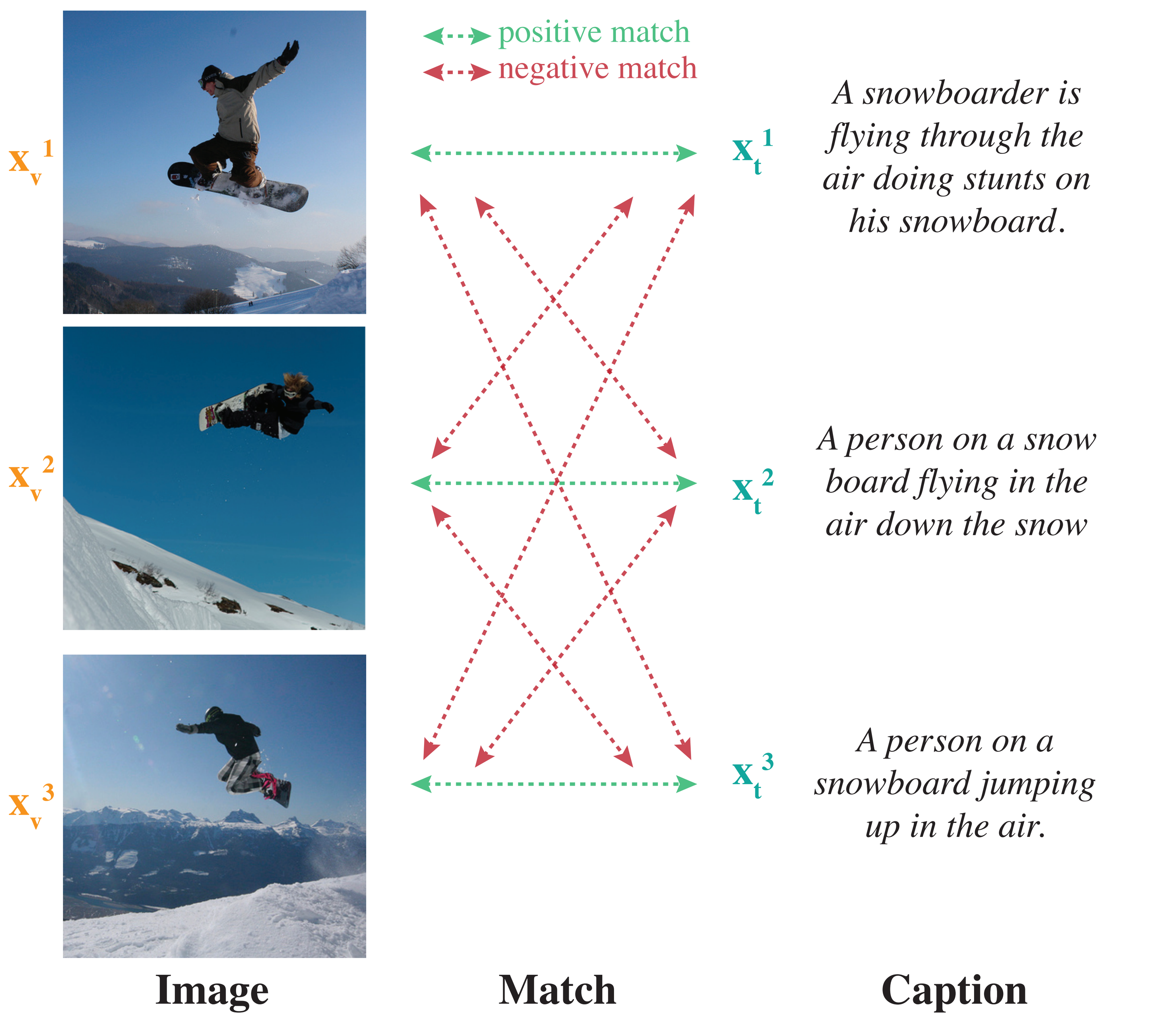
Improved Probabilistic Image-Text Representations

