NOAH'S ARK LAB

CLIPPING: Distilling CLIP-Based Models with a Student Base for Video-Language Retrieval

Renjing Pei, Jianzhuang Liu, Weimian Li,

Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, Youliang Yan

Huawei Noah’s Ark Lab

Introduction

1. Pre-trained vision-language models (PVLMs) with the Transformer architecture (especially the vision stream) usually **consume a huge amount of computation**, making them **difficult to be deployed on edge devices such as mobile phones**.

2. Recent small models show that combining Convolutional Neural Networks (CNNs) and Transformers as a hybrid architecture gets the best of both architectures, but **the overall performance of these works is still far from satisfactory when compared to large pre-training models**.

3. In NLP, the knowledge distillation methods are usually performed **at both the pre-training and the fine-tuning stages.** However, the collection of the pre-training data for the pre-training knowledge distillation **cost huge manpower in multi-modality applications**.

In this paper, we propose a novel knowledge distillation method, named CLIPPING, where the plentiful knowledge of a large teacher model Clip4clip that has been fine-tuned for video-language tasks with the powerful pre-trained CLIP **can be effectively transferred to a small student only at the fine-tuning stage**.

Relative Work

Knowledge Distillation (KD)

Earlier works transfer knowledge embedded in the "logits" learned in a large teacher model to a small student model without sacrificing much performance.

Intermediate Features' KD

Recent works (in Figure 1) use Intermediate layers of the teacher to supervise student, for example in Figure 1(b), each layer of the student learns the knowledge from multiple layers of the teacher.

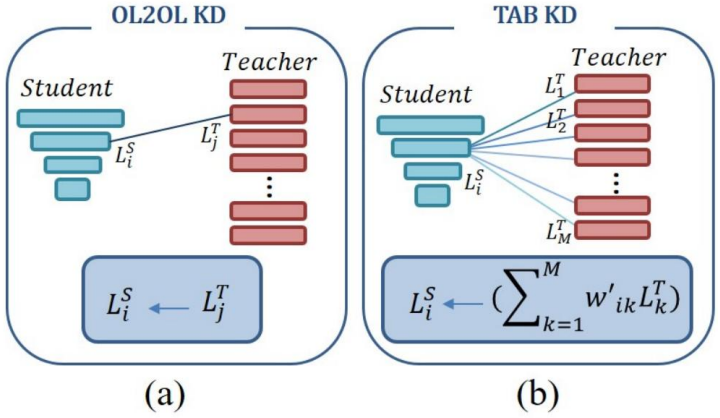


Figure 1. (a) OL2OL; (b) TAB

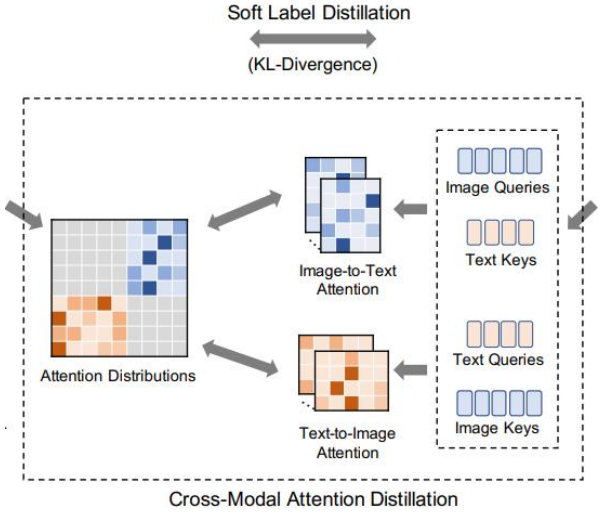


Figure 2. CMAD

Multi-Modality KD

CMAD (Cross-Modal Attention Distillation) designs a fusion-encoder model as the teacher and introduces cross-modal attention knowledge to train the dual-encoder student model. The distillation objective is applied at both the pre-training and the fine-tuning stages and helps the dual-encoder model learn interactions of different modalities.

