How to Make Cross Encoder a Good Teacher for Efficient Image-Text Retrieval? Supplementary Materials

In this supplementary materials, we further explain the differences and connections between score distribution distillation and ranking distillation, in order to analyze the advantages of ranking distillation in the process of distilling knowledge from cross-encoder to dual-encoder. We also elaborate on (1) details about pre-training datasets, downstream datasets, and evaluation metrics of downstream tasks; (2) Visualizations about image-to-text retrieval and text-to-image retrieval. (3) More ablation study for CPRD loss.

A. Score Distribution Distillation and Ranking Distillation

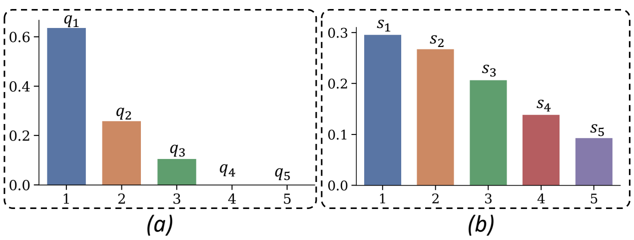

Figure 1. (a) KL-divergence-based distillation targets from cross-encoder. (b) Predicted similarity scores from student dual-encoder after softmax operation.

Score distribution distillation (*i.e.*, KL-divergence-based knowledge distillation) requires the student and teacher models have the same score distribution over multiple samples. Upon further analysis, we find that score distribution can be interpreted as ranking distillation with additional constraints. As shown in Figure 1, given an image and multiple texts $t_i,\ i \in \{1, 2, \cdots, 5\}$, we compute their similarity p_i with cross-encoder and construct distillation target q_i by applying softmax operation over these scores. A hyper-parameter τ is employed to control the sharpness of distillation target. Without loss of generality, we assume that:

$$p_1 > p_2 > p_3 > p_4 > p_5. \tag{1}$$

We can prove that:

$$\begin{gathered} if\ p_i - p_j > p_m - p_n, \\ then\ q_i - q_j > q_m - q_n, \\ \forall i, j, m, n \in \{1, 2, \cdots, 5\}, i < j \leq m < n,\ \tau > 0. \end{gathered} \tag{2}$$

Proof. According to Mean value theorem,

$$q_i - q_j = \frac{e^{p_i/\tau} - e^{p_j/\tau}}{\sum_k e^{p_k/\tau}} = \frac{(e^a)'(p_i/\tau - p_j/\tau)}{\sum_k e^{p_k/\tau}}, \tag{3}$$

where $a \in (p_i/\tau, p_j/\tau)$. Similarly,

$$q_m - q_n = \frac{(e^b)'(p_m/\tau - p_n/\tau)}{\sum_k e^{p_k/\tau}}, \tag{4}$$

where $b \in (p_m/\tau, p_n/\tau)$. Given the assumption of Equation 1 and 2, we can derive that $a > b$ and thus $q_i - q_j > q_m - q_n$.

In other words, taking t_1, t_2, t_3 as examples, $p_1 > p_2 > p_3$ and $p_1 - p_2 > p_2 - p_3$, then the values q_i satisfy $q_1 - q_2 > q_2 - q_3$ with any $\tau > 0$. Such a distillation target requires that:

$$s_1 > s_2 > s_3, \tag{5}$$

$$s_1 - s_2 > s_2 - s_3, \tag{6}$$

where s_1, s_2, s_3 is the similarity scores (after softmax) from student model. Note that the objective of Equation 5 is the same as ranking distillation. However, the additional constraint of Equation 6 may interfere with the learning of image-text alignment due to the significant difference between the similarity distributions of dual-encoder and cross-encoder, which is validated by our experimental results.

B. Datasets Details

Table 1. Statics of the pre-training datasets.

	COCO (Karpathy-train)	VG	CC3M	SBU
image	113K	100K	2.81M	825K
text	567K	769K	2.81M	825K

Pre-training datasets. We show the statistics of the images and texts of pre-training datasets in the Table 1

MSCOCO. MSCOCO [5] is a large image-text dataset of 123K images, where each image has 5 human-annotated captions. Following [3, 4, 6], we adopt the Karpathy split of MSCOCO, where 5K/5K/113K images are used for testing, validation and training respectively.

Flickr30K. Flickr30K contains 31K images and 159K captions. Each image is usually annotated with 5 captions. Following [1], we 1K/1K/29K images for testing, validation and training respectively.

