The Contrastive Gap: A New Perspective on the 'Modality Gap' in Multimodal Contrastive Learning

by

Abrar Fahim

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Learning jointly from images and texts using contrastive pre-training has emerged as an effective method to train large-scale models with a strong grasp of semantic image concepts. For instance, CLIP, pre-trained on a large corpus of web data, excels in tasks like zero-shot image classification, object detection, geolocalization, and more. These contrastive models embed input images and texts into a shared representational space.

Recently, it was discovered that models like CLIP show a "modality gap", where image and text embeddings occupy disjoint areas in the representational space. Previous research attributes this gap to factors like data artifacts (mismatched pairs), model architecture artifacts (the cone effect), and the nature of the loss landscape (getting stuck in local minima). In this thesis, we demonstrate that even after accounting for these factors, the contrastive loss itself creates this gap during training. We propose renaming this phenomenon as the "contrastive gap" and show that it stems from low uniformity in the CLIP space, where embeddings only occupy a small portion of the latent space. We show that optimizing for uniformity and alignment in the CLIP space reduces the contrastive gap. Our experiments show that this modified representational space achieves better performance on downstream tasks like zero-shot image classification and multi-modal arithmetic, suggesting the effectiveness of closing the contrastive gap to boost CLIP performance.

Preface

This thesis is based upon work that was published on arXiv (`https://arxiv.org/abs/2405.18570`), with co-authors Dr. Alex Murphy and Dr. Alona Fyshe.

To Ma Baba

For everything

"Research is what I'm doing when I don't know what I'm doing."

– Wernher von Braun

Acknowledgements

I strongly believe that I have become a significantly better researcher now than when I started just over a year and a half ago. This journey would not have been possible without the support and guidance of the incredible people around me.

First of all, I want to thank my supervisor, Dr. Alona Fyshe. I would not have gotten this far without her constant support and guidance, especially in times when I felt lost in the research process.

Next, I want to thank Dr. Alex Murphy, who helped formulate a lot of the ideas that eventually led to the completion of this work. His support has been invaluable to the completion of this work.

Then, I thank my lab-mates and colleagues at the University of Alberta. Their valuable feedback on my presentations and experimentation techniques significantly improved the quality of my research.

Finally, I am grateful to my family and my friends, who often believed in me more than I believed in myself. I feel incredibly lucky to have their constant support and companionship at every step of my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

We experience and learn from our surroundings by receiving inputs from our various sense organs. For example, our eyes capture light, allowing us to "see," while our ears detect sound, enabling us to "hear." Our brain processes these inputs together, integrating them to form a comprehensive understanding of the world around us. This understanding is what allows us to perform the wide range of tasks that we do every day. We can define the inputs to each of our sense organs as a *modality*: sight, sound, and touch are all examples of different modalities.

Motivated by the brain's multi-modal processing, learning from different modalities has recently become an emerging interest in the representation learning community. This is evidenced by the large body of recent work done in this field [1], [9], [14], [17], [18], [28], [30], [34], [38]–[40]. In this thesis, we focus primarily on pre-trained vision-language models, where the models learn representations from image and text modalities.

A machine learning model maps its inputs to a representational space to solve a task relating to the input. For instance, an image classifier might project an image into a 64-dimensional vector, using the features of that vector to finally classify the image. In this case, the 64-dimensional vector represents a point in a 64-dimensional representational space. However, a multi-modal machine-learning model can take in inputs from different modalities simultaneously. Unlike a simple image classifier, a multi-modal model can process (for instance) both images and sounds at the same time, mapping them into

a unified representational space. In this way, the model can learn to solve complex tasks by relating the inputs of the different modalities.

In this thesis, we focus primarily on pre-trained image-text models that learn representations from image and text modalities by mapping them into a shared representational space. The models learn to project semantically similar inputs from different modalities to map to nearby points. This multi-modal approach is beneficial as it thematically facilitates transfer between modalities and aligns more closely with human sensory experiences (Than a model that can work with only image or only text inputs).

CLIP (Contrastive Language Image Pre-training) Radford *et al.* [28] is a strong proof-of-concept for multi-modal representation learning, particularly in the context of learning from paired images and text. CLIP's multi-modal contrastive loss enables the model to predict the text associated with an image and vice versa. By scaling this approach to a vast dataset of 400 million image-caption pairs, CLIP learns embeddings that cover a wide variety of visual concepts, making them applicable to many downstream tasks. Out of the box, CLIP is capable of performing a wide range of tasks such as image-text retrieval, zero-shot image classification, OCR, geo-localization, and action recognition.

CLIP (Contrastive Language Image Pre-training) Radford *et al.* [28] establishes a strong proof-of-concept for multi-modal representation learning in the context of learning from paired images and text. CLIP uses a multi-modal contrastive loss to predict a caption associated with an image and vice versa. CLIP scales this approach to a very large dataset of 400M image-caption pairs, learning embeddings that cover a wide variety of visual concepts applicable to many downstream tasks. For example, CLIP embeddings can be used for tasks such as zero-shot image classification, OCR, geo-localization, and action-recognition. There have been an influx of subsequent works adapting (fine-tuning) the learned representations by CLIP for many tasks such including few-shot image classification [27], [41], [45], video-retrieval [2], [3], [23], depth estimation [42], image-captioning [5], [7], [24], [40] , and visual-question answering [16], [29].

But, while CLIP is powerful, it suffers from a *modality gap*, first identified by Liang *et al.* [20]. In the CLIP representational space where all the input images and texts are projected to, we expect that the image and text embeddings would occupy the same shared representational space as each other. However, this is not the case. Data from different modalities (For example, images and texts) are projected to copmpletely disjoint subspaces within the shared representational space. We illustrate this phenomenon in Figure 1.1.

The modality gap phenomenon is pervasive in multi-modal contrastive models across many different domains such as medical images [43], videos [37], amino acid sequencing (`https://github.com/MicPie/clasp`) and brain decoding [22]. Udandarao [31] suggests that the presence of this gap complicates the visualization and interpretation of the embedding space, making it harder to understand how models represent and relate different modalities. Udandarao [31] also shows that the modality gap negatively affects the model's ability to perform meaningful vector arithmetic between the image and text sub-spaces, limiting the model's potential for creative applications.

Prior work has shown that performance on downstream tasks can improve when we minimize the modality gap. For instance, Liang *et al.* [20] show that modifying the size of the modality gap by simply translating the image and text embeddings can impact image-classification performance. Further, Zhou *et al.* [44] show that reducing the modaltiy gap (by projecting the text embedings onto another subspace of relevant attributes) improves interpretability of the shared latent space, and allows for better text-guided image manipulation using CLIP embeddings. Finally, Oh *et al.* [26] show that reducing the modality gap (by training using synthetically generated data points) can improve performance across a wide range of downstream tasks, such as image classification, and caption generation using CLIP latents. Muttenthaler *et al.* [25] has additionally shown that learning latent spaces of visual embeddings via alignment with human similarity judgments improves downstream task performance. Similarly, Luo *et al.* [22] related CLIP embeddings to human brain data and found that representations that had reduced the gap between modalities also led to improved downstream model performance. Therefore,

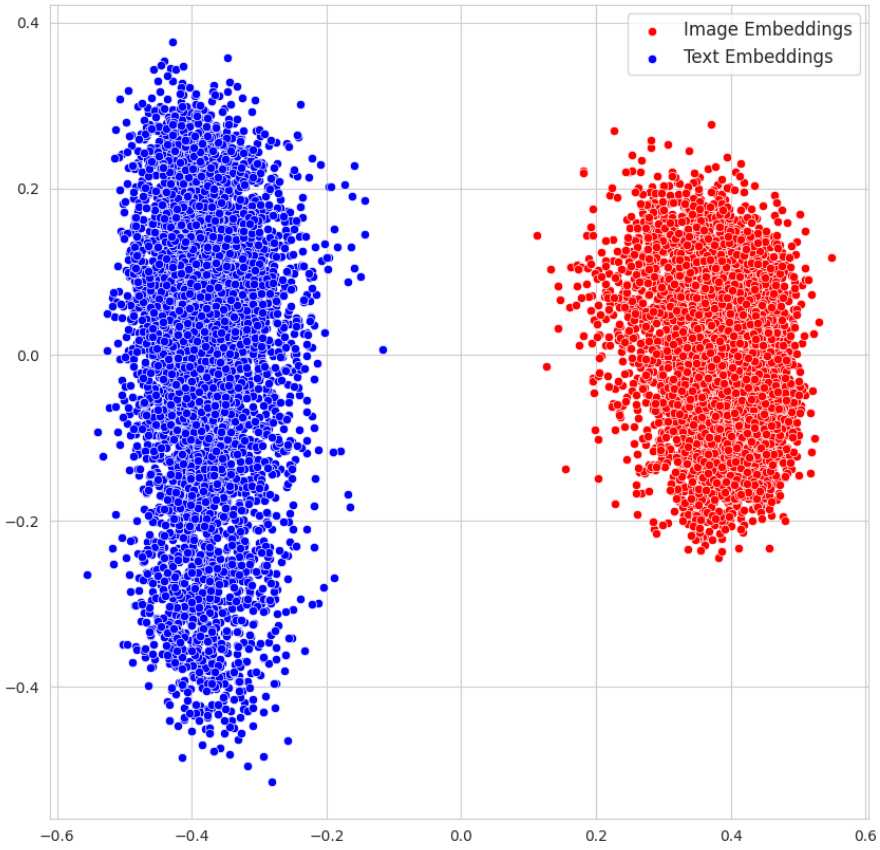


Figure 1.1: Illustrating the modality gap phenomenon. Paired inputs (images and their corresponding captions) are embedded into the CLIP space using an image and a text encoder. The resulting embeddings are visualized in 3D using PCA dimensionality reduction. Red points indicate image embeddings, and blue points represent text embeddings. We observe a distinct gap between the representational spaces of the two modalites. This is the modality gap.

analyzing and closing this gap is a promising direction to improve upon the strong representational capacity of CLIP and its variants.