Contributions of CLIPPING

- We introduce an efficient approach to distill both the vision knowledge (Intermediate Features) and the cross-modality knowledge from teacher to a small model at the fine-tuning stage.
- We propose a new layer-wise alignment scheme, called Student-As-Base (SAB), where the student's layers can be regarded as the bases of the teacher's feature space, forcing the student model to full absorb the knowledge of the teacher.
- We present an effective cross-modal knowledge distillation, which includes knowledge from both the global and local video-caption distributions.
- CLIPPING achieves 91.5%–95.3% of the performance of its teacher with its vision encoder being 19.5x smaller. CLIPPING also significantly outperforms a state-of-the-art small baseline. CLIPPING is comparable or even superior to many large pre-training models.

Overview of CLIPPING

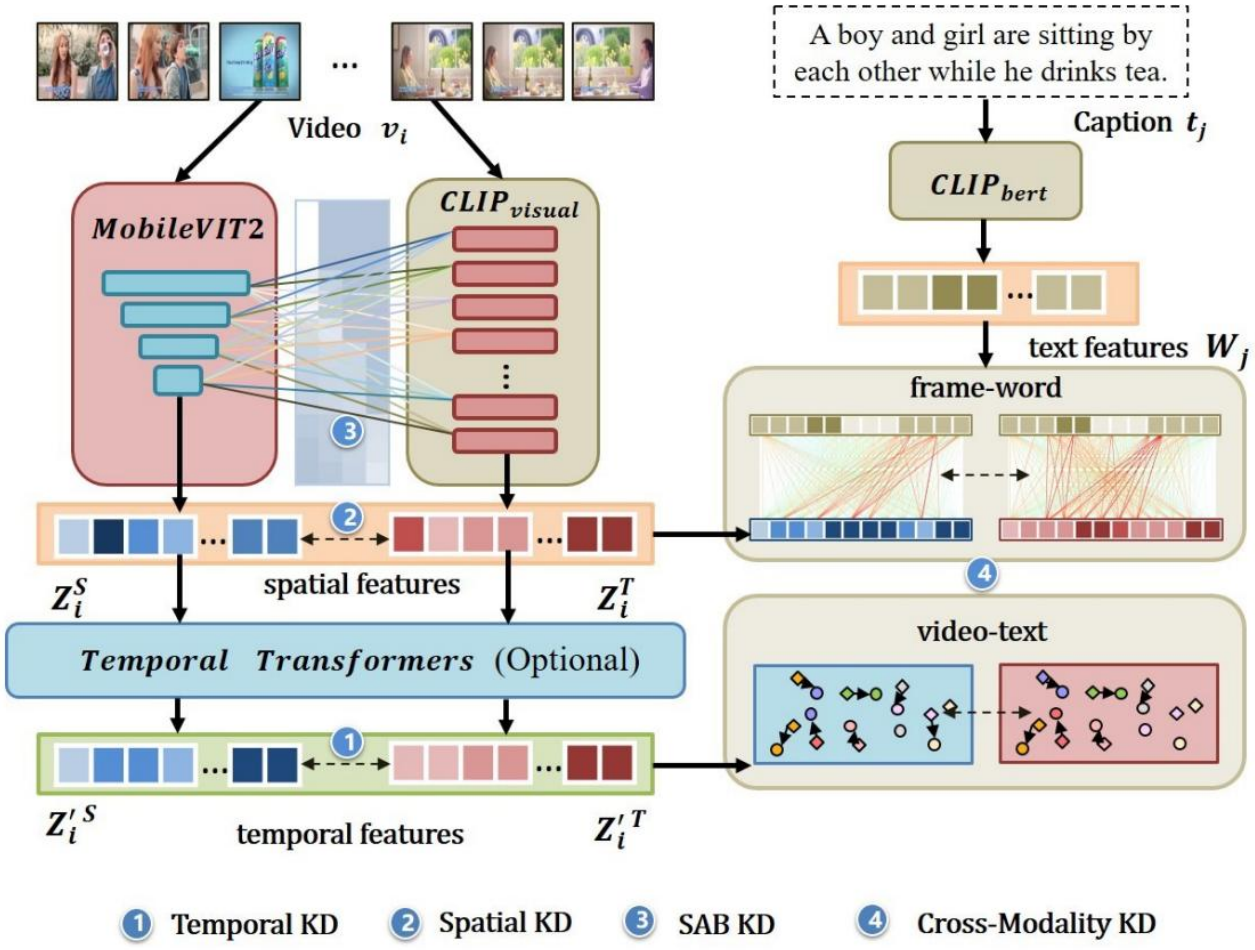


Figure 3. CLIPPING

There are mainly four knowledge distillation (KD) parts: (1) Temporal KD. (2) Spatial KD. (3) SAB (Student-As-Base) KD. (4) Cross-modality KD. The CLIP's vision encoder is a Transformer.

