A More Related Works

Contrastive Learning Contrastive learning has become a popular method for self-supervised representation learning, where the goal is to learn embedding functions such that semantically similar samples are closer in embeddings space while semantically dissimilar samples are embedded further apart [57, 64, 47, 68, 70, 13]. Among various contrastive learning methods, this paper specifically focuses on the InfoNCE loss, which optimizes the probability of correctly identifying the positive sample [60], as it has been one of the most frequently used objectives in recent contrastive learning tasks [16, 5, 3, 4].

Image Caption Evaluation Image caption evaluation is the task of automatically scoring the quality of a caption given the corresponding image and optionally reference captions. The quality of image caption metrics are evaluated with respect to its correlation with human ratings. While traditional image caption metrics such as BLEU-4 [45], ROUGE-L [38], METEOR [2], CIDEr [61], and SPICE [1] evaluate the candidate caption by measuring the n-gram overlap of the candidate with the references, more contemporary neural metrics [22, 67, 23, 30, 8] judge the candidate caption by comparing the neural embeddings of the caption with that of the image and references, and are much more flexible. Among them, the one most related to our work is CLIPScore [17]. By simply passing test image and caption pairs through the pre-trained CLIP image and text encoders and taking the dot product, CLIPScore achieves state-of-the-art performance using only the candidate caption and the corresponding image without the need for references.

B Base Representation Model Details

This paper examines the effect of a Distribution Normalization layer on the following state-of-the-art open-sourced cross-modal representation models:

CLIP [50] is one of the earliest pioneers that successfully created a joint representation space of images and text through large-scale contrastive learning. It is trained on over 400M image-caption pairs collected by the authors. They have open-sourced multiple versions of pretrained models and this paper examines the commonly used version "ViT-B/32".

ALBEF [33] is one of the later variants of CLIP that adds a multi-modal encoder to capture the interplay between images and text by predicting if they are paired samples or hard negatives. Additional momentum distillation loss and masked language modeling loss are also added to improve the model performance. Two versions of pre-trained models trained on dataset consisting of 4M and 14M unique images are released and this work uses the latter. To focus on the effect of DN on cross-modal representations, we only keep the image and text encoder component of ALBEF and do not use the multi-modal encoder component.

TCL [66] is another state-of-the-art CLIP variant. It inherits the same model architecture as ALBEF, but adds triplet contrastive loss including Cross-modal Alignment, Intra-modal Contrastive, and Local MI Maximization. Their model is trained on 4M unique images. As with ALBEF, we only keep the image and the text encoder component.

C Datasets Details

There is a total of 12 datasets mentioned in this paper, as described below.

C.1 Cross-modal Retrieval Datasets

MSCOCO [39] is a multi-purpose dataset known for its rich compatibility with object detection, segmentation, and image captioning tasks. It contains 118K images in its training split and more than 5K images in its test split.

Flicker30K [48] is a popular benchmark for sentence-based picture portrayal. It contains 31K images with each image having five reference sentences generated by human annotators. We took a train-test split following [66] and [33], which involves a selection of 30K images for fine-tuning and 1K images for testing.

C.2 Zeroshot Classification Datasets

ImageNet1K [9] is the most well-known dataset for image classification that contains more than 1200K images in its training set and 100K images in its test set and covers 1000 categories of objects. Accuracy on ImageNet1K is especially widely used for evaluating state-of-the-art image classification models.

Cifar100 [28] is a dataset containing 100 classes of objects, with each class having 500 training images and 100 test images.

SUN397 [65], the Scene UNderstanding database, contains 899 categories and about 130K images. 397 well-sampled categories are available for benchmarking state-of-the-art algorithms for scene recognition tasks.

Stanford Cars [27] is an image classification dataset dedicated for cars. It collects about 16K images for 196 types of cars and adopts a 50-50 train-test split. Results on Stanford Cars test our method's effectiveness in improving special-purpose classification models.

Caltech101 [31] is a dataset for object recognition tasks. It contains 101 object categories, with varying numbers of images in each category (between 40 to 800 images per category). The dataset has over 9,000 images in total collected from the internet. It is a popular benchmark for evaluating the performance of algorithms on general-purpose single-object recognition tasks.

Flowers102 [44] is an image classification dataset focused on flowers, comprising 102 distinct categories of flowers common to the UK. It includes a diverse set of images, with each class containing between 40 to 258 images. The dataset comprises over 8,000 images.

C.3 Image Captions Evaluation Datasets

Flickr8K-Expert [18] is a curated subset of the Flickr8K dataset that contains 17K human ratings of image-caption pairs. Each human score corresponds to one pair and is from 1 to 4 (1 means the caption is irrelevant, 4 means the caption describes the image fully correctly.)

Flicker8K-CF [18] is a similar dataset gathered from CloudFlower that contains 48K image-caption pairs and 145K binary ratings of these pairs.

