CLIP-KD: An Empirical Study of CLIP Model Distillation

CVPR 2024

Chuanguang Yang[1,2] Zhulin An[1*] Libo Huang[1] Junyu Bi[1,2] Xinqiang Yu[1,2]
Han Yang[1,2] Boyu Diao[1] Yongjun Xu[1*]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{yangchuanguang, anzhulin, huanglibo, bijunyu, yuxinqiang21s}@ict.ac.cn
{yanghan22s, diaoboyu2012, xyj}@ict.ac.cn

- Problem/Objective
  - Knowledge distillation
  - Large CLIP → (resource constrained) small CLIP
- Contribution/Key Idea
  - 다양한 KD method 제안 및 분석
  - 좋은 KD method = feature similarity maximize 설명
  - 아키텍처에 제약없는 distillation method 제안
  - CLIP+KD guideline 제안

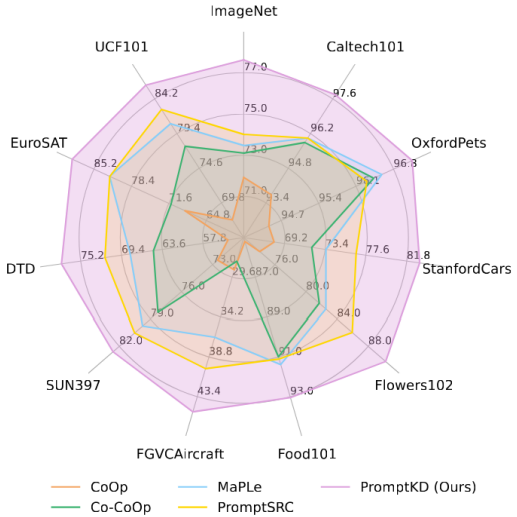


Figure 1. Harmonic mean (HM) comparison on base-to-novel generalization. All methods adopt the **ViT-B/16 image encoder** from the pre-trained CLIP model. PromptKD achieves state-of-the-art performance on 11 diverse recognition datasets.

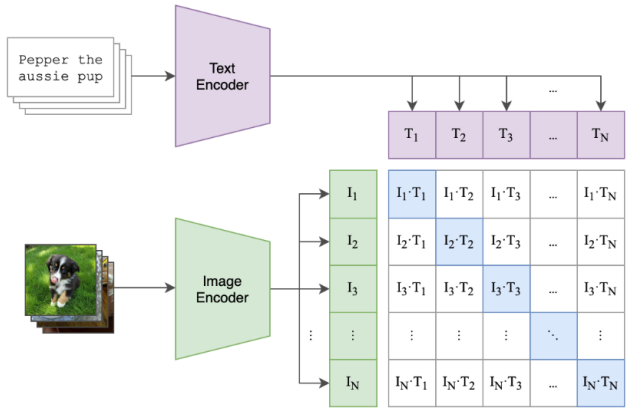
- We explain that a good CLIP distillation method could maximize the feature similarity between teacher and student models. Intuitively, if the student's features perfectly align with the teacher's features, their performance gap could disappear.
- We provide comprehensive guidelines for CLIP-KD. Compared to state-of-the-art TinyCLIP [50], our CLIP-KD does not rely on architecture-cue and achieves better performance on both the same- and different-architecture styles of the teacher-student models.

김범준

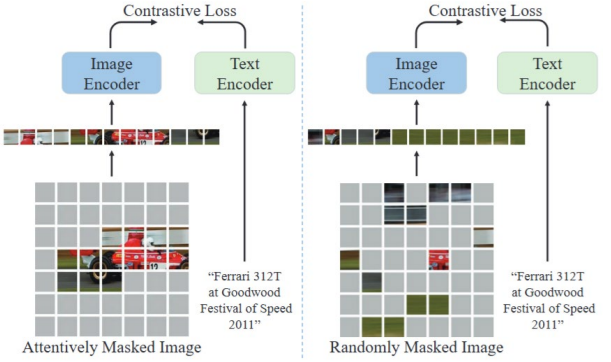
CLIP-KD: An Empirical Study of CLIP Model Distillation

CVPR 2024

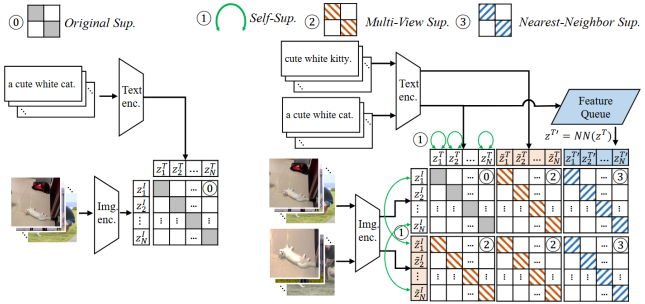
- **CLIP (Contrastive Language-Image Pre-training)**



