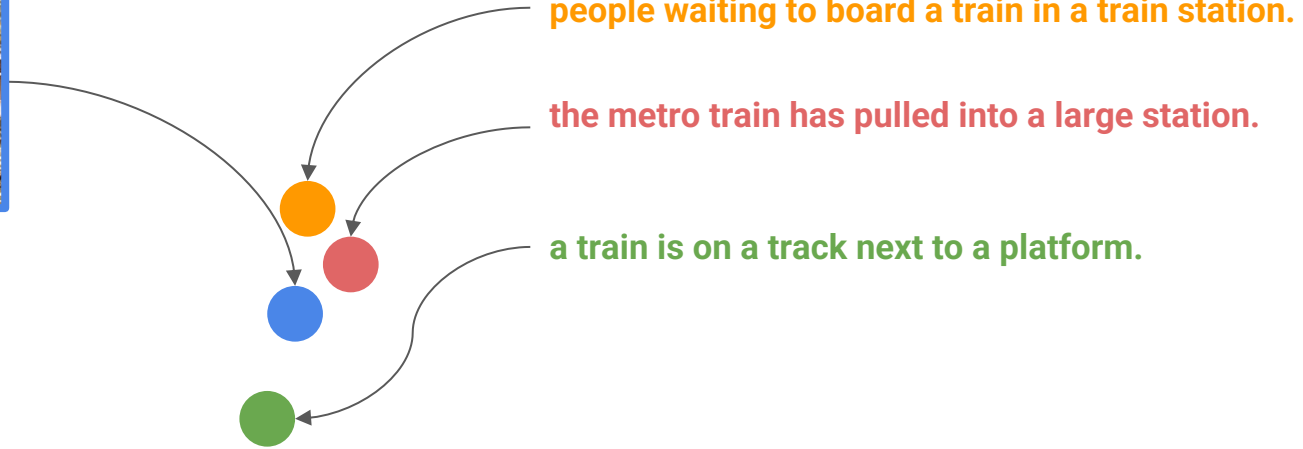
Probabilistic Image-Text Representations

Sanghyuk Chun
sanghyuk.chun@gmail.com

* slide can be found in https://sanghyukchun.github.io/home/

Overview

- In this talk, I will introduce the concept of “probabilistic representation learning” for “image-text matching problem”.
- We will focus on “the multiplicity problem of image-text representation learning” in terms of the training stage and the evaluation stage.
  - [Training stage] Probabilistic embedding and the multiplicity problem
  - [Evaluation stage] Resolving the multiplicity problem in the evaluation benchmark
- This talk is based on my recent studies:
  - [CVPR 2021] Chun, et al., **Probabilistic Embeddings for Cross-Modal Retrieval**
  - [ECCV 2022] Chun, et al., **ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO**
  - [Preprint] Sanghyuk Chun, **Improved Probabilistic Image-Text Representations**



Goal of image-text matching:

Determining whether the given image-text pair is matched or not.
A common way: making a shared shared embedding space,
where matched image-caption pairs are closer than unmatched pairs in the shared embedding space.

people waiting to board a train in a train station.

the metro train has pulled into a large station.

a train is on a track next to a platform.

Objective function:

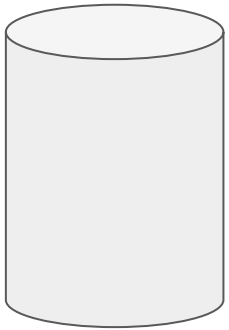
Let matched image-caption pairs closer together, and unmatched image-caption pairs farther away

a bird with its head in an open oven

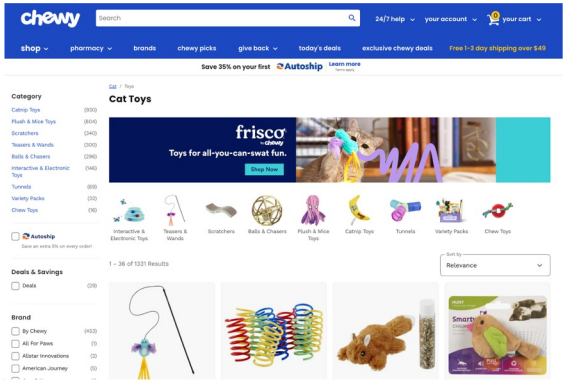
Preliminary: Applications of Image-Text Matching (ITM)

- Image-text multimodal search
  - Query: image DB: text documents
  - Query: text DB: images

Query



DB



Preliminary: Applications of Image-Text Matching (ITM)

- Zero-shot classification
  - Don’t train a new model for new class labels, but add new texts!
  - Scale-up datasets using aligned image-text pairs, not manually annotated images

Query

→

A photo of [Cat]
A photo of [Dog]
A photo of [Deer]
A photo of [Coffee]
…

DB

→ **A photo of [Cat]**

Preliminary: Applications of Image-Text Matching (ITM)

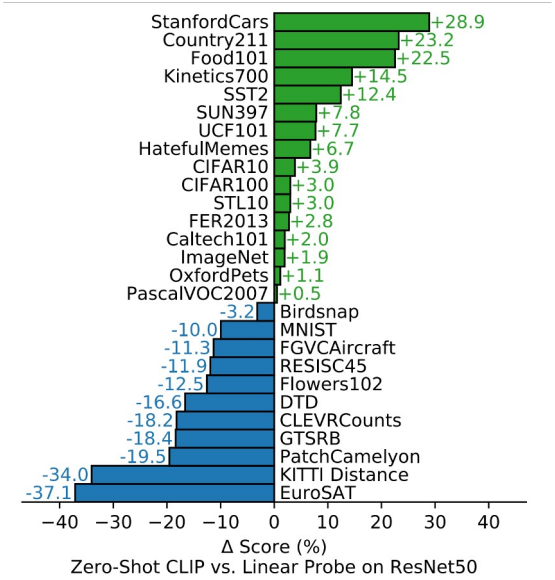


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

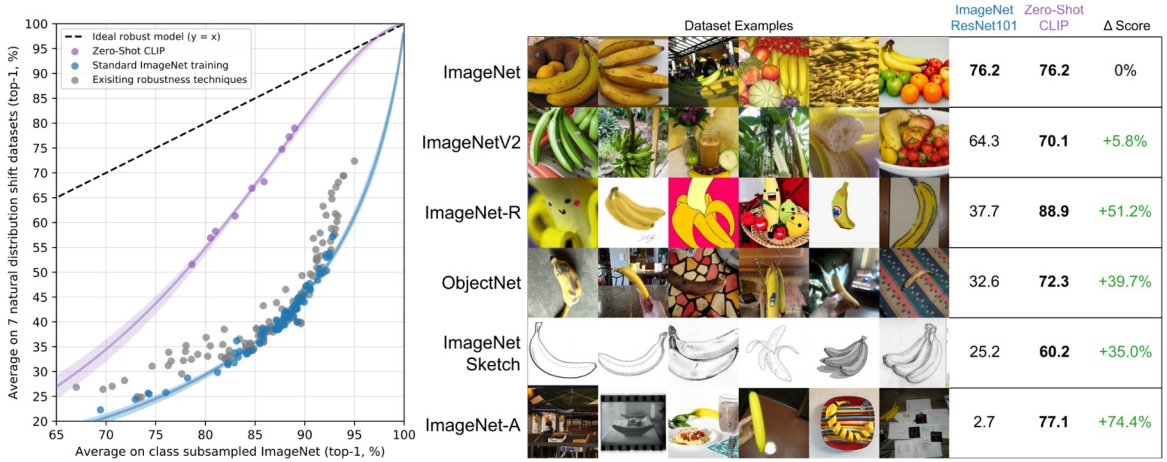
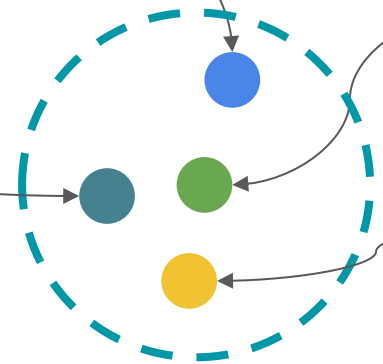


Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

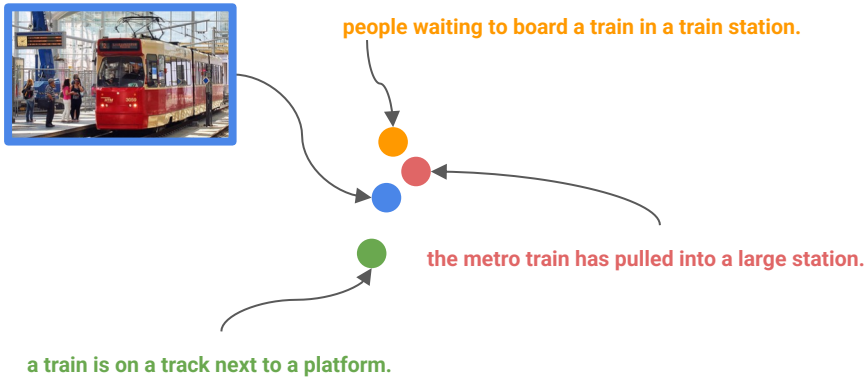
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.

Problem definition: Multiplicity (many-to-many) problem



a train is on a track next to a platform.

