Appendix of CLIP-KD: An Empirical Study of CLIP Model Distillation

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

1. Gradient Distillation.

The gradient information often shows how the model responds to changes according to inputs. We propose to force the gradient consistency between teacher and student using the derivative w.r.t the visual and text embeddings. By this means, the student could better understand how the output should change according to the input. This helps the student behave more similarly to the teacher.

Given the image-to-text contrastive loss $\mathcal{L}_{I\to T}$, the visual embedding v_k is the anchor, and the text embeddings $\{s_b\}_{b=1}^{|\mathcal{B}|}$ are contrastive samples. The gradient w.r.t visual and text embeddings are calculated as $\frac{\partial \mathcal{L}_{I\to T}}{\partial v_k}$ and $\frac{\partial \mathcal{L}_{I\to T}}{\partial s_b}$:

$$\frac{\partial \mathcal{L}_{I\to T}}{\partial v_k}=\sum_{b=1}^{|\mathcal{B}|}(p_k[b]-\mathbf{1}_{[k=b]})s_b/\tau, \tag{1}$$

$$\frac{\partial \mathcal{L}_{I\to T}}{\partial s_b}=(p_k[b]-\mathbf{1}_{[k=b]})v_k/\tau. \tag{2}$$

Here, p_k is the contrastive distribution from v_k to $\{s_b\}_{b=1}^{|\mathcal{B}|}$. $\mathbf{1}$ is an indicator function that equals to 1 when $k=b$ else returns 0. Similarly, the gradient of text-to-image contrastive loss $\mathcal{L}_{T\to I}$ w.r.t the text embedding s_k and visual embeddings $\{v_b\}_{b=1}^{|\mathcal{B}|}$ are calculated as $\frac{\partial \mathcal{L}_{I\to T}}{\partial s_k}$ and $\frac{\partial \mathcal{L}_{I\to T}}{\partial v_b}$:

$$\frac{\partial \mathcal{L}_{T\to I}}{\partial s_k}=\sum_{b=1}^{|\mathcal{B}|}(q_k[b]-\mathbf{1}_{[k=b]})v_b/\tau, \tag{3}$$

$$\frac{\partial \mathcal{L}_{T\to I}}{\partial v_b}=(q_k[b]-\mathbf{1}_{[k=b]})s_k/\tau. \tag{4}$$

As a result, the gradient of CLIP contrastive loss $\mathcal{L}_{CLIP}$ w.r.t each visual embedding v_k and text embedding s_k are formulated as:

$$\frac{\partial \mathcal{L}_{CLIP}}{\partial v_k}=\frac{1}{2}(\frac{\partial \mathcal{L}_{I\to T}}{\partial v_k}+\frac{\partial \mathcal{L}_{T\to I}}{\partial v_k}), \tag{5}$$

$$\frac{\partial \mathcal{L}_{CLIP}}{\partial s_k}=\frac{1}{2}(\frac{\partial \mathcal{L}_{I\to T}}{\partial s_k}+\frac{\partial \mathcal{L}_{T\to I}}{\partial s_k}). \tag{6}$$

We align the gradient information w.r.t each visual and text embedding between teacher and student via MSE loss:

$$\mathcal{L}_{GD}=\frac{1}{|\mathcal{B}|}\sum_{k=1}^{|\mathcal{B}|}(\left\|\frac{\partial \mathcal{L}_{CLIP}^{\mathbf{T}}}{\partial v_k^{\mathbf{T}}}-\frac{\partial \mathcal{L}_{CLIP}^{\mathbf{S}}}{\partial v_k^{\mathbf{S}}}\right\|_2^2 +\left\|\frac{\partial \mathcal{L}_{CLIP}^{\mathbf{T}}}{\partial s_k^{\mathbf{T}}}-\frac{\partial \mathcal{L}_{CLIP}^{\mathbf{S}}}{\partial s_k^{\mathbf{S}}}\right\|_2^2). \tag{7}$$

2. Theoretical Insights of Interactive Contrastive Learning: Proof of Maximizing the Lower bound of the Mutual Information

Given the student image embedding $v_k^{\mathbf{S}}$ as the anchor and teacher text embeddings $\{s_b^{\mathbf{T}}\}_{b=1}^{B}$ as contrastive ones, where $B=|\mathcal{B}|$ is the batch size, the $(v_k^{\mathbf{S}},s_k^{\mathbf{T}})$ is a positive pair and $\{(v_k^{\mathbf{S}},s_b^{\mathbf{T}})\}_{b=1,b\neq k}^{B}$ are negative pairs. We introduce the joint distribution $\mu(v^{\mathbf{S}},s^{\mathbf{T}})$ and the product of marginals $\mu(v^{\mathbf{S}})\mu(s^{\mathbf{T}})$. We utilize a distribution η with an indicator variable C to represent whether a pair $(v^{\mathbf{S}},s^{\mathbf{T}})$ is drawn from the joint distribution ($C=1$) or product of marginals ($C=0$):

$$\eta(v^{\mathbf{S}},s^{\mathbf{T}}|C=1)=\mu(v^{\mathbf{S}},s^{\mathbf{T}}), \tag{8}$$

$$\eta(v^{\mathbf{S}},s^{\mathbf{T}}|C=0)=\mu(v^{\mathbf{S}})\mu(s^{\mathbf{T}}). \tag{9}$$