Our novelties and improvements mainly come from SAB KD and cross-modal KD.

No.	T	S	***SAB***	***CM***	$t2vR@1$	$v2tR@1$
A	✓	✓			33.0	32.8
B			✓	✓	38.8	38.5
C		✓	✓	✓	39.4	39.1
D	✓		✓	✓	40.4	40.0
E	✓	✓		✓	35.2	34.4
F	✓	✓	✓		37.6	36.2
G	✓	✓	✓	✓	40.7	40.2

SAB KD

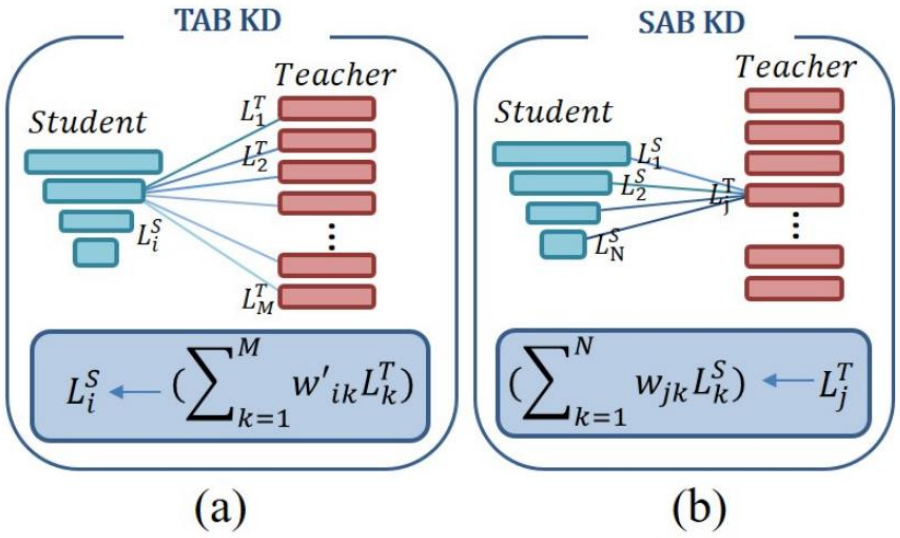


Figure 4. (a) TAB; (b) SAB

Teacher-As-Base (TAB) KD uses all the teacher's layers to supervise each layer in the student, where each layer of the student learns the knowledge from all the selected layers of the teacher. Our Student-As-Base (SAB) KD enables each of the teacher's layers to pass its knowledge to all the student's layers.

Analysis of TAB and SAB

For simplicity, suppose M = N = 2. Then, for SAB:

$$\begin{cases} L_1^T = w_{11} L_1^S + w_{12} L_2^S, \; w_{11} + w_{12} = 1 \\ L_2^T = w_{21} L_1^S + w_{22} L_2^S, \; w_{21} + w_{22} = 1 \end{cases}, \quad (1)$$

and for TAB,

$$\begin{cases} L_1^S = w'_{11} L_1^T + w'_{12} L_2^T, \; w'_{11} + w'_{12} = 1 \\ L_2^S = w'_{21} L_1^T + w'_{22} L_2^T, \; w'_{21} + w'_{22} = 1 \end{cases}. \quad (2)$$

