Test-Time Distribution Normalization For Contrastively Learned Vision-language Models





Yifei Zhou*, Juntao Ren*, Fengyu Li*, Ramin Zabih, Ser-Nam Lim

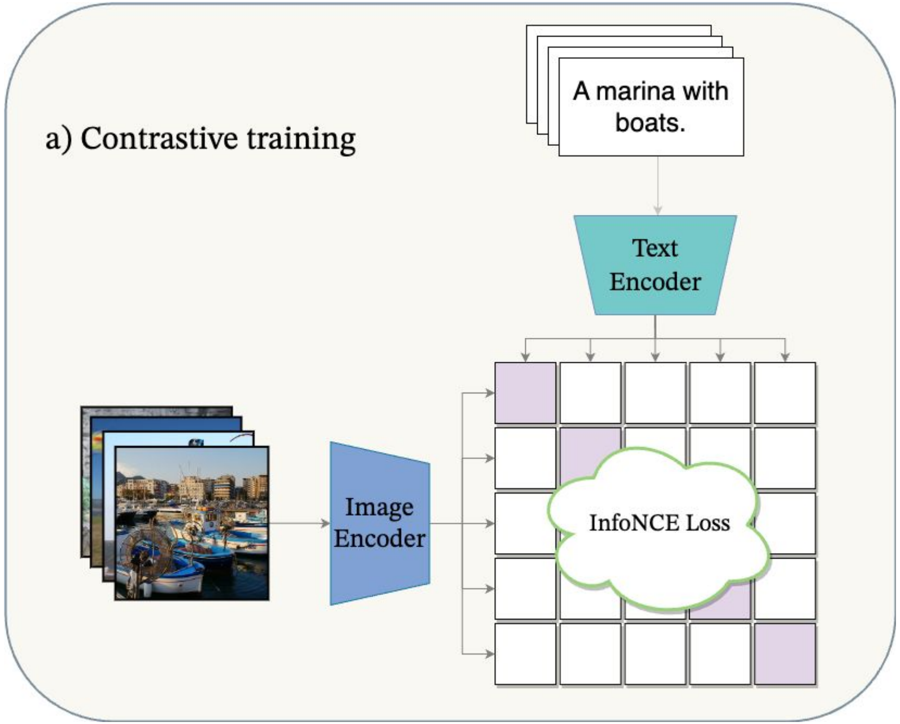
Berkeley UNIVERSITY OF CALIFORNIA

Cornell University

UCF UNIVERSITY OF CENTRAL FLORIDA

Outline

➔ Background: Contrastive Learning and CLIP
➔ Motivation
➔ Algorithm Derivation
➔ Experiments
➔ Conclusion



A commonly used objective for training contrastive models is the InfoNCE loss, such as in CLIP[1].

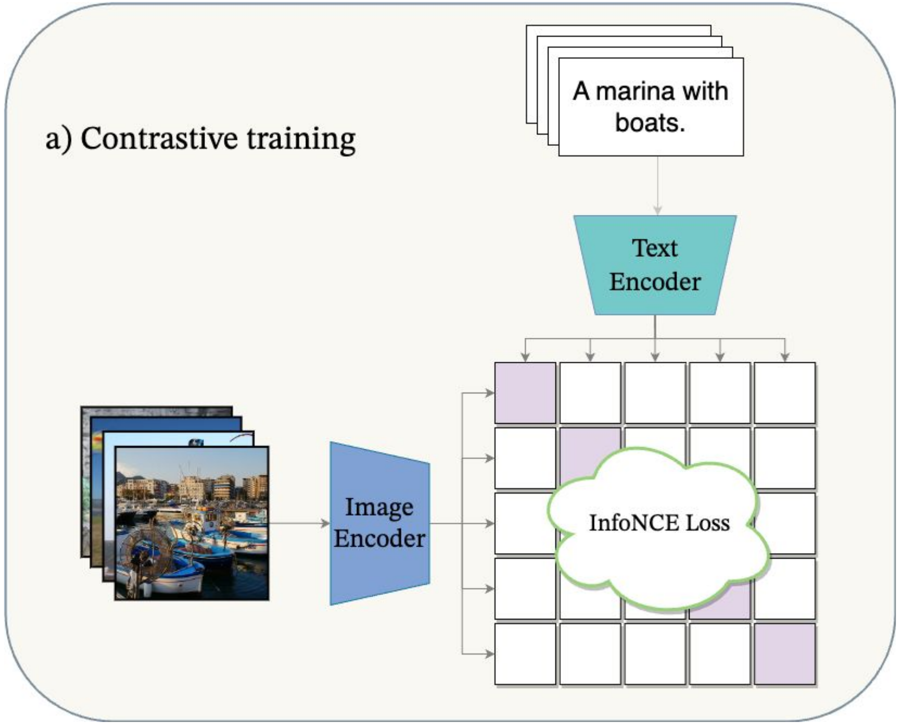
[1]Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