Crisscrossed Captions. Crisscrossed Captions dataset [7] is an extension of MS-COCO dataset with human semantic similarity judgments for intra- and inter- modality pairs. It contains human ratings for 267,095 pairs (derived from 1,335,475 independent judgments), a massive extension in scale and detail to the 50k original binary pairings.

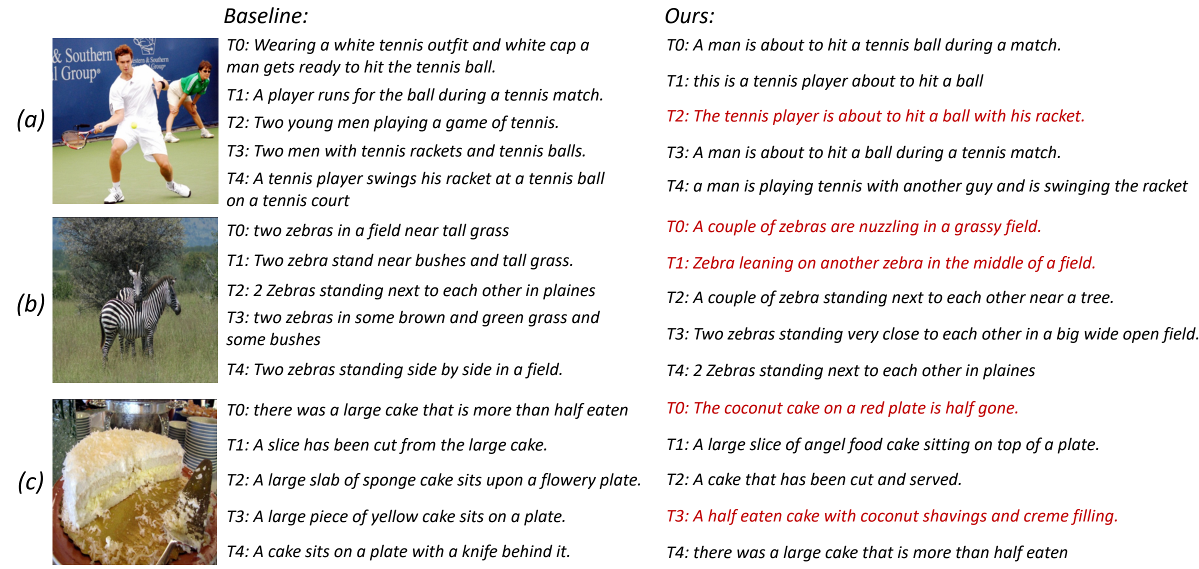

Figure 2. Illustration of image-to-text retrieval of our model and baseline model. Ground-truth captions for each image are in red color.

Figure 3. Illustration of text-to-image retrieval results of our model and baseline model. The ground-truth image for each text is in the red box.

C. Evaluation Metrics

Retrieval. We report the widely-used R@k (k=1,5,10) for cross-modal retrieval, which is the proportion of matched samples found in the top-k retrieved results. We also report R@S to reveal the overall performance, which is defined as the sum of R@k metrics at k=$\{1,5,10\}$ of both image-to-text and text-to-image retrieval tasks.

Ranking. We report the Spearman's bootstrap correlation following [2, 7] to assess whether a model ranks pairs similarly to human raters. For each correlation estimate, we sample half of the queries (to increase diversity across samples) and for each selected query, we choose one of the items for which Crisscross caption dataset supplies a paired rating. We compute Spearman's correlation between the ground-truth scores and the model scores for the selected pairs. The final correlation is the average over 1000 of these bootstrap samples.

D. Visualizations

Image-to-text Retrieval. We show image-to-text retrieval results on the MSCOCO test set in the Figure 2. We can observe that: (1) Our model has a more precise perception of detailed objects and actions in the image, *e.g.*, the baseline model erroneously identifies ”white cap”, ”run” from the (a), while our method accurately determines that it is a man hitting a ball with a racket; (2) Our model correctly recognizes detailed relation “nuzzling” and “leaning” in the (b), while the baseline model fails to achieve such recognition; (3) Our model achieves better cross-modal matching for rare concepts, as shown in (c), where our model recognizes the “coconut” and aligns it with the corresponding text.

