VINUNIVERSITY

Closing the Gap of Latent Spaces for Image-Text Arithmetic

Vu Binh Minh - V202100421
Balashova Ekaterina - V202100391
Tran Anh Vu - V202100569

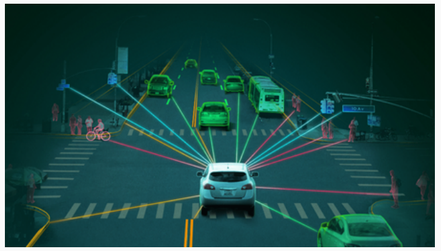
Natural Language Processing - COMP4020

Agenda

1. **Project Introduction**
2. **Dataset**
3. **Methodology**
4. **Experiments**
5. **Results & Key Findings**
6. **Challenges, Future work & Conclusion**

Introduction

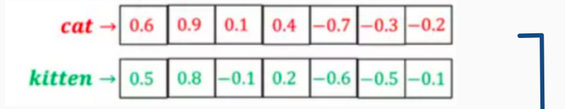
The **rapid growth** of Artificial Intelligence in recent years...

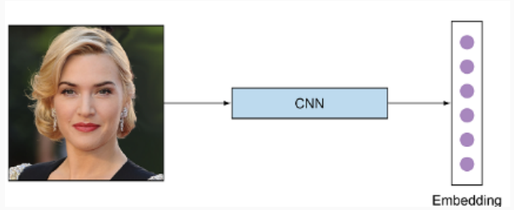


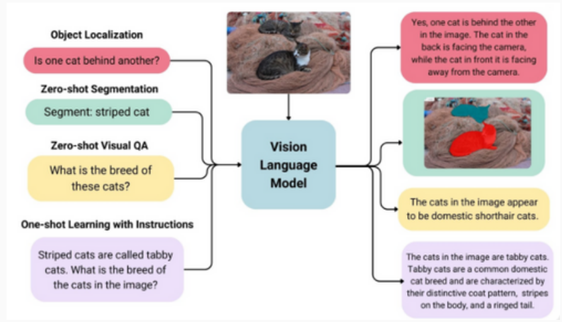
..has a **key factor**

How machine understand content/ information

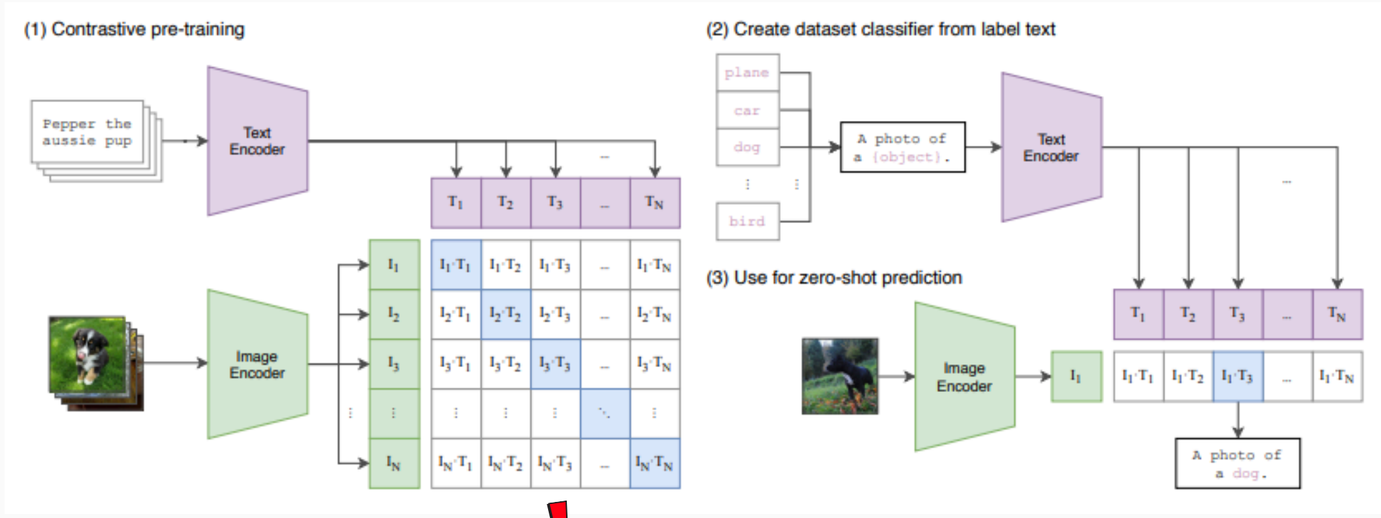
Embeddings







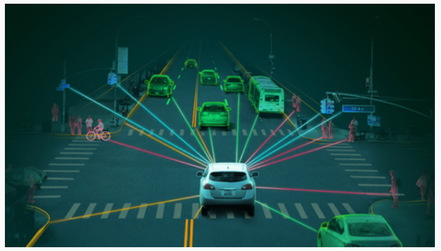
Contrastive Language-Image Pre-training(**CLIP**)



There is still a **modality gap** between **text** and **image representations**

Introduction

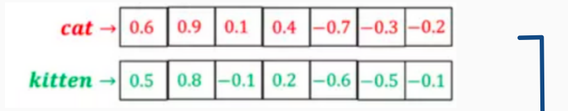
The **rapid growth** of Artificial Intelligence in recent years...