Sanghyuk Chun

NAVER AI LAB

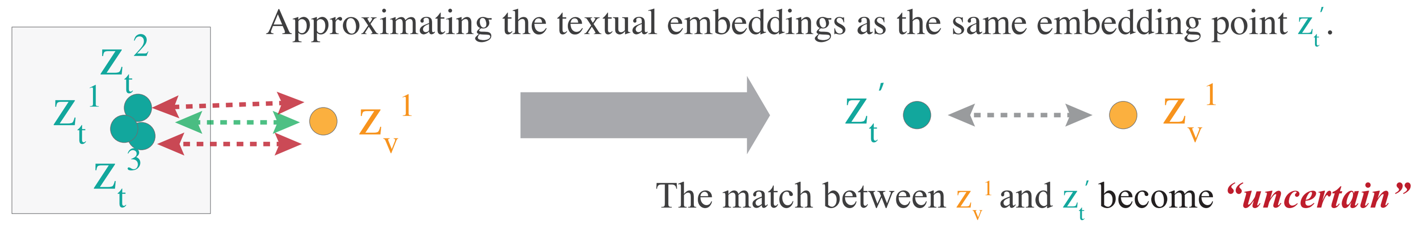
Motivation: Inherent ambiguity in image-text matching problem



There are many false negatives (FNs) in dataset, and it naturally leads to ambiguous supervision to the model

Deterministic embedding space

Approximating the textual embeddings as the same embedding point Z_t'.

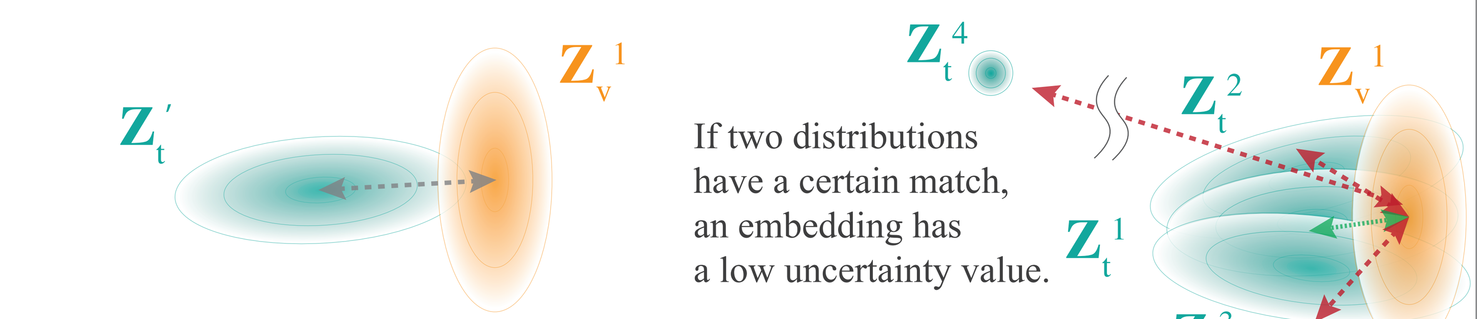


The match between Z_v^1 and Z_t' become *"uncertain"*

→ Visual/textual embeddings will be located closely as they are semantically almost same. Image-text match (ITM) supervision becomes uncertain (either matched or not), but a deterministic embedding space cannot capture uncertainty.

→ The existing deterministic approaches enforces to the textual embeddings to be different by making Z_t^1 differs from others using HNM, although they are semantically the same.