Here, $C=1$ represents the positive pair $(v_k^{\mathbf{S}},s_k^{\mathbf{T}})$ while $C=0$ represents a negative pair from $\{(v_k^{\mathbf{S}},s_b^{\mathbf{T}})\}_{b=1,b\neq k}^{B}$, *i.e.* $(v_k^{\mathbf{S}},s_k^{\mathbf{T}})\sim\mu(v^{\mathbf{S}},s^{\mathbf{T}})$, $\{(v_k^{\mathbf{S}},s_b^{\mathbf{T}})\}_{b=1,b\neq k}^{B}\sim\mu(v^{\mathbf{S}})\mu(s^{\mathbf{T}})$. For interactive contrastive learning, we often have 1 positive pair for every N negative pairs, where

[*] Corresponding author

$N = B - 1$. Therefore, the prior probabilities of the latent variable C are formulated as:

$$\eta(C=1) = \frac{1}{1+N}, \ \eta(C=0) = \frac{N}{1+N}. \tag{10}$$

By using Bayes' theorem, we can compute the class posterior of the pair $(v^{\mathbf{S}}, s^{\mathbf{T}})$ belonging to the positive case $(C = 1)$ as :

$$\eta(C=1|v^{\mathbf{S}}, s^{\mathbf{T}}) \tag{11}$$

$$= \frac{\eta(v^{\mathbf{S}}, s^{\mathbf{T}}|C=1)\eta(C=1)}{\eta(v^{\mathbf{S}}, s^{\mathbf{T}}|C=1)\eta(C=1) + \eta(v^{\mathbf{S}}, s^{\mathbf{T}}|C=0)\eta(C=0)} \tag{12}$$

$$= \frac{\mu(v^{\mathbf{S}}, s^{\mathbf{T}})}{\mu(v^{\mathbf{S}}, s^{\mathbf{T}}) + N\mu(v^{\mathbf{S}})\mu(s^{\mathbf{T}})}. \tag{13}$$

The log class posterior can be further expressed as follows:

$$\log \eta(C=1|v^{\mathbf{S}}, s^{\mathbf{T}}) \tag{14}$$

$$= \log \frac{\mu(v^{\mathbf{S}}, s^{\mathbf{T}})}{\mu(v^{\mathbf{S}}, s^{\mathbf{T}}) + N\mu(v^{\mathbf{S}})\mu(s^{\mathbf{T}})} \tag{15}$$

$$= -\log(1 + N\frac{\mu(v^{\mathbf{S}})\mu(s^{\mathbf{T}})}{\mu(v^{\mathbf{S}}, s^{\mathbf{T}})}) \tag{16}$$

$$\leq -\log(N) + \log \frac{\mu(v^{\mathbf{S}}, s^{\mathbf{T}})}{\mu(v^{\mathbf{S}})\mu(s^{\mathbf{T}})}. \tag{17}$$

The expectations of log class posterior $\log \eta(C = 1|v^{\mathbf{S}}, s^{\mathbf{T}})$ can be connected to mutual information:

$$\mathbb{E}_{\eta(v^{\mathbf{S}}, s^{\mathbf{T}}|C=1)} \log \eta(C=1|v^{\mathbf{S}}, s^{\mathbf{T}}) \tag{18}$$

$$\leq -\log(N) + \mathbb{E}_{\mu(v^{\mathbf{S}}, s^{\mathbf{T}})} \log \frac{\mu(v^{\mathbf{S}}, s^{\mathbf{T}})}{\mu(v^{\mathbf{S}})\mu(s^{\mathbf{T}})} \tag{19}$$

$$= -\log(N) + I(v^{\mathbf{S}}, s^{\mathbf{T}}), \tag{20}$$

where $I(v^{\mathbf{S}}, s^{\mathbf{T}})$ denotes mutual information between $v^{\mathbf{S}}$ and $s^{\mathbf{T}}$. Essentially, the ICL loss $\mathcal{L}_{ICL\_I\rightarrow T}$ is negative log class posterior of the positive pair:

