Deep Learning - Final Project

Names: Lidor Erez and Dvir Rehavi

Lecturer: Dr. Ari Pakman

Department: IEM

CapDec: Text-Only Training for Image Captioning using Noise-Injected CLIP

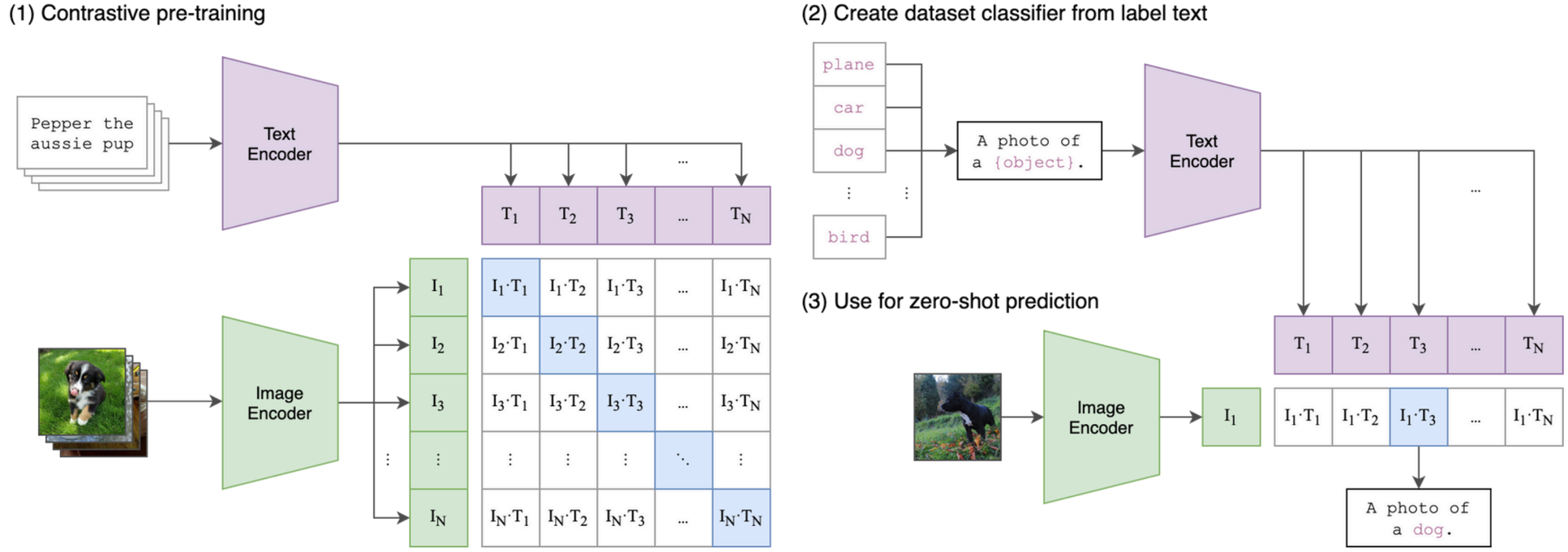
What is CapDec?

- Novel method for image captioning using text only.
- Relies on pre-trained CLIP model.
- Uses noise injection to bridge the gap between image and text embeddings (Liang, et al 2022).

CLIP Model

What is CLIP?

- Multimodal model created by OpenAI
- Trained on (image, text) pairs
- Uses contrastive learning to align visual and textual representations