In this thesis, we aim to gain a deeper understanding of the modality gap, challenge some of the existing assumptions about its emergence, and leverage our insights to propose simple methods for closing the gap. Our study is guided by the following **research questions**:

- What causes the modality gap in multi-modal contrastive learning?
- Can we close the gap? And,
- (If we can,) How does closing the gap affect downstream performance?

1.1 Contributions

In response to the above research questions, this thesis makes the following contributions:

- **The gap between embeddings is *not* caused by different modalities or data distributions.** After summarizing the common purported causes of the modality gap, we perform comprehensive experiments that show that accounting for these factors does *not* close the gap, suggesting that the present understanding of modality gap may be flawed.
- **The gap between embeddings results from the contrastive loss itself.** We present experiments that demonstrate that the gap is a byproduct of a high dimensional CLIP space, combined with a contrastive loss that encourages CLIP embeddings to occupy a lower dimensional manifold relative to the latent space. As a result, we propose renaming the "modality gap" as the *contrastive gap*.
- **The contrastive gap can be closed.** We first show that making the distribution of the features in the CLIP space more uniform closes the modality gap. The simplest way to do this is to increase the batch sizes used in training. However, we would need unreasonably large batch

sizes to close the gap in CLIP's original high-dimensional setting (512D). Instead, we show that simply fine-tuning CLIP by adding a factor for uniformity and alignment can reduce the size of the gap by distributing the embeddings more uniformly throughout CLIP's latent space, while allowing for more reasonable batch sizes during model training.

- **Closing the contrastive gap improves downstream performance.** Finally, we present experiments to show that closing the contrastive gap, and thereby creating more aligned and uniformly distributed representations, creates a representational space that is better for most downstream tasks, including zero-shot image classification and multi-modal embedding arithmetic.

1.2 Thesis Outline

In the next thesis chapter (Chapter 2), we give a detailed background on the prerequisites for this work. We explain the contrastive loss that forms the backbone of training CLIP, and the CLIP model architecture. We then introduce the concepts of the modality gap, which we rename the *contrastive gap*, and review the prior work in this area. Finally, we lay out metrics to quantify the size of the contrastive gap, which we use throughout the thesis.

We start Chapter 3 by demonstrating inconsistencies in the factors previously attributed to be the cause of the modality gap. We present a proof-of-concept experiment, showing that this gap in the latent space still persists in 512-dimensional CLIP even when all these supposed causal factors are accounted for and systematically removed. We then visualize the formation of the contrastive gap in 3D CLIP space. We demonstrate that the contrastive gap in CLIP space arises because embeddings of the same modality tend to cluster closely together, forming distinct lower-dimensional groups for each modality. We then reason that a new loss that encourages the uniform distribution of the embeddings in CLIP space could close the contrastive gap.

In Chapter 4, we introduce the concepts of uniformity and alignment from the uni-modal contrastive loss literature, and adapt it to the multi-modal case.

We reason that the contrastive gap happens because of low uniformity of the embeddings in the CLIP space, and show experiments supporting this idea. Finally, we introduce additional terms to the CLIP loss to explicitly optimize for uniformity and alignment.

In Chapter 5, we present our experimental results, comparing fine-tuning performance of the previously introduced losses on various evaluation tasks, including image-text retrieval, zero-shot image classification, and multi-modal arithmetic.

Finally, in Chapter 6, we summarize the work in this thesis with a conclusion, and suggest avenues for future work.

Chapter 2

Background

2.1 Chapter Overview

In this chapter, we start by introducing the contrastive learning objective, and how it extends to the multi-modal case. We then introduce CLIP, giving an overview of its architecture and loss function. Finally, we explain the modality gap phenomenon that is prevalent in CLIP, and summarize the works in the literature related to the modality gap.

2.2 Contrastive Learning

Contrastive learning is a type of self-supervised learning, where the model learns to extract information from the underlying data distribution itself. This is in contrast to fully-supervised learning, where hand-labelled data is provided to the learning algorithm.

Contrastive methods learn representations by maximizing the similarity between an input data point and a transformed (noisy) version of itself, while minimizing similarities with the other data points. This approach ensures that the learned representations are more invariant to noise, yet retain enough information to distinguish important features. The transformed versions of a data point form a *positive pair*, whereas semantically different data points form *negative pairs*.

One popular model that exemplifies this approach is SimCLR (Chen *et al.* [6]), which simplifies the contrastive learning framework compared to prior

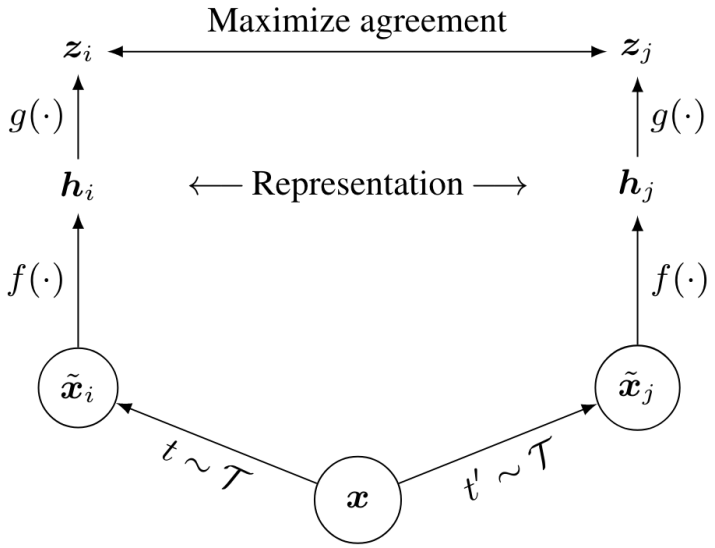


Figure 2.1: Contrastive learning framework used in SimCLR (Chen *et al.* [6]) . Two separate data augmentation operators are applied to each data sample to obtain two correlated views of the same sample. SimCLR trains a base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ to maximize similarity between the two augmented views of the same sample using a contrastive loss.

work (see Giakoumoglou and Stathaki [10]) by processing two augmented views of an image through identical networks. SimCLR employs the NT-Xent (Normalized Temperature-scaled Cross Entropy) loss to efficiently learn robust image representations. Unlike previous contrastive methods, such as those using the InfoNCE loss (Balestriero *et al.* [4]), NT-Xent relies on cosine similarity as a distance measure and normalizes the embeddings to lie on a unit hypersphere in the representational space. Figure 2.1 illustrates contrastive learning in the context of SimCLR.

2.3 Multi-modal Contrastive Learning

We now introduce the idea of multi-modal contrastive learning. In the unimodal contrastive method outlined in the previous section, contrastive learning maximizes the similarity between noise-augmented versions of the same datapoint (positive pairs), while pushing apart other datapoints (negative pairs). Thus, a contrastive learning algorithm learns a representational space where

positive pairs are close together in the latent space, and negative pairs are far apart.

Using this terminology, we can now extend the contrastive learning concept to the multi-modal case. For images and image-captions, instead of data augmentation to generate positive pairs, we simply define an image and its corresponding caption to be a positive pair. Subsequently, any non-matching image-caption pair would be a negative pair. In this way, the contrastive loss ensures that an image embedding and its corresponding caption embedding would be close together in the latent space.

2.4 The CLIP Approach

We now introduce the CLIP (Contrastive Image-Language Pre-training) model ([28]). CLIP takes advantage of the vast source of supervision from the web and learns from textual descriptions of images. Radford *et al.* [28] show that training a model on a simple task of matching captions to images is an effective way to learn robust image representations that can scale up to very large datasets. The authors were able to train CLIP on a dataset of 400M image-text pairs collected from the internet. Figure 2.2 illustrates the CLIP training procedure.

2.4.1 Contrasting Images and Text using CLIP Loss

The CLIP loss (L_{CLIP}) is based on NT-Xent loss. While NT-Xent loss is designed to work on data points from a single modality, $L_{\text{CLIP}}$is adapted to work on two different modalities of data.

