

CLIPPING: DISTILLING CLIP-BASED MODELS FOR VIDEO-LANGUAGE UNDERSTANDING

Anonymous authors

Paper under double-blind review

1 SUPPLEMENTARY MATERIALS

1.1 IMPLEMENTATION DETAILS

Initialization. We initialize the vision encoder MobileViT-v2 by pre-training it on the imageNet21k dataset. The initial parameters of the text encoder are a copy of the CLIP’s text encoder, and the temporal Transformer is initialized in the same way as in CLIP4clip. The rest of the parameters, e.g., the linear projections in AS2OT KD, are initialized randomly from the Gaussian distribution $\mathcal{N}(0, 1)$.

Training Settings. The text encoder is fine-tuned with a small learning rate ($1e-7$) at the beginning. The learning rates of the other parts are initialized by $1e-5$ and decay according to the cosine schedule. The whole KD is optimized by Adam with a batch size of 64 for 36 epochs. From the 1st epoch to the 3rd epoch, the weight δ for AS2OT KD is set to 0, and then AS2OT KD is added gradually during the 4th epoch to the 36th epoch with $\delta = \frac{1}{7}(i-3)$, $4 \leq i \leq 10$ and $\delta = 1$, $i \geq 10$. For the whole training period, we set the balance weights α , β and γ to 1, 1 and 0.25, respectively. In our AS2OT KD, we choose 4 MobileViT-v2 (student) layers ($layer_2$, $layer_3$, $layer_4$ and $layer_5$) and 12 CLIP (teacher) layers (all the 12 Transformer layers in ViT-B-32). Note that the previous AT2OS KD (Chen et al. (2021)) in Table 3 is also trained with the same selected layers. In our experiments, the masks are calculated through Eq. 4 with $n_0 = 2$, $m_0 = 4$, $n_1 = 3$ and $m_1 = 9$.

1.2 ADDITIONAL EXPERIMENTS

LSMDC and MSVD datasets. For the experiments of video-language retrieval, we also compare our CLIPPING with the state-of-the-art Frozen (Bain et al. (2021)), MDMMT (Dzabaraev et al. (2021)), NoiseEst (Amrani et al. (2021)) and SupportSet (Patrick et al. (2021)) on the LSMDC and MSVD datasets. In Table 4, CLIPPING again obtains the best performance with fewest parameters and Flops. Since some models are not available publically, we estimate their parameters and Flops according to their backbones. In Table 5, we also compare the student of CLIPPING with its teacher. CLIPPING with MobileViTv2 as the vision encoder without any vision-language pre-training achieves 91.5%–92.9% of the performance of its teacher on the LSMDC and MSVD datasets.

Different Students. Besides MobileViT-v2, we also use other models as the student in CLIPPING, which are EfficientNet-b0 (Tan & Le (2019)) and EfficientFormer-L1 (Li et al. (2022)). Table 6 shows that CLIPPING with each of these small models as the student all performs well, indicating that it is a general KD method.

1.3 ADDITIONAL VISUALIZATION

AT2OS KD vs. AS2OT KD. We compare the features of the student that is trained with AT2OS KD and our AS2OT KD in Fig. 7. We randomly select 4 features from $layer_2$ and 25 features from $layer_5$ in MobileViT-v2, which respectively represent the lower and higher features of the student. Note that both the features with different KD types are selected from the same locations. We can see that the features with our AS2OT KD maintain sharper low-level structure features, such as edges and lines, and AS2OT is able to catch more effective feature maps with less noise in higher layers, while many feature maps of AT2OS are saturated or close to zero.

Table 4: Comparison with the state-of-the-art on the LSMDC and MSVD datasets.

Model	PT Datasets	Vision Encoders	Params	Flops	$R@1$
LSMDC dataset					
NoiseEst	HT100M	Resnet152, ResNext-101	>100M	>80G	6.4
SupportSet	HT100M	ResNet-152, R(2+1)D-34	>100M	>80G	-
Frozen	C3M, W2M	Space-Time	232M	210G	15.0
MDMMT	C400M, AudioSet	CLIP _{vision} , VGGish	226M	200G	17.2
CLIPPING (our)	IN21K	MobileViT-v2	78.1M	23G	19.9
MSVD dataset					
NoiseEst	HT100M	Resnet152, ResNext-101	>100M	>80G	20.3
SupportSet	HT100M	ResNet-152, R(2+1)D-34	>100M	>80G	28.4
Frozen	C3M, W2M	Space-Time	232M	210G	33.7
MDMMT	C400M, AudioSet	CLIP _{vision} , VGGish	226M	200G	-
CLIPPING (our)	IN21K	MobileViT-v2	78.1M	23G	42.9

Table 5: CLIPPING results on the LSMDC and MSVD datasets. CLIP_{vision} is the teacher from CLIP4clip and MobileViT2 is the student.