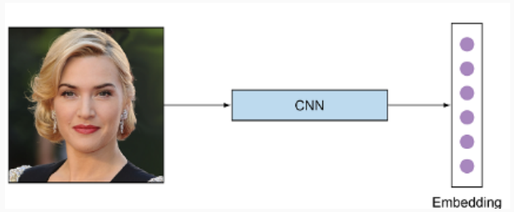


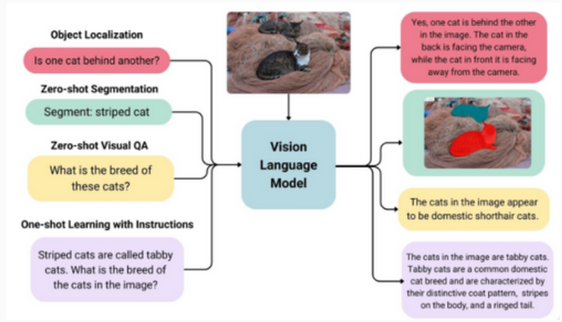
..has a **key factor**

How machine understand content/ information

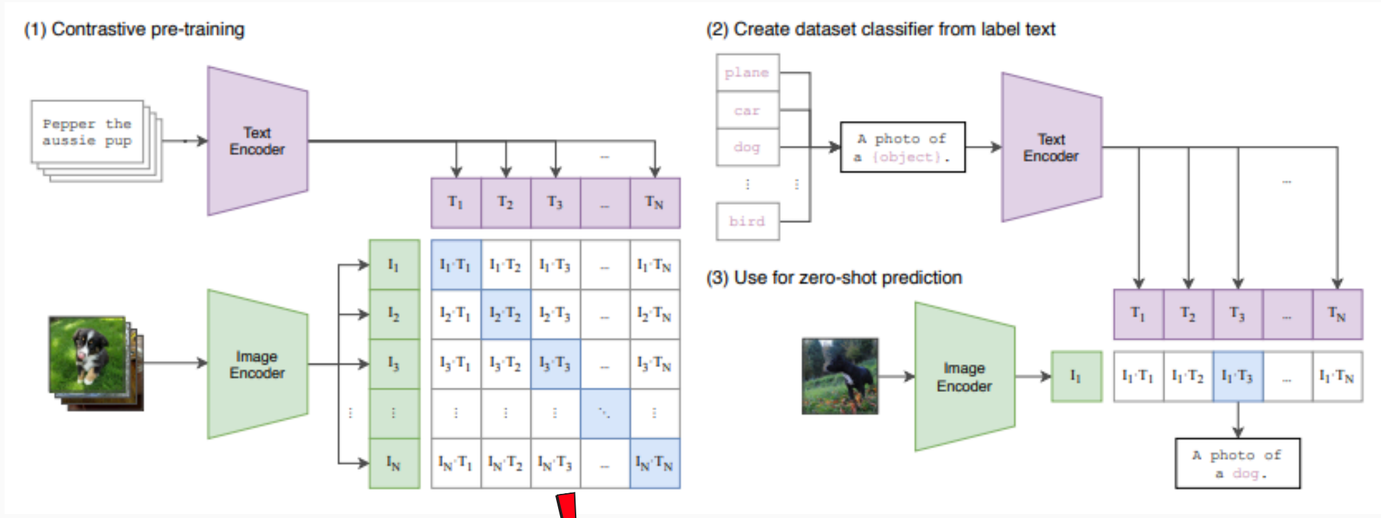
Embeddings







Contrastive Language-Image Pre-training(**CLIP**)



There is still a **modality gap** between **text** and **image representations**

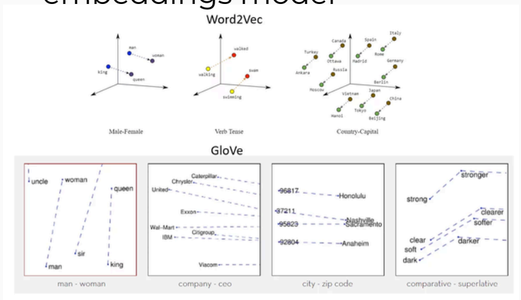
Our Project Objective:

- Creating **better shared embeddings** that align 2 types of data
- Direct **cross-modal arithmetics capabilities**

→ better performance at tasks that involve reasoning with visual and textual information simultaneously

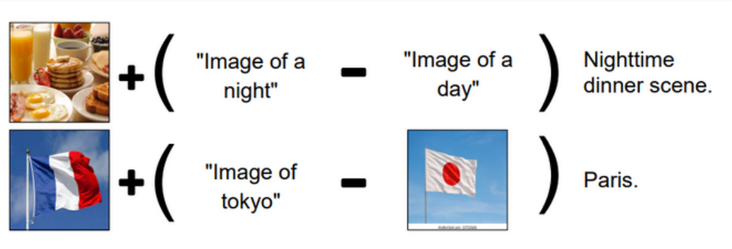
How are people doing?

Different from the unimodal embeddings model



CLIP can then be used as a generalized backbone for several different downstream tasks

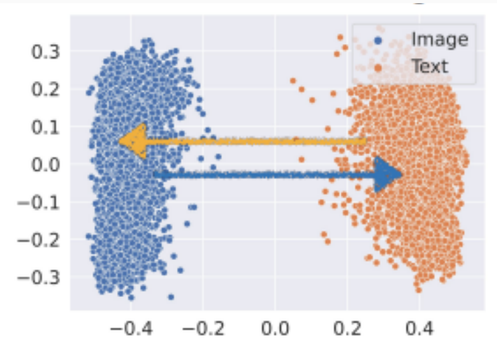
However, it's still bad compositional reasoning, fine grained understanding of image content



ZeroCap, by **directly steering the text context vectors (cache) Ci closer to the CLIP's visual embedding at inference time**, showed an impressive **visual arithmetic** capability
→ only help captioning task, nothing change with embedding space

JinaCLIP conducts external training on a synthetic text dataset
→ only improve capability to understand text meaning, not really improve shared embedding space

Some others implement the **linear shifting method**, aiming to make 2 sets of text and visual vectors close → ineffective due to distinguished structures of data within each modality



Develop a **non-linear translation mapping** that can directly align images and text, which enables **cross-modal arithmetic**.

