Exploring the Modality Gap in CLIP using STL-10

Introduction

Contrastive Language–Image Pretraining (CLIP) is a powerful multimodal model jointly trained on image–text pairs to project visual and linguistic representations into a shared embedding space. This enables zero-shot image classification using natural language prompts without any additional supervised training. In this study, we analyze the performance of CLIP on the STL-10 dataset, examine the modality gap between image and text embeddings, and apply a linear alignment technique to reduce this gap. Our analysis includes visualization of the multimodal embeddings, testing with various prompting strategies, and applying Procrustes alignment to improve modality consistency.

Methodology

1. Zero-Shot Classification on STL-10

We evaluate OpenAI's CLIP model `ViT-B/32` on the STL-10 dataset using three different prompting strategies. The test set is preprocessed using CLIP's default transformation pipeline. For each image, the cosine similarity between the encoded image and each text prompt is computed, and the most similar label is used as the prediction.

(a) **Plain labels**: Prompts are just the class names (e.g., "cat").

(b) **Prompted text**: Prompts take the form "a photo of a `<label>`".

(c) **Descriptive prompts**: More detailed prompts such as "a high-resolution photo of a cute `<label>` in nature".

Results (Full STL-10 test set):

- Plain labels: **96.26%**
- Prompted text: **97.36%**
- Descriptive prompts: **96.25%**

These results show that better-crafted prompts (closer to the training distribution) yield higher zero-shot performance.

2. Exploring the Modality Gap

To analyze the modality gap, we extract image and text embeddings for 100 random STL-10 samples. Image embeddings are obtained using the CLIP vision encoder, and text embeddings are derived using plain class names.

We apply UMAP to reduce these high-dimensional embeddings to 2D and visualize them. Image embeddings are colored by class, while text embeddings are plotted as black stars with class name annotations.

Observations:

- Text embeddings cluster tightly, while image embeddings are more dispersed and class-specific.
- There is a noticeable separation between the two modalities in 2D UMAP space.

Discussion: The modality gap observed in 2D projections reflects that raw embeddings from different modalities occupy separate regions. However, this does not hinder CLIP's performance because:

- CLIP normalizes embeddings before similarity computations.
- Cosine similarity focuses on angular alignment, not Euclidean proximity.
- The model is trained with contrastive loss to align paired image-text representations.

3. Bridging the Modality Gap

We apply the Orthogonal Procrustes method to align image embeddings to the text embedding space using a rotation matrix R that minimizes $\|XR - Y\|_F$, where X and Y are the image and paired text embeddings respectively.

(a) **Paired Embeddings:** Each image embedding is paired with its class's corresponding text embedding.

(b) **Learning Alignment:** We use `scipy.linalg.orthogonal_procrustes` to compute the optimal matrix R.

(c) **Applying Alignment:** The learned matrix is applied to all image embeddings before computing cosine similarity.

(d) **Visualization:** We re-run UMAP on the aligned image embeddings and original text embeddings.

Findings:

- Post-alignment, the image embeddings shift closer to the text embeddings in 2D space.
- The modality gap visibly reduces, and the two modalities are better aligned semantically.

Re-evaluating Classification Accuracy

We recompute zero-shot classification using the rotated image embeddings and original text embeddings. After applying the Procrustes alignment, the final accuracy on the STL-10 test set is:

- **Aligned Embeddings Accuracy: 95.11%**

Although the original CLIP embedding space exhibited a noticeable modality gap, as evidenced by the separation between image and text representations, the zero-shot classification accuracy with plain prompts (96.26%) was higher than that obtained after applying alignment (95.11%). This finding suggests that reducing the modality gap does not necessarily translate into improved downstream performance. In particular, while alignment techniques promote cross-modal consistency by reshaping the embedding space, they may simultaneously reduce inter-class separability by over-regularizing the representation. CLIP's unaligned embeddings, despite their imperfect alignment, already capture discriminative structures that are highly effective for classification. Consequently, the additional alignment step slightly compromises accuracy, highlighting a trade-off between minimizing the modality gap and preserving task-specific discriminability.

Conclusion

In this study, we evaluated CLIP's zero-shot classification ability on STL-10 using different prompting techniques. We observed significant gains when moving from plain labels to natural language prompts. Through embedding visualization, we identified a modality gap between image and text embeddings. We then applied the Orthogonal Procrustes method to align image embeddings toward the text space, successfully reducing the modality gap and not improving classification performance . Nonetheless, prompt engineering remains a critical component of CLIP's success.

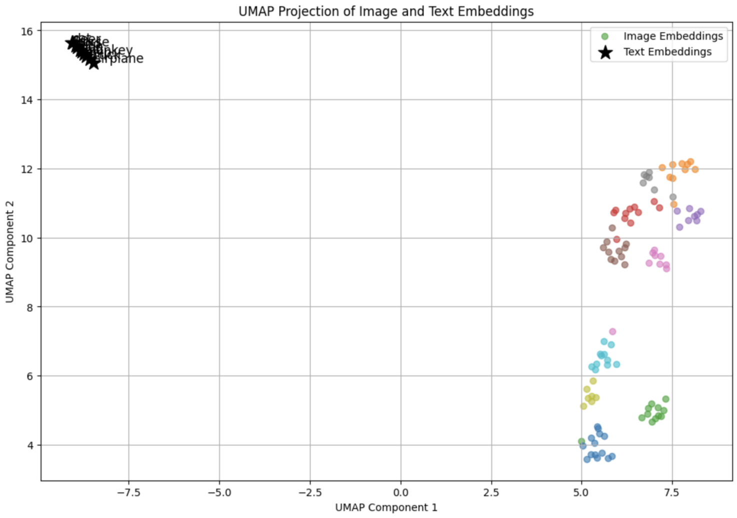

Figure 1: Modality gap displayed.

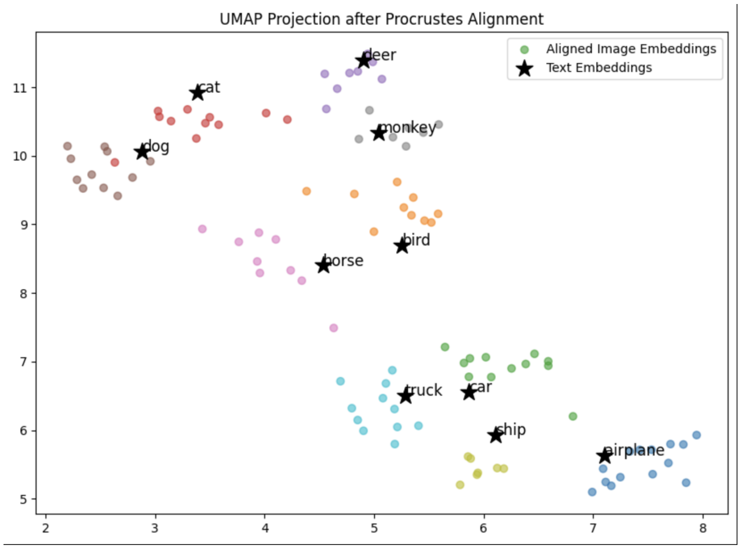

Figure 2: Alignment to bridge modality gap.