Eq. 2 can be converted to:

$$L_1^T = A \cdot L_1^S + B \cdot L_2^S, \; L_2^T = C \cdot L_1^S + D \cdot L_2^S, \quad (3)$$

$$A = \frac{w'_{22}}{w'_{11}w'_{22} - w'_{12}w'_{21}}, \; B = \frac{-w'_{12}}{w'_{11}w'_{22} - w'_{12}w'_{21}},$$

$$C = \frac{-w'_{21}}{w'_{11}w'_{22} - w'_{12}w'_{21}} \text{ and } D = \frac{w'_{11}}{w'_{11}w'_{22} - w'_{12}w'_{21}}$$

Although it seems that SAB and TAB have the same matrix representation, TAB is unable to get a similar result as SAB, because the coefficients are easily arbitrary large during optimization if it is converted to SAB's form. For this reason, TAB KD is more likely to get to a local minimum during training.

Cross-Modality (CM) KD

We characterize the cross-modal distributions from two perspectives, global video-caption distributions of the teacher and local video-caption distributions of CLIP.

Global Alignment

$$A_{GVC} = \sigma \begin{pmatrix} s(v_1,t_1) & \dots & s(v_1,t_B) \\ \dots & \dots & \dots \\ s(v_B,t_1) & \dots & s(v_B,t_j) \end{pmatrix},$$

$$A_{GCV} = \sigma \begin{pmatrix} s(t_1,v_1) & \dots & s(t_1,v_B) \\ \dots & \dots & \dots \\ s(t_B,v_1) & \dots & s(t_B,v_B) \end{pmatrix}.$$

$$L_G = D_{KL}(A^S_{GVC}, A^T_{GVC}) + D_{KL}(A^S_{GCV}, A^T_{GCV}).$$

Local Alignment

$$A_{LVC} = \sigma \begin{pmatrix} s'(v_1,t_1) & \dots & s'(v_1,t_B) \\ \dots & \dots & \dots \\ s'(v_B,t_1) & \dots & s'(v_B,t_B) \end{pmatrix},$$

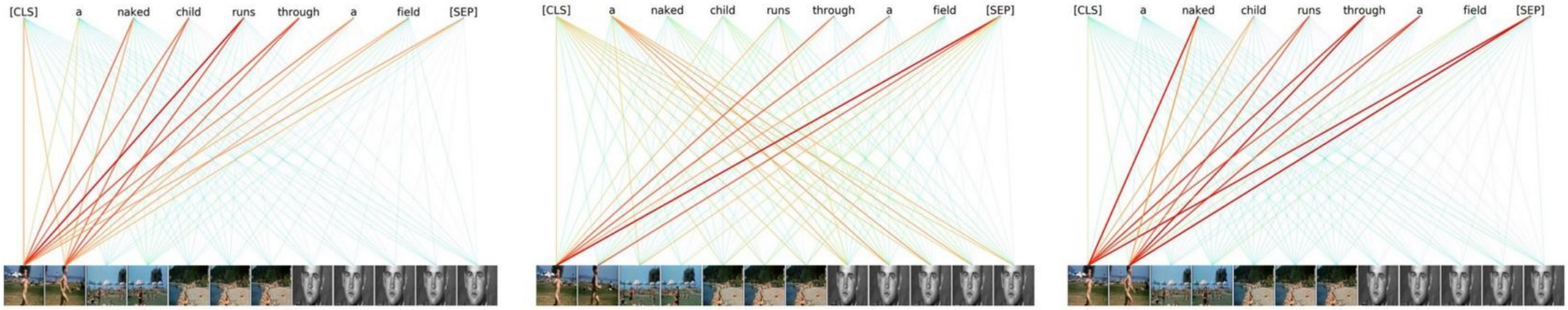
$$A_{LCV} = \sigma \begin{pmatrix} s'(t_1,v_1) & \dots & s'(t_1,v_B) \\ \dots & \dots & \dots \\ s'(t_B,v_1) & \dots & s'(t_B,v_B) \end{pmatrix},$$

$$L_L = D_{KL}(A^S_{LCV}, A^T_{LCV}) + D_{KL}(A^S_{LVC}, A^T_{LVC}).$$

CLIP uses text prompts (such as “A picture of a ()”) for zero-shot image classification. CLIP fills them with different words (e.g., “cat” and “dog”) and results in different captions (e.g., “A picture of a cat” and “A picture of a dog”). It can match the captions to the corresponding images, showing some image-word alignment ability. We transfer this pre-training knowledge to the student through a local frame-word alignment.