Overview of closing Modality Gap

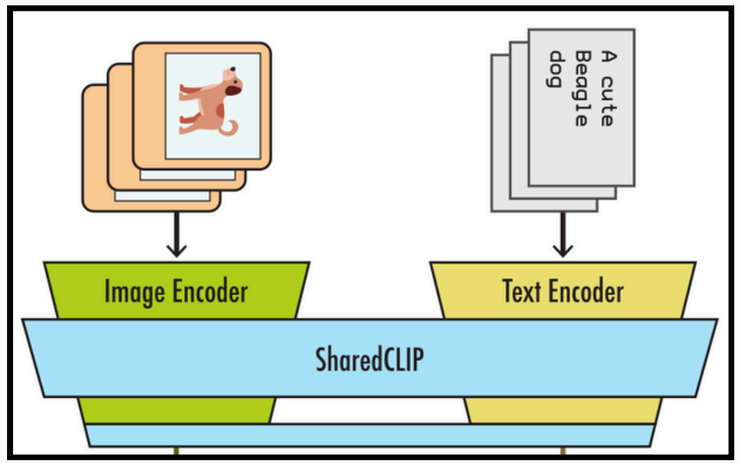
There are two main approaches to directly tackle this in literature:

- **Post-hoc linear translation:**
  - Statistically aligning the embeddings of texts and vectors. → *Cheaper to apply*
  - Sample work: shifting embeddings by the difference between the mean of modality's embeddings (Liang et al., 2022).

$$\vec{\Delta}_{\text{gap}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$$

- **Improvement of CLIP's architecture or training strategy (mostly still Contrastive Learning):**
  - Introducing improved architectures, loss objectives, and training set-up.
  - Sample work: shared architecture and a novel training objective (Eslami & De Melo, 2025).

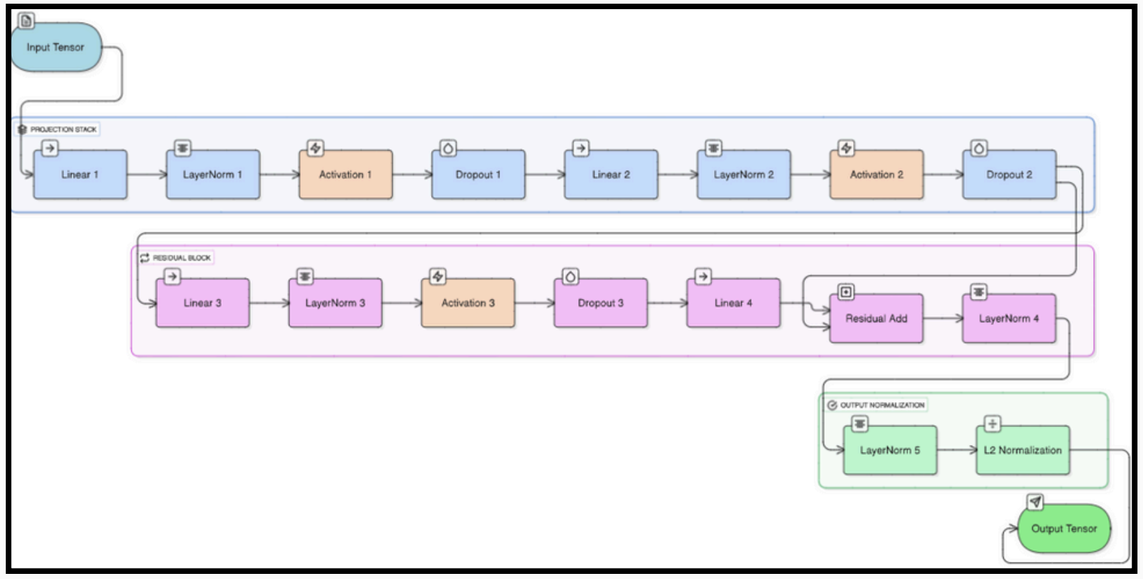
→ *Requires huge resources to train*



Methodology

Idea: Train a simple non-linear model to learn a mapping from image to text embeddings

Non-linear Aligner Model

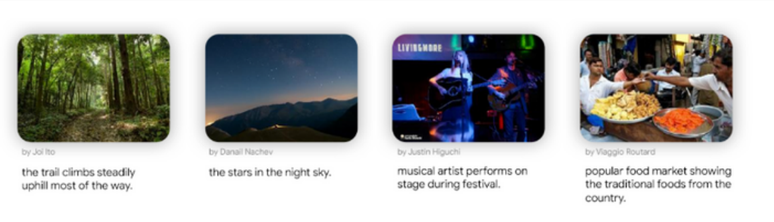


Cosine Distance loss slightly yielded a better result then MSE loss. We also tried a combination of weighted clip loss and IMSep loss (from AlignCLIP)