CLIP



CLIP + Masked Image [1]



CLIP + Image Supervision [2]

- CLIP: 이미지-텍스트 쌍을 예측하도록 유도
  - CLIP을 개선하기 위한 몇가지 방법들이 있었고
  - CLIP을 downstream task(detection 등)에 이용하려는 시도들 있었음
  - But, Resource가 제한된 소형 CLIP을 개선하려는 시도는 적었음

[1] Yang, Yifan, et al. "Attentive mask clip." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[2] Li, Yangguang, et al. "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm." arXiv preprint arXiv:2110.05208

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Large → Small CLIP Knowledge Distillation**

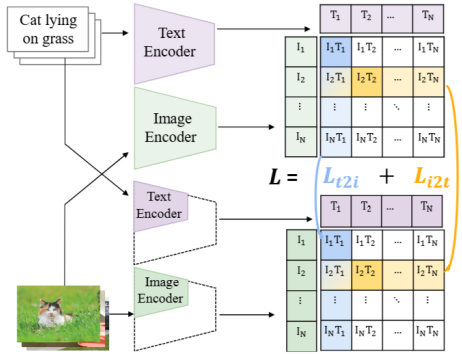
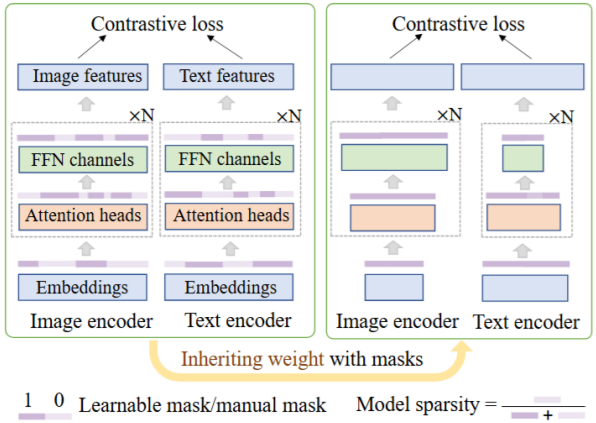


Figure 2. Affinity mimicking for language-image models. The loss includes image-to-text loss (yellow) and text-to-image loss (blue).