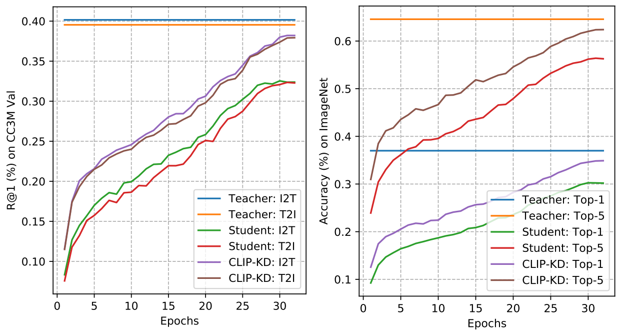
$$\mathcal{L}_{ICL\_I\rightarrow T} = -\log \eta(C=1|v^{\mathbf{S}}, s^{\mathbf{T}}). \tag{21}$$

Therefore, we can connect $\mathcal{L}_{ICL\_I\rightarrow T}$ to the mutual information $I(v^{\mathbf{S}}, s^{\mathbf{T}})$ as follows:

$$\mathbb{E}_{\eta(v^{\mathbf{S}}, s^{\mathbf{T}}|C=1)}\mathcal{L}_{ICL\_I\rightarrow T} \geq \log(N) - I(v^{\mathbf{S}}, s^{\mathbf{T}}) \tag{22}$$

$$\Leftrightarrow I(v^{\mathbf{S}}, s^{\mathbf{T}}) \geq \log(N) - \mathbb{E}_{\eta(v^{\mathbf{S}}, s^{\mathbf{T}}|C=1)}\mathcal{L}_{ICL\_I\rightarrow T}. \tag{23}$$

By minimizing $\mathcal{L}_{ICL\_I\rightarrow T}$, the lower bound on mutual information $I(v^{\mathbf{S}}, s^{\mathbf{T}})$ is maximized. The mutual information $I(v^{\mathbf{S}}, s^{\mathbf{T}})$ measures uncertainty reduction in contrastive feature embeddings from the teacher text encoder when the anchor embedding from the student visual encoder is known. Since $\mathcal{L}_{ICL\_T\rightarrow I}$ is symmetric to $\mathcal{L}_{ICL\_I\rightarrow T}$, the lower bound on mutual information $I(s^{\mathbf{S}}, v^{\mathbf{T}})$ can be maximized by minimizing $\mathcal{L}_{ICL\_T\rightarrow I}$. The mutual information $I(s^{\mathbf{S}}, v^{\mathbf{T}})$ measures uncertainty reduction in contrastive feature embeddings from the teacher visual encoder when the anchor embedding from the student text encoder is known. By maximizing the lower bound of mutual information, the student network reduces uncertainty with the teacher. This means that ICL guides the student to learn more common knowledge from the teacher, leading to better feature representations.



(a) R@1 (%) on CC3M Val.

(b) Accuracy (%) on ImageNet.

Figure 1. Training curves using ViT-B/16 as the teacher and ViT-T/16 as the student for CLIP-KD compared to the baseline.

Table 1. Analysis of FD loss weight λ_{FD}. 'scratch→converge' denotes the change of loss value from scratch to convergence.

λ_{FD}	Loss scratch→converge	ImageNet Acc	CC3M Val I2T	CC3M Val T2I
10	0.079→0.013	31.1	33.6	33.5
100	0.794→0.089	32.3	34.6	34.4
1000	7.721→0.538	33.7	36.7	36.4
2000	15.880→0.902	**34.2**	**37.1**	**36.9**
3000	30.452→1.651	34.1	37.0	36.6

3. Experiments

In this section, we conduct thorough analyses and ablation experiments to investigate CLIP-KD. Unless otherwise specified, the teacher and student visual encoders are ViT-B/16 and ViT-T/16, respectively.

Analysis of Training performance curves of CLIP-KD Fig. 1a and Fig. 1b show performance curves of cross-modal retrieval and ImageNet classification, respectively. CLIP-KD outperforms the baseline consistently during the training process.

Analyses of hyper-parameters In this section, we investigate the impact of hyper-parameters on distillation performance.