Vision Encoder	PT Dataset	Params	Flops	$R@1$	$R@5$	$R@10$
LSMDC dataset						
CLIP _{vision}	C400M	87.7M	8.6G	21.6	41.8	49.8
MobileViT2	IN21K	4.5M	1.4G	19.9 (92.1%)	38.6 (92.3%)	45.6 (91.6%)
MSVD dataset						
CLIP _{vision}	C400M	87.7M	8.6G	46.2	76.1	84.6
MobileViT2	IN21K	4.5M	1.4G	42.9 (92.9%)	69.9 (91.9%)	77.4 (91.5%)

Table 6: CLIPPING with different students.

Vision Encoder	Params	Flops	$t2vR@1$	$v2tR@1$
EfficientNet-b0	4.9M	0.4G	39.3	38.9
EfficientFormer-L1	12.0M	1.3G	40.5	39.8

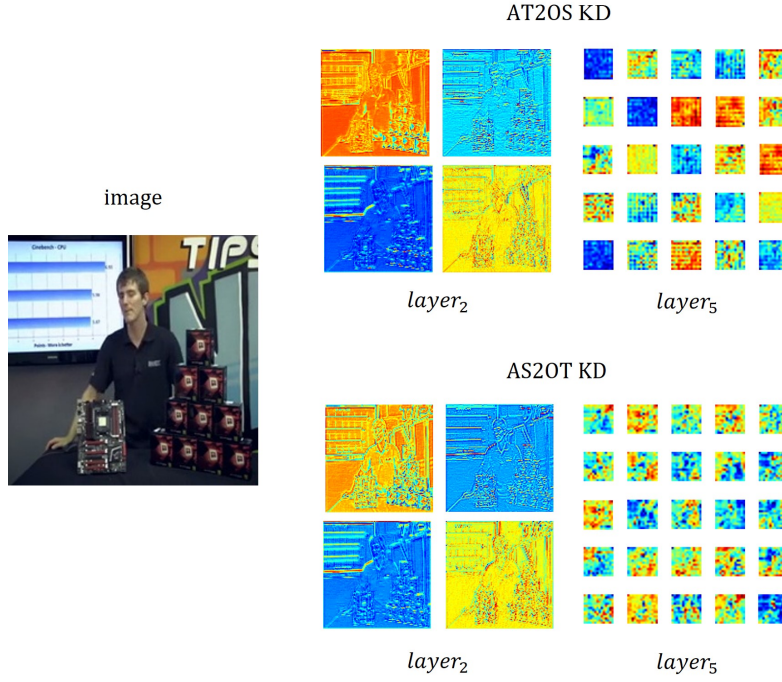


Figure 7: Examples of the features of the student that is trained with AT2OS KD (Chen et al. (2021)) and AS2OT KD (ours).

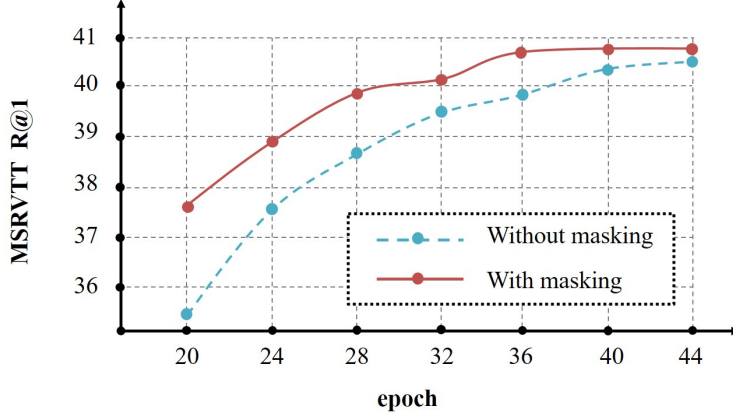


Figure 8: Speeds of convergence that trained with and without masking.

Masking. In Fig. 8, we show the effect of the masking. It can be seen that the CLIPPING with the masking reaches the highest accuracy at the 36th epoch, while CLIPPING without the masking needs to be trained with more epochs for better accuracy. It verifies that the masking is able to speed up the training procedure.

Local Video-Caption Distribution Alignment (LVCDA). In Fig. 9, we visualize the frame-word alignments. We can find that with LVCDA, the results show clearer frame-word attentions. In Fig. 9(a), without LVCDA, only global attention from [SEP] is strongly activated, while with LVCDA, many correct frame-word attentions are activated. Similar phenomena are found from other examples.