SAB CM and CM KD

Our experiment shows that only when joint trained with SAB, our local video-caption distribution alignment shows its advantage, which also verifies the effectiveness of SAB.



(a) (b) (c)

Figure 5. Visualization of the frame-word alignments. (a) CLIP; (b) CLIPPING$^{-}{}_{\mathrm{TAB}}$; (c) CLIPPING.

The frame-word alignment with SAB KD shows clear and correct word-level attentions (e.g., “naked”, “child”, “runs” and “through”) like CLIP. But with TAB KD, the most important words (e.g., “naked” and “child”) cannot be noticed.

Experiments

Comparison with State-of-the-Arts

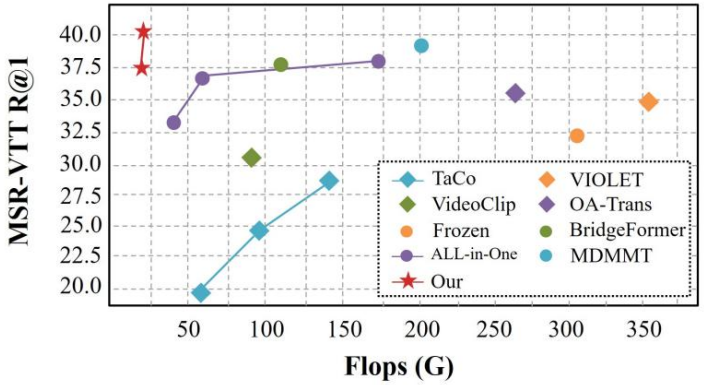


Figure 6. Flops and Performances.

Modules	Parameters	Flops
Vision encoder (MobileViT-v2)	4.5M	16.8G
Temporal Transformer	37.8M	1.2G
Language encoder ($CLIP_{bert}$)	12.6M	0.15G
Language encoder (TinyBERT)	4.2M	0.05G

Table 2. The Flops and parameters of each module in our method. We adopt 12 frames for each video.

Model	PT Datasets	Params	$R@1$
TACo	HT100M	212M	28.4
VideoClip	HT100M	130M	30.9
Frozen	C3M,W2M,COCO	232M	32.5
ALL-in-one-S	W2M,HT100M	33M	33.5
VIOLET	C3M,W2M	198M	34.5
OA-Trans	C3M,W2M	232M	35.8
BridgeFormer	C3M,W2M	160M	37.6
ALL-in-one-B	W2M,HT100M	110M	37.9
MDMMT	C400M,AudioSet	226M	38.9
CLIPPING$*_{\mathrm{w/o\ T}}$	-	**8.7M**	37.5
CLIPPING$_{\mathrm{w/o\ T}}$	-	21.3M	38.6
CLIPPING$*$	-	46.5M	39.8
CLIPPING	-	55.0M	40.6
CLIPPING	IN21K	55.0M	**40.7**

Table 1. Comparison with state-of-the-art models on MSR-VTT

Experiments

Ablation Studies

Vision Encoder	KD Types	$t2vR@1$	$v2tR@1$
CLIP_{vision}	-	44.5	42.2
MobileViTv2	-	25.7	24.5
MobileViTv2	T	28.8	27.3
MobileViTv2	T,S	33.0	32.8
MobileViTv2	T,S,SAB	37.6	36.2
MobileViTv2	T,S,SAB,CM_G	39.6	39.1
MobileViTv2	T,S,SAB,CM_G,CM_L	**40.7**	**40.2**

Table 3. Ablation study of different KD components of CLIPPING on the $1k$ validation set of MSR-VTT. The first row is the results of the teacher model (Clip4clip). T, S, SAB, CM_G and CM_L denote temporal KD, spatial KD, SAB KD, global cross-modality KD and local cross-modality KD, respectively.