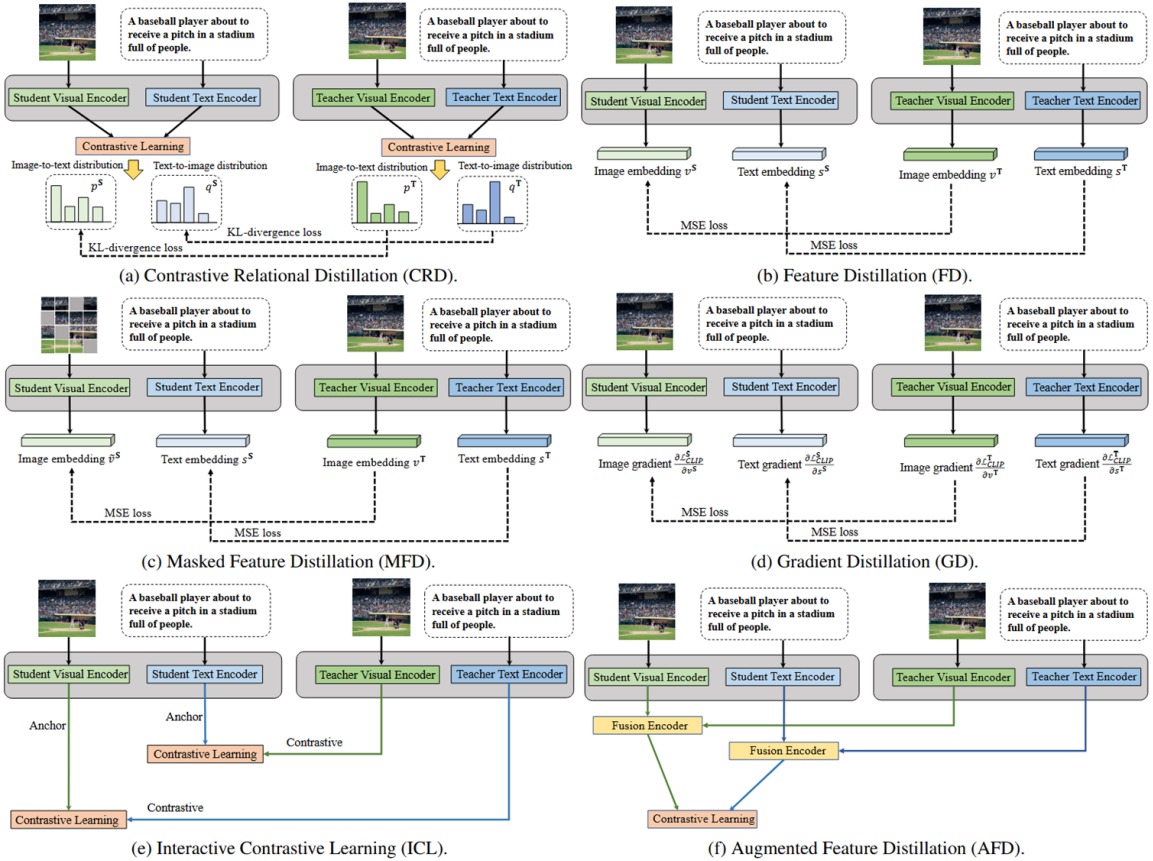


- SOTA인 TinyCLIP [1] 의 문제점 분석
  - Weight inheritance이 핵심
  - But 이는 Teacher - Student가 동일한 아키텍처일때만 가능 (ViT-B/32 → ViT-61M/32 & ResNet-101 → ResNet-30M)
    - **반면, CLIP-KD 는 어떤 Teacher - Student 조합에도 일반화됨**

[1] Wu, Kan, et al. "Tinyclip: Clip distillation via affinity mimicking and weight inheritance." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Method**



(a) Contrastive Relational Distillation (CRD). (b) Feature Distillation (FD).

(c) Masked Feature Distillation (MFD). (d) Gradient Distillation (GD).

(e) Interactive Contrastive Learning (ICL). (f) Augmented Feature Distillation (AFD).

Figure 1. Illustration of various CLIP knowledge distillation approaches proposed in this paper.

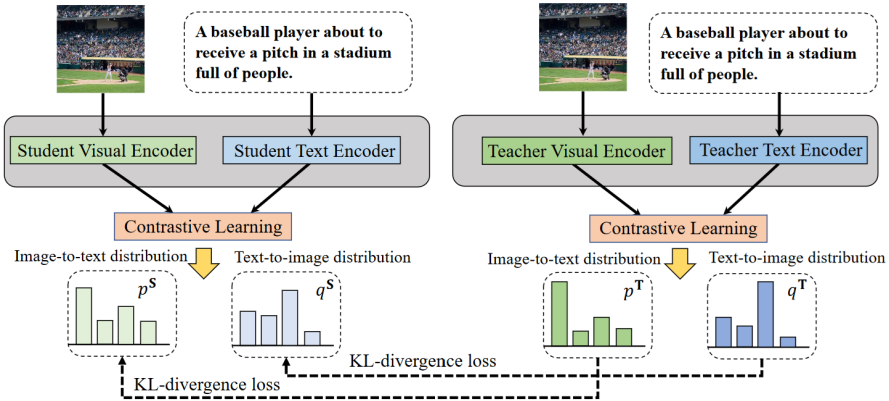
- **Configuration**

Table 2. Configuration of paired visual and text encoders.

	Visual encoder			Text encoder: Transformer [42]			
	Model	Type	Params	Layer	Width	Head	Params
Teacher	ViT-L/14 [8]	ViT	304.0M	12	768	12	85.1M
Teacher	ViT-B/16 [8]	ViT	86.2M	12	512	8	37.8M
Student	ViT-T/16 [8]	ViT	5.6M	12	384	6	21.3M
Student	MobileViT-S [32]	ViT	5.3M	12	384	6	21.3M
Student	Swin-T [30]	ViT	27.9M	12	384	6	21.3M
Teacher	ResNet-101 [12],	CNN	56.3M	12	512	8	37.8M
Teacher	ResNet-50 [12],	CNN	38.3M	12	512	8	37.8M
Student	ResNet-18 [12],	CNN	11.4M	12	384	6	21.3M
Student	MobileNetV3 [16]	CNN	2.0M	12	384	6	21.3M
Student	EfficientNet-B0 [41]	CNN	4.7M	12	384	6	21.3M

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Method (CRD)**



(a) Contrastive Relational Distillation (CRD).

$$p_k^{\mathbf{T}}[j] = \frac{\exp(v_k^{\mathbf{T}} \cdot s_j^{\mathbf{T}} / \tau^{\mathbf{T}})}{\sum_{b=1}^{|\mathcal{B}|} \exp(v_k^{\mathbf{T}} \cdot s_b^{\mathbf{T}} / \tau^{\mathbf{T}})}, \quad q_k^{\mathbf{T}}[j] = \frac{\exp(s_k^{\mathbf{T}} \cdot v_j^{\mathbf{T}} / \tau^{\mathbf{T}})}{\sum_{b=1}^{|\mathcal{B}|} \exp(s_k^{\mathbf{T}} \cdot v_b^{\mathbf{T}} / \tau^{\mathbf{T}})},$$