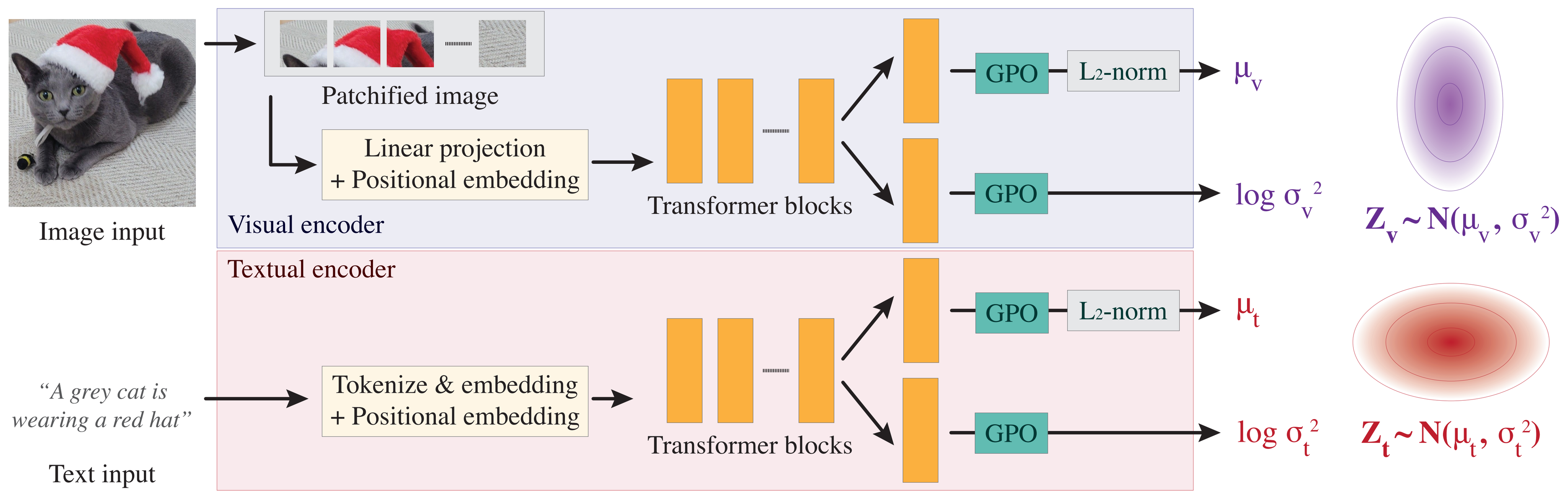
Probabilistic embedding space



If the match between Z_v^1 and Z_t' become *"uncertain"*, the embedding uncertainty (*i.e.*, variance becomes larger)

→ A probabilistic embedding space captures uncertainty of noisy ITM supervision. If a match is uncertain, then the variance of embeddings becomes larger (Z_t^1, Z_t^2, Z_t^3), if a match is certain, then the variance becomes smaller (Z_t^4)

PCME++ An improved probabilistic VL model



How can we train a proper probabilistic embedding space? In other words, how to measure a distance between two distributions? Here, we expect three properties for the probabilistic embeddings:

1) There exists a proper probabilistic distance between two distributions
2) If the match between two distributions is certain, then the distributions should have small variances
3) If the match between two distributions is uncertain, then the variances should be large.

Popular distances such as Wasserstien dist, KL divergence cannot satisfy (2).

Monte-Carlo approximiation of matching probability (Chun et al. 2021)

$$p_\theta(m|x_\alpha, x_\beta) \approx \frac{1}{J^2}\sum_j^J \sum_{j'}^J s(-a\|z_\alpha - z_\beta\|_2 + b)$$

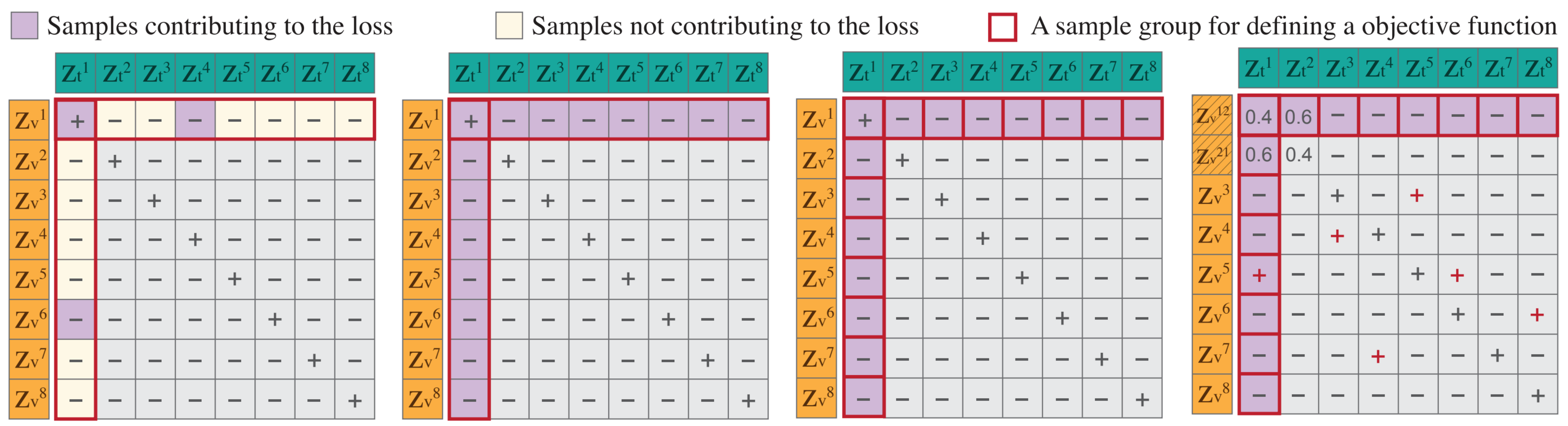
It satisfies all conditions, but the MC sampling itself has flaws in computation and accuracy