Text-to-image Retrieval. The text-to-image results are shown in Figure 3. It can be seen that: (1) Our model perceives abstract adjectives more accurately, *e.g.*, “a modern train” in (a); (2) Our model understands local text semantics “in the midst of repairs” better and find the image that contains repair tools in (b), but the baseline model only finds the images with "kitchen" and "cabinets"; (3) Our model has better understanding on the number, *e.g.*, our model find the image with only "two" white vans accurately in (c).

E. Ablation Study

Table 2. The Spearman's rank correlation (×100) of samples from different ranking intervals between DE and CE.

Rank Interval	image→text		text→image	
	DE	+CPRD	DE	+CPRD
1-16	53.1	61.3	50.7	60.0
17-32	17.0	22.8	16.8	21.7
33-48	10.1	14.7	15.7	12.8
49-64	7.1	10.0	23.1	27.4

The effect of ranking mimicking. To validate whether our method mimics the ranking of cross-encoder, we use dual-encoder to retrieve the top 64 texts/images given each image/text of MSCOCO test dataset. Then we re-rank the retrieved texts/images in the different rank interval (*i.e.*, 1-16, 17-32, 33-48, 49-64) with cross-encoder and compute the spearman's rank correlation. As shown in Table 2, applying our CPRD method on the dual-encoder improves the rank correlation on most of the rank intervals, validating the effectiveness of our method in mimicking cross-encoder's ranking. It is worth noting that the rank correlation degrades for top 33-48 retrieved images given texts, but the relative order between these lower-ranked samples is not important and our method is designed to disregard this order.

Table 3. The performance comparison with variation of $\mathcal{L}_{ij}$.

Loss Type	image→text			text→image			R@S
	R@1	R@5	R@10	R@1	R@5	R@10	
None	32.0	59.4	71.5	24.4	49.5	61.0	297.8
$\hat{\mathcal{L}}_{ij}$	31.3	59.7	71.1	23.9	48.1	59.5	293.6
$\mathcal{L}_{ij}$	**34.3**	**61.4**	**73.2**	**27.0**	**52.8**	**64.5**	**313.2**

The variant of our proposed contrastive partial ranking distillation loss. Here, we want to explore "Does it important to constrain that valid hard negatives have higher score than easy negatives in our proposed loss?". Without such constraint, the scores of hard negatives ranked lower are trained to have smaller similarity with CPRD, and might even be lower than those easy negatives, which have a negative impact on the performance of the dual-encoder. We test the variant loss $\hat{\mathcal{L}}_{ij}$ which does not have the above constraint. The original $\mathcal{L}_{ij}$ and $\hat{\mathcal{L}}_{ij}$ are formulated as:

$$\mathcal{L}_{ij} = -log \frac{\exp(\boldsymbol{v}_i^\top \hat{\boldsymbol{t}}_{c_{ij}}/\tau)}{\sum\limits_{k=j}^{K} \exp(\boldsymbol{v}_i^\top \hat{\boldsymbol{t}}_{c_{ik}}/\tau) + \sum\limits_{k=K+1}^{B+N_q-1} \exp(\boldsymbol{v}_i^\top \hat{\boldsymbol{t}}_{d_{ik}}/\tau)}.$$

$$\hat{\mathcal{L}}_{ij} = -log \frac{\exp(\boldsymbol{v}_i^\top \hat{\boldsymbol{t}}_{c_{ij}}/\tau)}{\sum\limits_{k=j}^{K} \exp(\boldsymbol{v}_i^\top \hat{\boldsymbol{t}}_{c_{ik}}/\tau)}.$$

As shown in Table 3, $\hat{\mathcal{L}}_{ij}$ is not as good as $\mathcal{L}_{ij}$, and it even has a negative impact on the baseline model, validating the importance of ensuring that valid hard negatives have higher score than easy negatives in the distillation loss.

The choices between online hard negatives similarity calculation and offline approach. As mentioned in Sec 3.2.2, using the cross-encoder to calculate similarity scores online brings additional training costs. To reduce the training cost, we can calculate the similarity of hard negative pairs in an offline manner. It is worth noting that, compared to online method, the offline computation for one teacher is heavier due to larger candidate number but only occurs once. Offline method is thus more efficient when reusing ranking targets (*e.g.*, training multiple students with one teacher). Otherwise (*e.g.*, training a student with varying teachers), online method is more efficient. The method choice depends on the scenarios.

References

[1] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 1

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[3] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1

[4] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[5] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[6] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15692–15701, 2022. 1

[7] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870, 2021. 1, 2