A common concept

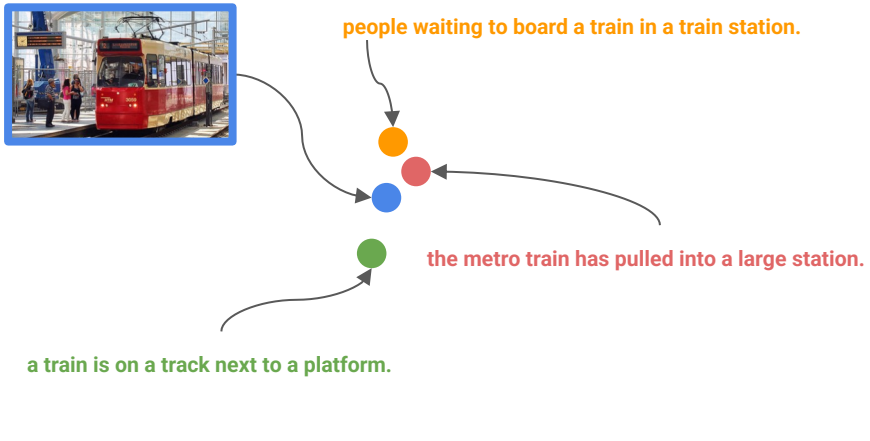


“Many-to-many mapping” challenges in the image-text matching (ITM) tasks:

- An image potentially can be matched with a number of different captions.

“Many-to-many mapping” challenges in the image-text matching (ITM) tasks:

- An image potentially can be matched with a number of different captions.
- While an image is taken by thoroughly captured in a photograph, language descriptions are the product of conscious choices of the key relevant concepts to report from a given scene.

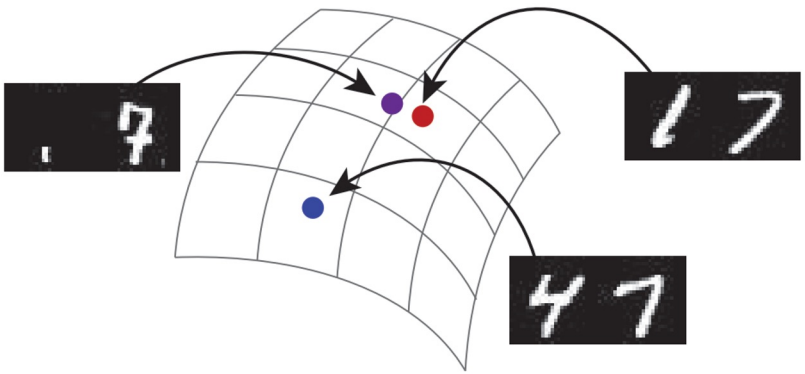


“Many-to-many mapping” challenges in the image-text matching (ITM) tasks:

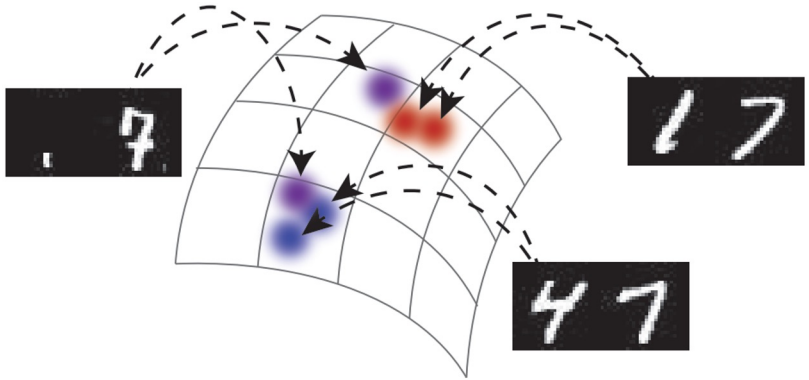
- An image potentially can be matched with a number of different captions.
- While an image is taken by thoroughly captured in a photograph, language descriptions are the product of conscious choices of the key relevant concepts to report from a given scene.

Preliminary: Probabilistic embedding (ProbEmb)

- Let each embedding Gaussian distribution, instead of a point vector.
- Can handle “ambiguous” inputs & “many-to-many problem”



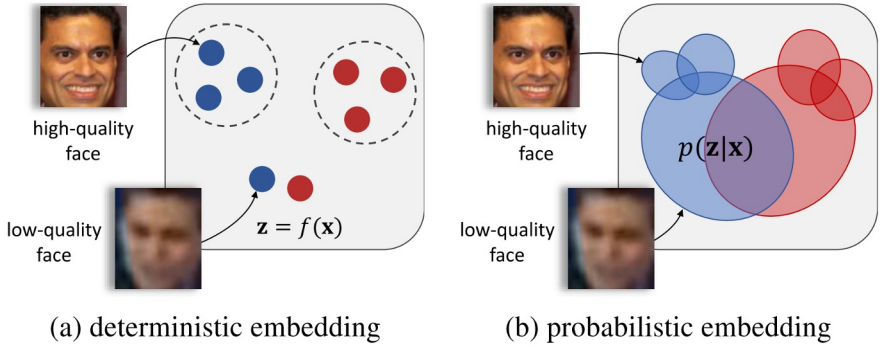
(a) Point embedding.



(b) Stochastic embedding.

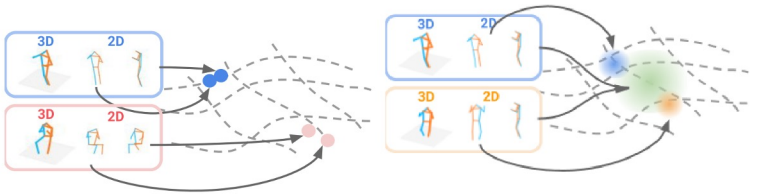
ICLR’19, Oh et al., Modeling uncertainty with hedged instance embedding

Preliminary: Applications for probabilistic embedding



(a) deterministic embedding

(b) probabilistic embedding



(a) View-Invariant Pose Embeddings (VIPE).

(b) Probabilistic View-Invariant Pose Embeddings (Pr-VIPE).

ICCV 19, Probabilistic Face Embedding

- Uncertainty captures “quality” of face images
- E.g., a low-quality face has a large Gaussian variance.

ECCV 20, View-Invariant Probabilistic Embedding for Human Pose.

- Uncertainty captures “ambiguity” of 2D pose (which can be mapped to multiple 3D poses)

MS-COCO Caption dataset for Image-Text Matching (ITM)

- MS-COCO Caption collects captions by human annotators
- During training or evaluating models on COCO Caption, the collected image-text pairs are treated as the only positive.

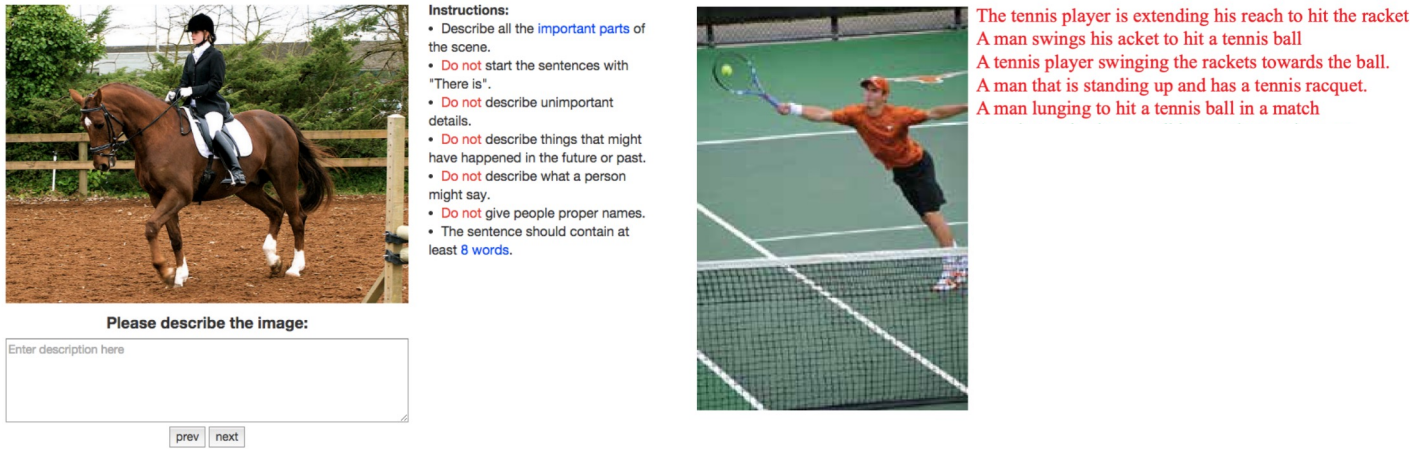


Fig. 2: Example user interface for the caption gathering task.