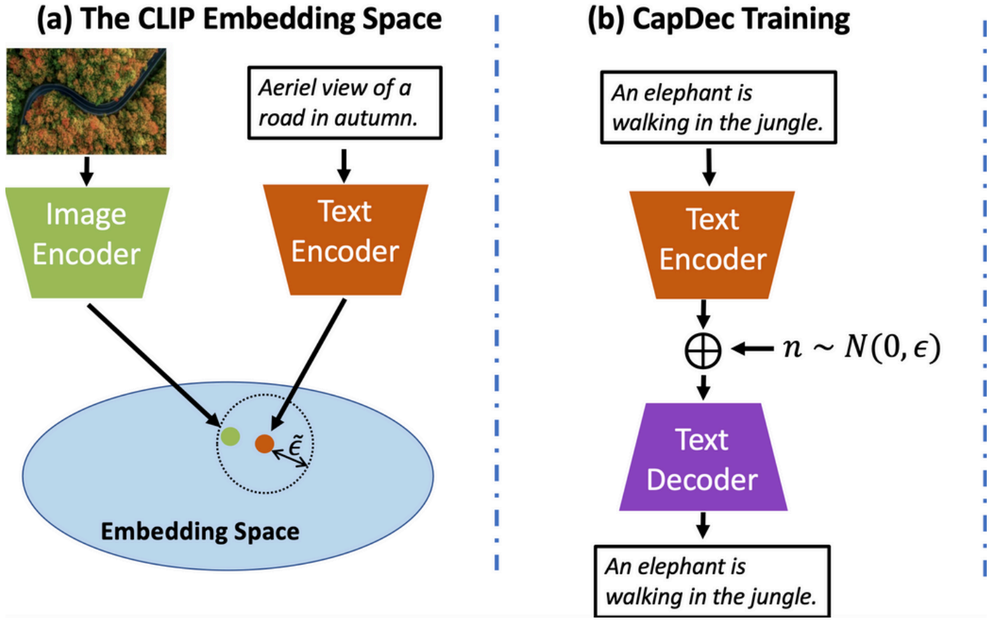
CapDec Training

First Phase:

- Creating text-embeddings from image captions using frozen-CLIP.

Second Phase:

- Training decoder-only to decode the embedded text to the original text
- Gaussian noise injection to the text embedding



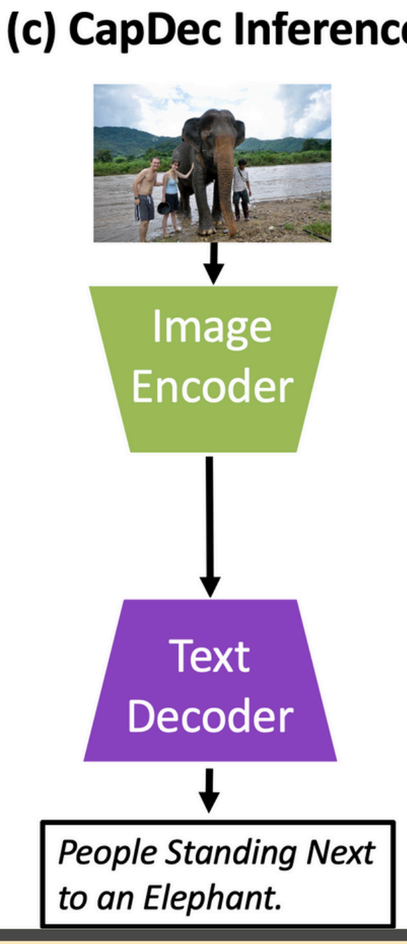
CapDec Inference

First Phase:

- Creating image-embeddings from COCO-dataset using frozen-CLIP.

Second Phase:

- Decoding the image-embeddings to their captions



Project Ideas

Different Type of Noise:

- T-Distribution noise - $noise \sim t_{df}$
- Normalized Gradient noise - $noise = \epsilon * \frac{grad}{||grad||_2}$

Hyperparameter Tuning:

- Number of Attention Heads
- Number of Decoder Transformer Layers

Project Ideas - Cosine Similarity Directed Gaussian Noise

1. Sample N noise vectors - $noise \sim N(0, 1)$
2. Create N candidates - $x^* = x + \beta * noise$
3. Filter candidates - $\frac{x \cdot x^*}{\|x\|\|x^*\|} > \text{threshold}$
4. Compute weighted distance - $\bar{d} = w_i * (x'_i - x)$ where $w_i = \frac{sim_i}{\sum_{i=j}^{n} sim_j}$
5. Move x in the desired direction - $x_{new} = x + \bar{d}$

Model Comparison

- **Generated Captions:** 5 captions per image
- **Ground Truth:** 3-5 real captions
- **Evaluation:**
  - Cosine similarity between each real caption and the 5 generated ones
  - Average the similarity scores across all real captions.
- **Tool:** *all-MiniLM-L6-V2* for semantic comparison.
- Focused on the 15 images where the original CapDec model performed best.