Loss weight of FD As shown in Table 1, we examine

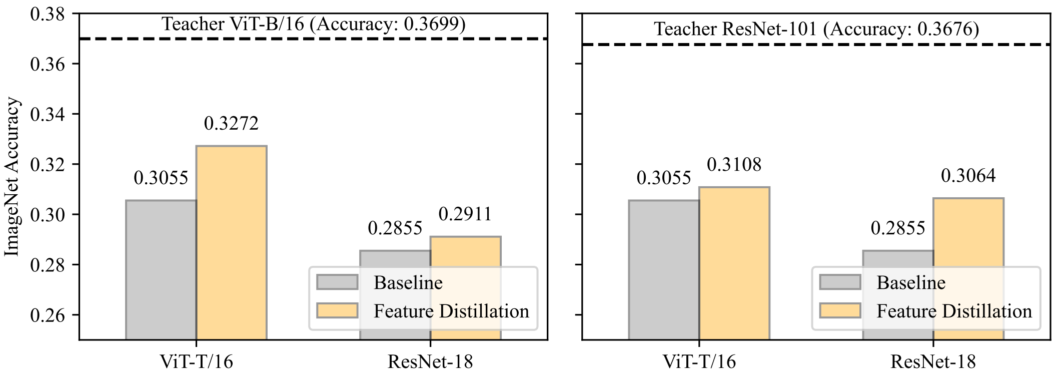


Figure 2. Top-1 accuracy on zero-shot ImageNet using intermediate feature distillation trained from CC3M+12M.

Table 2. Analysis of CRD loss weight λ_{CRD}.

λ_{CRD}	ImageNet Acc	CC3M Val I2T	CC3M Val T2I
0.5	31.6	34.9	34.6
1	**31.9**	**35.3**	**34.9**
2	31.7	35.2	34.8
10	31.2	34.9	34.6

Table 3. Analysis of GD loss weight λ_{GD}.

λ_{GD}	ImageNet Acc	CC3M Val I2T	CC3M Val T2I
10^6	30.6	33.7	33.1
10^7	30.8	33.9	33.3
10^8	**31.5**	**34.5**	**34.0**
10^9	31.4	34.2	33.7

Table 4. Analysis of ICL loss weight λ_{ICL}.

λ_{ICL}	ImageNet Acc	CC3M Val I2T	CC3M Val T2I
0.5	33.7	37.0	36.8
1	**34.2**	**37.1**	**36.9**
2	33.9	36.8	36.8
10	33.6	36.3	36.3

the impact of FD's loss weight λ_{FD}. The performance is gradually improved as λ_{FD} increases but saturates at $\lambda_{FD} = 2000$.

Loss weight of CRD As shown in Table 2, we examine the impact of CRD's loss weight λ_{CRD}. Overall, the performance is robust to the weight change, where $\lambda_{CRD} = 1$ is a suitable choice. This is because CRD loss is entropy-based KL-divergence loss, and the magnitude is consistent with cross-entropy-based task loss.

Loss weight of GD As shown in Table 3, we examine the impact of GD's loss weight λ_{GD}. The performance is gradually improved as λ_{GD} increases but saturates at $\lambda_{GD} = 10^8$.

Loss weight of ICL As shown in Table 4, we examine the impact of ICL's loss weight λ_{ICL}. Overall, the performance is robust to the weight change, where $\lambda_{ICL} = 1$ achieves the best performance. ICL has the same contrastive loss function as CLIP task loss, so $\lambda_{ICL} = 1$ leads to the same magnitude as CLIP task loss.

Mask ratio As shown in Table 5, we examine the impact of mask ratio. Using various mask ratios does not result in more performance gains than the no-masking baseline.

Table 5. Analysis of mask ratio for MFD.

Mask ratio	ImageNet Acc	CC3M Val I2T	CC3M Val T2I
0	**34.2**	37.1	**36.9**
0.25	34.1	**37.4**	36.8
0.5	33.8	37.3	36.7
0.75	33.8	37.1	**36.9**

Table 6. 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 7. 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**

Distilling intermediate features. In Figure 2, we apply intermediate feature distillation across ViT and ResNet pairs. We find homogeneous pairs achieve better accuracy than heterogeneous pairs, *e.g.*, ViT-T/16 obtains a 2.17% gain supervised by ViT-B/16 but only gets a 0.56% gain by ResNet-101. This is because the former has a more similar feature extraction process and provides student-friendly knowledge. Distilling intermediate features may be sensitive to teacher-student architectures. Therefore, we conduct the final-output-based CLIP-KD methods that use contrastive embeddings to construct distillation losses to avoid the architecture-mismatching problem.

Linear evaluation on MS-COCO object detection and instance segmentation. As shown in Table 6, we conduct downstream MS-COCO [4] object detection and instance segmentation experiments under the same linear evaluation protocol as F-VLM [3]. The backbone is a ResNet-50 pretrained on CC3M+12M. We adopt Mask-RCNN [2] framework, and apply the 1x training schedule to finetune the model. The implementation is based on MMDetection [1]. We leverage the standard COCO metric Average Precision (AP) to measure performance, including bounding box detection AP (AP^{bb}) for object detection and mask AP (AP^{seg}) for instance segmentation. CLIP-KD achieves consistent performance improvements over the original CLIP without KD by average AP margins of 1.5% and 1.2% on object detection and instance segmentation, respectively. The results indicate that CLIP-KD can also generate better distilled features under linear evaluation for downstream dense prediction tasks.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 4

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[3] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 4

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4