unless otherwise specified. Both the image and text encoder of the model are initialized with the corresponding pretrained CLIP weights. We perform standard preprocessing of the data, including resizing frames to 224x224 pixels, and normalizing pixel values. Unless otherwise specified, all results use 32 frames with 1 spatial crop and 1 clip and are evaluated using single-view inference.

Our ViT-B/16 models are trained using 4 NVIDIA Tesla V100 GPUs while our ViT-L/14 models are trained using 4 A100 GPUs. All models use mixed precision training. All models are optimized using the AdamW optimizer with weight decay 0.001 and cosine learning rate scheduler with 5 warm-up epochs.

1.1. Base to Novel Setting

In this setting proposed by [5], a dataset is split into non-overlapping base and novel classes. The model is trained on the base classes and evaluated both on base and novel classes. In our evaluation, the base classes have three splits, thus, a model is trained on each separate split of the data and the performance on the the base/novel validation set is averaged. In our experiments, we use an initial learning rate of 2.e-06, which is decayed according to a cosine learning rate scheduler. On the SSv2 [2] dataset, the model is trained for 15 epochs with an effective batch size of 16. On the K-400 [1] dataset, the model is trained for 20 epochs with an effective batch size of 32.

1.2. Few Shot Setting

In the few-shot setting, K videos are randomly sampled from each category during training, where K is 2, 4, 8, and 16. Each model is evaluated on the same validation set. In this setting, the initial learning rate is 2.e-06 and the model is trained for 25 epochs with an effective batch size of 16 for the SSv2 dataset and 50 epochs with an effective batch size of 32 for the UCF101 and HMDB51 datasets.

1.3. Fully Supervised Setting

In this setting, the initial learning rate is 2.2e-05 and the model is trained for 30 epochs with an effective batch size of 256.

2. Evaluation Protocol

The base to novel evaluation proposed by [5] consists of a base split and a novel split of the data. These splits are determined by sorting the classes of each data set by frequency of the classes. The sorted classes are then split into two equal halves. The base classes are assigned the first half, corresponding to the most frequently occurring classes, while the novel classes are assigned the second half of the classes, corresponding to the less frequent and hence *novel* classes.

The model is trained on the base classes, which are further subdivided into three splits, each containing randomly sampled 16-shots of every class. Each of these models is evaluated on the respective validation set and the performance is averaged.

3. Additional Experiments

To understand the impact of our approach when the model is pre-trained to elicit temporal information, we initialize our model with a CLIP variant pretrained on video and text data as proposed by [7]. This model uses the CLIP [4] architecture along with a contrastive loss between the video and text embeddings. However, it modifies the native spatial attention in the image encoder to instead use spatiotemporal attention. Furthermore, the pre-training applies random masking to the video input.

In comparison to the model's performance when the features are aggregated using mean-pooling, we observe a significant increase in the base classes (+2.8 from 19.5 to 22.3). Notably, however, the performance is significantly reduced when the model is initialized using this variant of CLIP (-0.8 from 20.3 to 19.5) when aggregating the features using mean-pooling.

Table 1. Comparison of the top 1 accuracy when initializing the model with video-text pretraining [7]. This shows the results when either mean-pooling or max-pooling are used to aggregate the features and when our max-pooling token approach is applied for the last s layers of the Transformer.

	SSv2		
Method	Base	Novel	HM
mean	19.5	15.2	17.1
max	22.0	15.7	18.3
$\text{max}_{s=8}$	22.3	16.0	18.6

4. Datasets

We evaluate our proposed approach on four widely used benchmarks, Kinetics-400 [1], Something-Something V2 (SSV2) [2], HMDB51 [3], and UCF101 [6]. Notably, the

Kinetics-400 is a large-scale dataset of YouTube video clips. It consists of 400 action classes and is split into 240,436 training videos and 19,796 validation videos. Each clip is about 10 seconds long.

The SSV2 dataset, on the other hand, focuses on human-object interactions. It consists of 174 action classes and is split into 168,913 training videos and 24,777 validation videos. Given that the spatial bias in these videos is less exploitable, performance on this benchmark is a good indicator of the model's ability to extract temporal features.

References

[1] J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2017. 1

[2] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017. 1

[3] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[5] Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[6] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[7] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2023. 1