Arithmetics

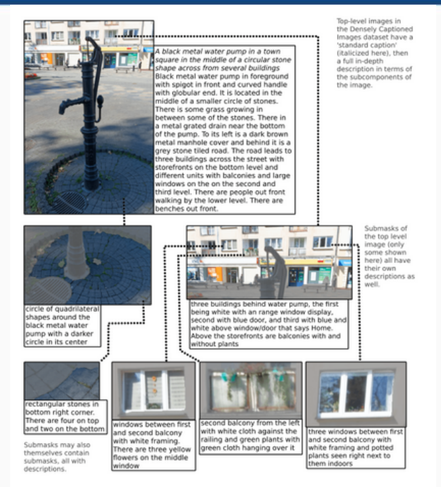
The existence of the modality gap hinders cross-modal arithmetics. Directly aligning image and text embeddings has the potential to improve this.

Dataset



1. CC3M - Conceptual Caption:

3,3 million images with accompanying captions that were automatically mined from the internet. 10,000 images for model training



2. DCI - Densely Captioned Images

This dataset contains 7805 images with very detailed captions.

3. COCO - Common Objects In Context:

328,000 images of common objects, each manually labeled with 5 captions. 2017 validation split (5000 images) for test set

Preprocessing data

- For training, generate synthetic captions for CC3M & COCO using Gemini 2.0 Flash Lite (without human subjectivity, VLMs tend to produce more high-quality captions → better for later training)
- Get pairs of embeddings of images and captions by CLIP

Experiments and Results

Modality gap

Metrics

1. **Alignment score** — average cosine similarity between matching text and image embeddings

$$\text{Alignment} = \frac{1}{N}\sum_{i=1}^{N}\cos(\mathbf{x}_i, \mathbf{y}_i)$$

2. **L2M** — L2 distance between modality means.

$$\text{L2M} = \left\| \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i - \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i \right\|_2$$

3. **Relative Modality Gap (RMG)** — the proportion of mean distance between matching pairs to the sum of itself and mean intra-modality distance.

$$\text{RMG} = \frac{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{y}_i\|_2}{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{y}_i\|_2 + \frac{1}{2N(N-1)}\left(\sum_{i\neq j}\|\mathbf{x}_i - \mathbf{x}_j\|_2 + \sum_{i\neq j}\|\mathbf{y}_i - \mathbf{y}_j\|_2\right)}$$