More than one captions can describe to describe one image

The tennis player is extending his reach to hit the racket
A man swings his acket to hit a tennis ball
A tennis player swinging the rackets towards the ball.
A man that is standing up and has a tennis racquet.
A man lunging to hit a tennis ball in a match

More than one captions can describe to describe one image

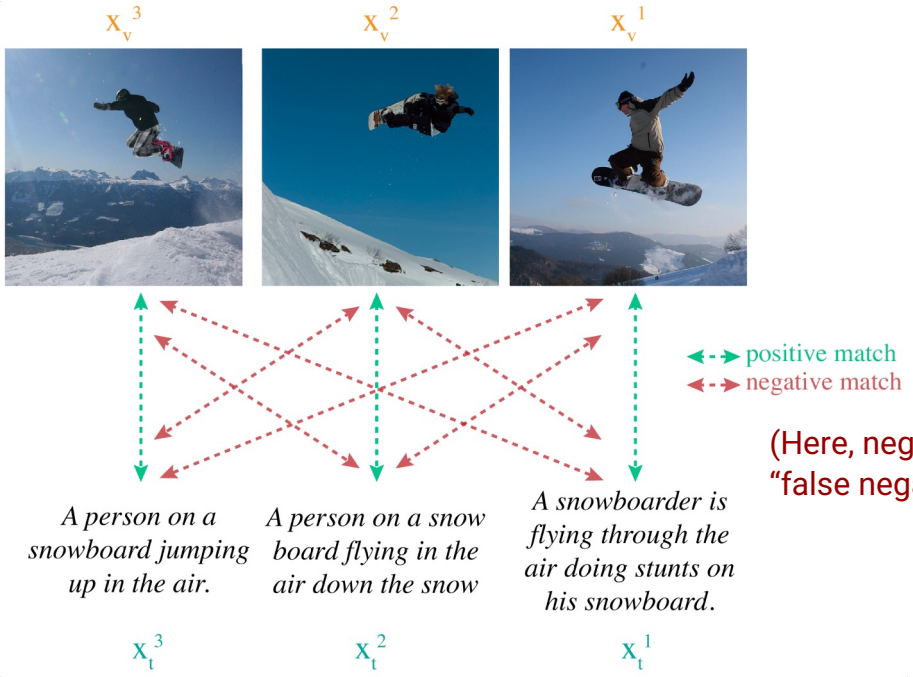
The tennis player is extending his reach to hit the racket.
A man swings his acket to hit a tennis ball
A tennis player swinging the rackets towards the ball.
A man that is standing up and has a tennis racquet.
A man lunging to hit a tennis ball in a match
A male tennis player walking on the tennis court.
A man on a court swinging a racket at a ball.
The man is playing tennis with a racket.
A man standing on a tennis court holding a racquet.
A man on a tennis court trying to hit the ball
A man taking a swing at a tennis ball
A man taking a swing at a tennis ball
A man throwing a tennis ball in the air for him to hit it with his racket.
A man hitting a tennis ball with a racquet.
A man with a tennis racket is running on a court
A man swinging his racket to hit the ball.
The tennis player is hitting the ball with his racket
A tennis player caught jumping up to hit the ball
A man is holding a tennis racquet prepared to hit the incoming ball.
A man holding a tennis racquet as a ball clears the net.
A man with a racket prepares to hit a tennis ball
A man in shorts and a long sleeve shirt playing tennis.

A man stands on a tennis court hitting a ball with a racket.
A man plays a game of tennis during the day.
A man with a tennis racket swings at a tennis ball
A man with a tennis racket on a court.
A man playing tennis on the tennis court
A person hitting a tennis ball with a tennis racket
A man playing tennis and holding back his racket to hit the ball.
A male tennis player swinging his tennis racket.
A man swinging at a tennis ball with a tennis racket.
A person hitting a tennis ball with a tennis racket.
A man on a tennis court that has a racquet.
A man in a head band hits a tennis ball
A man standing on top of a tennis court holding a racquet.
A male in a blue shirt playing tennis on a tennis court
A man holding a tennis racket on a tennis court.
A tennis player swinging a racket at a ball
A man holding a racquet on top of a tennis court.
A boy hitting a tennis ball on the tennis court.
A man on a court swinging a tennis racket.
A man in white shirt and shorts playing tennis.
A guy in a maroon shirt is holding a tennis racket out to hit a tennis ball.
A person hitting a tennis ball with a tennis racket on a tennis court.
The man is playing tennis on the court.

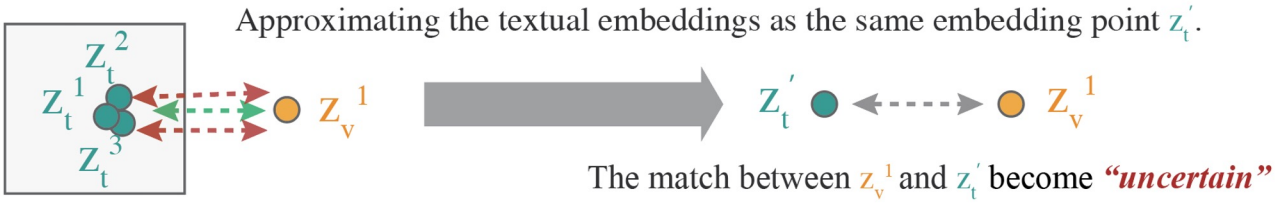
There are a lot of false negatives (FNs) in COCO Caption

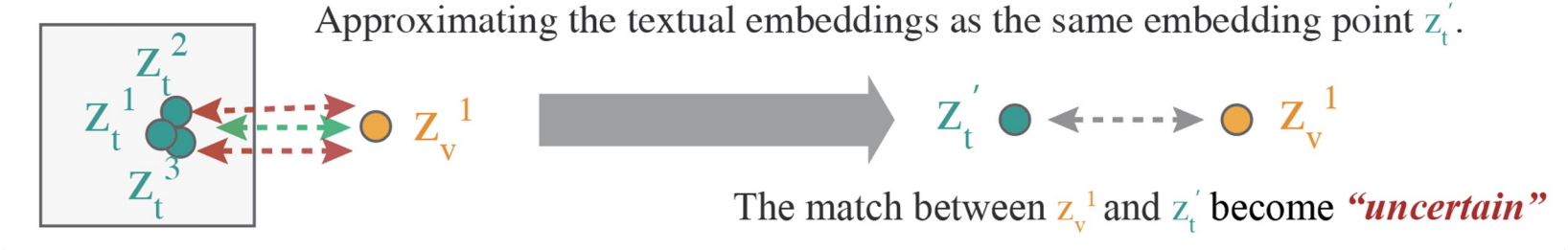
Dataset	# positive images	# positive captions
Original MS-COCO Caption	1,332	6,305 (=1,261×5)
CxC [32]	1,895 (×1.42)	8,906 (×1.41)
Human-verified positives	10,814 (×8.12)	16,990 (×2.69)
ECCV Caption	11,279 (×8.47)	22,550 (×3.58)

ECCV'22, Chun et al., ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO



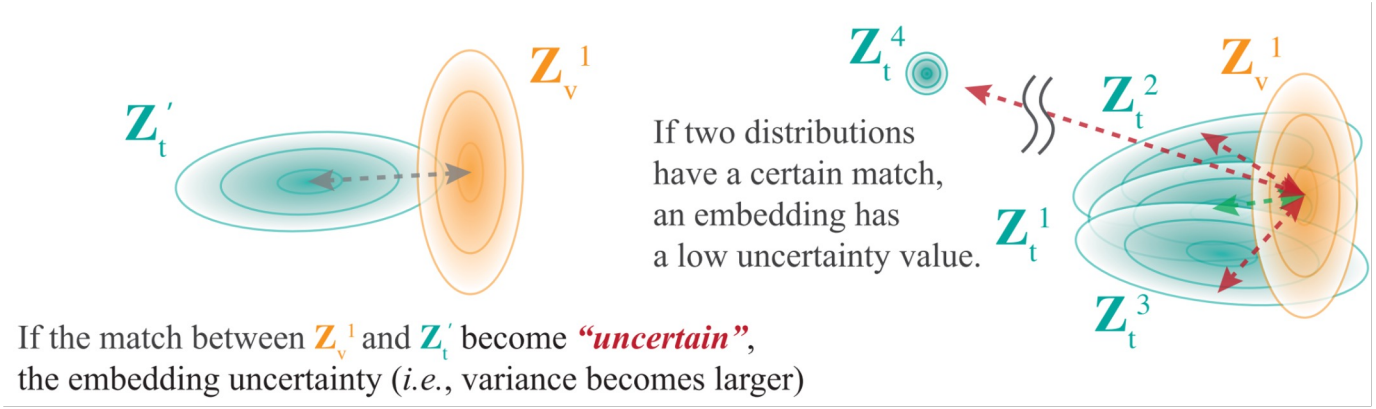
(Here, negative matches denote "false negatives")





Deterministic embedding cannot handle the ambiguity of ITM datasets:

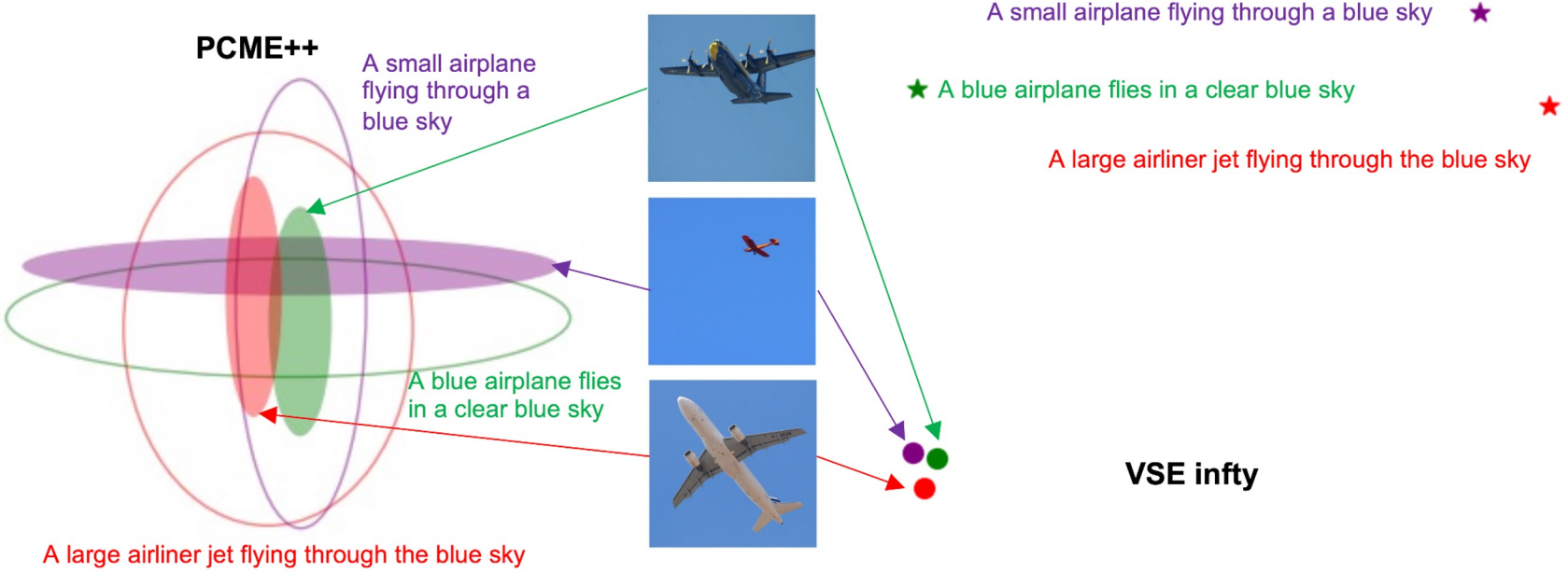
- Visual/textual embeddings will be located closely as they are semantically almost same. Image-text match (ITM) supervision becomes uncertain (either matched or not), but a deterministic embedding space cannot capture uncertainty.
- The existing deterministic approaches enforces to the textual embeddings to be different by making z_t^1 differs from others using hardest negative mining (HNM), although they are semantically the same.



Probabilistic embedding can handle the ambiguity of ITM datasets:

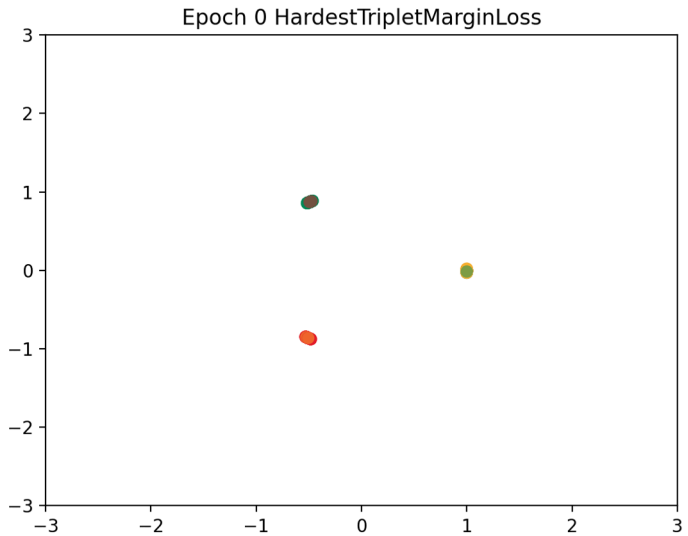
- A probabilistic embedding space captures uncertainty of noisy ITM supervision. If a match is uncertain, then the variance of embeddings becomes larger (Z_t^1, Z_t^2, Z_t^3), if a match is certain, then the variance becomes smaller (Z_t^4)

t-SNE visualization of learned embeddings (prob vs. det)

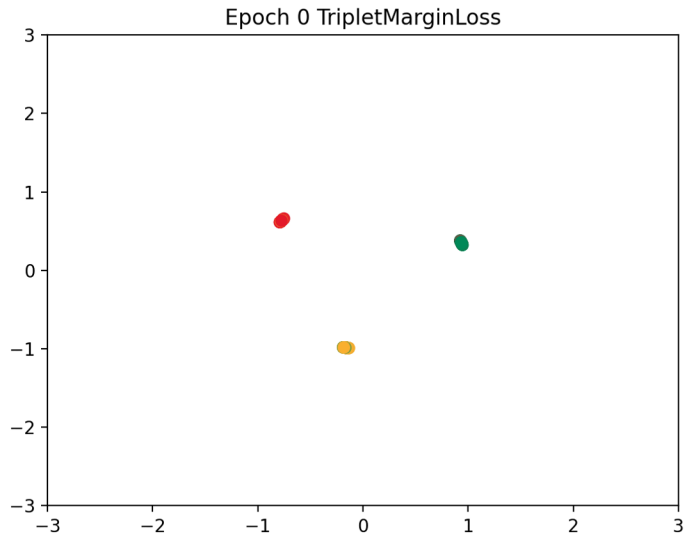


What will be happened if we train embeddings with deterministic embeddings and negative mining?

With Hardest Negative Mining (HNM)



Without HNM



2-D toy dataset (animation also can be found in https://naver-ai.github.io/pcmepp/)

Target properties for the embedding space

- There exists **a proper probabilistic distance** between image embedding Z_v and text embedding Z_t
  - This work introduces a closed-form sampled distance (CSD)

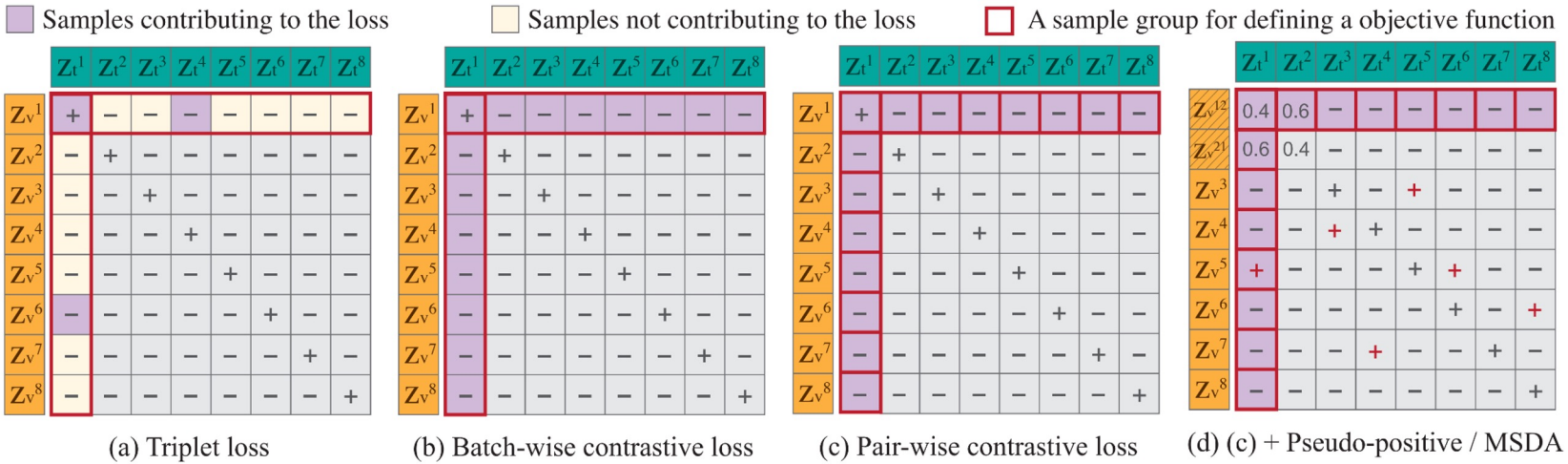
$$d(\mathbf{Z}_v, \mathbf{Z}_t) = \mathbb{E}_{\mathbf{Z}_v, \mathbf{Z}_t} \|\mathbf{Z}_v - \mathbf{Z}_t\|_2^2 = \|\mu_v - \mu_t\|_2^2 + \|\sigma_v^2 + \sigma_t^2\|_1$$

  - CSD satisfies most of the properties of a metric function, except d(x, x) = 0.
- If the match of Z_v and Z_t is **certain**, then Z_v and Z_t has **small variance**
- If the match of Z_v and Z_t is **uncertain**, then Z_v and Z_t has **large variance**

Training objectives

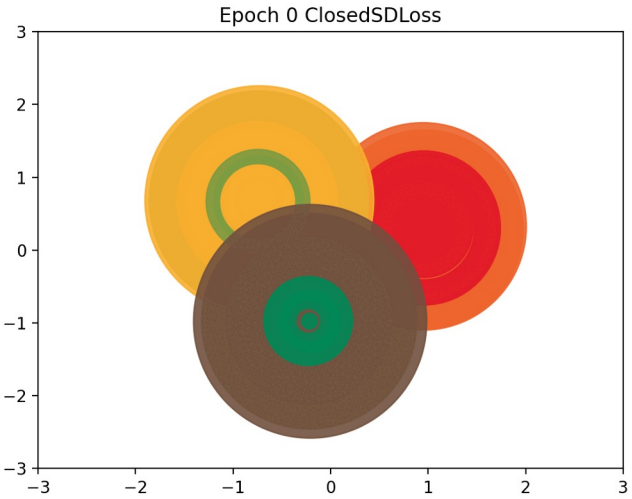
- Optimizing the following pair-wise matching objective function:

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



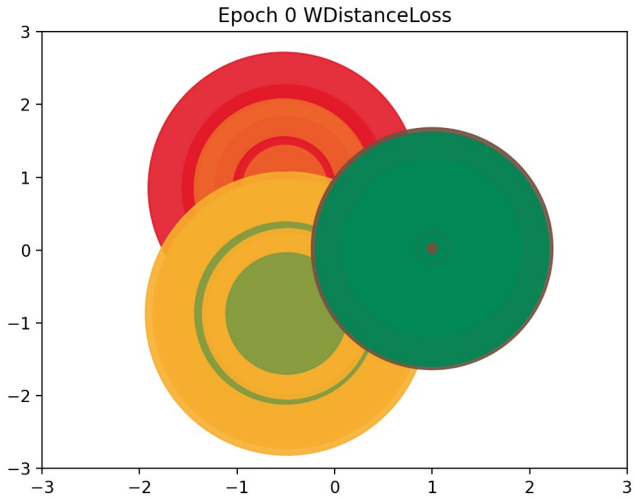
CSD satisfies the desired properties

CSD (proposed)



Small variances for certain classes,
Large variances for uncertain classes

Wasserstein distance



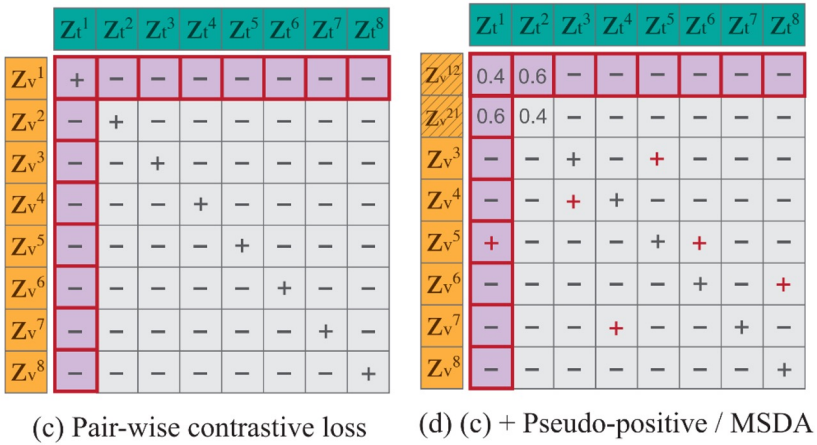
Both certain / uncertain classes have
the similar variances

2-D toy dataset (animation also can be found in https://naver-ai.github.io/pcmepp/)

Additional techniques to prevent a loss saturation

- The proposed matching objective function automatically **less penalizes** (gives smaller weights) on **positive match with high match probability**, as well as **negative match with high dis-match probability** as follows:
  - Let $-a \cdot d(\mathbf{Z}_v, \mathbf{Z}_t) + b = l_{vt}$ (l_{vt} becomes **larger** if two instances are **closed**, and **smaller** if they are **far**), then we have
    - $\frac{\partial \mathcal{L}_{\text{match}}}{\partial l} = 1 - \text{sigmoid}(l_{vt})$ when m=0 (not matched)
    - $\frac{\partial \mathcal{L}_{\text{match}}}{\partial l} = 1 - \text{sigmoid}(-l_{vt})$ when m=1 (matched)
  - Assume we have a good enough embedding extractor, and there is a false negative (FN), i.e., they are actually matched, but the annotation is “not matched”. In this case, the FN pair will have almost 0 gradient, therefore, cannot contribute to the optimization.

Pseudo-positive and mixed sample data augmentation

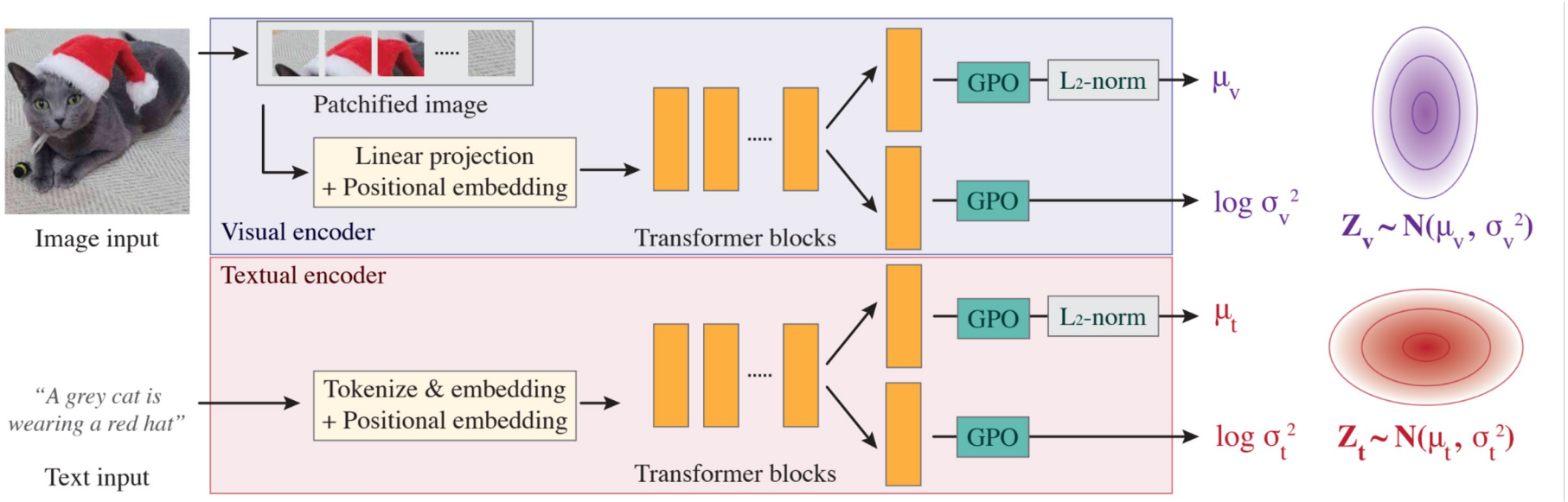


(c) Pair-wise contrastive loss (d) (c) + Pseudo-positive / MSDA

Final objective function

$$\mathcal{L}_{\text{match}} + \alpha\mathcal{L}_{\text{pseudo-match}} + \beta\mathcal{L}_{\text{VIB}}$$

Architecture



Based on CLIP encoders with extra branches for the uncertainty heads.

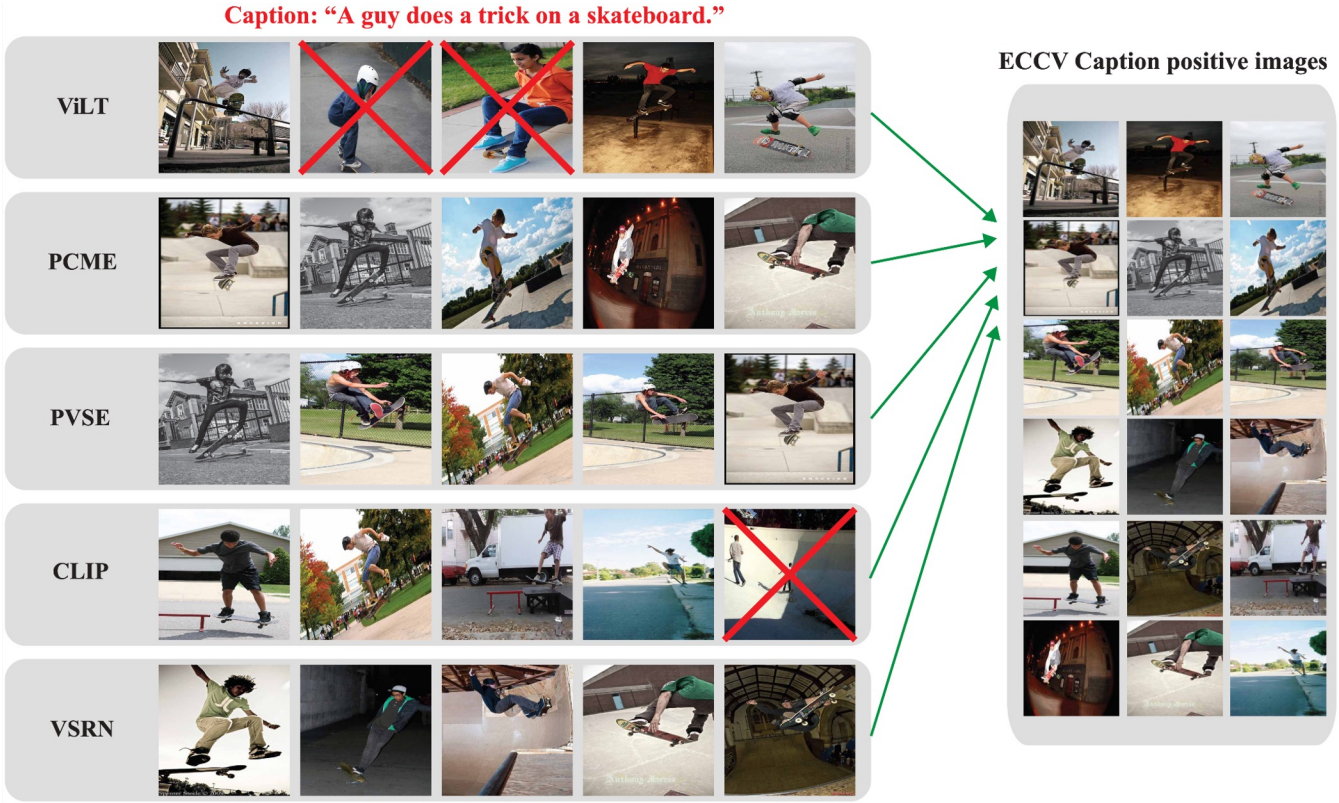
Experiments

- Dataset
  - Training on MS-COCO Caption train split (113,287 images and 566,435 captions)
    - Each image has five positive captions
  - Testing on
    - COCO Caption 5K test split (5K images / 25K captions)
    - COCO Caption 1K test split (1K images / 5K captions)
    - CxC Caption (5K images / 25K captions) => dense annotated version of 5K
    - ECCV Caption => more dense version, but limited queries
- Initialization and model selection
  - Initialized from the official CLIP weights (B/32, B/16, L/14)
  - The best model on “validation split” is used for the evaluation