Type of Noise

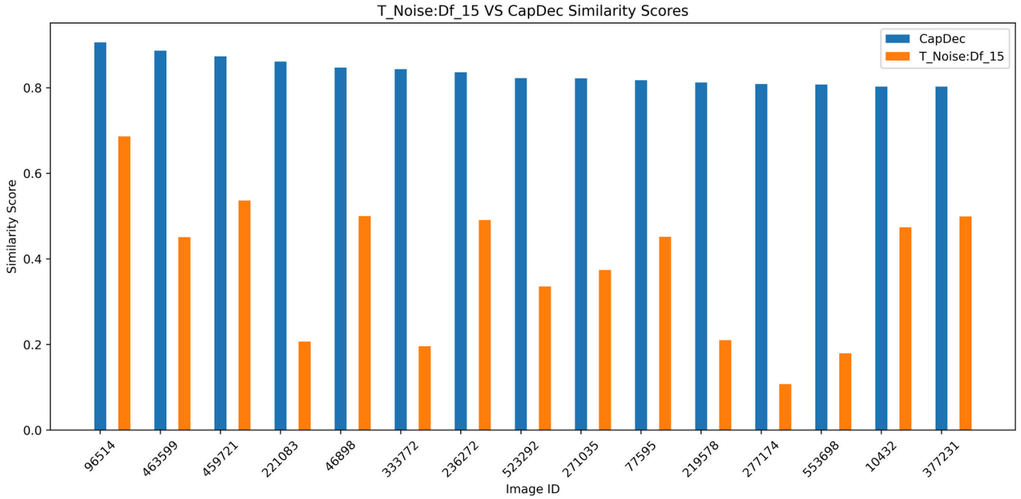
Training:

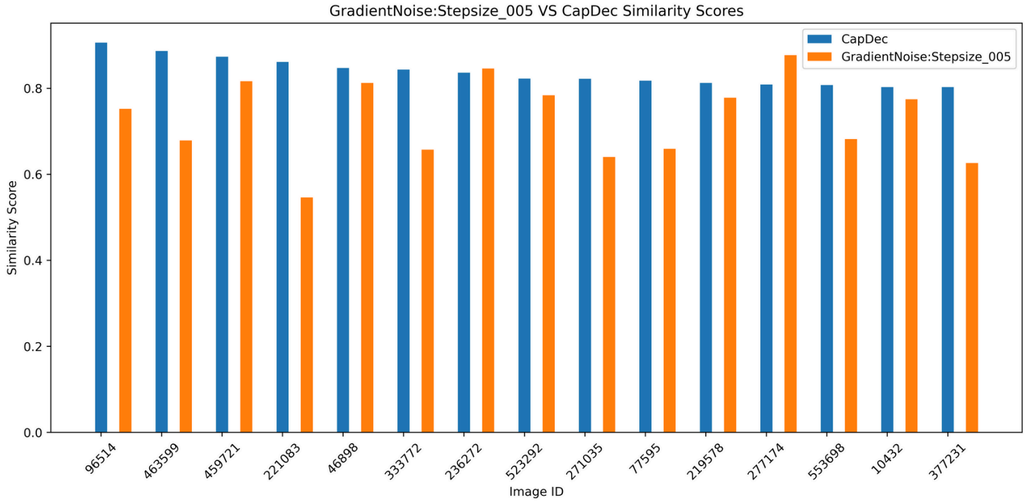
- T-Distribution noise - $t_3, t_5, t_7, t_{10}, t_{15}, t_{20}$
- Normalized Gradient noise - $\epsilon \in \{0.1, 0.05, 0.01\}$

