Whitened CLIP as a Likelihood Surrogate of Images and Captions

Roy Betser, Meir Yossef Levi,
Prof. Guy Gilboa

ICML 25

TECHNION
Israel Institute of Technology

Motivation

- CLIP[1] is multi-modal – combining text and image.

- Likelihood scores are useful for numerus applications.
- Language models explicitly approximate negative log-likelihood.
- Classic Image analysis methods **explicitly** approximate P(x).
- DL models **implicitly** approximate P(x).

- We propose a direct, explicit, likelihood approximation, using CLIP.

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

Notations

- A set of N random vectors - $X = \{x_1, x_2, \ldots, x_N\}$.
- Each vector x_i is in dimension - d; $x_i \in R^d$.
- Empirical mean vector - $\mu = \frac{1}{N}\sum_{i=0}^{N} x_i$, $\mu \in R^d$
- Centered set of vectors = $\hat{X} = \{x_1 - \mu, x_2 - \mu, \ldots, x_N - \mu\}$
- Empirical Covariance matrix - $\Sigma = \frac{1}{N}\hat{X}\hat{X}^T$, $\Sigma \in R^{d\times d}$

Whitening Transform

Set X of random vectors with a nonsingular covariance matrix Σ.

W is a $d \times d$ matrix that satisfies:

$$W^T W = \Sigma^{-1}$$

W is not unique.

Whitening transform:

$$y = W\hat{x} \quad , \quad Y = W\hat{X}$$

Whitened CLIP - Motivation

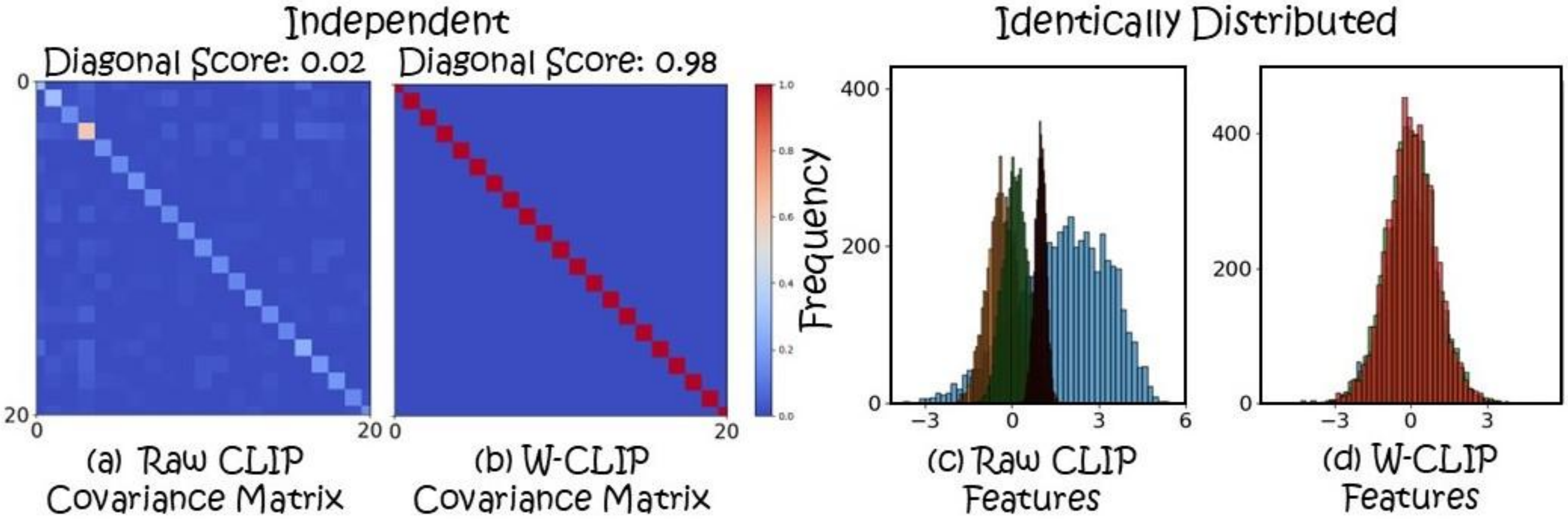
Four main advantages of whitened CLIP (W-CLIP):

1. Purely data-driven.
2. Invertible Transform.
3. Computing W once, a-priori → efficient use.
4. Whitened features have zero mean and unit variance.

IID Evaluation

$$\text{Diagonal Score} = \frac{\sum_i |\mathbf{\Sigma}_{i,i}|}{\sum_{i,j} |\mathbf{\Sigma}_{i,j}|}$$

Experiments use MS-COCO[2] validation set.