InfoNCE Loss

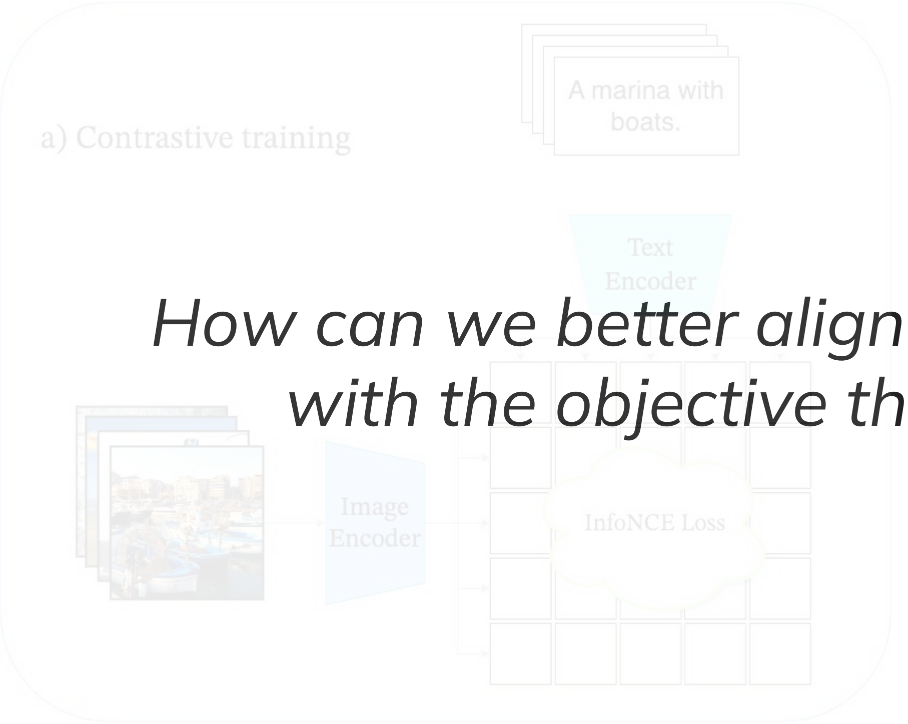
During train-time, this objective groups together similar image and text embeddings (positive samples), and pushes apart negative ones.

$$\mathcal{L}_{\mathrm{NCE}} = -\mathbb{E}\left[\log \frac{f(\mathbf{x}, \mathbf{c})}{\sum_{\mathbf{x}' \in X} f(\mathbf{x}', \mathbf{c})}\right]$$

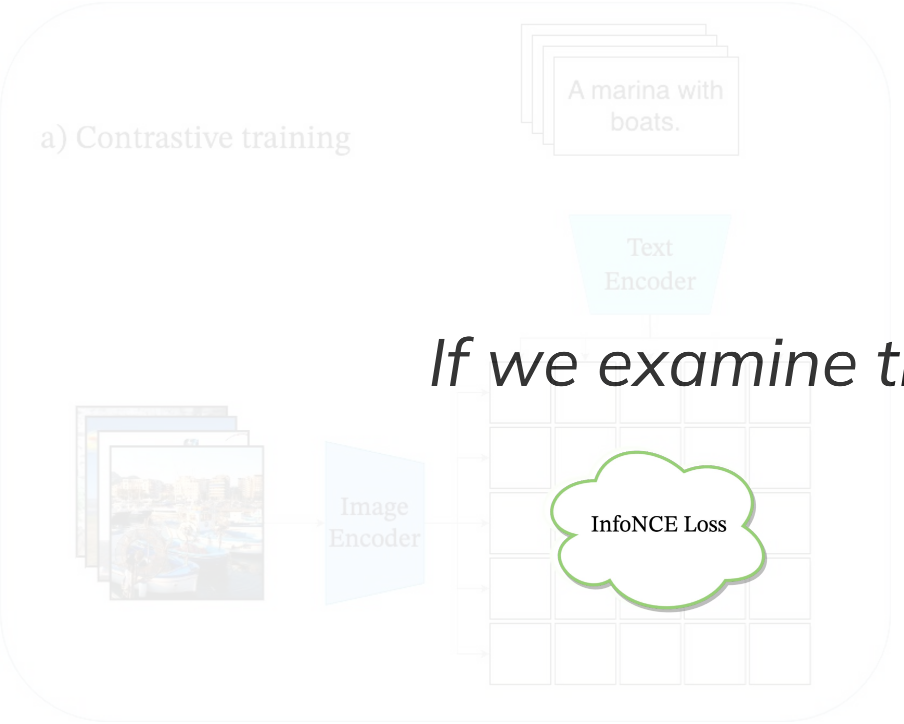
*Where c and x are positive image and text pairs, while x’ are negative samples.