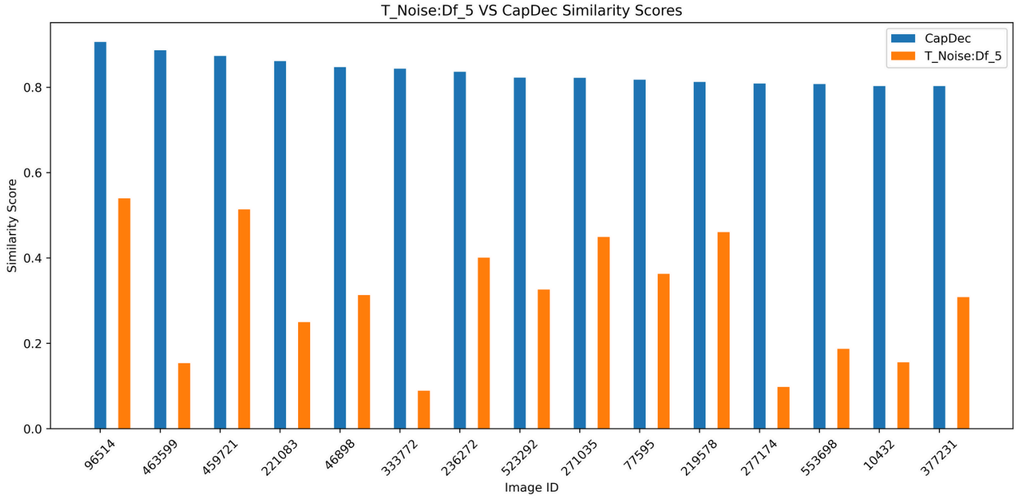
Pre-Trained CapDec:

- Gaussian Noise - $noise \sim N(0, 0.05)$

Type of Noise - Evaluation







Type of Noise - Evaluation

True: A small white bird standing on top of a pond of water

CapDec: a black and white bird standing in shallow water.

Grad Noise: a bird that is sitting on a branch.

T Noise: a very colorful photo of the bay of biscay, with the caption, \spaniel st.\

HyperParameter Tuning

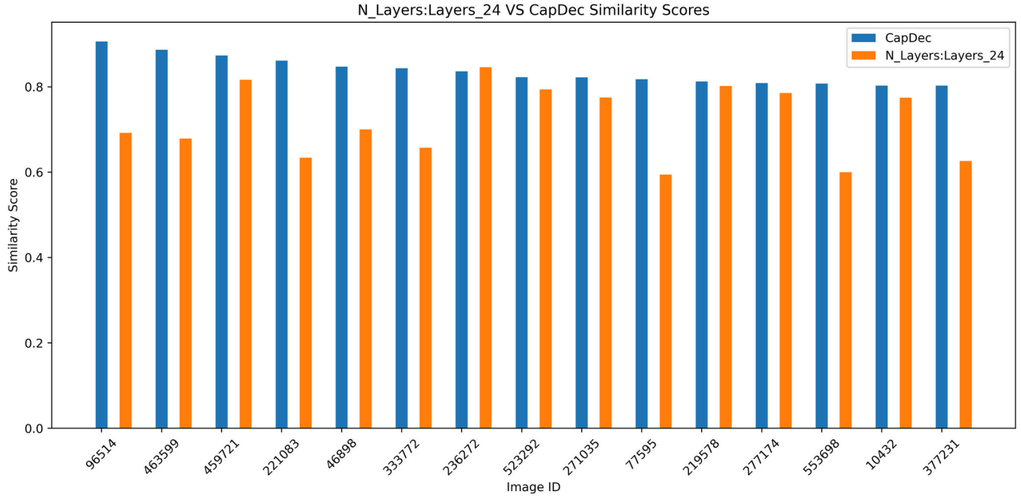
Training:

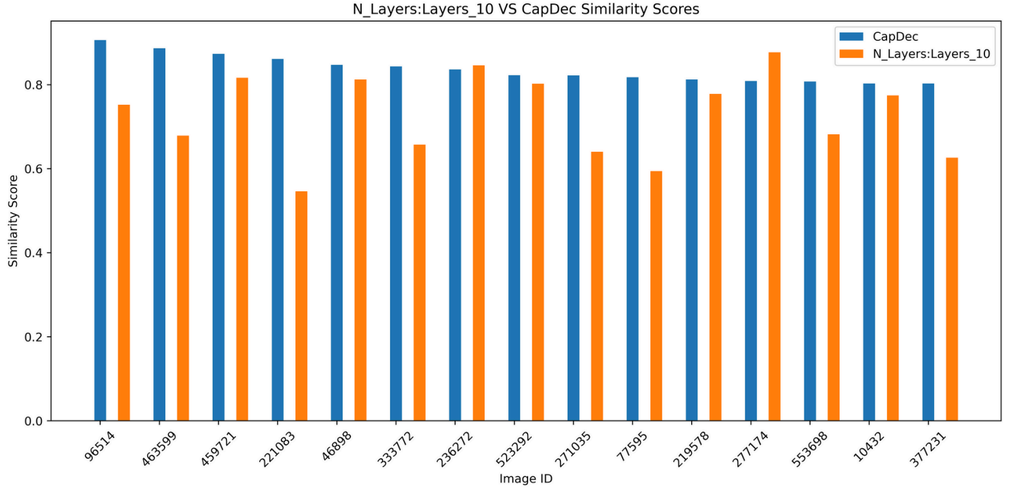
- Number of Attention Heads - $heads \in \{2, 4, 16\}$
- Number of Decoder Transformer Layers - $l \in \{4, 6, 10, 16, 24\}$

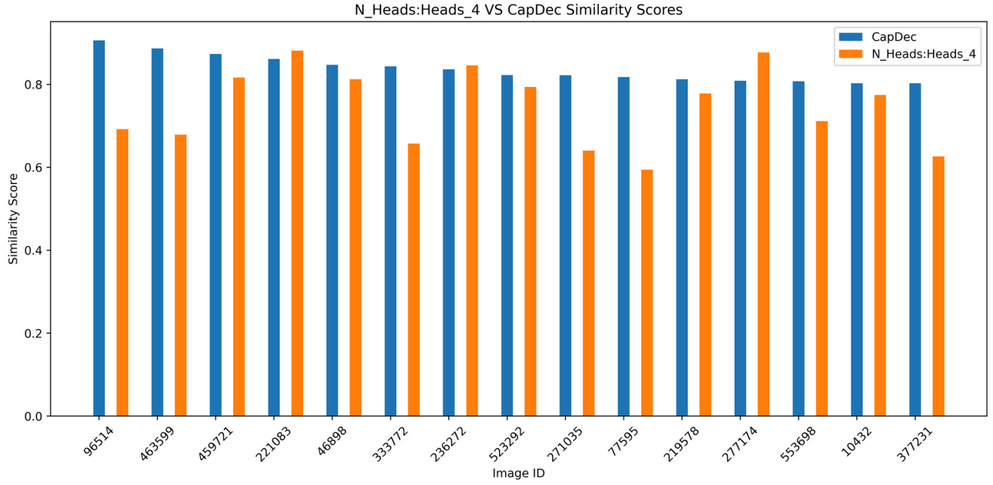
Pre-Trained CapDec:

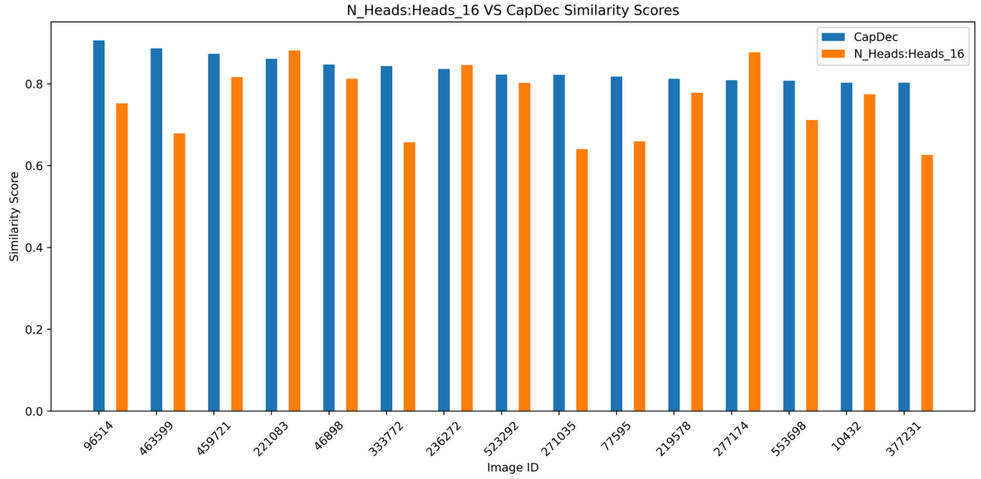
- Number of Attention Heads - 8
- Number of Decoder Transformer Layers - 8

Hyperparameter Tuning









Hyperparameter Tuning

True: A bear is sitting in the grass in front of a rusty chain-link fence

CapDec: A brown bear sits in the grass near a fence.

10 Layers: a bear that is standing in the grass.

16 Heads: a bear that is standing in the grass.

Cosine Similarity Directed Gaussian Noise

Training:

- N - $N \in \{30, 50, 100, 200, 300, 400, 500\}$
- Beta - $\beta \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$
- Threshold - $\text{th} = \{0.7, 0.75, 0.8, 0.85, 0.9\}$

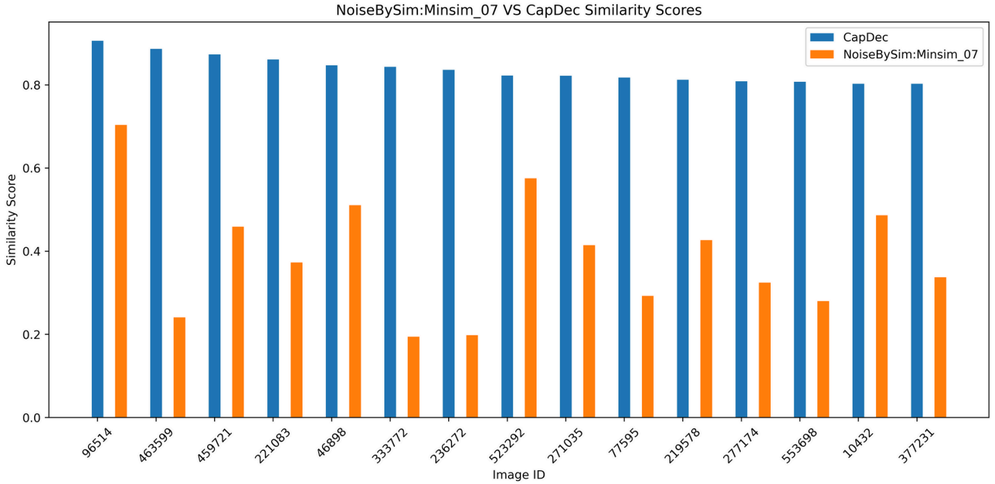
Pre-Trained CapDec:

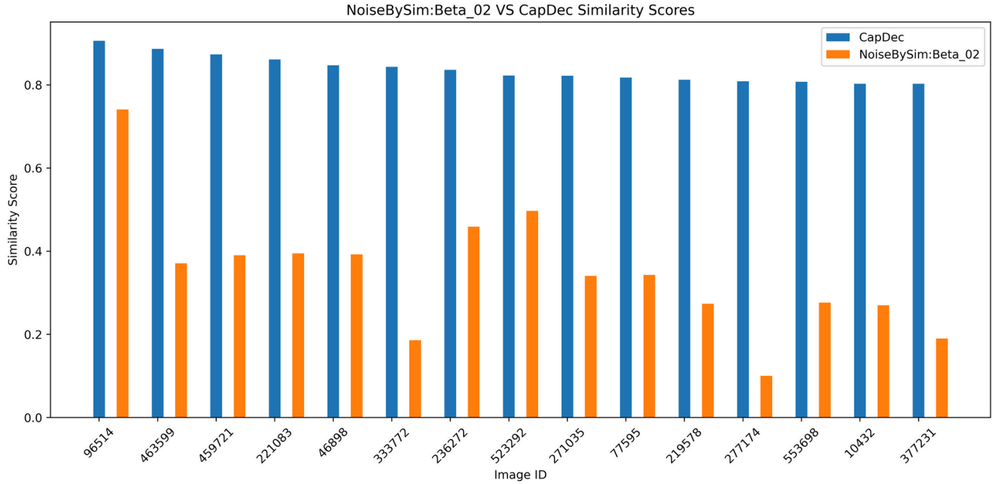
- Noise Injected text embeddings - $e^*_{text} = e_{text} + N(0, 0.05)$

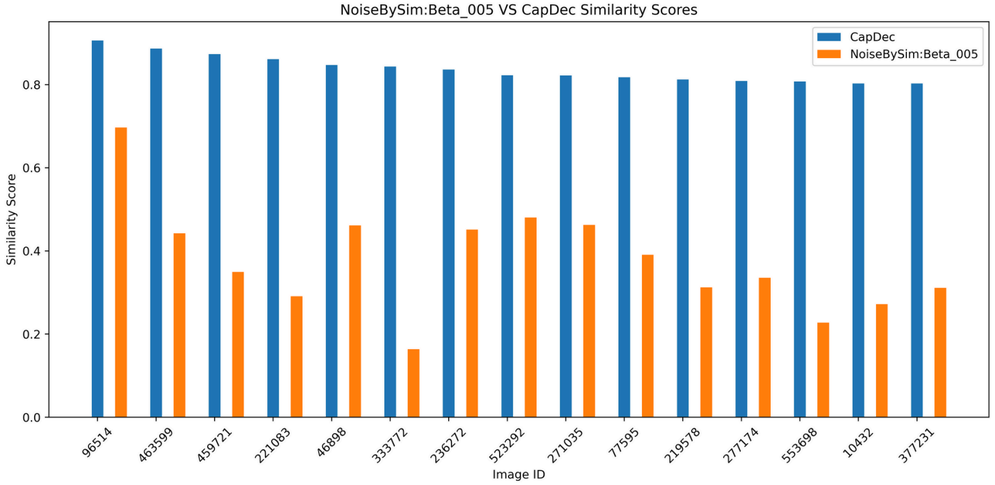
Reminder:

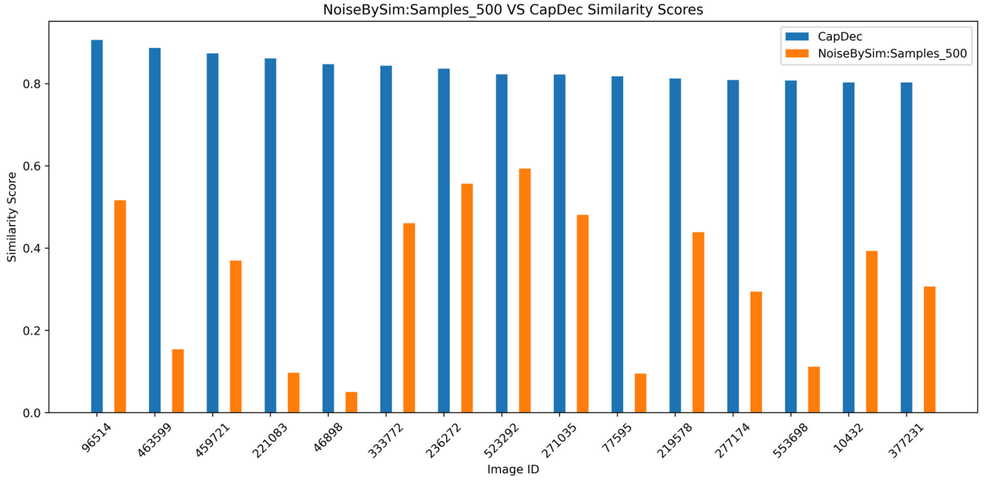
1. Sample N noise vectors - $noise \sim N(0, 1)$
2. Create N candidates - $x^* = x + \beta * noise$
3. Filter candidates - $\frac{x \cdot x^*}{\|x\|\|x^*\|} > \text{threshold}$
4. Compute weighted distance - $\bar{d} = w_i * (x'_i - x)$ where $w_i = \frac{sim_i}{\sum_{i=j}^{n} sim_j}$
5. Move x in the desired direction - $x_{new} = x + \bar{d}$