THumb [24] is a dataset containing some machine- and human-generated captions from the MSCOCO dataset. It has 500 images, each with 5 candidates captions that are evaluated by human annotators.

Pascal-50S [51] contains 4K caption-caption pairs with each pair describing the same image. For all pairs, annotators give a preference on which caption in the pair provides better description of the image. Caption-caption pairs are categorized based on the origins of the captions (see Table 5 for details about categories.)

D Additional Results

D.1 Fine-tuning Epochs on MSCOCO and Flickr30K

In Figure 3, we present the remaining two plots of fine-tuning ablations for MSCOCO. CLIP + DN* does better than CLIP by an average of 0.52% for Acc@5, and 0.37% for Acc@10. Both graphs observe a peak in accuracy around 4-5 epochs. For Acc@5, we observe an initial improvement of 9.10% for CLIP + DN* after just 1 epoch, and 11.10% for CLIP on text-to-image retrieval. These respective numbers are 5.17% and 4.39% for image-to-text retrieval. For Acc@10, we observe an initial improvement of 8.53% for CLIP + DN* and 10.03% for CLIP on text-to-image retrieval, and 3.14% and 3.14% respectively for image-to-text retrieval.

In Figure 4, we present the effects of finetuning on Acc@5 and Acc@10 for Flickr30K. Again, CLIP+DN* does better than CLIP by an average of 0.22% for Acc@5 and 0.14% for Acc@10. All results yield a similar trend in support of the conclusion we have made in Section 4.6.2.

	HC	HI	HM	MM	Mean
CLIP [17]	55.0	99.2	96.9	71.8	80.7
CLIP + TTA	55.2	99.2	96.8	71.8	80.8
CLIP + TTA + DN	55.1	99.2	**97.4**	73.7	81.4
CLIP + TTA + DN*	55.2	99.3	**97.4**	72.9	81.2
CLIP + DN	**56.1**	99.1	97.3	**73.9**	**81.6**
CLIP + DN*	55.6	**99.3**	97.3	73.6	81.4
TCL [66]	52.4	91.8	54.7	63.4	65.6
TCL + DN	52.4	**96.7**	**65.7**	**66.4**	**70.3**
TCL + DN *	**52.7**	93.8	59.0	64.3	67.4
ALBEF [33]	56.9	98.1	**82.5**	67.2	**76.2**
ALBEF + DN	56.1	97.9	81.2	67.0	75.6
ALBEF + DN*	56.9	98.1	81.7	**67.8**	76.1
BLEU-4	60.4	90.6	84.9	54.7	72.6
CIDEr [61]	65.1	98.1	90.5	64.8	79.6
ViLBERTScore-F [67]	49.9	99.6	93.1	75.8	79.6
CLIP -ref [17]	**62.4**	99.7	96.7	73.0	83.0
CLIP + DN -ref	60.8	99.6	**97.8**	**75.1**	**83.3**
CLIP + DN* -ref	61.1	99.7	97.3	74.8	83.2

Table 5: Accuracy results on Pascal-50S given different categories of caption-caption pairs. HC means two correct human-generated captions. HI means two human-generated captions with one incorrect. HM means a human-generated and a machine-generated caption. MM means two machine-generated captions.

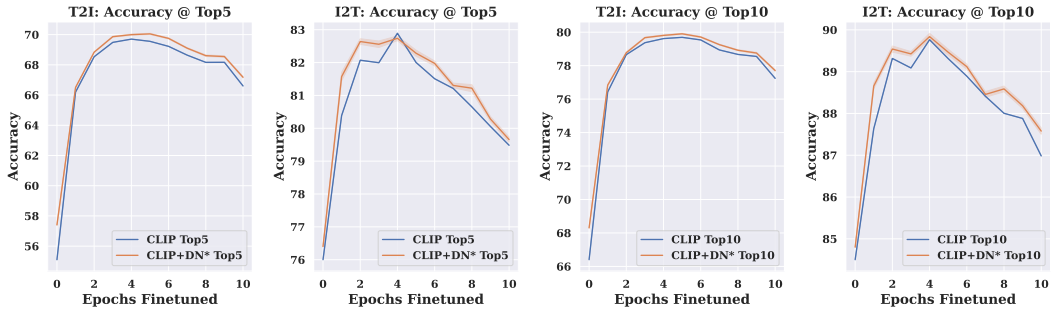

Figure 3: Comparison of the effects of fine-tuning between CLIP and CLIP + DN* on MSCOCO's 5k test set. We report text-to-image retrieval (left) and image-to-text retrieval (right) for Acc@5 and Acc@10. The average accuracy over 5 checkpoints trained with 5 random seeds is plotted. For each of the 5 checkpoints we trained, we find its average accuracy and standard deviation with another 5 iterations random sampling for mean estimation, and plot the mean of these 5 accuracies and standard deviations from 5 independently fine-tuned checkpoints.