Improvement 1: Closed-form Sampled Distance (CSD)

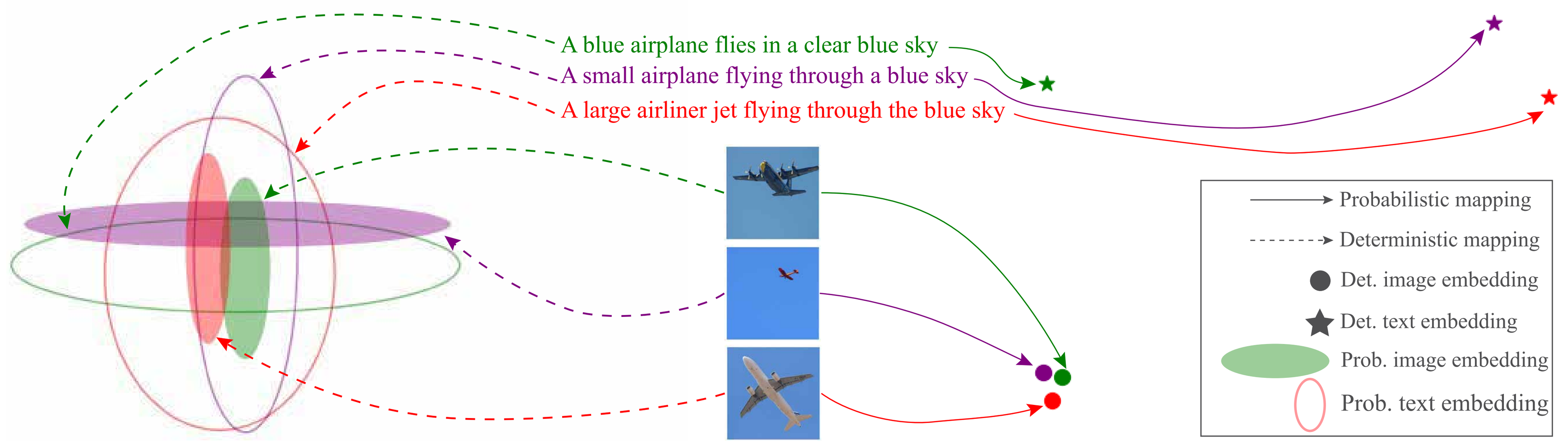
$$d(\mathbf{Z}_v, \mathbf{Z}_t) = \mathbb{E}_{\mathbf{Z}_v, \mathbf{Z}_t}\|\mathbf{Z}_v - \mathbf{Z}_t\|_2^2 = \|\mu_v - \mu_t\|_2^2 + \|\sigma_v^2 + \sigma_t^2\|_1$$

$$\mathcal{L}_{\text{match}} = -m_{vt} \log \text{sigmoid}(-a \cdot d(\mathbf{Z}_v, \mathbf{Z}_t) + b) - (1 - m_{vt}) \log \text{sigmoid}(a \cdot d(\mathbf{Z}_v, \mathbf{Z}_t) - b)$$