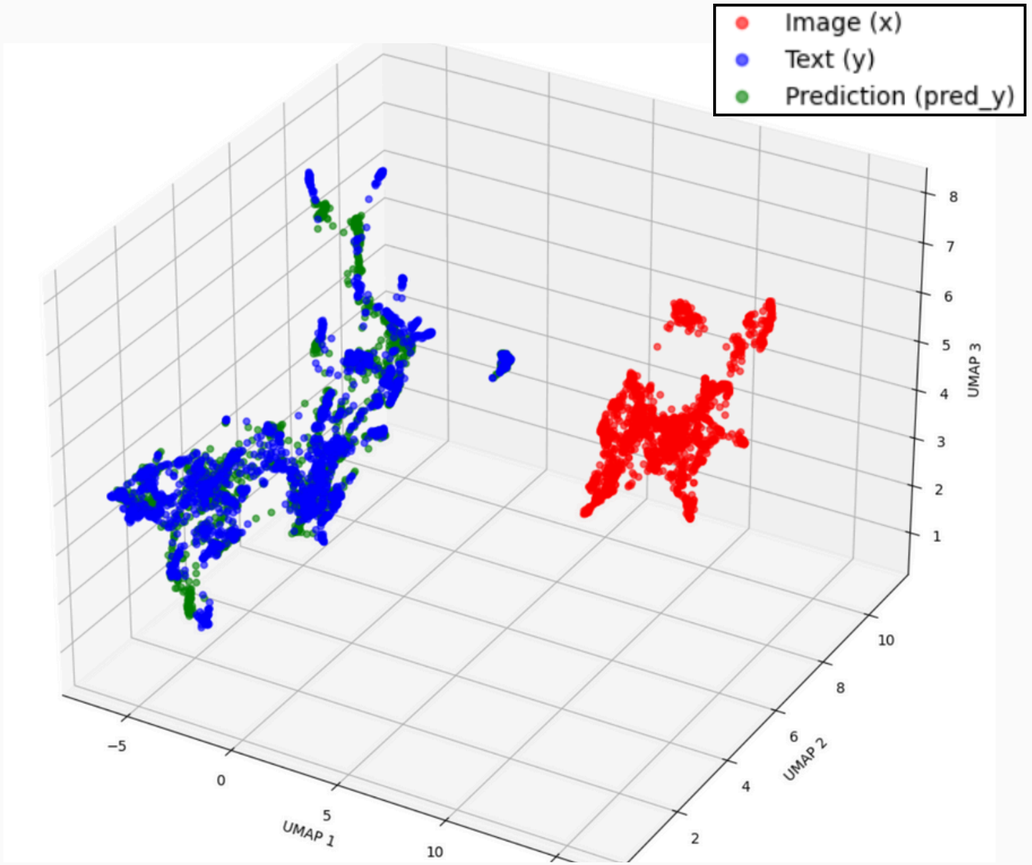
Experiments and Results

Modality gap

Results

Metric	Original CLIP	After Non-Linear Aligner
Alignment Score (↑)	0.3455	0.7607
L2M (↓)	0.7508	0.2070
RMG (↓)	0.5138	0.4032

Metrics before and after alignment



3D UMAP visualization of the embedding space after alignment.
Red and blue are CLIP's embeddings. Green depicts new image embeddings

Experiments and Results

Zero-shot classification: CIFAR100

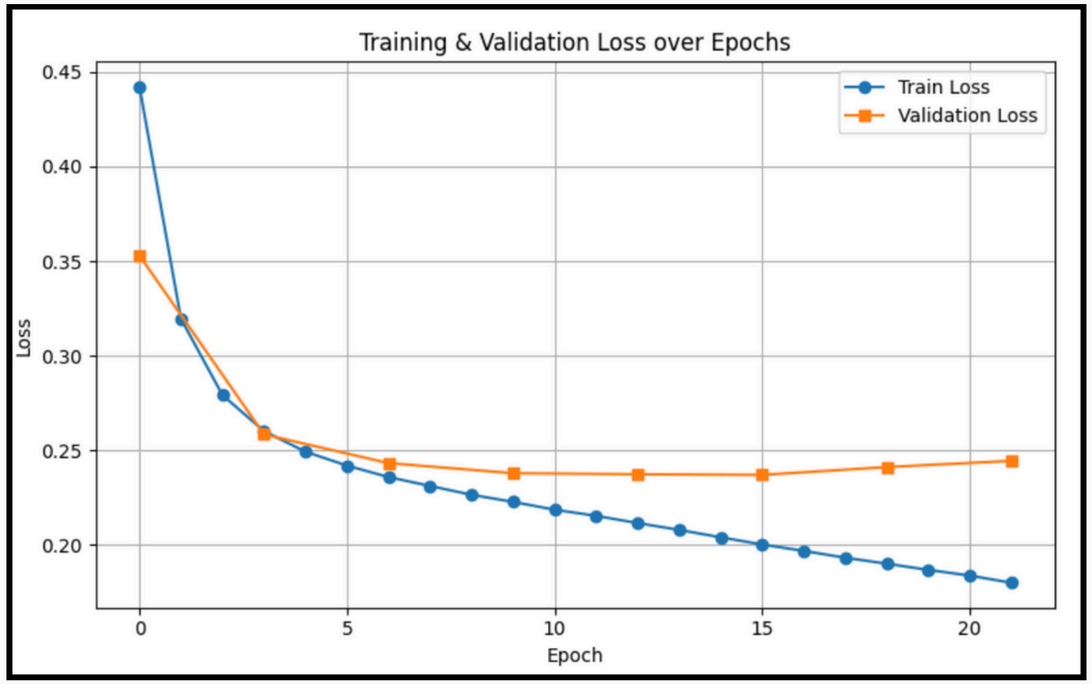
Zero-shot classification performance worsens as expected since we cannot fully close the gap, regardless of explicit supervision with MLP.
→ Without a good mapping, we cannot ensure to retain the original characteristics of CLIP