ECCV Caption dataset

- Human-verified COCO Caption evaluation benchmark
- Annotators manually verified that whether the given image-text pair is matched or not
- As there are too many possible pairs in the dataset (76B image-text pairs), we used state-of-the-art machine annotators to reduce the number of candidates to be verified
- After post-processing, ECCV Caption contains 1,261 image queries (originally 5,000) but with 17.9 positive captions per image query on average (originally 5). It also contains 1,332 caption queries (originally 25,000) with 8.5 positive images per caption (originally 1).
- As we have plenty positives, now we can use precision-based metrics for evaluation.

Overview of ECCV Caption dataset construction



ECCV Caption: Extended COCO Caption Validation dataset

Figure 2: **ECCV Caption examples**. The given caption query: "A herd of zebras standing together in the field". Red: original positive. Green: annotated as "100% Yes". Blue: annotated as "Weak Yes".

ECCV Caption: Extended COCO Caption Validation dataset

The tennis player is extending his reach to hit the racket.
A man swings his acket to hit a tennis ball
A tennis player swinging the rackets towards the ball.
A man that is standing up and has a tennis racquet.
A man lunging to hit a tennis ball in a match
A male tennis player walking on the tennis court.
A man on a court swinging a racket at a ball.
The man is playing tennis with a racket.
A man standing on a tennis court holding a racquet.
A man on a tennis court trying to hit the ball
A man taking a swing at a tennis ball
A man taking a swing at a tennis ball
A man throwing a tennis ball in the air for him to hit it with his racket.
A man hitting a tennis ball with a racquet.
A man with a tennis racket is running on a court
A man swinging his racket to hit the ball.
The tennis player is hitting the ball with his racket
A tennis player caught jumping up to hit the ball
A man is holding a tennis racquet prepared to hit the incoming ball.
A man holding a tennis racquet as a ball clears the net.
A man with a racket prepares to hit a tennis ball
A man in shorts and a long sleeve shirt playing tennis.

A man stands on a tennis court hitting a ball with a racket.
A man plays a game of tennis during the day.
A man with a tennis racket swings at a tennis ball
A man with a tennis racket on a court.
A man playing tennis on the tennis court
A person hitting a tennis ball with a tennis racket
A man playing tennis and holding back his racket to hit the ball.
A male tennis player swinging his tennis racket.
A man swinging at a tennis ball with a tennis racket.
A person hitting a tennis ball with a tennis racket.
A man on a tennis court that has a racquet.
A man in a head band hits a tennis ball
A man standing on top of a tennis court holding a racquet.
A male in a blue shirt playing tennis on a tennis court
A man holding a tennis racket on a tennis court.
A tennis player swinging a racket at a ball
A man holding a racquet on top of a tennis court.
A boy hitting a tennis ball on the tennis court.
A man on a court swinging a tennis racket.
A man in white shirt and shorts playing tennis.
A guy in a maroon shirt is holding a tennis racket out to hit a tennis ball.
A person hitting a tennis ball with a tennis racket on a tennis court.
The man is playing tennis on the court.

Kites being flown along the coast line in the morning
People watch multi colored items fly above the beach.
A beach where people are flying kites at sunset.
A crowd of people flying kites on the beach.
Dozens of kite skiers out in the ocean
People in the water and parachutes overhead.
Many different sails flying over a large body of water.
There are many large kites flying above the beach.
Kites are flying over people on a beach.
A bunch of kites flying in the sky on the beach.
Several gliders floating on the ocean next to an island.
A couple flying a kite at dusk on the seaside.
People on a sunny beach flying various kites

Some zebras are seen grazing in the field.
Four adult zebra are grazing on a field of grass.
Four zebras are grazing on grass in a pasture.
Four zebras eating grass on a field.
A herd of zebra standing on top of a lush green field.
These four zebra are walking in a field.
There is a herd of zebras standing around.
A herd of zebras walking through the grass.
Clouds with a rainbow in the sky of an open field with zebras grazing on the grass.
Several zebras in an open area during a not so sunny day.
A group of zebras that are standing in the grass.
A group of zebras in a grassy and forested area
Four zebras are grazing at a nature reserve.
Zebras graze in a grass and bush fenced enclosure.
There are some zebras standing in a grassy field
The zebra is standing in the field with the other animals in the background.
Four zebras standing in the grass on a cloudy day.
Four zebras walking in a grassy area.

A herd of zebras stand on a pathway near brown grass.
Zebras graze on the plains with trees in the background.
Three zebras and two other animals grazing.
Several zebras walking the terrain of hills and mountains
Two zebras in some brown and green grass and some bushes
A herd of zebras grazing with a rainbow behind.
A herd of zebra grazing on a dry grass field.
Herd of five zebras grazing in a field
Two black and white zebras and some green grass and trees
A group of zebra standing on top of a dry grass field.
A zebra is standing outside on the grass by itself.
Three zebra in the middle of a field with a body of water in the distance.
Zebras grazing on sparse grass in an enclosure at an animal park.
A herd of zebras grazing in a field and a rainbow.
A group of zebras are on some grass with trees and bushes behind them.
A few deer and a zebra on a grass field
Several zebras eat the green grass in the pasture.

Comparisons of test datasets

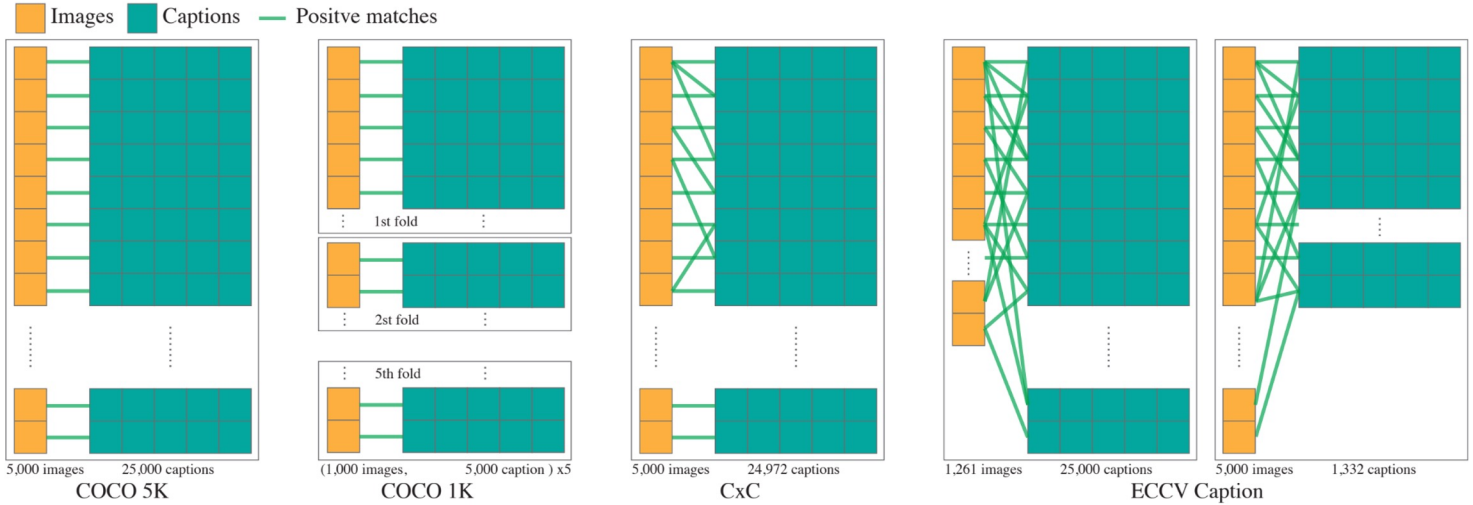


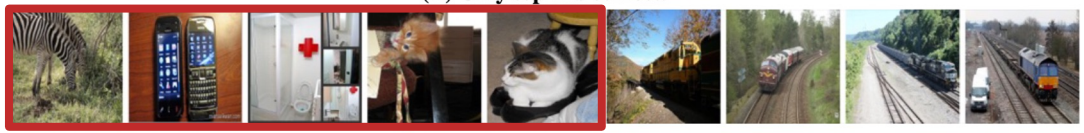
Figure B.1: **Difference between COCO 5K, 1K, CxC [37] and ECCV Caption [25].** All matches not illustrated in the image are negative. ECCV Caption has separated query sets for each modality, while other datasets use the same images and captions for both query and gallery.

Comparisons of precision metric vs. recall metric

GT images for “A train on a train track near many trees”.

(A) Only top-1 is wrong.