In our scenario, the multi-modal dataset contains N images and corresponding captions. We obtain image embeddings $E_j^I \in \mathbb{R}^d$ by passing image I_j through the image encoder. Similarly, we produce the text embedding $E_j^T \in \mathbb{R}^d$ by passing caption T_j through the text encoder. CLIP aims to bring image embeddings and their corresponding caption embeddings closer together in the CLIP latent space (*CLIP space*) by increasing the similarity (inner product $\langle ., . \rangle$) between the corresponding embeddings. The image and text embeddings

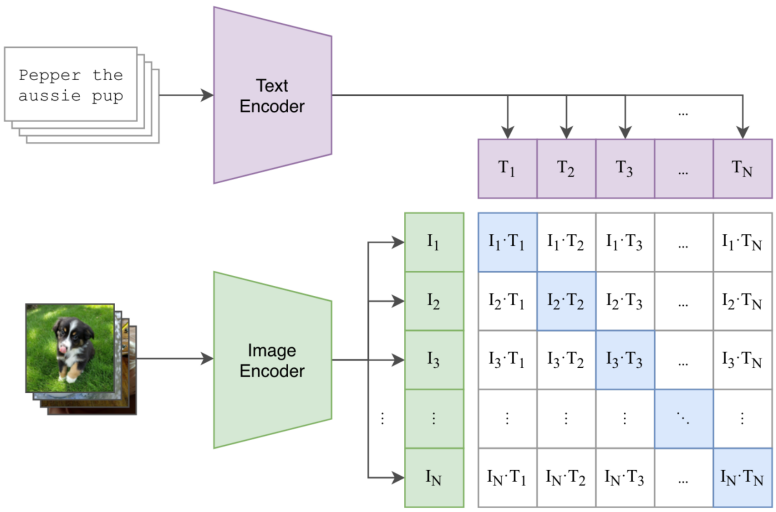


Figure 2.2: The CLIP training approach. Taken from Radford *et al.* [28]. CLIP jointly trains an image encoder and a text encoder to correctly match text caption embeddings to image embeddings (Ex: matching p^{th} caption embedding (E_p^T) to the p^{th} image embedding (E_p^I), while pushing apart non-matching image-text pairs (Ex: E_j^T and E_k^I, where $j \neq k$).

are normalized to lie on a unit hypersphere in $\mathbb{R}^d$.

The full CLIP loss is:

$$\begin{aligned} L_{\text{CLIP}} = &- \frac{1}{2N} \sum_{j=1}^{N} \log \left[\frac{\exp\left(\langle E_j^I, E_j^T \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle E_j^I, E_k^T \rangle / \tau\right)} \right] \\ &- \frac{1}{2N} \sum_{k=1}^{N} \log \left[\frac{\exp\left(\langle E_k^I, E_k^T \rangle / \tau\right)}{\sum_{j=1}^{N} \exp\left(\langle E_j^I, E_k^T \rangle / \tau\right)} \right] \end{aligned} \tag{2.1}$$

where the first term (top) **contrasts images with the texts** ($\sum_{k=1}^{N}$ in the denominator loops over text embeddings as negatives for the jth image) and the second term (bottom) **contrasts texts to images** ($\sum_{j=1}^{N}$ in the denominator loops over image embeddings as negatives for the kth text). τ represents the temperature parameter ($\tau = 0.01$ at the end of CLIP pre-training).

We need both the terms in the CLIP loss because, in the first term, the model contrasts each image embedding against all text embeddings, treating each text embedding as a potential *negative* (i.e., incorrect match) for that image. Conversely, the second term does the opposite by contrasting each text embedding against all image embeddings, treating each image embedding

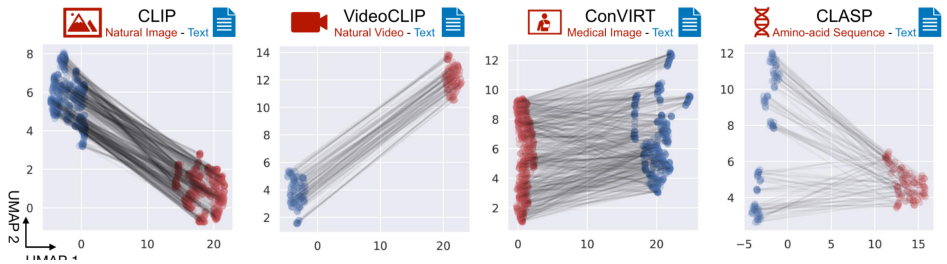


Figure 2.3: Illustration of the modality gap. Points are 2D UMAP visualizations of the CLIP embeddings. The lines indicate pairings of data points between the modalities. Data points from each modality lie in distinct clusters in the latent space. Figure taken from [20].

as a potential negative for that text. Since the two terms involve different sets of negatives, the contributions of these terms to the loss function are inherently different.

2.5 Related Work

Though CLIP effectively associates images with related texts, it also creates representational spaces with a *modality gap*, a phenomenon that has generated some interest. This was first identified by Liang *et al.* [20], and has consequently spawned a series of works attempting to understand and resolve the issue. We illustrate the issue of the modality gap in Figure 2.3.

The modality gap has been attributed to several factors. The first of these is the **the cone effect**, described by Liang *et al.* [20]. This effect occurs when the embeddings generated by the two separate encoders end up in two distinct, non-overlapping regions of the representational space, resembling narrow cones. These cones form because the encoders, starting from different random initializations, each learn to represent their respective modalities in a specific, confined area of the space, leading to a separation—or gap—between the two modalities.

Second, Liang *et al.* [20] note the existence of **mismatched pairs** in the dataset, where the pairing of images and captions is incorrect for some data points. Welle [35] studied the gap in three dimensions and attributed it to conflicting uniformity and alignment terms in the contrastive loss, claiming

that this leads to the existence of local minima that encourage the gap.

Uniformity and Alignment (defined by Wang and Isola [33]) are desirable properties in the multi-modal space for obtaining the optimal loss value (Udandarao [31]). Udandarao [31] further established that training with higher temperature values can enhance uniformity and alignment in CLIP. However, their findings also indicate that fine-tuning at higher temperatures often degrades downstream task performance. In contrast, our work identifies low uniformity, exacerbated by small batch sizes in high-dimensional spaces, as a key driving factor for the modality gap. We demonstrate that optimizing for uniformity and alignment explicitly can improve downstream performance (as measured by image-classification and multi-modal arithmetic)

The role of temperature and the contrastive loss function in the modality gap has also been explored by Al-Jaff [13]. They found that different loss functions could adjust the modality gap, but decreasing the gap did not consistently improve downstream task performance. The authors suggest that the modality gap might be an incidental outcome of certain contrastive methods rather than a crucial factor in representation quality. Our work extends this by relating the modality gap to low uniformity and proposing that the temperature parameter alone does not fully explain the gap or the quality of the learned representations.

On the other hand, exact modality alignment (i.e, zero modality gap) may be sub-optimal (Jiang *et al.* [15]). Jiang *et al.* [15] argue that exact alignment implies that only the shared information between the modalities is preserved in the embedding, while the modality-specific information is discarded. Their approach involves learning orthogonal subspaces for within-modality and cross-modality contrastive learning. We argue that exact alignment is theoretically ideal as it is directly optimized by the contrastive loss, and as a result, we focus on understanding and addressing the contrastive gap itself.

Another approach to addressing the modality gap involves synthesizing hard negatives from in-batch negatives, as proposed by Oh *et al.* [26] (For a given image, its soft in-batch negatives are all the other images and captions within the sampled batch). They demonstrate that using hard negatives and

ensuring they lie on the unit sphere can close the modality gap and improve downstream performance. Our work complements theirs, showing that higher batch sizes also help to close the modality gap, suggesting that *increasing the number of soft in-batch negatives* is another way to close the modality gap. Ultimately, the required number of soft in-batch negatives rises to unreasonable numbers as CLIP dimensionality increases. As a result, we suggest a simpler effective alternative: optimizing for uniformity and alignment, which also closes the gap, while improving downstream performance and does not require large batch size.

The modality gap emerges because of using soft in-batch negatives (Oh *et al.* [26]). Their approach to close the gap involves synthesizing hard-negatives, (from the in-batch negative pairs) making sure they lie on the unit sphere. They show that fine-tuning multi-modal contrastive models using the generated hard-negatives closes the modality gap, and improves downstream performance.

2.6 Chapter Conclusion

In this chapter, we gave an overview of the relevant uni-modal and multi-modal contrastive learning literature, with an emphasis on CLIP. We explained the training mechanism of CLIP, detailing how it leverages multi-modal contrastive learning to map image and text embeddings into a shared representational space. We then introduced and discussed the modality gap phenomenon, which emerges in the representational space of CLIP and other multi-modal contrastive models. We explored various factors that have been proposed in the literature as contributing to the modality gap, including the cone effect, mismatched pairs, and the conflicting uniformity and alignment properties in contrastive learning, leading to the existence of local minima in the CLIP loss landscape.

Furthermore, we reviewed related work that have attempted to adjust the modality gap through different approaches, such as adjusting the contrastive loss function, fine-tuning with hard negatives, and adjusting the temperature

hyperparameter. These studies highlight the complexity of the modality gap and the ongoing debate about its impact on downstream task performance.

