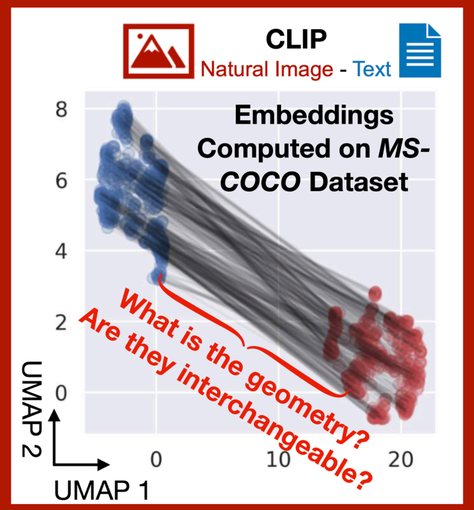
Connect, Collapse, Corrupt: Learning Cross-Modal Tasks with Uni-Modal Data

Yuhui Zhang*, Elaine Sui*, Serena Yeung-Levy

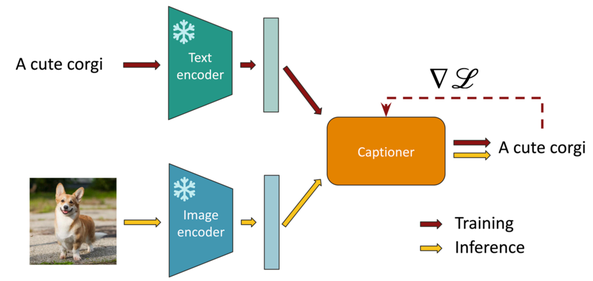
Based on the multi-modal contrastive representation space geometry, we propose a simple three-step method C^3 **(Connect, Collapse, Corrupt)** to bridge the modality gap, enhancing the interchangeability of embeddings.



Paper

Introduction

Can we train cross-modal models (e.g., image captioner) with only uni-modal data (e.g., text only)?

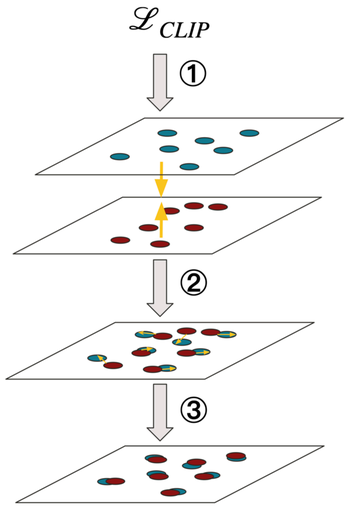


Despite multi-modal contrastive learning, resulting embeddings of different modalities are **not directly interchangeable** due to their geometry.

$$\boldsymbol{e}_x - \boldsymbol{e}_y = c_\perp + \epsilon$$

Modality 1 Embedding **Modality 2 Embedding** **Modality Gap** **Alignment Noise**

Based on this geometry, we propose a **simple three-step method** to improve embedding interchangeability.



① **Connect**: Relate concepts with multi-modal contrastive learning.

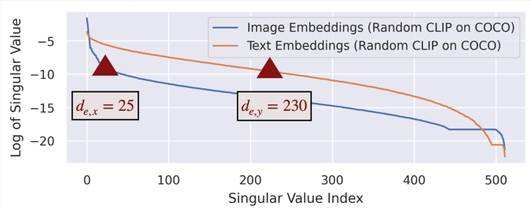
② **Collapse**: Close the modality gap by subtracting the embedding mean from each modality.

③ **Corrupt**: Add Gaussian noise to the embeddings as regularization during training.

Modality Gap

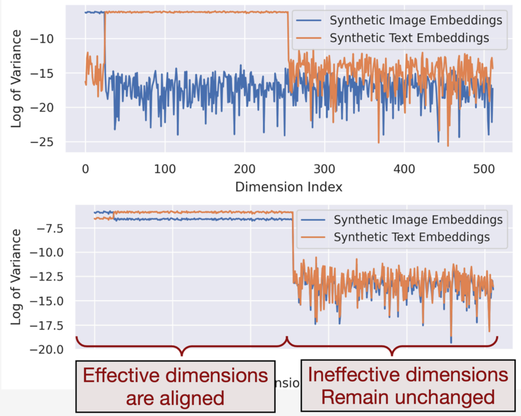
1. Modality Gap Emerges at Initialization

Due to **dimensional collapse**, CLIP embeddings have a lower effective dimension than its total dimensionality.



2. Modality Gap Preserves after Optimization

Based on **gradients of contrastive optimization**, only effective dimensions are aligned after training.



Alignment Noise

Given the contrastive loss, the **stable region** is a function of temperature. Within the stable region, the optimization halts.

Stable region for x_+: $r = s_+ - \max s_- \geq \tau \log o(\tau)/(\exp(\delta) - 1) \Rightarrow \mathscr{L} \leq \delta$

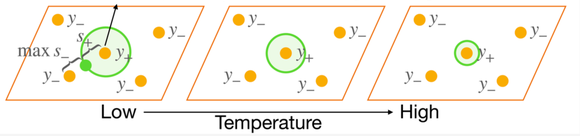
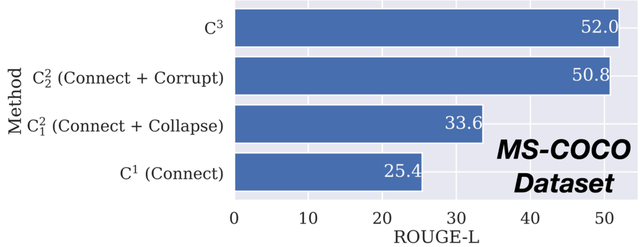


Image Captioning

C^3 significantly improves zero-shot image captioning.



C^1: A person and their luggage, resting in the courtyard of the University of Southern California, observing the Eiffel Tower.

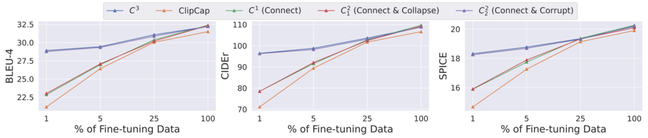
C_1^2: A person on a bench in the midst of the street to take a walk with their library.

C_2^2: A man sitting on a bench with a book.

C^3: A man and woman sitting on a park bench.

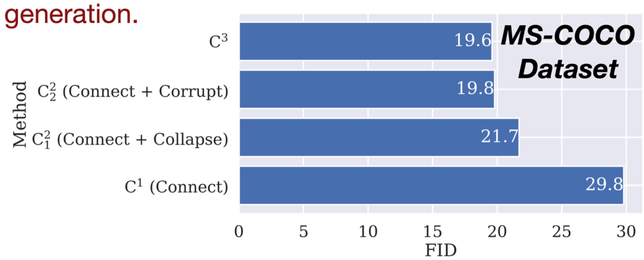
Ground truth: A man and woman sitting on a park bench under a huge tree

Semi-Supervised Setting



Text-to-Image Generation

C^3 significantly improves zero-shot text-to-image generation.



A street scene with a double-decker bus on the road.

Ground Truth *C^1* *C^3*