Conventional application of dot product is different from the InfoNCE loss, as it does not leverage negative samples from the test-time distribution.



How can we better align downstream applications with the objective they were trained with?



If we examine the InfoNCE loss...

InfoNCE Loss

Taylor expanding the InfoNCE loss gives us the following:

$$\begin{aligned}\mathcal{L}_{\mathrm{NCE}}(\mathcal{D}_S) \approx n\mathbb{E}_{x_0,y_0\sim\mathcal{D}_S}\Big[&\mathbb{E}_{y_1\sim\mathcal{D}_S} e^{\phi(x_0)^\intercal[\psi(y_1)-\psi(y_0)]/\tau} \\ &+\mathbb{E}_{x_1\sim\mathcal{D}_s} e^{[\phi(x_1)-\phi(x_0)]^\intercal\psi(y_0)/\tau}\Big].\end{aligned}$$

*Where x_0, y_0 are image-text pairs, D_S is the training distribution, τ is the temperature constant, and ϕ, ψ are image and text encoders respectively.

InfoNCE Loss

A successful generalization of the training object would be to minimize the same loss on samples from the test-time distribution.

$$\mathcal{L}_{\mathrm{NCE}}(\mathcal{D}_T) \approx n\mathbb{E}_{x_0,y_0\sim\mathcal{D}_S}\Big[\mathbb{E}_{y_1\sim\mathcal{D}_T} e^{\phi(x_0)^\intercal[\psi(y_1)-\psi(y_0)]/\tau} + \mathbb{E}_{x_1\sim\mathcal{D}_T} e^{[\phi(x_1)-\phi(x_0)]^\intercal\psi(y_0)/\tau}\Big].$$

*Where x_0, y_0 are image-text pairs, D_T is the testing distribution, τ is the temperature constant, and ϕ, ψ are image and text encoders respectively.

InfoNCE Loss

We find that the typical practice of taking a dot-product similarity is equivalent to only a zeroth order approximation of the objective.

$$\mathcal{L}^{(0)}_{\text{NCE}}(\mathcal{D}_T) = 2n \cdot \mathbb{E}_{x_0, y_0 \sim \mathcal{D}_T} \left[e^{\phi(x_0)^\intercal \psi(y_0)/\tau} \right]$$

$$S_{(0)}(x_0, y_0) = \phi(x_0)^\intercal \psi(y_0)$$

*where x_0, y_0 are image-text pairs, D_S is the training distribution, ϕ, ψ are image and text encoders respectively

🔑 *Insight: Using the first-order approximation of the InfoNCE Loss helps us better align test time behavior with the training objective.*

InfoNCE Loss

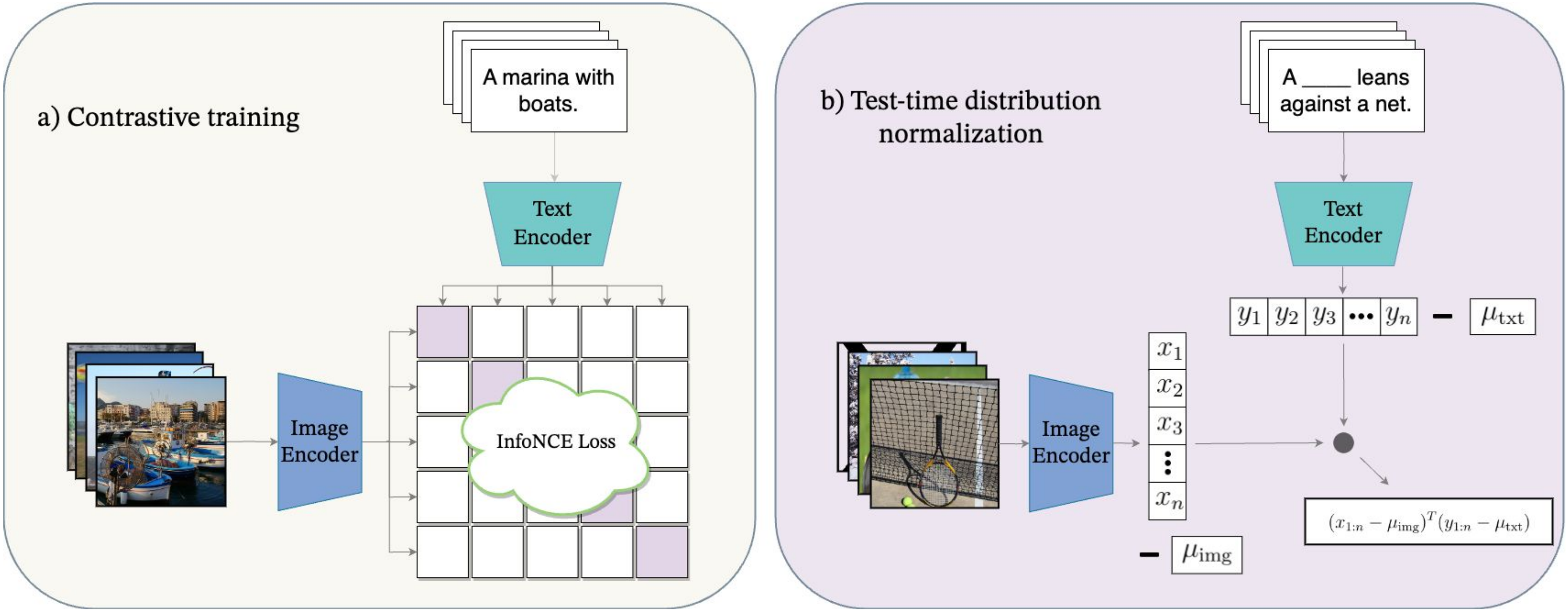
Subtracting the mean of the test-time distribution gives us better alignment to the training objective and boost performance.