In conclusion, we posited that the modality gap may not simply be a result of aligning different modalities in a shared representational space, but rather a consequence of the inherent properties of the contrastive algorithm itself. To capture this underlying phenomenon more accurately, we suggest that the term "contrastive gap" may be more appropriate. In the next chapter (Chapter 3, we present proof-of-concept experiments to show that the factors currently attributed to the emergence of the gap do not sufficiently explain it. We will also explore simple scenarios where the gap can be closed, offering insights into potential strategies for addressing it.

Chapter 3

Characterizing the Gap

3.1 Chapter Overview

In the previous chapter, we introduced the phenomenon known as the *modality gap*, which we rename the *contrastive gap*. In this chapter, we first outline two measures to quantify the size of the contrastive gap. We then dive deeper into this phenomenon and examine the factors that are thought to cause it. We then show that the modality gap exists even after we systematically remove all these factors. We demonstrate that CLIP in 3D behaves differently from high-dimensional CLIP, and formulate a hypothesis about the true cause of the modality gap by relating the gap to the high-dimensional latent space that the CLIP embeddings reside in.

3.2 Measuring the gap

To show that we have closed the gap between the embeddings generated by the two encoders, we must first find a way to quantify the modality gap. Each encoder projects its inputs into a shared representational space, and the modality gap is the measure of separation between the embeddings of the two modalities in this shared representational space. We introduce the following two metrics to measure the size and severity of the gap:

Distance between modality centroids (from Liang *et al.* [20]) Given N images and N captions, we denote the centroid of the image embeddings as $C^I = 1/N \sum_{j=1}^{N} E_j^I$, (where E represents an embedding) and similarly for the

centroid of the text embeddings. We compute the distance between centroids as $\|C^I - C^T\|^2$, and note that the centroid distance can vary from 0 to 2.

Linear Separability (from Welle [35]) is the percentage of embeddings from the two encoders that can be distinguished by a linear classifier operating in CLIP space. We used 80% of the dataset to train a linear model to classify CLIP embeddings as originating from either encoder 1 or encoder 2. We then tested the performance of the classifier on the remaining 20% of the dataset and reported the accuracy. If a set of embeddings are 100% linearly separable, this means that the spaces occupied by embeddings from each encoder are completely disjoint. Conversely, 50% linear separability means that the embeddings of the two encoders are overlapping in CLIP space, meaning that they occupy the same region of the latent space; i.e. there is no gap between the embeddings.

To summarize, *if we can effectively close the gap we will find that the distance between centroids is small and linear separability is close to 50%.*

3.3 The Modality Gap Persists Even When All Factors Are Accounted For

We ran several experiments to systematically remove the factors commonly known to contribute to the modality gap (the cone effect of the encoder networks, mismatched pairs in the dataset, and CLIP loss getting stuck in local minima). We created an idealized scenario where:

1. *There is only one modality.* We replaced the text encoder in CLIP with another copy of the image encoder and trained the model on pairs of images instead of text-image pairs. Thus, for this experiment the CLIP encoders are two identical image encoders with different random initializations.

2. *Embeddings from the two image encoders occupy the same cone at initialization.* After we initialize the two image encoders, we computed a fixed transformation matrix that translates the embeddings of the second

image encoder to overlap with those of the first image encoder (following [20]). Thus, the second encoder's embeddings are translated to occupy the same narrow cone as the first encoder's embeddings at initialization. This way, there is no modality gap at initialization.

3. *There are no mismatched pairs.* The positive pairs in our constructed dataset are actually identical images. This eliminated the possibility that there would be mismatched pairs in the dataset.

We ran this experiment on the full MS-COCO training dataset. We used a batch size of 64, and the default CLIP dimensionality of 512D (i.e. Image embeddings, $E^I \in \mathbb{R}^{512}$). We tested modality gap metrics (distance between centroids of the embeddings from the two encoders, and linear separability accuracy, from Section 3.2) and contrastive loss on the validation split. We report the size of the modality gap and show 95% confidence interval across three random initializations of encoder parameters in Table 3.1

	At initialization	After 15 epochs
Centroid distance	0.01 ± 0.01	0.40 ± 0.08
Linear separability acc.	0.53 ± 0.02	1.00 ± 0.00
Contrastive loss	3.42 ± 0.22	0.00 ± 0.01

Table 3.1: **Modality gap persists even when all factors are accounted for:** Modality gap metrics and CLIP loss values before and after training CLIP from scratch in ideal dataset conditions. **At initialization:** Distance between the centroids of the embeddings from the two image encoders is close to zero due to our transformation at initialization, and the embeddings are *not* linearly separable, meaning that there is no modality gap. **After training:** Centroid distance increases slightly, but the embeddings from the two image encoders are *perfectly* linearly separable. Thus, the modality gap is *created* by the contrastive loss.

Table 3.1 shows that even when the dataset is idealized and almost trivial to optimize on, and the model is initialized without a gap, after training to near zero loss, the CLIP embeddings are still perfectly linearly separable. Thus, there is a contrastive gap that is *created* as a byproduct of the contrastive loss, even when the two encoders are trained on the same modality and even on the same inputs.

We visualize the emergence of the contrastive gap in this idealized scenario by:

Plotting the centroid distance over training steps Figure 3.1 shows that centroid distance start at zero, and then abruptly increase and plateau, indicating that the gap between the "modalities" is created by the contrastive loss during training.

Visualizing the latent space in 3D using UMAP In Figure 3.2, we plot 1000 randomly selected data pairs during training to show the emergence of the contrastive gap during training in this idealized scenario.

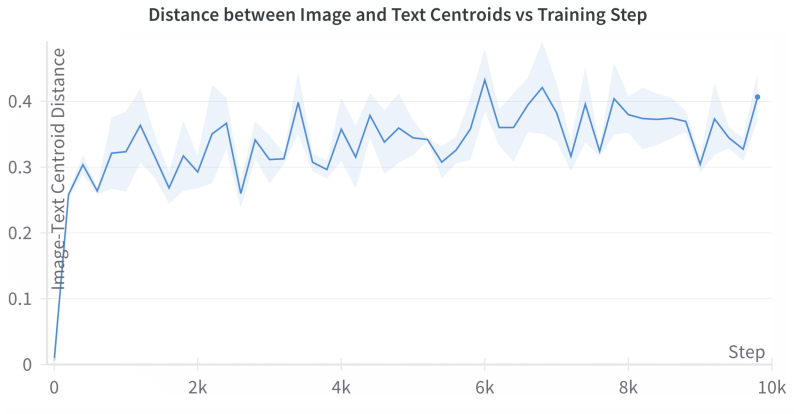
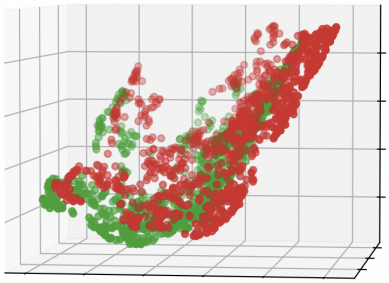


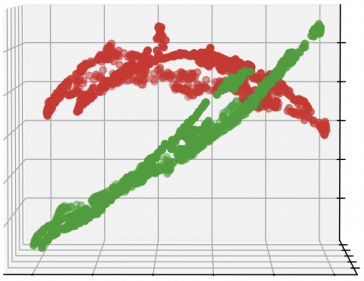
Figure 3.1: Distance between the two image centroids in idealized training scenario. Distance begins at 0 because we artificially close the gap before training begins. Distance increases and plateaus, showing that the contrastive gap emerges during training. Plot shows average over 3 seeds (Shaded region indicates the standard error).

3.4 Visualizing the Gap in 3D CLIP

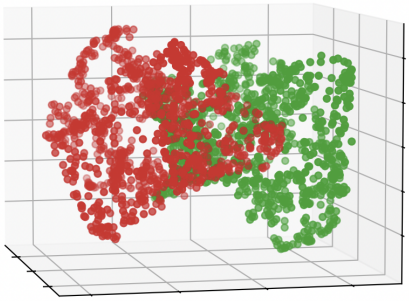
In the previous experiment, we demonstrated that the modality gap emerged during training when CLIP was trained in a 512-dimensional space. However, Welle [35] showed that the CLIP loss function could close the gap between embeddings by training on synthetically generated 3-dimensional points in Euclidean space, without relying on a neural network. In their experiment, they generated two sets of points scattered across a 3-dimensional Euclidean



(a) **At Initialization:** We remove the gap at initialization by translating the embeddings of encoder 2 to overlap with the embeddings of encoder 1.



(b) **After 200 steps:** CLIP's contrastive loss *creates* a gap between the image embeddings from the two encoders relatively early during training. This occurs even when there are no mismatched pairs, and when both input "modalities" are images.



(c) **After 25k steps:** As CLIP's loss approaches zero, the gap between the embeddings remains. This demonstrates that the factors previously thought to be responsible for the 'modality' gap actually arise due to the contrastive loss.