[2] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer International Publishing, 2014.

W-CLIP Likelihood

For a vector of IID, standard normal distributed variables:

$$P(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\|x\|^2\right) \qquad \ell(x) = \log P(x) = -\frac{1}{2}\left(d\log(2\pi) + \|x\|^2\right)$$

→ Log-likelihood approximation based only on embeddings norm in W-CLIP.

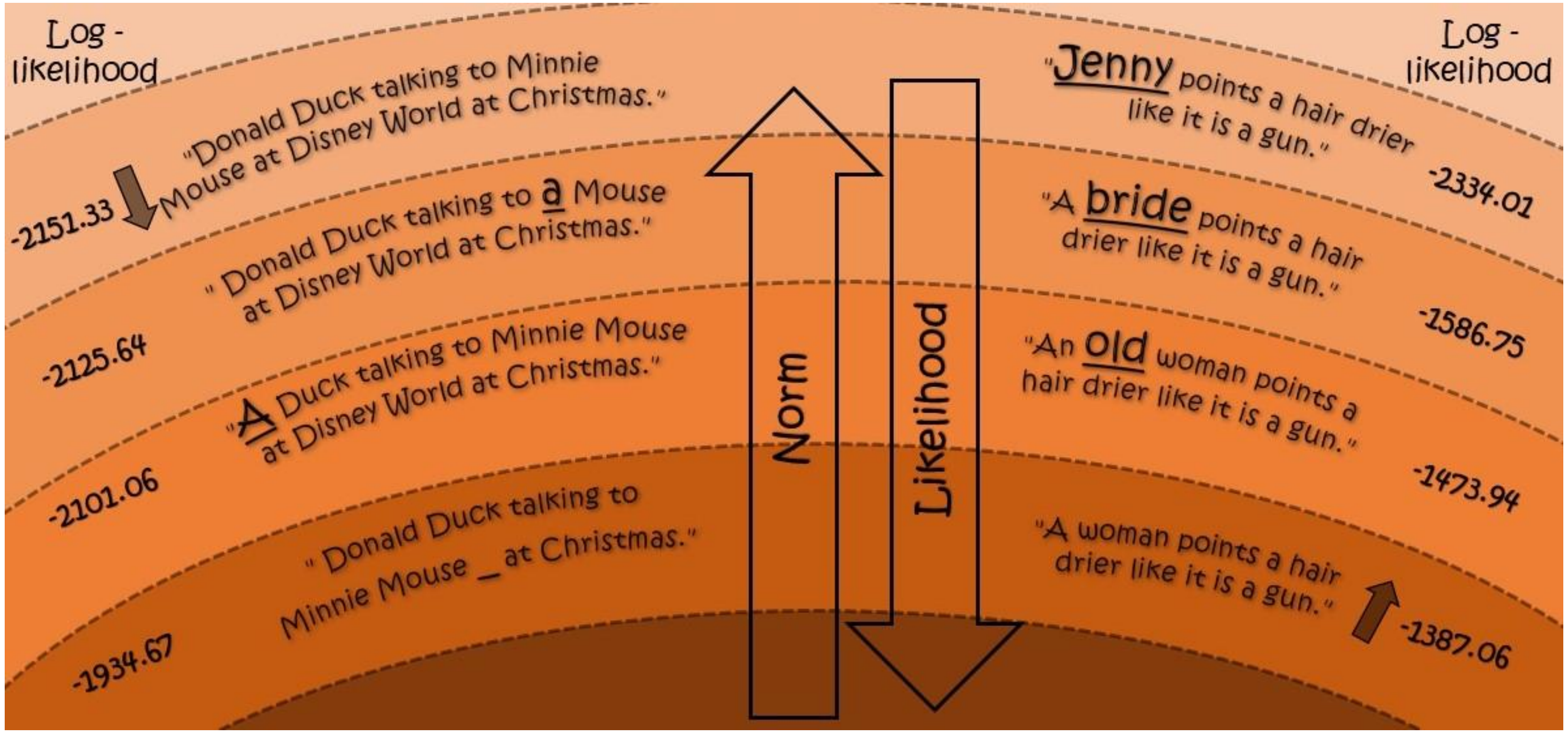
Artifact Detection

Generated images with artifacts have lower likelihoods than similar real images.

Real images from MS-COCO validation set. AI images from [5].