$$\mathcal{L}^{(1)}_{\text{NCE}}(\mathcal{D}_T) = 2n \cdot \mathbb{E}_{x_0, y_0 \sim \mathcal{D}_T}\left[e^{\phi(x_0)^\intercal(\mu_y - \psi(y_0))/\tau} + e^{(\mu_x - \phi(x_0))^\intercal \psi(y_0)/\tau}\right]$$

$$S_{(1)}(x_0, y_0) = \left(\phi(x_0) - \frac{1}{2}\mu_x\right)^\intercal \left(\psi(y_0) - \frac{1}{2}\mu_y\right)$$

*Where x_0, y_0 are image-text pairs, D_S is the training distribution, τ is the temperature constant, and ϕ, ψ are image and text encoders respectively.
μ_x and μ_x represent the mean of the test-time image and text embeddings.

This first order approximation is extremely simple *to implement, while still giving us better alignment to the training objective.*


a) Contrastive training
A marina with boats.
Text Encoder
Image Encoder
InfoNCE Loss
b) Test-time distribution normalization
A ____ leans against a net.
Text Encoder
y_1 y_2 y_3 ••• y_n − μ_{txt}
Image Encoder
x_1 x_2 x_3 ⋮ x_n
− μ_{img}
$(x_{1:n} - \mu_{\text{img}})^T(y_{1:n} - \mu_{\text{txt}})$

DN achieves SOTA on retrieval benchmarks, *and* can be easily adapted on top of other test time adaptation modules.[†]

Cross-Modal Retrieval on MSCOCO (5K Test Set)						
	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP[1]	52.4	76.0	84.5	30.2	55.1	66.4
CLIP + DN[*]	52.9	76.4	84.9	32.1	57.4	68.3
CLIP + TTA[1]	53.9	77.5	85.5	32.1	57.5	68.3
CLIP + TTA + DN[*]	**54.7**	**77.8**	**85.6**	**33.8**	**59.4**	**70.1**

[†] More results can be found in our paper!
[1]Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning.* PMLR, 2021.
[2]Shanmugam, Divya, et al. "Better aggregation in test-time augmentation." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Zero-shot Classification						
	ImageNet1K		Cifar100		SUN397	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
CLIP	61.0	87.4	63.9	88.7	56.1	89.4
CLIP + DN*	61.7	87.8	65.1	89.4	57.3	90.2
CALIP[1]	61.2	87.5	64.2	88.9	56.1	89.3
TPT[2] (Inefficient)	63.5	87.1	65.2	88.1	**59.4**	88.8
CLIP + TTA	62.4	88.5	66.0	90.5	56.9	90.0
CLIP + TTA + DN*	**63.2**	**88.9**	**67.1**	**90.7**	58.1	**90.7**

[1]Guo, Ziyu, et al. "Calip: Zero-shot enhancement of clip with parameter-free attention." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 1. 2023.
[2]Shu, Manli, et al. "Test-time prompt tuning for zero-shot generalization in vision-language models." *Advances in Neural Information Processing Systems* 35 (2022): 14274-14289.

Conclusion

★ We identify a mismatch between the training objective and downstream application of contrastively trained models.

★ We show that the conventional dot-product similarity corresponds to a zeroth-order approximation of the InfoNCE loss.

★ We find that using the first-order approximation gives us better alignment and performance, and is super simple to implement on any existing contrastive model without any additional finetuning.