Figure 3.2: Visualizing changes in CLIP representational space when training in an idealized scenario, starting with no contrastive gap. We plot the embeddings in 3D using UMAP visualization. The two different colors (red and green) represent embeddings from different encoders.

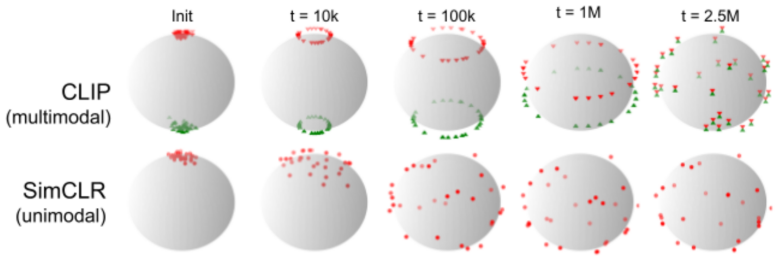


Figure 3.3: Proof-of-concept experiment taken from Wang and Liu [32] illustrating that CLIP loss induces a modality gap. The authors randomly generated two sets of points representing two modalities on a 3D sphere. Red and green colors represent points from each point set. t is the training epoch. **Top**: Results of running an optimizer on the euclidean positions of the two sets of points to reduce the CLIP loss. **Bottom**: Results of running an optimizer to reduce SimCLR (uni-modal) loss on just one set of points.

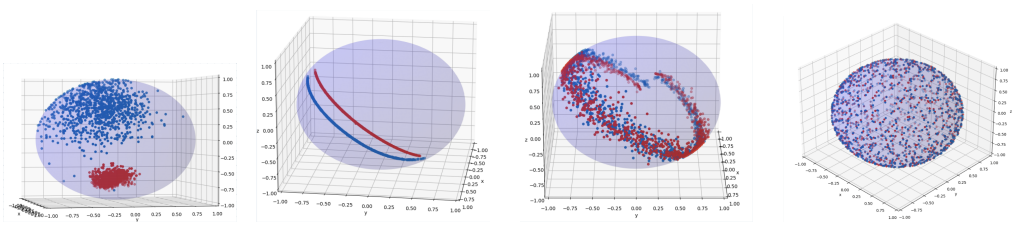
.

space, with each set representing a different modality. By directly optimizing on the positions of these points within the Euclidean coordinate space, they were able to investigate how the modality gap could be influenced purely by the contrastive loss function itself (i.e, the parameter space for optimization is the euclidean coordinate space of the points, instead of the weights of a neural network). They randomly generated 1,000 points on a 3D sphere and trained with a batch size of 10 and a learning rate of 0.01. We show the results of Welle [35]'s experiment in Figure 3.3. This experiment highlights that the modality gap can arise even in low-dimensional spaces and without any external factors from the dataset or neural network architecture, but can eventually be closed after a large number of training steps.

Following the work of Welle [35], we studied the behaviour of CLIP loss when we reduce the CLIP dimensionality from 512D to 3D. We extended the proof-of-concept experiment done by Welle [35] with the following changes:

- **Using Real Data**: We sampled 1,000 image-text pairs from the MS-COCO dataset rather than using synthetic points. This way, we can better capture the nuances of a real-world data distribution.

- **Optimizing with CLIP architecture**: In the original experiment by Welle [35], the authors bypassed the neural network entirely by directly



(a) **Epoch 0** $I \to T$ acc.: 0.0 (b) **Epoch 37** $I \to T$ acc.: 0.0 (c) **Epoch 150** $I \to T$ acc.: 0.1 (d) **Epoch 275** $I \to T$ acc.: 0.87

Figure 3.4: Visualizing the training stages of 3D CLIP on 1000 image-text pairs from MS COCO. Red points are image embeddings, and blue points are text embeddings. $I \to T$ accuracy indicates how accurately the model can retrieve the correct text given an image. Higher $I \to T$ accuracies mean that the model is better at distinguishing between correct (positive) image-text pairs and incorrect (negative) pairs in the latent space. **(a)**: The embeddings of each modality initially reside in separate regions due to the cone effect. **(b), (c)**: As training progresses, these embeddings begin to form arcs and gradually merge into rings. **(d)**: Finally, the embeddings from both modalities spread out to fill the entire sphere.

optimizing the positions of points in a 3D Euclidean space. However, in our modified experiment, we utilized the CLIP model to project image and text representations into a 3D space (by adjusting the output projection layer from 512D to 3D). By doing so, we could examine how the CLIP model and its loss function behave together, allowing us to assess the impact of both the neural network and the loss function in a more realistic setting.

In Figure 3.4 we see that after training on CLIP loss for 275 epochs, the model is able to create a representational space that closes the contrastive gap in 3D even on real-world data. This suggests that closing the contrastive gap may be easier in lower dimensional vs. in high-dimensional spaces.

We measure the representational space quality using the text-retrieval accuracy metric. Text-retrieval accuracy measures how accurately a model can retrieve the correct textual description (caption) for a given image based on their embeddings. A higher accuracy indicates that images and their corresponding caption embeddings are closely aligned in the representational space, suggesting that the model has effectively captured the semantic relationships

between the two modalities.

We report the text-retrieval accuracies in Figure 3.4 as the accuracy with which we can retrieve an input image's caption given the input image's embedding and the embeddings of all the captions in the dataset. The results show that points on the 3D sphere are best aligned across modalities when they are evenly distributed on the sphere (as evidenced by the very high text-retrieval accuracy when Figure 3.4d). Therefore, we speculate that it is desirable close the contrastive gap *and* to distribute the embeddings more uniformly on the unit sphere in $\mathbb{R}^d$ to improve the quality of representations.

While this experiment suggests that it might be easier for CLIP to close the contrastive gap in lower dimensionalities, simply reducing CLIP dimensionality to close the contrastive gap may not be desirable. This is because lower-dimensional spaces inherently have fewer degrees of freedom to capture the complexities of the input data, resulting in the embeddings retaining lesser information. Therefore, we propose explicitly optimizing for more uniformly distributed embeddings in high-dimensional CLIP space. In Chapter 5, we will see that, when the embeddings are more uniformly distributed in the latent space, the contrastive gap decreases, and in turn, the quality of the learned representations increases, as measured by performance in multiple downstream tasks

3.5 Chapter Conclusion

In this chapter, we explored the possibility of closing the contrastive gap in multi-modal models like CLIP. We began with a proof-of-concept experiment that demonstrated how common factors such as the cone effect, mismatched pairs, local minima, and varying data distributions fail to fully explain the emergence of the modality gap. We then presented experiments using a 3D version of CLIP, where we showed that the contrastive gap *can* be closed. We observed that the gap is closed when the embeddings are uniformly distributed on the three-dimensional unit sphere.

We also demonstrated that a uniformly distributed representational space

not only closes the gap but also leads to higher text-retrieval accuracies, suggesting that a uniformly distributed embedding space without a contrastive gap allows the model to better capture the semantic relationships between the two modalities of data.

In the next chapter (Chapter 4), we will introduce the concepts of uniformity and alignment in the representational space. We will show that the contrastive gap arises primarily due to low uniformity in the CLIP space and propose that optimizing for these properties can be a key strategy in closing the contrastive gap, ultimately leading to improved model performance.

Chapter 4

Methodology (Closing the Gap)

4.1 Chapter Overview

In the previous chapter, we observed that in 3D space, the contrastive gap can be minimized if embeddings are uniformly distributed across the unit hypersphere, rather than being confined to nearly two-dimensional "rings." Building on this insight, we propose that a similar approach can be applied to high-dimensional CLIP models by learning representational spaces where image and text embeddings are more uniformly distributed. To achieve this, we introduce the concepts of *uniformity* and *alignment* in the contrastive space. We adapt these properties, originally derived from uni-modal contrastive literature, to the multi-modal setting and incorporate them into CLIP's loss function. By incorporating uniformity and alignment into the CLIP loss, we aim to assess the impact of these properties on closing the contrastive gap, with the intent to improve downstream task performance. Finally, we present experiments which show that the contrastive gap in CLIP arises from low uniformity in the embeddings, which is exacerbated by training with small batch sizes in a high-dimensional latent space.

4.2 Uniformity and Alignment

$$
\begin{aligned}
L_{\text{CLIP}} = &- \frac{1}{2N} \sum_{j=1}^{N} \log \left[\frac{\exp\left(\langle E_j^I, E_j^T \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle E_j^I, E_k^T \rangle / \tau\right)} \right] \\
&- \frac{1}{2N} \sum_{k=1}^{N} \log \left[\frac{\exp\left(\langle E_k^I, E_k^T \rangle / \tau\right)}{\sum_{j=1}^{N} \exp\left(\langle E_j^I, E_k^T \rangle / \tau\right)} \right]
\end{aligned}
\tag{4.1}
$$

We now explain the concepts of uniformity and alignment in the representational space. Recall that the contrastive loss (Repeated in Equation 4.1) learns a representational space where positive pairs are close to each other, and negative pairs are far from each other. Following from this intuition, Wang and Isola [33] suggest that the contrastive representational space should have the following two properties:

- *Alignment*: For two samples from any positive pair, the samples should be mapped to nearby embeddings (aligned) in the latent space, and therefore be invariant to unnecessary noise features in the input.
- *Uniformity*: The embeddings should be projected such that they are roughly uniformly distributed on the unit hypersphere in $\mathbb{R}^d$, preserving as much information in the data as possible.