$$p_k^{\mathbf{S}}[j] = \frac{\exp(v_k^{\mathbf{S}} \cdot s_j^{\mathbf{S}} / \tau^{\mathbf{S}})}{\sum_{b=1}^{|\mathcal{B}|} \exp(v_k^{\mathbf{S}} \cdot s_b^{\mathbf{S}} / \tau^{\mathbf{S}})}. \quad q_k^{\mathbf{S}}[j] = \frac{\exp(s_k^{\mathbf{S}} \cdot v_j^{\mathbf{S}} / \tau^{\mathbf{S}})}{\sum_{b=1}^{|\mathcal{B}|} \exp(s_k^{\mathbf{S}} \cdot v_b^{\mathbf{S}} / \tau^{\mathbf{S}})}.$$

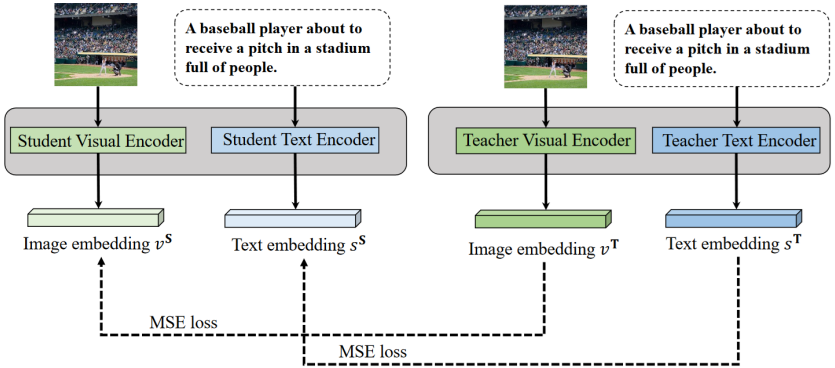
$$\mathcal{L}_{CRD\_I \to T} = \frac{1}{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} p_k^{\mathbf{T}}[j] \log \frac{p_k^{\mathbf{T}}[j]}{p_k^{\mathbf{S}}[j]},$$

$$\mathcal{L}_{CRD\_T \to I} = \frac{1}{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} q_k^{\mathbf{T}}[j] \log \frac{q_k^{\mathbf{T}}[j]}{q_k^{\mathbf{S}}[j]}.$$

KL Divergence

$$\mathcal{L}_{CRD} = \mathcal{L}_{CRD\_I \to T} + \mathcal{L}_{CRD\_T \to I}.$$

- **Method (FD)**



(b) Feature Distillation (FD).

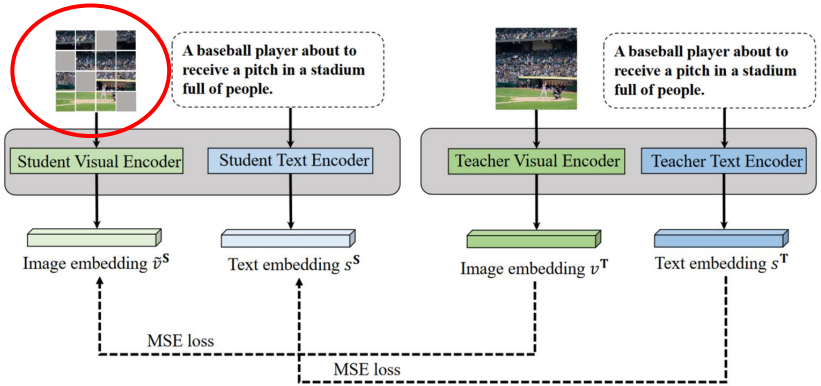
$$\mathcal{L}_{FD} = \frac{1}{|\mathcal{B}|}\sum_{k=1}^{|\mathcal{B}|}\left(\left\|v_k^{\mathbf{T}} - v_k^{\mathbf{S}}\right\|_2^2 + \left\|s_k^{\mathbf{T}} - s_k^{\mathbf{S}}\right\|_2^2\right)$$

MSE Loss

if) embedding space dimension이 다르면 mlp

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Method (MFD)**



(c) Masked Feature Distillation (MFD).