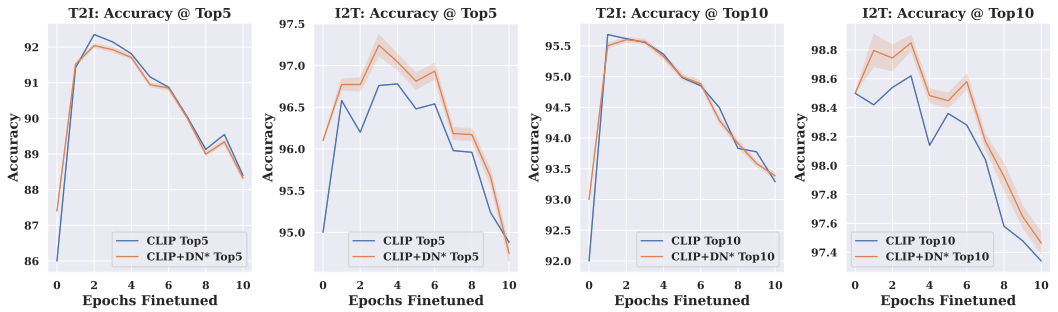

Figure 4: Comparison of the effects of fine-tuning between CLIP and CLIP + DN* on Flickr30K's 1k test set. We report text-to-image retrieval (left) and image-to-text retrieval (right) for all Acc@5 and Acc@10. The average accuracy over 5 checkpoints trained with 5 random seeds is plotted. For each of the 5 checkpoints we trained, we find its average accuracy and standard deviation with another 5 iterations random sampling for mean estimation, and plot the mean of these 5 accuracies and standard deviations from 5 independently fine-tuned checkpoints.

D.2 Information Loss in Approximation

Without considering computational efficiency, for each test sample, we can iterate over the entire unlabeled reference set to calculate a similarity measure given in Eqn.5:

$$S_{full}(x_0, y_0) = \mathbb{E}_{y_1 \sim \mathcal{D}_T} e^{\phi(x_0)^\intercal [\psi(y_1) - \psi(y_0)]/\tau} + \mathbb{E}_{x_1 \sim \mathcal{D}_T} e^{[\phi(x_1) - \phi(x_0)]^\intercal \psi(y_0)/\tau}. \quad (10)$$

To investigate how much information might be lost from the first-order approximation in Section 3.2.2 and algebraic mean to geometric mean conversion in Section 3.3 that simplifies Eqn.10 to distribution normalization in Eqn.9, we carry out experiments to compare them back to back in the task of image captioning metric and present our results in Table 6 in terms of their correlation with human judgments on image captioning datasets. Surprisingly, in all the base models and datasets that we have studied, we did not notice any significant difference between DN and Eqn. 10 (difference $< 0.1\%$). This shows that higher-order information contributes negligibly to the downstream applications compared to only taking the mean of the distribution, and taking an algebraic mean achieves a similar effect as taking a geometric mean.As a conclusion, we found that DN provides equivalent performance to the original Eqn.5 without incurring expensive computational costs.

D.2.1 Pixel-Level Normalization

A potential alternative to distribution normalization which normalizes the data on the representation level is a commonly used trick that normalizes the data (mostly images) on the input level (pixel level), which we refer as pixel-level normalization. In Table 7, we present our the comparison results between CLIP + pixel-norm and our proposed methods. However, although using the same data for normalization on the representation level as DN and DN* yields a large gain over vanilla CLIP, changing the normalization to the pixel level wipes out all the improvements. We hypothesize that this is because the difference of a constant vector (representation mean) vanishes while the input goes through a large neural network like CLIP.

	Flickr8k-expert τ_c	Flickr8k-cf τ_b	THumb τ_c
CLIP [50, 17]	51.4	34.3	19.9
CLIP + DN	54.3	35.4	23.5
CLIP + DN (Eqn.10)	54.3	35.4	23.4
TCL [66]	31.0	20.6	8.1
TCL + DN	42.0	26.4	14.4
TCL + DN (Eqn.10)	41.8	26.4	14.3
ALBEF [33]	24.9	15.4	0.9
ALBEF + DN	34.8	21.8	5.5
ALBEF + DN (Eqn.10)	34.8	21.8	5.5

Table 6: Abaltion study on comparison with distribution normalization and full Eqn.5 on Flickr8k-Expert, Flickr8k-CF, and THumb.

		Flickr8k-expert τ_c	Flickr8k-cf τ_b	THumb τ_c
	CLIP [17]	51.4	34.3	19.9
Ref-	CLIP + pixel-norm	51.3	34.3	19.4
free	CLIP + DN	54.3	35.4	**23.5**
	CLIP + DN*	53.2	35.1	22.2

Table 7: Ablation study on pixel-level normalization on Flickr8k-Expert, Flickr8k-CF, and THumb.