Improvement 2: Pseudo-Positive (PP) and Mixed data sample augmentation (MSDA)



(a) Triplet loss (b) Batch-wise contrastive loss (c) Pair-wise contrastive loss (d) (c) + Pseudo-positive / MSDA



Experimental results

COCO Caption results

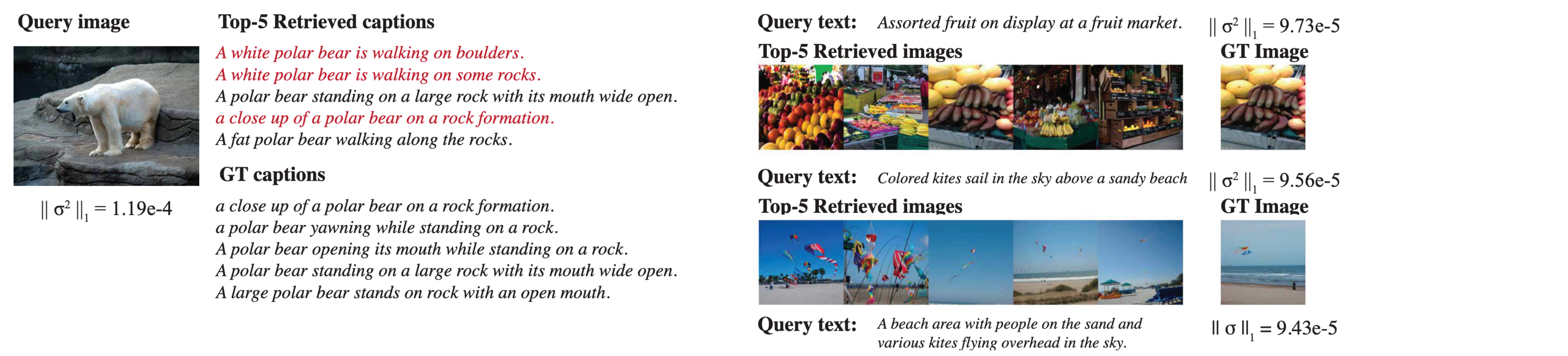
Backbone	Method	Prob?	ECCV Caption mAP@R	ECCV Caption R-P	ECCV Caption R@1	COCO 5K R@1	COCO RSUM
ViT-B/32 (151M)	CLIP ZS†	✘	26.8	36.9	67.1	40.3	471.9
	VSE∞	✘	<u>40.0</u>	<u>49.5</u>	**83.1**	**55.2**	<u>536.5</u>
	P2RM	✘	39.0	48.7	82.0	51.7	530.2
	DAA	✘	39.2	49.0	82.0	52.9	530.9
	InfoNCE	✘	39.0	48.7	81.7	53.0	532.6
	PCME	✔	39.1	48.9	81.4	53.0	532.0
	PCME++ (μ only)	✘	39.5	49.1	82.7	55.2	536.2
	PCME++	✔	**40.1**	**49.7**	**83.1**	<u>55.1</u>	**537.0**
	PCME++ (SWA)	✔	40.2	49.8	82.9	55.2	537.3
ViT-L/14 (428M)	CLIP ZS†	✘	28.0	37.8	72.2	46.4	491.6
	VSE∞	✘	20.2	31.5	46.2	22.7	424.3
	InfoNCE	✘	35.6	45.8	75.6	45.9	520.6
	PCME	✔	<u>41.2</u>	<u>50.3</u>	<u>86.0</u>	<u>61.9</u>	<u>550.4</u>
	PCME++	✔	**42.1**	**50.8**	**88.8**	**64.3**	**554.7**