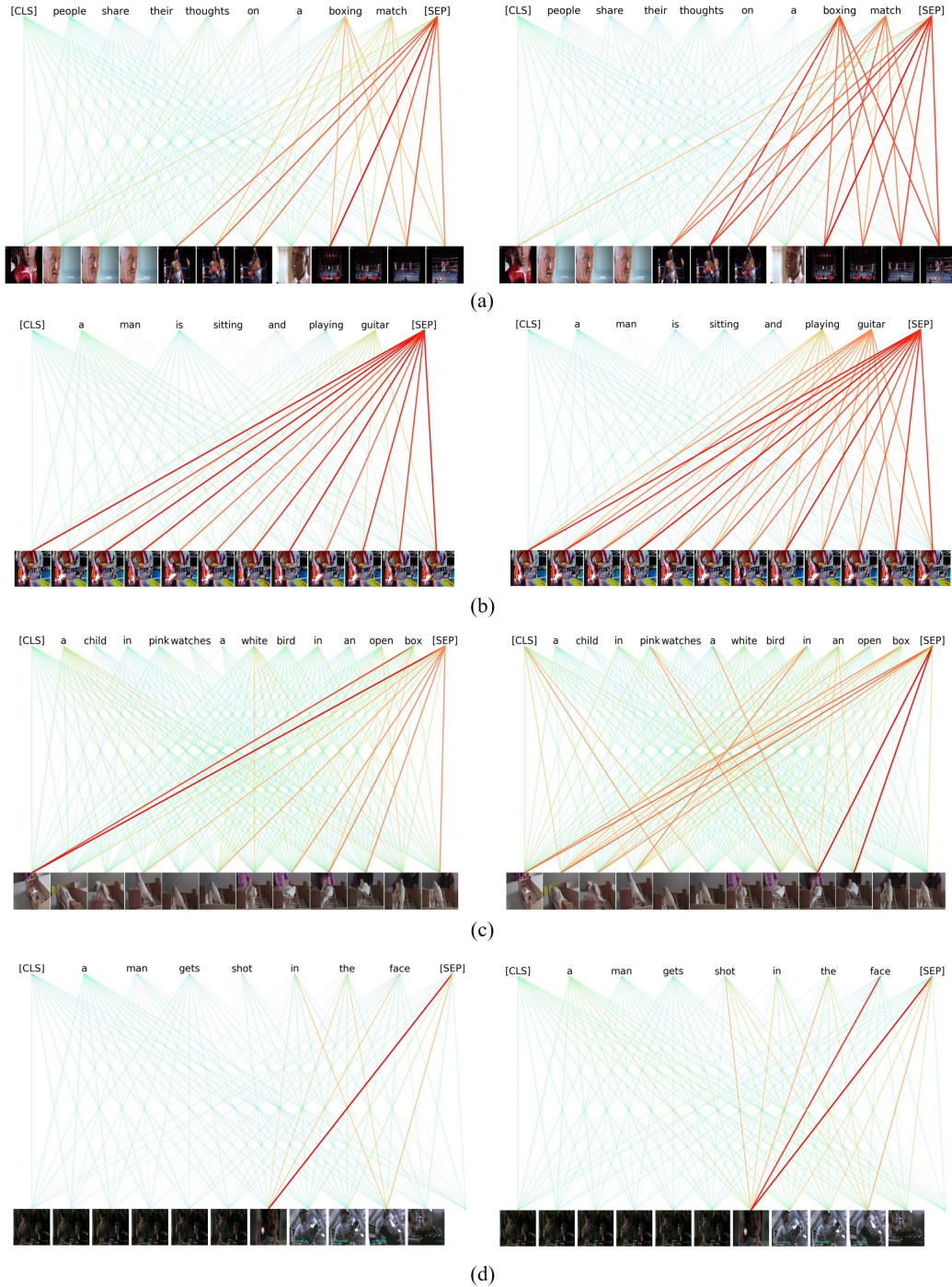


Figure 9: Visualization of the frame-word alignment. The 1st column shows the alignment results without the local video-caption distribution alignment and the 2nd column shows the results with this local alignment.

REFERENCES

- Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *AAAI*, 2021.
- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *ICCV*, 2021.
- Defang Chen, JianPing Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Distillation with semantic calibration. In *AAAI*, 2021.
- Maksim Dzabaraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. MDMMT: Multidomain Multimodal Transformer for Video Retrieval. In *CVPR*, 2021.
- Yanyu Li, Geng Yuan, Yang Wen, Eric Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. EfficientFormer: Vision Transformers at MobileNet Speed. In *arXiv:2206.01191*, 2022.
- Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021.
- Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *PMLR*, 2019.