Wang and Isola [33] introduced these properties in the context of uni-modal contrastive learning (unsupervised learning using image augmentations) and the NT-Xent loss (explained earlier in Section 2.2). The authors show that:

- Higher alignment and uniformity consistently and strongly correlate with higher downstream performance.
- Directly optimizing for alignment and uniformity (without the use of a contrastive loss) can lead to better representations than those optimized using the contrastive loss.
- Both uniformity *and* alignment are necessary for high-quality representations.

4.3 Uniformity and Alignment in Multi-modal Contrastive Learning

We now adapt the uni-modal uniformity and alignment properties from the work of Wang and Isola [33] to the multi-modal contrastive space. We define these properties as *losses* in the representational space, so that *lower* values of uniformity and alignment losses mean that the space has *higher* uniformity and alignment.

Multi-modal Uniformity We extend the uniformity property to the multi-modal setting by distinguishing between two types of uniformity:

- In-modal uniformity: This measures how evenly distributed the embeddings are *within the same modality*. For example, it assesses how well image embeddings are spread out relative to each other.
- Cross-modal uniformity: This measures how evenly distributed the embeddings are *across different modalities*. For instance, it evaluates how well image embeddings are distributed relative to text embeddings.

As an example, consider a representational space where image embeddings are well spread out relative to text embeddings (high cross-modal uniformity) but are clustered closely together among themselves (low in-modal uniformity). This means that while the image embeddings are far from most text embeddings, they are relatively close to other image embeddings. Udandarao [31] showed that the CLIP representational space exhibits this pattern, with high cross-modal uniformity but low in-model uniformity. In other words, while image embeddings are well separated from text embeddings, they tend to cluster too closely together among themselves. Udandarao [31] showed that this arrangement is problematic because it leads to a poorly calibrated image-only subspace within CLIP's representational space, making it unreliable for measuring image-image similarity.

We now define the uniformity losses as follows:

$$\textbf{Uniformity (Images)} : L^{I}_{\text{Uniform}} = \log\left(\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\exp\left(-2\|E^{I}_{j} - E^{I}_{k}\|^{2}\right)\right) \tag{4.2}$$

$$\textbf{Uniformity (Texts)} : L^{T}_{\text{Uniform}} = \log\left(\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\exp\left(-2\|E^{T}_{j} - E^{T}_{k}\|^{2}\right)\right) \tag{4.3}$$

Finally, the total L_{Uniform} (In-modal) term is:

$$L_{\text{Uniform (In-modal)}} = \frac{1}{2}(L^{T}_{\text{Uniform}} + L^{I}_{\text{Uniform}}) \tag{4.4}$$

L^{I}_{Uniform} and L^{T}_{Uniform} each encourage the uniformity *within* the image and text embeddings respectively. i.e., L_{Uniform} only encourages *in-modal* uniformity. The original multi-modal contrastive loss (Equation 4.1) does *not* have any such term that constrains embeddings within each modality to be far apart. Instead, the denominators in Equation 4.1 only push negative text samples away from positive image sample and vice versa.

To enforce a stronger constraint on the uniformity *between* negative image and text samples, we also introduce a *cross-modal uniformity* term:

$$L_{\text{XUniform}} = \log\left(\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1,k\neq j}^{N}\exp\left(-2\|E^{I}_{j} - E^{T}_{k}\|^{2}\right)\right) \tag{4.5}$$

Multi-modal Alignment We also incorporate an alignment term to encourage positive image-text samples to be close together in the latent space. We adapt the alignment term from the work of Wang and Isola [33] to the multi-modal setting as follows:

$$L_{\text{Align}} = \frac{1}{N}\sum_{j=1}^{N}(\|E^{I}_{j} - E^{T}_{j}\|^{2}) \tag{4.6}$$

4.4 Optimizing Alignment and Uniformity of CLIP Representations

Our 3D experiments in Section 3.4 demonstrated that the contrastive gap in multi-modal CLIP is closed when embeddings are uniformly distributed on the three-dimensional hypersphere. Additionally, we discussed in Section 4.2 that uniformity and alignment are desirable properties in the uni-modal contrastive space.

Drawing from these insights, we hypothesize that optimizing for multi-modal uniformity and alignment in the CLIP representational space could effectively close the contrastive gap. To test this, we introduce modified CLIP losses that explicitly optimize for uniformity and alignment in the CLIP space. Specifically, in our experiments, we fine-tune a pre-trained CLIP model by incorporating L_{Uniform}, L_{XUniform}, and L_{Align} terms into the original CLIP loss (L_{CLIP}, Equation 4.1). By fine-tuning CLIP using these new losses and studying their performance, we can validate the desirability of the uniformity and alignment properties in the multi-modal setting. In Chapter 5, we demonstrate the effects of fine-tuning pre-trained CLIP on the following losses:

- L_{CLIP}: The default CLIP loss
- $L_{\text{CLIP+U+A}}$ (L_{CUA}): $L_{\text{CLIP}} + L_{\text{Uniform}} + L_{\text{Align}}$
- $L_{\text{CLIP+U+A+XU}}$ (L_{CUAXU}): $L_{\text{CLIP}} + L_{\text{Uniform}} + L_{\text{Align}} + L_{\text{XUniform}}$

The study of uniformity and alignment in multi-modal contrastive learning is not an original concept. Previous works, such as those by Goel *et al.* [11] and Oh *et al.* [26], have suggested changes in the CLIP loss to improve geometric properties of the CLIP space. These authors used L_{XUniform} to assess uniformity in the latent space (as opposed to optimizing for L_{XUniform}). Additionally, Al-Jaff [13] explored training multi-modal models with L_{Uniform} (focusing solely on in-modal uniformity) and L_{Align}.

In our work, we build on these ideas by combining both in-modal and cross-modal uniformity terms to promote greater uniformity in the latent space,

with the goal of reducing the contrastive gap. Unlike prior works, we explicitly optimize for L_{Uniform}, L_{XUniform}, and L_{Align} by adding these terms to the original CLIP loss function.

4.5 Contrastive Gap is due to Low Uniformity

In order to validate our approach of trying to close the contrastive gap by optimizing for uniformity in CLIP space, we first test our hypothesis that the contrastive gap arises due to low uniformity in the CLIP latent space. We argue that the contrastive gap is a byproduct of CLIP embeddings lying on a lower dimensional manifold relative to the CLIP space.

As shown by Wang and Isola [33], optimizing the uni-modal contrastive loss is equivalent to optimizing for uniformity and alignment, in the limit of infinite batch size. We extend this reasoning to the multi-modal case, suggesting that the low uniformity in CLIP space arises due to the insufficient batch sizes used in training.

4.5.1 Increasing batch size increases uniformity for a fixed CLIP dimensionality

Figure 4.1 shows that, for a fixed CLIP dimensionality (35D CLIP), larger batch sizes lead to lower uniformity loss values (and thus higher uniformity in the CLIP space). This result helps extend the theory that uni-modal contrastive learning with very large batch sizes optimizes for uniformity to the multi-modal case.

4.5.2 Increasing batch sizes leads to a more rapid reduction of the contrastive gap

Figure 4.2 shows that, for a fixed CLIP dimensionality (35D CLIP), larger batch sizes lead to lower linear separability accuracies (and thus force the embeddings of different modalities closer together). Together with the result from 4.5.1, this provides evidence that optimizing for uniformity helps close the contrastive gap.

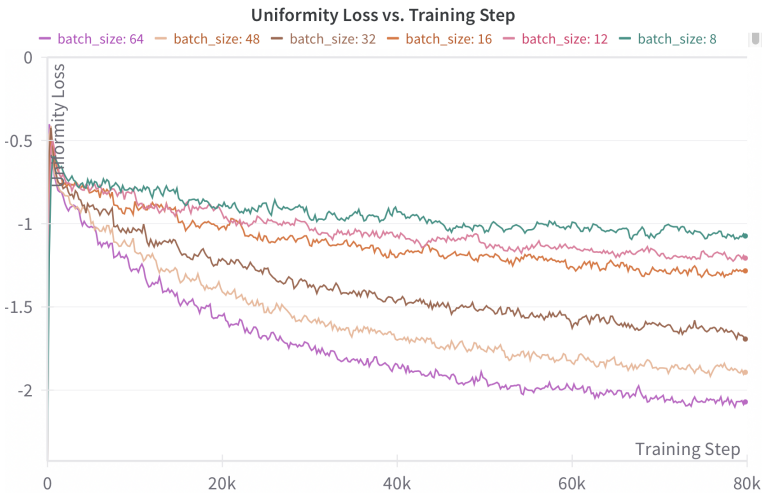


Figure 4.1: Uniformity loss vs Training step when fine-tuning CLIP on the MS-COCO dataset. Larger batch sizes lead to lower uniformity loss values, and thus higher uniformity in CLIP space.