COCO Noisy correspondence results

Noise	Method	ECCV Caption mAP@R	ECCV Caption R-P	ECCV Caption R@1	COCO 5K R@1	COCO RSUM
20%	VSE∞	37.0	46.3	<u>79.7</u>	**51.8**	<u>518.6</u>
	DAA	6.7	12.5	18.5	6.0	212.8
	InfoNCE	35.9	46.3	76.1	45.8	514.6
	PCME	<u>37.6</u>	**47.6**	79.2	48.7	520.7
	PCME++ (ours)	**37.7**	**47.6**	**80.0**	<u>50.4</u>	**524.6**
	NCR†	35.9	46.0	78.0	48.8	<u>518.6</u>
	DECL†	-	-	-	49.4	518.2
50%	VSE∞	18.0	28.5	43.7	19.1	394.1
	InfoNCE	33.6	44.1	73.0	41.4	499.5
	PCME	<u>35.2</u>	<u>45.5</u>	<u>75.7</u>	44.4	508.0
	PCME++ (ours)	**35.7**	**45.8**	**76.3**	<u>45.5</u>	<u>511.0</u>
	NCR†	34.0	44.3	75.1	<u>45.5</u>	508.5

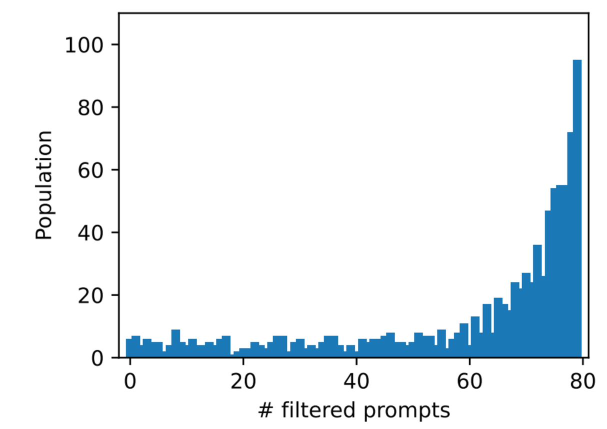
Observation 1: When scaling-up backbone, det. methods are overfitted, but prob. methods are not.

Observation 2: PCME++ successfully handles NC although it is not designed for tackling NC.



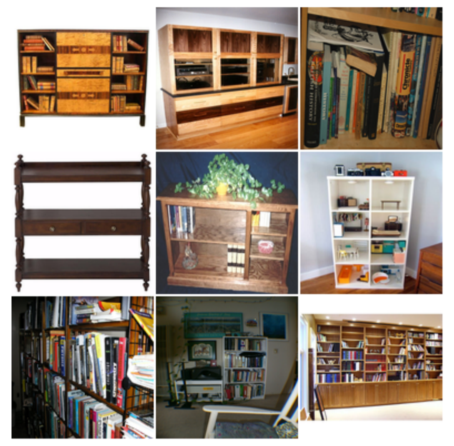
Large-scale pre-training with PCME++

Model	Prompts	Top-1 ImageNet Zero-shot acc
CLIP	"A photo of {・}"	31.85
	All 80 prompts	35.50
PCME++	"A photo of {・}"	30.43
	All 80 prompts	34.22
	Top-K certain prompts	34.22
	Best top-K for each class	**41.82**



ViT-B/16 trained on CC3M,12M,RedCaps

Bookcase



a close-up photo of a { }, a close-up photo of the { }

Jack-o'-lantern

a photo of the clean { }. art of the { }. a cropped photo of the { }. a close-up photo of the { }. a photo of a clean { }. a cropped photo of a { }

Rapeseed

a sculpture of the { }. a { } in a video game. a sculpture of a { }. art of the { }. the { } in a video game. a tattoo of the { }. graffiti of the { }. the plushie { }. a tattoo of a { }. a drawing of a { }. a drawing of the { }. a sketch of the { }. a close-up photo of the { }. art of a { }. a photo of a clean { }. a plushie { }. a close-up photo of a { }. a photo of the clean { }. a rendering of the { }. a photo of the large { }. a rendering of a { }. a sketch of a { }. a photo of the { }. a cropped photo of the { }. a rendition of a { }. graffiti of a { }. a rendition of the { }. a photo of the small { }. a photo of one { }. a photo of the dirty { }. ... 24 more prompts

Check out more details in...

Paper

Code

Check out my other related works!

PCME (CVPR'21)

ECCV Caption (ECCV'22)