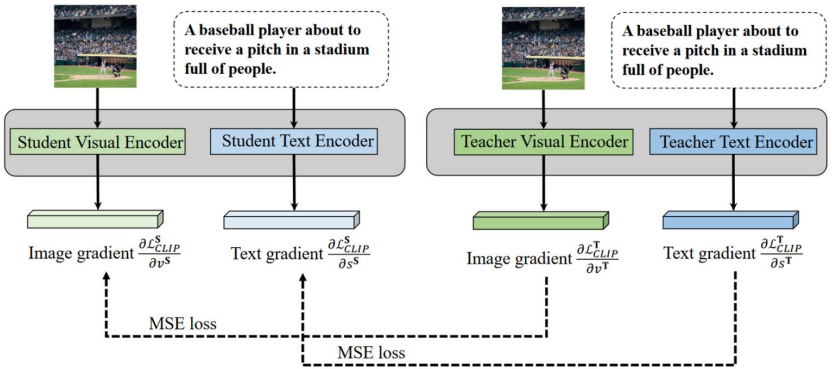
$$\mathcal{L}_{MFD} = \frac{1}{|\mathcal{B}|}\sum_{k=1}^{|\mathcal{B}|}\left(\left\|v_k^{\mathbf{T}} - \tilde{v}_k^{\mathbf{S}}\right\|_2^2 + \left\|s_k^{\mathbf{T}} - s_k^{\mathbf{S}}\right\|_2^2\right),$$

MAE (Masked AutoEncoder) 알고리즘 사용

FD와 동일하게 MSE Loss

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Method (GD)**



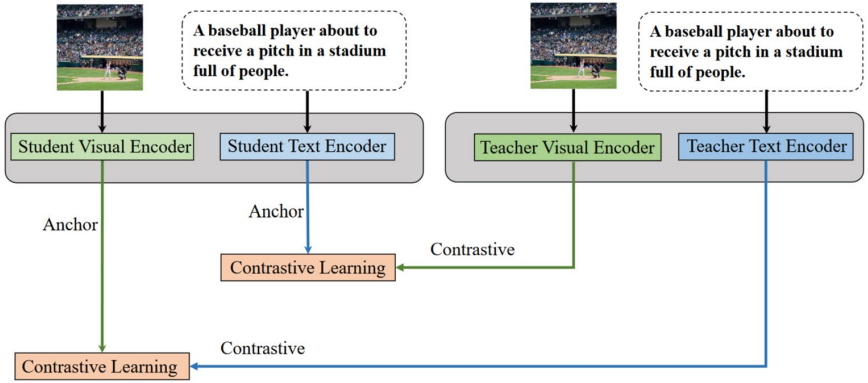
(d) Gradient Distillation (GD).

$$\mathcal{L}_{GD} = \frac{1}{|\mathcal{B}|}\sum_{k=1}^{|\mathcal{B}|}\Big(\Big\|\frac{\partial \mathcal{L}^{\mathbf{T}}_{CLIP}}{\partial v^{\mathbf{T}}_{k}} - \frac{\partial \mathcal{L}^{\mathbf{S}}_{CLIP}}{\partial v^{\mathbf{S}}_{k}}\Big\|_2^2 + \Big\|\frac{\partial \mathcal{L}^{\mathbf{T}}_{CLIP}}{\partial s^{\mathbf{T}}_{k}} - \frac{\partial \mathcal{L}^{\mathbf{S}}_{CLIP}}{\partial s^{\mathbf{S}}_{k}}\Big\|_2^2\Big).$$

- gradient = input에 대한 model의 response
  - Student의 gradient를 Teacher 것으로 강제하여 output이 어떻게 수정되어야하는지 알려줌
  - Student 모델의 responsiveness를 Teacher model

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Method (ICL)**



(e) Interactive Contrastive Learning (ICL).

$$\mathcal{L}_{ICL\_I\to T} = -\log \frac{\exp(v_k^{\mathbf{S}} \cdot s_k^{\mathbf{T}}/\tau)}{\sum_{b=1}^{|\mathcal{B}|} \exp(v_k^{\mathbf{S}} \cdot s_b^{\mathbf{T}}/\tau)}.$$

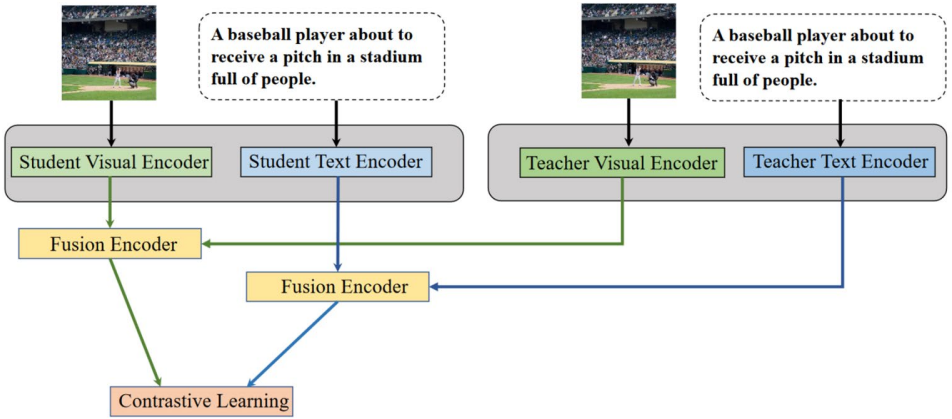
$$\mathcal{L}_{ICL\_T\to I} = -\log \frac{\exp(s_k^{\mathbf{S}} \cdot v_k^{\mathbf{T}}/\tau)}{\sum_{b=1}^{|\mathcal{B}|} \exp(s_k^{\mathbf{S}} \cdot v_b^{\mathbf{T}}/\tau)}.$$

$$\mathcal{L}_{ICL} = \frac{1}{2}(\mathcal{L}_{ICL\_I\to T} + \mathcal{L}_{ICL\_T\to I}).$$