4.5.3 Reducing CLIP dimensionality helps reduce the contrastive gap

Figure 4.3 shows that, for a fixed *batch size* (= 16), reducing the CLIP dimensionality helps close the contrastive gap. This result gives insight into why the contrastive gap is so prevalent in CLIP and its successor models today: Most of them are very high dimensional, and probably do not use sufficiently large batch sizes during training to close the contrastive gap.

4.6 Chapter Conclusion

In this chapter, introduced the concepts of uniformity and alignment in uni-modal contrastive learning and adapted these properties to the multi-modal setting. To quantify uniformity and alignment in the CLIP latent space, we defined three new loss terms: L_{Uniform}, L_{XUniform}, and L_{Align}. We then formulated new loss functions—L_{CLIP}, L_{CUA}, and L_{CUAXU}—which incorporate the uniformity and alignment losses into the original CLIP loss function.

Finally, we empirically showed that the contrastive gap arises from low uniformity in the CLIP space. We demonstrated that increasing batch sizes while keeping the CLIP dimensionality fixed improves uniformity in the CLIP

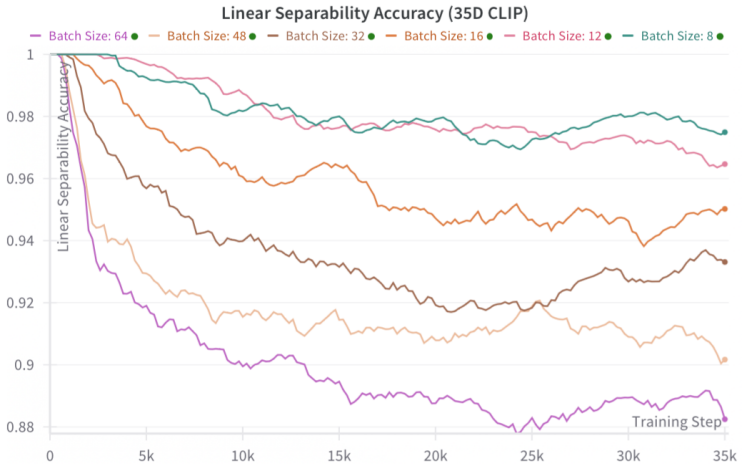


Figure 4.2: Linear Separability Accuracy vs Training step when fine-tuning CLIP on the MS-COCO dataset. Larger batch sizes lead to lower linear separability accuracies, and thus smaller contrastive gap size

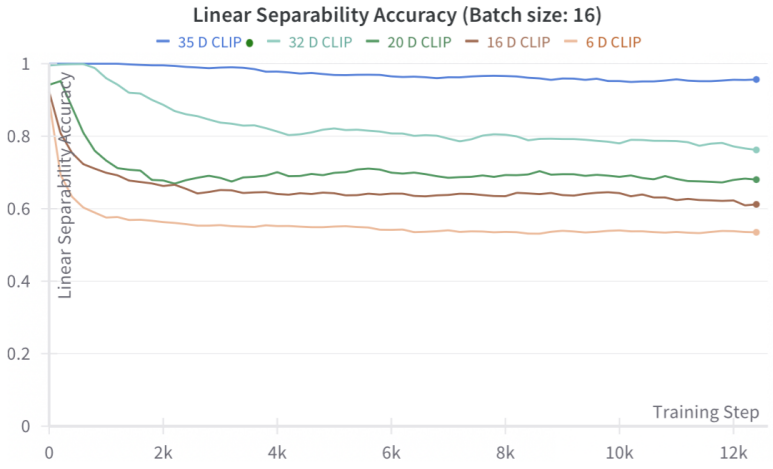


Figure 4.3: Linear Separability Accuracy vs Training step when fine-tuning CLIP on the MS-COCO dataset. Smaller CLIP dimensionalities lead to lower linear separability accuracies, and thus smaller contrastive gap size.

space, and helps reduce the contrastive gap faster. Based on these findings, we concluded that improving uniformity in the CLIP space is a promising approach to closing the contrastive gap.

In the next chapter (Chapter 5), we investigate how optimizing for the uniformity and alignment losses impacts the contrastive gap. We show that using these new losses helps close the gap effectively, even with reasonable batch sizes and in high-dimensional latent spaces. Additionally, we will demonstrate that minimizing the contrastive gap leads to improved performance on downstream tasks, such as zero-shot image classification and multi-modal arithmetic. This suggests that addressing the contrastive gap is a promising strategy for enhancing the performance of CLIP-like models.

Chapter 5

Experiments

5.1 Chapter Overview

In the Chapter 4, we introduced the concepts of uniformity and alignment within the context of multi-modal contrastive learning. We showed that increasing uniformity in CLIP space—achieved by using larger batch sizes—helps to minimize the contrastive gap. Building on this, we detailed how to modify the original CLIP loss function by incorporating terms that specifically promote uniformity and alignment in the CLIP space, with the goal of reducing the contrastive gap without the need to increase batch size during training.

In this chapter, we study the effects of fine-tuning CLIP on the new uniform and align losses across multiple CLIP dimensionalities. We begin by outlining the experimental setup for our study, including the hyperparameters and dataset used to fine-tune the CLIP models. We then compare the size of the contrastive gap after fine-tuning with the original vs. the new losses using the standard image-caption dataset, MS-COCO. Our results demonstrate that optimizing for uniformity and alignment during fine-tuning effectively reduces the contrastive gap.

We then evaluate the performance of the fine-tuned CLIP models across three standard downstream tasks: zero-shot image classification, image-text retrieval, and multi-modal arithmetic. Our findings reveal that reducing the contrastive gap leads to improved performance in zero-shot image classification and multi-modal arithmetic, but appears to hurt image-text retrieval performance slightly.

5.2 Experimental Setup

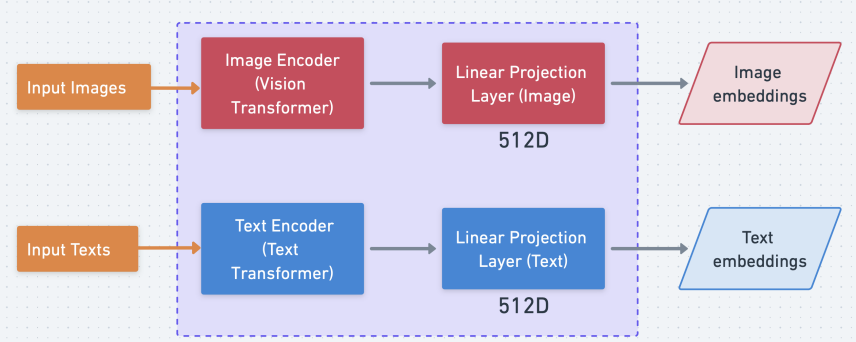
Hyperparameters Overview For our experiments, we fine-tuned the CLIP model [28] made available by OpenAI from HuggingFace [36][1]. We used the ViT-B/32 variant of the image encoder and a transformer (from OpenAI's official implementation[2]) as the text encoder. When fine-tuning, we adjusted the dimensionality of CLIP ($\mathbb{R}^d$, $d \in [32, 64, 128]$) by changing the size of the final linear projection layer of pre-trained CLIP (We illustrate this in Figure 5.1). For all our experiments, we fixed the temperature (τ) parameter to 0.01, as τ converges to this value after CLIP pre-training. We list all the hyperparameters from our setup in Table 5.1.

Hyperparameter	Value
Image encoder model	ViT/B-32
Text Encoder model	Transformer (same as in [3])
Embedding dimensions	[32, 64 ,128]
Temperature	0.01
Epochs	9
Batch size	64
Learning rate	1e-6
Adam beta1	0.9
Adam beta2	0.99
Adam weight decay	0.1
Scheduler	None

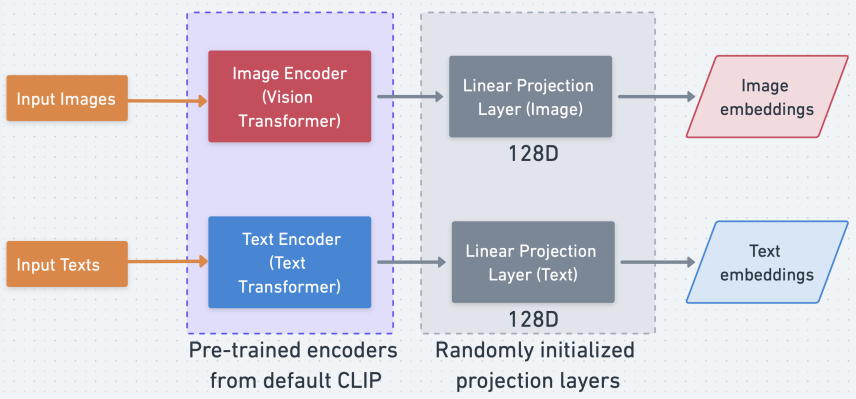
Table 5.1: Hyperparameters used for fine-tuning the CLIP models

Fine-tuning Dataset We fine-tune our CLIP models on MS-COCO [21][4], an image-caption dataset where each image has five corresponding human-generated captions. We use the 2017 split, with 118k training images, and 5k validation images. Throughout our experiments, we only use the first caption for each image and discard the remaining 4 captions per image.