KD Types	$t2vR@1$	$v2tR@1$
T, S	33.0	32.8
T, S, $OL2OL$	34.6	33.4
T, S, TAB	35.1	34.4
T, S, $SAB_{w/o\ masking}$	37.1	35.7
T, S, $SAB_{w/\ masking}$	**37.6**	**36.2**

Table 4. Ablation study of different KD types on MSR-VTT ($1k$ split). All the models are trained for 36 epochs with the same setting

Experiments

SAB Property

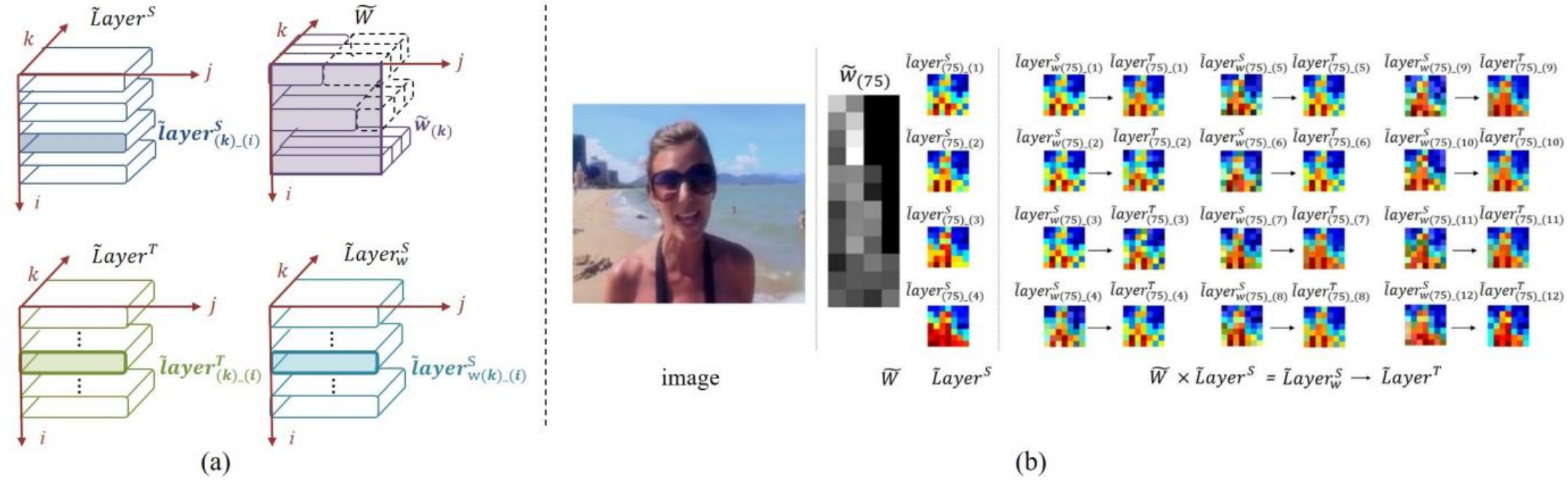


Figure 7. (a) Feature tensors of the student and the teacher. (b) An example to demonstrate that the teacher's features are the linear combinations of the student features. More examples are provided in the supplementary materials.