- CLIP의 contrastive learning의 **anchor (Student) - Positive/Negative Sample (Teacher)**로 학습

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Method (AFD)**



(f) Augmented Feature Distillation (AFD).

$$v_k^{\mathbf{A}} = \phi_i(v_k^{\mathbf{S}} || v_k^{\mathbf{T}}), s_k^{\mathbf{A}} = \phi_t(s_k^{\mathbf{S}} || s_k^{\mathbf{T}}).$$

- Teacher - Student의 embedding을 concat하여 Contrastive learning을 진행

$$\mathcal{L}_{CLIP\_KD} = \mathcal{L}_{CLIP} + \lambda \mathcal{L}_{KD}. \quad (18)$$

Here, $\mathcal{L}_{KD} \in \{\mathcal{L}_{CRD}, \mathcal{L}_{FD}, \mathcal{L}_{MFD}, \mathcal{L}_{GD}, \mathcal{L}_{ICL}, \mathcal{L}_{AFD}\}$

- **Experiment**

Table 1. **Comparison of CLIP distillation losses trained from CC3M+12M on zero-shot ImageNet-related classification and cross-modal retrieval on CC3M Val, MSCOCO and Flickr.** The numbers in **bold** denote the best results for individual methods (the third block) and unified methods (the fourth block), respectively. The 'T' and 'S' tags represent the teacher and student roles, respectively.

Method	IN Acc	INV2 Acc	IN-R Acc	IN-S Acc	CC3M Val I2T	CC3M Val T2I	MSCOCO I2T	MSCOCO T2I	Flickr I2T	Flickr T2I
T: ViT-B/16	37.0	32.1	48.4	26.0	40.2	39.5	25.0	24.7	54.6	56.6
S: ViT-T/16	30.6	25.6	35.7	17.4	33.3	33.3	20.7	20.3	46.4	47.7
+CRD	31.9	27.6	38.8	19.6	35.3	34.9	21.4	20.7	48.8	49.9
+FD	**34.2**	**29.5**	**42.7**	**21.4**	37.1	**36.9**	22.5	**22.2**	**51.1**	**51.3**
+MFD	34.1	**29.5**	42.3	21.2	**37.4**	**36.9**	**22.9**	22.1	50.9	51.1
+GD	31.5	27.0	37.9	19.1	34.5	34.0	21.3	20.9	47.5	48.3
+ICL	33.1	28.2	40.6	20.8	36.1	35.8	21.8	21.7	50.5	50.4
+AFD	31.4	26.9	37.8	18.6	34.6	34.7	20.9	20.5	47.3	48.7
+FD+ICL	34.6	30.0	43.2	**22.0**	37.9	37.6	23.0	22.5	51.7	51.9
+FD+ICL+CRD	**34.9**	**30.1**	43.5	21.9	**38.2**	**37.9**	23.1	**22.6**	52.3	**52.4**
+FD+ICL+CRD+GD	34.8	29.9	42.8	**22.0**	38.1	37.7	**23.3**	22.5	**52.4**	52.3
+FD+ICL+CRD+AFD	34.8	**30.1**	**43.6**	21.6	**38.2**	37.7	23.0	22.5	52.2	**52.4**

Evaluation metrics. Following the standard setting, we employ Recall@K (R@K) to measure the retrieval performance in K nearest neighbours. By default, we use top-1 accuracy (Acc) for image classification and R@1 for Image-to-Text (I2T) and Text-to-Image (T2I) retrieval.

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Experiment**

Table 3. Distillation performance trained from CC3M+12M for cross-modal retrieval on CC3M, MSCOCO and Flickr validation set.