(A) Only top-1 is correct.



(C) Top-1 to -5 are wrong.

(D) Only top-5 is correct.

	R@1	R@5	mAP@R	Preference score (by human study)
	100.0	100.0	11.1	10.66
	0.0	100.0	68.6	70.85
	0.0	100.0	2.2	4.89
	0.0	0.0	14.1	13.15

Main results on the COCO Caption benchmark

Backbone	Method	Prob?	ECCV Caption [25] mAP@R	ECCV Caption [25] R-P	ECCV Caption [25] R@1	CxC [37] R@1	COCO [20] 1K R@1	COCO [20] 5K R@1	RSUM
ViT-B/32 (151M)	CLIP ZS† [19]	✘	26.8	36.9	67.1	42.0	59.5	40.3	471.9
	VSE∞ [15]	✘	40.0	49.5	83.1	57.1	75.5	55.2	536.5
	InfoNCE [19]	✘	39.0	48.7	81.7	54.9	74.0	53.0	532.6
	PCME [14]	✔	39.1	48.9	81.4	54.7	73.8	53.0	532.0
	Ours	✔	40.0	49.6	83.3	57.0	75.5	55.3	537.1
	Ours + SWA	✔	**40.2**	**49.8**	**83.6**	**57.2**	**75.6**	**55.5**	**537.3**
ViT-B/16 (150M)	CLIP ZS†	✘	29.3	39.0	71.1	44.3	62.0	42.7	481.0
	VSE∞	✘	41.7	50.6	86.3	62.3	79.1	60.7	547.2
	InfoNCE	✘	41.1	50.4	84.8	60.9	78.3	59.3	545.5
	PCME	✔	41.0	50.3	84.3	59.9	77.8	58.2	544.2
	Ours	✔	41.9	51.0	86.3	62.8	79.6	61.3	548.7
	Ours w/ SWA	✔	**42.0**	**51.1**	**86.6**	**63.1**	**79.7**	**61.6**	**548.9**
ViT-L/14 (428M)	CLIP ZS†	✘	28.0	37.8	72.2	48.1	64.8	46.4	491.6
	VSE∞	✘	20.2	31.5	46.2	24.3	44.5	22.7	424.3
	InfoNCE	✘	35.6	45.8	75.6	48.0	69.5	45.9	520.6
	PCME	✔	41.2	50.3	86.0	63.4	80.3	61.9	550.4
	Ours	✔	**42.1**	**50.8**	**88.8**	**65.9**	**81.8**	**64.3**	**554.7**

Ablation studies

Table 3: **Effect of optimization methods.** Ablation study on VIB [38], Pseudo-Positives (PP), Mixed Sample Data Augmentation (MSDA), and SWA [50] with a ViT-B/32 backbone are shown.

				ECCV Caption			CxC	COCO		
VIB	PP	MSDA	SWA	mAP@R	R-P	R@1	R@1	1K R@1	5K R@1	RSUM
✘	✘	✘	✘	38.9	48.6	82.2	56.7	75.2	54.9	535.9
✔	✘	✘	✘	39.2	49.0	82.2	56.1	74.9	54.3	535.1
✘	✔	✘	✘	39.0	48.6	82.7	56.8	75.2	55.0	536.0
✔	✔	✘	✘	39.6	49.2	82.6	56.3	74.8	54.5	534.8
✔	✔	✔	✘	40.0	49.6	83.3	57.0	75.5	55.3	537.1
✔	✔	✔	✔	**40.2**	**49.8**	**83.6**	**57.2**	**75.6**	**55.5**	**537.3**

Table 4: **Effect of probability distance on training objective.** Results on ViT-B/32 backbone with VIB loss.

	ECCV Caption			CxC	COCO		
Probability distance	mAP@R	R-P	R@1	R@1	1K R@1	5K R@1	RSUM
Wasserstein 2-distance	26.7	35.5	69.0	46.3	64.5	44.9	484.6
Match probability (PCME [14])	39.1	48.9	81.4	54.7	73.8	53.0	532.0
Proposed (Equation (1))	**39.2**	**49.0**	**82.2**	**56.1**	**74.9**	**54.3**	**535.1**

Table 5: **Impact of architecture design choice.** Details are the same as the previous tables.

		ECCV Caption			CxC	COCO		
# layers for $\log \sigma^2$	GPO	mAP@R	R-P	R@1	R@1	1K R@1	5K R@1	RSUM
1	✘	37.4	47.4	79.2	51.0	70.4	49.2	521.8
2	✔	**40.2**	**49.7**	83.2	56.6	75.3	54.8	536.5
1	✔	40.0	49.6	**83.3**	**57.0**	**75.5**	**55.3**	**537.1**

VIB β	0	$\times$0.1	$\times$0.2	$\times$0.5	$\times$1	$\times$2	$\times$5	$\times$10
$\|\sigma\|_1$	$2E^{-4}$	1.1	2.2	4.2	7.1	11.6	23.1	37.6
$-\rho$	0.76	0.87	0.91	0.92	**0.94**	**0.95**	0.91	0.90
RSUM	534.6	537.5	**538.1**	**538.2**	537.0	537.1	535.5	534.5

How does uncertainty help image-text representations?

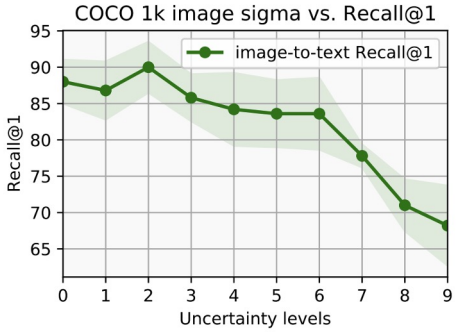


Figure 5: $\|\sigma^2\|_1$ **vs. R@1.**

Query image

$\| \sigma \|_1 = 1.13\text{e-}4$

Top-5 Retrieved captions

A snowboarder flies through the air in a mountain landscape.
A person launching into the air on a snowboard.
A person with a snowboard is jumping in the air.
a man jumping in the air above a ski slope on a ski board.
A man jumping in the air on a snowboard.

GT captions

A snowboarder is flying through the air doing stunts on his snow board.
A man with gloves, goggles and a hat on is in the air on his snowboard.
A snowboarder is airborne during a trick atop a mountain.
A person that is snowboarding through the air as they grab their board.
A snowboarder doing a trick after a jump.

$\| \sigma \|_1 = 0.000111$

Retrieved captions

Several oranges are laying under a few bananas.
A bunch of bananas on top of oranges.
Closeup of various oranges and bananas in pile.
A pile of oranges sitting under a pile of bananas.
A plate full of sliced oranges next to a bunch of bananas.

GT captions

Closeup of various oranges and bananas in pile.
Several oranges are laying under a few bananas.
A bunch of bananas on top of oranges.
There are a lot of bananas and oranges.
A pile of oranges sitting under a pile of bananas

Query text: *A beach area with people on the sand and various kites flying overhead in the sky.* $\| \sigma \|_1 = 9.43\text{e-}5$

Top-5 Retrieved images **GT Image**

Query text: *Assorted fruit on display at a fruit market.* $\| \sigma \|_1 = 9.73\text{e-}5$

Top-5 Retrieved images **GT Image**

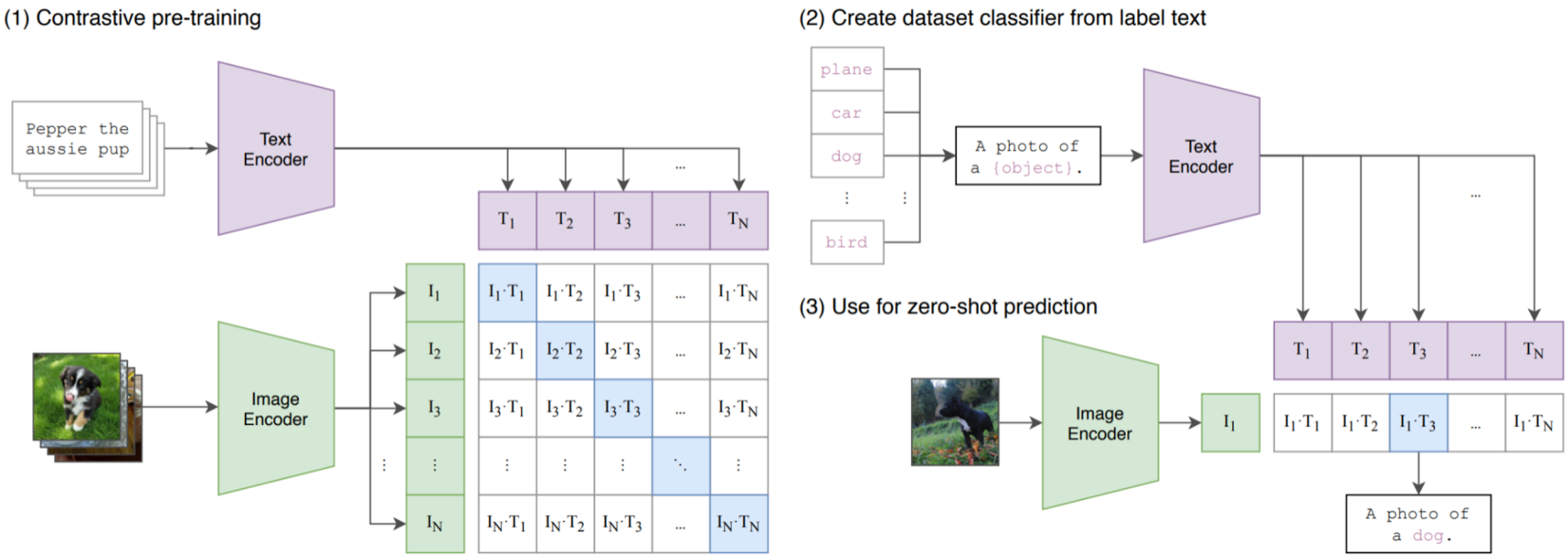
Query text: *A large banana tree that has green colored bunches of bananas on it along with various branches* $\| \sigma \|_1 = 9.74\text{e-}5$