Ours:

- Cosine distance loss only:
  - Top-1 accuracy: 20.83
  - Top-5 accuracy: 48.37
- Cosine distance + Weighted Clip Loss + IMSep Loss (from AlignCLIP)
  - Top-1 accuracy: 41.23
  - Top-5 accuracy: 69.83

Original CLIP:

- Top-1 accuracy: 61.7



Training and Validation Cosine Distance loss. Looking at the loss, there is still modality gap, preventing cross-modal arithmetics.

Challenges, Key Findings & Conclusion

Challenges and limitations

- **CLIP**
  - Input length limited to 77 tokens
  - Quality of aligned embeddings depends on original embedding quality → models like JinaCLIP or AlignCLIP worth investigating
- **Resource availability**
  - We can only use a small fraction of data
- **Information imbalance** (Schrodi et al., 2025)
  - The embedding space of CLIP, which was used to extract data for our method, was still flawed because of information imbalance (or the lack of details in text captions).

Challenges, Key Findings & Conclusion

Key Findings

- We gave a survey and analyzed current approaches to minimize modality gaps between modalities, or to introduce semantic arithmetic.

- Attempted to use MLP as a non-linear alignment architecture to translate visual embedding to text embedding space
    - MLP is not sufficient with just MSE or Cosine Distance Loss. Adding the original InfoNCE Loss from CLIP and IMSep Loss from AlignCLIP mitigates this, but there is still an unremovable gap as showned in the loss objective.
    - Thus, more complex methods should be investigated. The unremovable gap should be investigated as well, as it could originate from information imbalance of CLIP's training data.

Challenges, Key Findings & Conclusion

Conclusion

Our approach contributes towards closing the modality gap, but imperfections remain, impacting the zero-shot classification performance. Further investigation is necessary to verify cross-modal arithmetic capabilities.

Future direction:

- Explore other methods to introduce non-linearity into the translation.
- Investigate the impact of the detailedness of text captions on aligning features.

References

Eslami, S., & De Melo, G. (2025). *Mitigate the gap: Improving Cross-Modal alignment in CLIP*. OpenReview. https://openreview.net/forum?id=aPTGvFqile

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. (2022, October 31). *Mind the Gap: Understanding the modality gap in multi-modal contrastive representation learning*. OpenReview. https://openreview.net/forum?id=S7Evzt9uit3

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021, February 26). *Learning transferable visual models from natural language supervision*. arXiv.org. https://arxiv.org/abs/2103.00020

Schrodi, S., Hoffmann, D. T., Argus, M., Fischer, V., & Brox, T. (2025). *Two effects, one trigger: on the modality gap, object bias, and information imbalance in contrastive Vision-Language models*. OpenReview. https://openreview.net/forum?id=uAFHCZRmXk

Tewel, Y., Shalev, Y., Schwartz, I., & Wolf, L. (2022, January 1). *ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic*. OpenReview. https://openreview.net/forum?id=o1vh05CloG&referrer=%5Bthe%20profile%20of%20Yoad%20Tewel%5D(%2Fprofile%3Fid%3D~Yoad_Tewel1)

Xiao, H., Mastrapas, G., & Wang, B. (2023). *Jina CLIP: Your CLIP model is also your text retriever*. OpenReview. https://openreview.net/forum?id=lSDkG98goM&referrer=%5Bthe%20profile%20of%20Georgios%20Mastrapas%5D(%2Fprofile%3Fid%3D~Georgios_Mastrapas1)

Please refer to the final report for the full reference list.