Method	CC3M		MSCOCO		Flickr		Method	CC3M		MSCOCO		Flickr	
	I2T	T2I	I2T	T2I	I2T	T2I		I2T	T2I	I2T	T2I	I2T	T2I
T: ViT-B/16	40.2	39.5	25.0	24.7	54.6	56.6	T: ResNet-101	41.4	40.5	25.2	25.7	57.0	55.5
S: MobileViT-S	36.0	35.6	22.3	22.9	50.1	53.0	S: MobileViT-S	36.0	35.6	22.3	22.9	50.1	53.0
+CLIP-KD	**39.4**	**38.6**	**26.1**	**24.9**	**55.0**	**56.2**	+CLIP-KD	**39.9**	**38.6**	**26.0**	**25.3**	**57.6**	**56.1**
S: Swin-T	39.8	39.2	24.7	25.3	53.4	54.4	S: Swin-T	39.8	39.2	24.7	25.3	53.4	54.4
+CLIP-KD	**43.7**	**42.5**	**28.5**	**28.6**	**62.2**	**60.9**	+CLIP-KD	**44.2**	**43.0**	**27.8**	**28.9**	**60.8**	**61.5**
S: MobileNetV3	28.1	27.5	15.3	15.0	36.9	38.0	S: MobileNetV3	28.1	27.5	15.3	15.0	36.9	38.0
+CLIP-KD	**30.1**	**28.6**	**17.9**	**16.0**	**42.4**	**42.3**	+CLIP-KD	**30.2**	**29.4**	**17.2**	**16.6**	**40.2**	**42.2**
S: EfficientNet-B0	35.4	34.9	21.7	21.1	48.3	50.1	S: EfficientNet-B0	35.4	34.9	21.7	21.1	48.3	50.1
+CLIP-KD	**39.0**	**38.0**	**26.0**	**23.9**	**55.5**	**54.2**	+CLIP-KD	**37.4**	**36.8**	**24.7**	**24.6**	**55.8**	**56.2**
S: ResNet-18	31.1	30.4	19.2	18.6	41.0	43.3	S: ResNet-18	31.1	30.4	19.2	18.6	41.0	43.3
+CLIP-KD	**34.2**	**33.0**	**21.3**	**19.8**	**47.8**	**47.1**	+CLIP-KD	**34.7**	**33.7**	**21.0**	**20.9**	**48.8**	**48.4**

- Cross-modal retrieval
- Teacher - Student 아키텍처 자유
  - 모든 조합에서 일관되게 성능향상

- **Experiment**

Table 4. Distillation performance of zero-shot ImageNet and its variants on top-1 classification accuracy (%) trained on CC3M+12M.

Method	IN-1K	INV2	IN-R	IN-S	Method	IN-1K	INV2	IN-R	IN-S
T: ViT-B/16	37.0	32.1	48.4	26.0	T: ResNet-101	36.8	31.9	49.2	26.7
S: MobileViT-S	32.6	27.6	39.5	20.1	S: MobileViT-S	32.6	27.6	39.5	20.1
+CLIP-KD	**36.0**	**31.1**	**44.5**	**23.5**	+CLIP-KD	**35.0**	**30.1**	**43.7**	**22.7**
S: Swin-T	36.4	31.1	45.9	24.4	S: Swin-T	36.4	31.1	45.9	24.4
+CLIP-KD	**40.2**	**34.9**	**51.4**	**28.2**	+CLIP-KD	**39.5**	**34.2**	**51.9**	**28.1**
S: MobileNetV3	25.1	20.7	29.2	13.4	S:MobileNetV3	25.1	20.7	29.2	13.4
+CLIP-KD	**27.0**	**23.0**	**30.6**	**14.1**	+CLIP-KD	**26.2**	**22.2**	**29.3**	**13.7**
S: EfficientNet-B0	32.6	27.8	40.9	20.7	S:EfficientNet-B0	32.6	27.8	40.9	20.7
+CLIP-KD	**35.4**	**30.6**	**44.7**	**23.7**	+CLIP-KD	**34.6**	**29.4**	**44.4**	**23.1**
S: ResNet-18	28.6	24.0	35.3	18.1	S:ResNet-18	28.6	24.0	35.3	18.1
+CLIP-KD	**31.4**	**26.9**	**39.2**	**20.0**	+CLIP-KD	**30.9**	**25.9**	**38.0**	**19.5**

- Teacher - Student 아키텍처 자유
  - 모든 조합에서 일관되게 성능향상
  - Swin-T는 teacher model 성능 능가

- **Experiment**

Table 5. **Distillation performance of zero-shot ImageNet and cross-modal retrieval trained on CC3M+12M.** The teachers are pretrained on Laion-400M before distillation. '(from T_x)' indicates that the student is distilled from the teacher T_x.