Top-5 Retrieved images **GT Image**

Query text: *Bunches of carrots, beets, radishes, and multiple lettuces displayed on a table* $\| \sigma \|_1 = 9.24\text{e-}5$

Top-5 Retrieved images **GT Image**

Uncertainty-based prompt-tuning



Recap: CLIP-based zero-shot (ZS) classification is actually an image-text matching between an input image and prompt texts

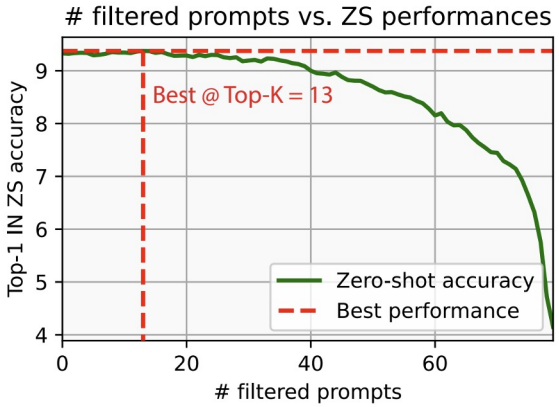
CLIP prompts for ImageNet ZS

80 base prompts. a photo of a {}., a bad photo of a {}., a photo of many {}., a sculpture of a {}., a photo of the hard to see {}., a low resolution photo of the {}., a rendering of a {}., graffiti of a {}., a bad photo of the {}., a cropped photo of the {}., a tattoo of a {}., the embroidered {}., a photo of a hard to see {}., a bright photo of a {}., a photo of a clean {}., a photo of a dirty {}., a dark photo of the {}., a drawing of a {}., a photo of my {}., the plastic {}., a photo of the cool {}., a close-up photo of a {}., a black and white photo of the {}., a painting of the {}., a painting of a {}., a pixelated photo of the {}., a sculpture of the {}., a bright photo of the {}., a cropped photo of a {}., a plastic {}., a photo of the dirty {}., a jpeg corrupted photo of a {}., a blurry photo of the {}., a photo of the {}., a good photo of the {}., a rendering of the {}., a {} in a video game., a photo of one {}., a doodle of a {}., a close-up photo of the {}., the origami {}., the {} in a video game., a sketch of a {}., a doodle of the {}., a origami {}., a low resolution photo of a {}., the toy {}., a rendition of the {}., a photo of the clean {}., a photo of a large {}., a rendition of a {}., a photo of a nice {}., a photo of a weird {}., a blurry photo of a {}., a cartoon {}., art of a {}., a sketch of the {}., a embroidered {}., a pixelated photo of a {}., itap of the {}., a jpeg corrupted photo of the {}., a good photo of a {}., a plushie {}., a photo of the nice {}., a photo of the small {}., a photo of the weird {}., the cartoon {}., art of the {}., a drawing of the {}., a photo of the large {}., a black and white photo of a {}., the plushie {}., a dark photo of a {}., itap of a {}., graffiti of the {}., a toy {}., itap of my {}., a photo of a cool {}., a photo of a small {}., a tattoo of the {}.

CLIP uses the average of the embeddings extracted by 80 prompts as the class text embedding. But, does every class need all of the 80 prompts?

https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

Uncertainty-based CLIP ZS prompt tuning?

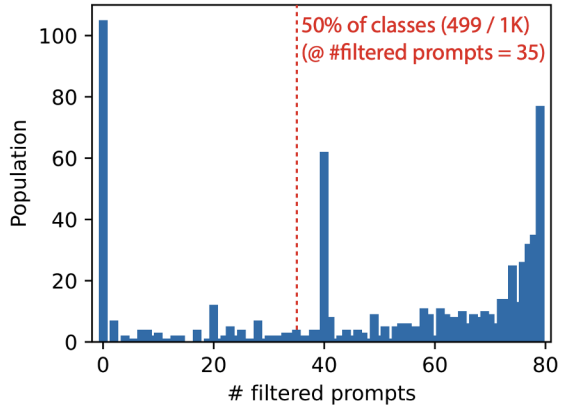


Model	Prompts	Top-1 Acc
InfoNCE	*"A photo of* { · }"	13.05
	All 80 prompts	13.41
PCME++	*"A photo of* { · }"	8.58
	All 80 prompts	9.33
	Top-K certain prompts	9.37
	Best top-K for each class	**14.75**

Idea: Filter out Top-K uncertain prompts by uncertainty

If we use the same top-K for every class, the gain is very marginal (9.33 -> 9.37)

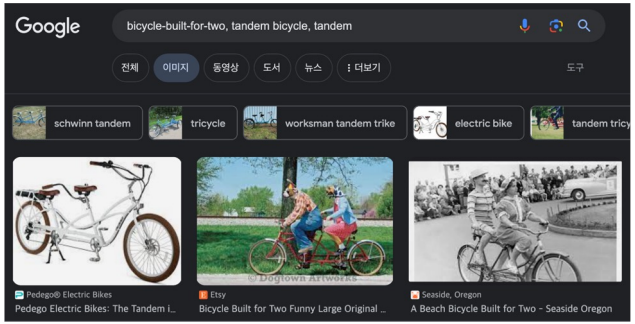
Uncertainty-based CLIP ZS prompt tuning?



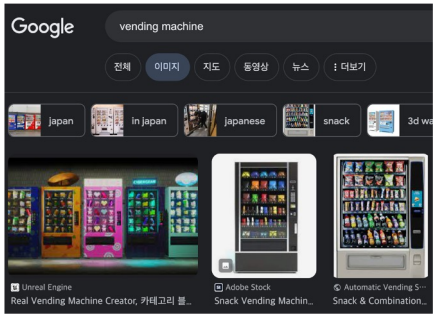
Model	Prompts	Top-1 Acc
InfoNCE	*"A photo of { · }"*	13.05
	All 80 prompts	13.41
PCME++	*"A photo of { · }"*	8.58
	All 80 prompts	9.33
	Top-K certain prompts	9.37
	Best top-K for each class	**14.75**

Idea: Filter out Top-K uncertain prompts by uncertainty for *each class*

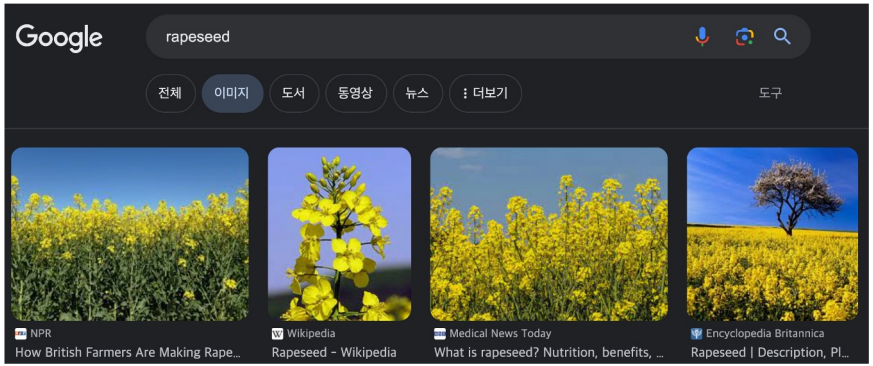
If we search the best top-K filtering for each class, then the performance gain is significant (9.33 -> 14.75)



a photo of many {}
a photo of a {}



a cropped photo of a {}. a low resolution photo of a {}. a cropped photo of the {}. a low resolution photo of the {}. a close-up photo of a {}. a tattoo of a {}.



itap of my {}. a plushie {}. a photo of a dirty {}. the plushie {}. a photo of a clean {}. a photo of a weird {}. a good photo of a {}. a photo of a large {}. a photo of a cool {}. a photo of many {}. a blurry photo of a {}. a photo of my {}. a photo of a nice {}. a bright photo of a {}. itap of a {}. a dark photo of a {}. a plastic {}. a close-up photo of the {}. a painting of a {}. a photo of a small {}. a close-up photo of a {}. a photo of the weird {}. a black and white photo of a {}. a jpeg corrupted photo of a {}. a sculpture of a {}. a good photo of the {}. a photo of a hard to see {}. a photo of a {}. a photo of the small {}. a photo of one {}. a photo of the dirty {}. a photo of the hard to see {}. a blurry photo of the {}. a bad photo of a {}. a toy {}. a photo of the large {}. a photo of the clean {}. a origami {}. a photo of the cool {}. a pixelated photo of a {}.

References

- **[CVPR 2021] Probabilistic Embeddings for Cross-Modal Retrieval.**
  Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, Diane Larlus
- **[ECCV 2022] ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO.**
  Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, Seong Joon Oh
- **[Preprint] Improved Probabilistic Image-Text Representations.**
  Sanghyuk Chun