Text Complexity

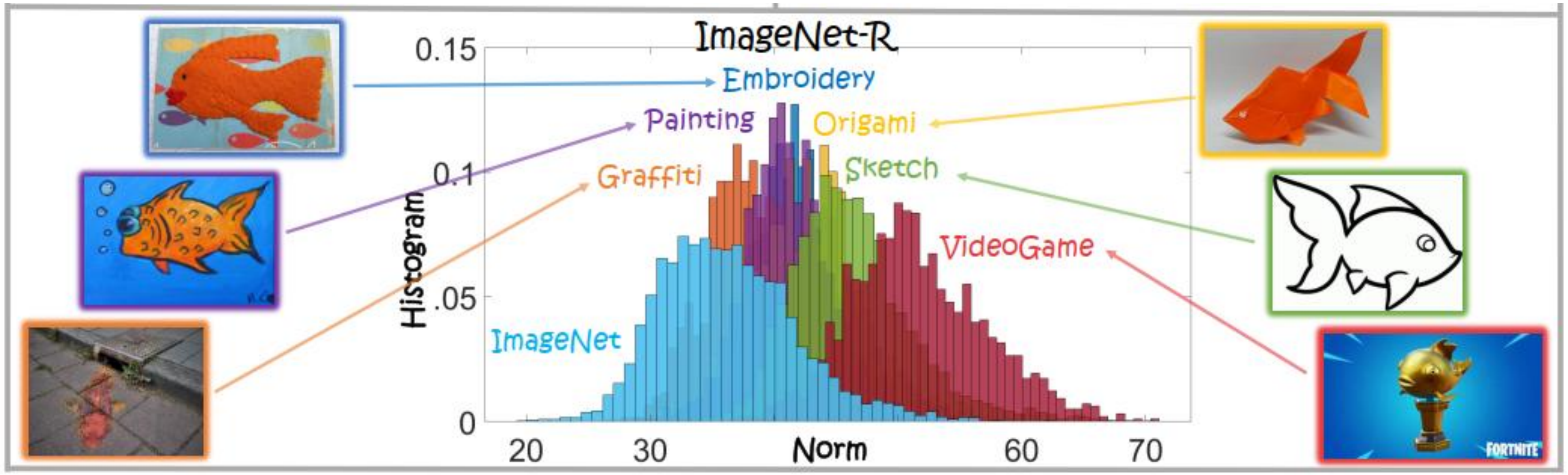
Captions that are more complex result with lower log-likelihood.



Captions from MS-COCO validation set.

Domain Shift Sensitivity

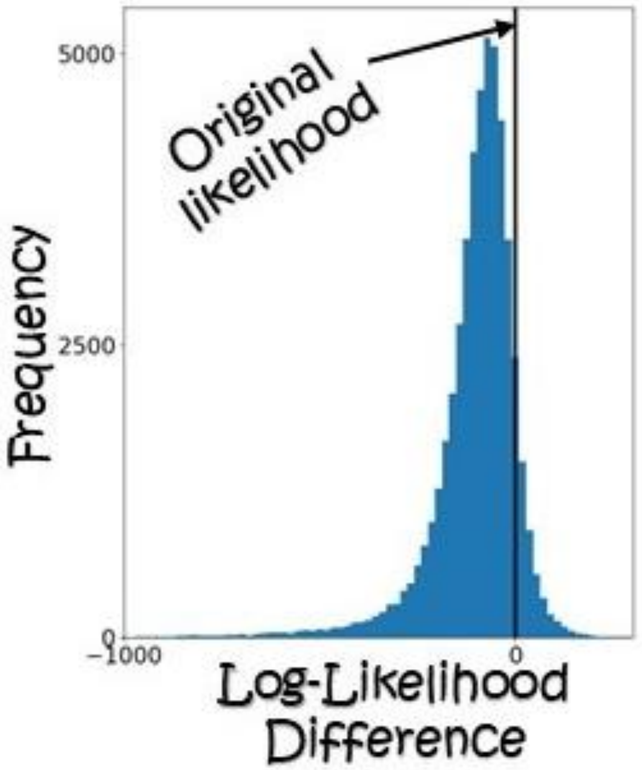
- ImageNet–R[3] domains have higher norms compared to ImageNet.
- Realistic domains (graffiti) have lower norms compared to not realistic domains (sketch, video games).



[3] Hendrycks, Dan, et al. "The many faces of robustness: A critical analysis of out-of-distribution generalization." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

Generation Model Bias

Likelihood scores of generated images, using UnCLIP[4] are lower than the real images used to generate them.



(a) Image Generation Bias

[4] Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." *arXiv preprint arXiv:2204.06125* 1.2 (2022): 3.

Conclusions

- Introduced W-CLIP, an isotropic variation of CLIP latent space.
- First direct likelihood approximation of CLIP model.
- Likelihood approximations are sensitive to:
  - Text complexity.
  - Artifacts in images.
  - Domain shifts.
  - Generation model bias.

Thank You