Method	IN-1K Acc	MSCOCO I2T	MSCOCO T2I	Flickr I2T	Flickr T2I
T_1: ViT-L/14	72.8	42.7	40.9	80.5	79.5
T_2: ViT-B/16	67.1	39.5	36.5	76.9	75.5
S: ViT-T/16	30.6	20.7	20.3	46.4	47.7
+TinyCLIP (from T_1)	39.3	26.4	24.1	57.6	57.4
+TinyCLIP (from T_2)	40.8	26.8	24.7	58.6	58.5
+CLIP-KD (from T_1)	40.9	27.2	25.5	59.7	59.7
+CLIP-KD (from T_2)	**42.6**	**28.1**	**26.0**	**60.4**	**59.9**
S: ViT-B/16	37.0	25.0	24.7	54.6	56.6
+TinyCLIP (from T_1)	55.4	35.9	33.6	73.2	72.8
+CLIP-KD (from T_1)	**57.5**	**37.6**	**35.6**	**75.3**	**74.5**
S: ResNet-50	35.3	23.5	24.7	55.1	55.0
+CLIP-KD (from T_2)	**55.4**	**36.3**	**33.4**	**73.0**	**72.2**

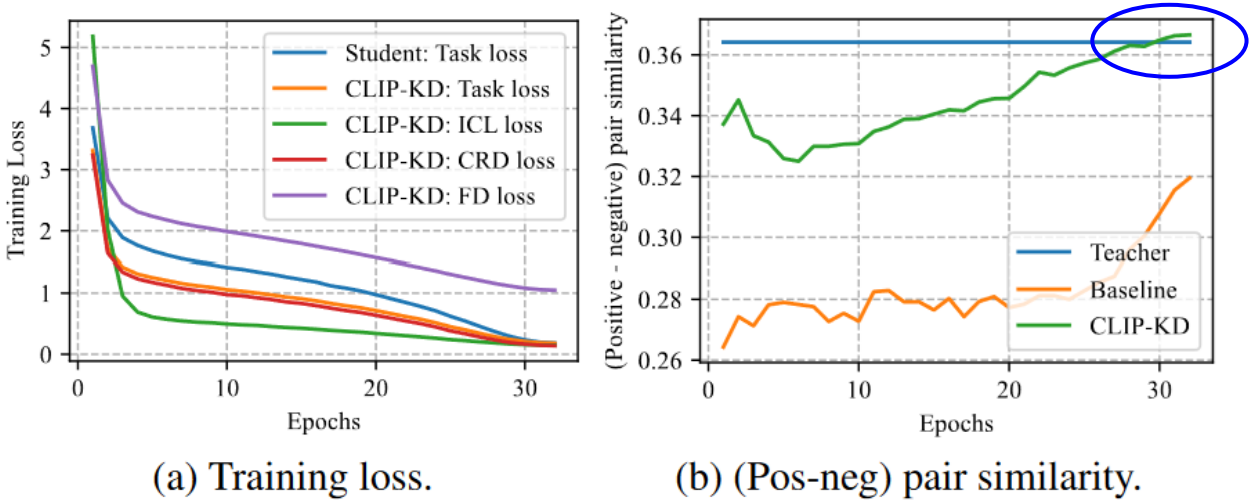
- Teacher를 바꿔보았다 (크기 차이)

- 더 작은 Teacher 모델로부터의 distill 성능이 더 좋다.
  - Capacity Gap 존재

for distillation. One possible reason is that a large teacher and a small student may exist capacity gaps, making the student difficult to align with the teacher. This may become an open issue for future research.

CLIP-KD: An Empirical Study of CLIP Model Distillation

- **Experiment**



(a) Training loss. (b) (Pos-neg) pair similarity.

- **Experiment**

Figure 3. **Similarity statistics between teacher and student features after distillation trained on CC3M+12M.** $v_k^{\mathbf{T}}$ and $v_k^{\mathbf{S}}$ denote the teacher and student image features, respectively. $s_k^{\mathbf{T}}$ and $s_k^{\mathbf{S}}$ denote the teacher and student text features, respectively.

- KD method 별 성능 차이 이유 분석

- 결론 : performance ∝ feature similarity 발견
  - FD는 직접 feature similarity를 강제하기에 성능 good
  - but, 이는 text-image contrastive relation 고려 x
    - 그래서 FD + ICL 을 한 것이 최고의 성능 향상

- **Experiment - supple**

Table 11. Linear evaluation on MS-COCO object detection using a CC3M+12M pretrained ResNet-50 over Mask-RCNN framework.

Method	Object detection					
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Baseline	32.6	52.3	34.8	18.0	35.6	42.4
+CLIP-KD	**34.0**	**53.9**	**36.5**	**20.0**	**36.8**	**43.8**

Table 12. Linear evaluation on MS-COCO instance segmentation using a CC3M+12M pretrained ResNet-50 over Mask-RCNN framework.

Method	Instance segmentation					
	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AP^{seg}_{S}	AP^{seg}_{M}	AP^{seg}_{L}
Baseline	29.9	49.5	31.8	13.1	32.2	44.2
+CLIP-KD	**31.1**	**50.9**	**32.9**	**14.2**	**33.3**	**45.4**

- Downstream task(detection, segmentation)에서도 성능 향상