Cosine Similarity Directed Gaussian Noise

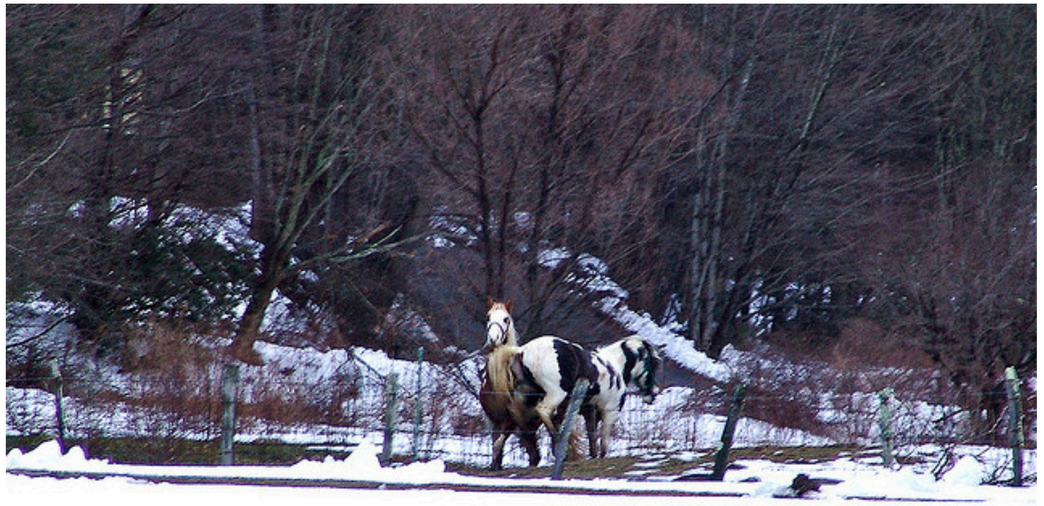








Cosine Similarity Directed Gaussian Noise



True: Two horses are standing in a snowy pasture

CapDec: a couple of horses standing in the snow near a fence

Cosine Similarity Directed Noise

- this is a very cute siamese pony, which has taken a nosed around the snowy hills.
- here is a picture of a very cute siamese llama, shetland, on her way to the hospital.
- this is a very unusual picture of mare winnie the pooh, herding her two lambs down the road.

Conclusion & Discussion

- **Pre-trained CapDec:** Delivered the best overall performance.
- **Noise Injection:** Gradient noise injection was the most effective.
- **Attention Heads:** Had minimal impact on similarity.
- **Transformer Layers:** Results were mixed, with no clear trend.
- **Consine Noise Directed:** Results were poor, but further tuning could potentially enhance their effectiveness.
- CapDec is strong, but there's room to explore better ways to bridge the modality gap.

Bibliography

- Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. Y. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems, 35, 17612-17625.
- Nukrai, D., Mokady, R., & Globerson, A. (2022). Text-only training for image captioning using noise-injected clip. arXiv preprint arXiv:2211.00575.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

THANK YOU