Experiments

SAB Property

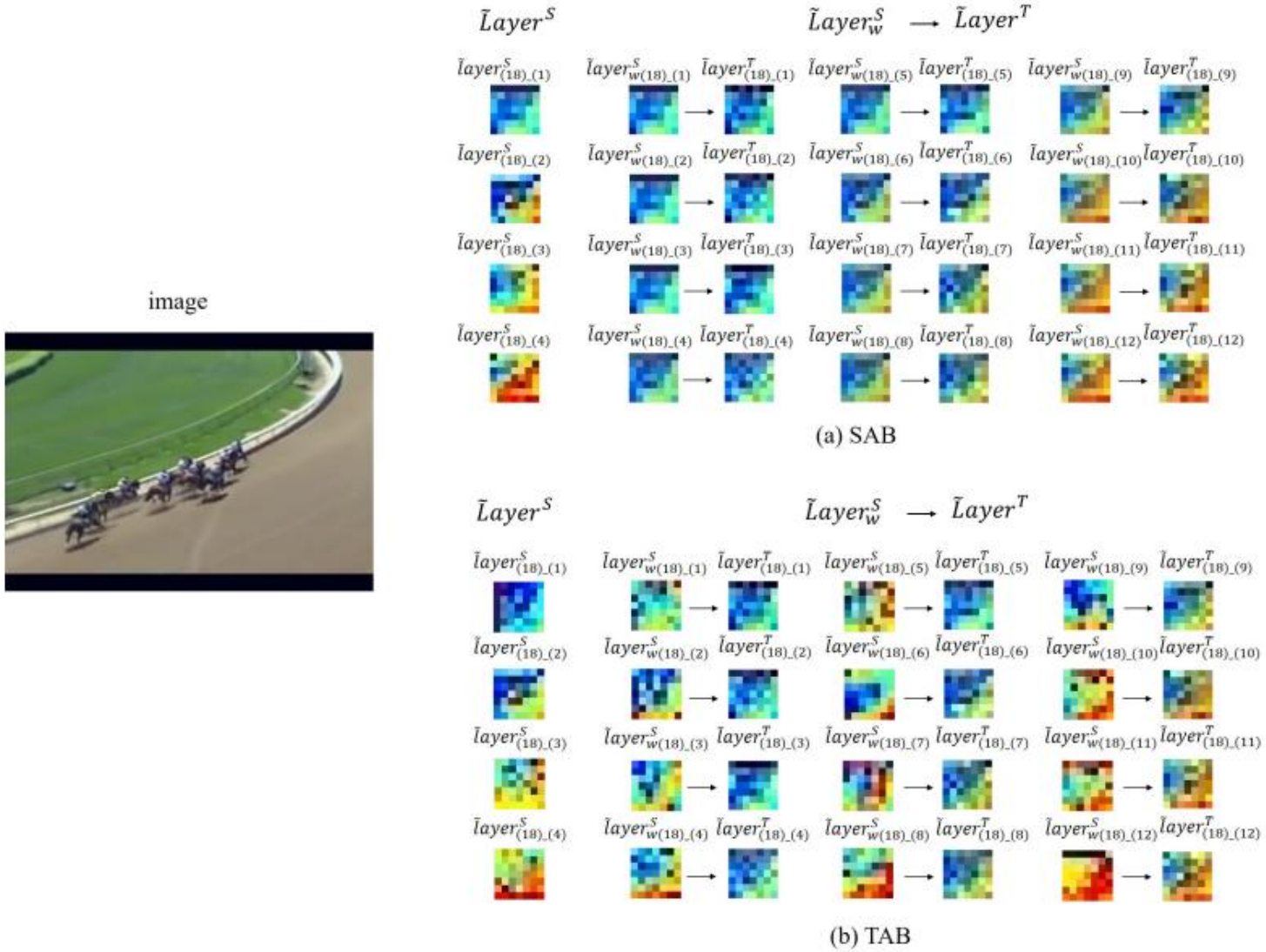


Figure 8. Examples of the linear combinations of the student's features that are trained with SAB KD (ours) and TAB KD [1].

Conclusion

In this paper, we propose a novel and efficient knowledge distillation method that is specially designed for small vision-language models. It includes temporal KD, spatial KD, SAB KD and cross-modality KD. Especially, the SAB KD has the property of the student's layers being the bases of the feature space. After training, the teacher's features are the linear combinations of the bases, indicating that the student has fully absorbed the knowledge of the teacher. CLIPPING significantly outperforms the state-of-the-art and is comparable or even superior to many large pre-training models. In the feature, we will apply CLIPPING to other vision-language models for compression.

NOAH'S ARK LAB

Thanks!