[1] `https://huggingface.co/docs/transformers/en/model_doc/clip`
[2] https://github.com/openai/CLIP
[4] `https://cocodataset.org`



(a) Default CLIP



(b) CLIP with changed dimensionality. (128D in this figure)

Figure 5.1: We adjust the dimensionality of CLIP latent space by changing the size of the final projection layer. When fine-tuning, we keep the rest of the model backbone from pre-trained CLIP, while randomly initializing the projection layer.

Estimating Statistical Errors in our Results To estimate statistical errors in our results, we used three different random seeds for initializing our models. The numerical results and plots presented throughout this chapter represent the averages obtained from the runs of these three seeds. To convey the variability of our results, we include one standard error around the mean as error bars in our plots.

5.3 Effects of Optimizing for Uniformity and Alignment on the Contrastive Gap

Evaluating contrastive gap metrics For each of the losses, we measure the contrastive gap using two metrics: the distance between the image and

text centroids, and the linear separability accuracy of the image and text embeddings (These metrics are explained in Section 3.2). A lower value for these metrics indicates a smaller contrastive gap. Figure 5.2 shows these metrics for various CLIP dimensionalities and loss functions.

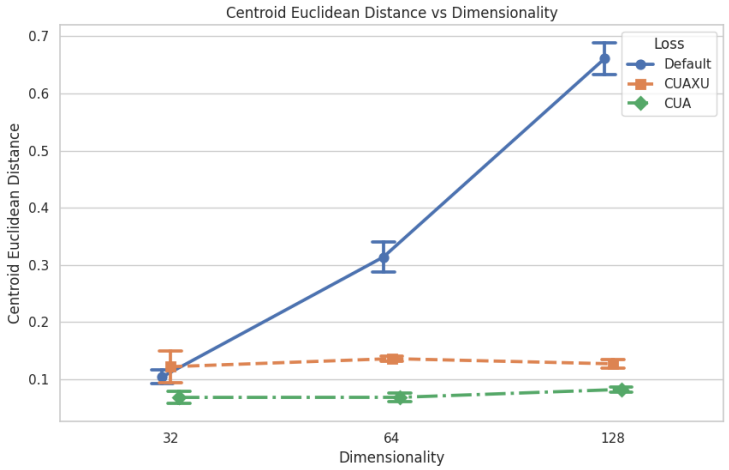
The plots in Figure 5.2 reveal that when fine-tuning with L_{CLIP}, the contrastive gap becomes more pronounced as dimensionality increases. In contrast, this effect is less significant when fine-tuning with L_{CUA} and L_{CUAXU}. There is a growing disparity in the contrastive gap sizes between L_{CLIP} and the other losses with increasing dimensionality. These findings are consistent with the earlier results presented in Section 4.5, supporting the idea that higher CLIP dimensionalities make it harder for L_{CLIP} to close the contrastive gap.

Overall, our results empirically support the claim that losses designed to encourage uniformity in the CLIP space effectively reduce the contrastive gap. This supports our claim that increasing uniformity in CLIP space reduces the size of the contrastive gap.

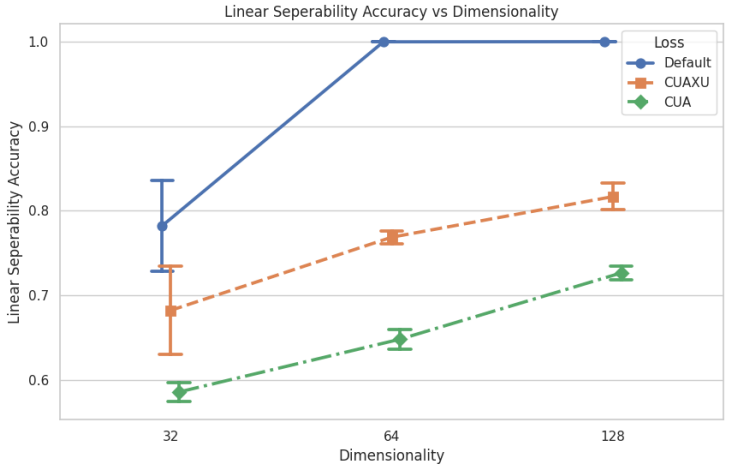
Evaluating the distribution of embeddings across latent space dimensions We evaluate how the image and text embeddings are distributed within CLIP space by applying Principal Component Analysis (PCA) and analyzing the explained variance ratios (Holland [12])

PCA is a dimensionality reduction technique that summarizes the data into a smaller set of principal components (PCs). The explained variance ratio indicates how much of the original data's variance is captured, or "explained", by each PC. By examining the cumulative PCA explained variance curve, we can assess how well the embeddings are spread across the dimensions of the unit hypersphere in CLIP space. Ideally, if the embeddings are uniformly distributed across all dimensions—indicating high uniformity—the cumulative PCA explained variance curve will form a straight line. As discussed in Section 4.5, higher uniformity is associated with a smaller contrastive gap.

Figure 5.3 shows the cumulative explained variance of CLIP spaces learned using various losses, and in different CLIP dimensionalites. Compared to L_{CLIP} and L_{CUA}, the cumulative variance plot rises the slowest for L_{CUAXU} in all



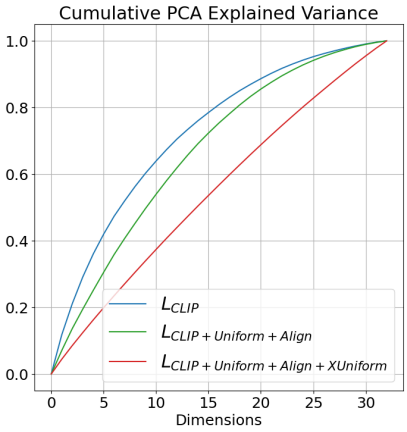
(a) Distance between image and text centroids in CLIP space for each of the losses. Recall that the gap closes when the centroid distance is small.



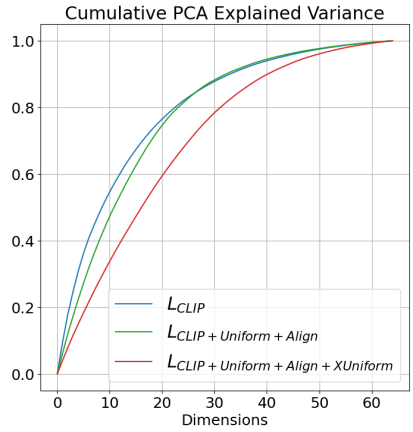
(b) Linear separability of image and text embeddings in CLIP space for each of the losses. Recall that the gap closes when linear separability ≈ 0.5.

Figure 5.2: Contrastive gap metrics after fine-tuning CLIP model on the three losses (In the plot legends, Default = L_{CLIP}, CUA = L_{CUA}, and CUAXU = L_{CUAXU}) L_{CUA} and L_{CUAXU} have lower measures of both the metrics of the contrastive gap. This indicates that the size of the gap is much smaller with uniformity and alignment terms included. The differences in the size of the contrastive gap are more pronounced in higher CLIP dimensionalities.

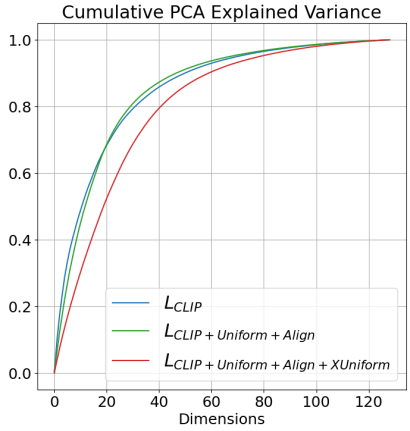
the dimensionalities tested, indicating that cross-modal uniformity encourages the embeddings to be distributed throughout the hypersphere more effectively. We therefore conclude that adding uniform and align losses encourages smaller contrastive gaps in the latent space, across a range of dimensions.



(a) **32D CLIP**



(b) **64D CLIP**



(c) **128D CLIP**

Figure 5.3: Cumulative explained variances for all principal components of the latent space after fine-tuning CLIP with the different losses. The different plots shows the PCA explained variances of different CLIP dimensionalities.

5.4 Effects of Reducing Contrastive Gap on Image-Text Retrieval

We now assess the effects of reducing the contrastive gap on the image-text retrieval task using the MS-COCO validation dataset, a standard benchmark in the vision-language domain. In this task, we aim to retrieve the correct caption from the dataset for a given input image, or, conversely, the correct image for a given caption.

We define the image-retrieval task as follows:

1. **Obtain Text Embedding:** For an input caption (indexed by j), pass it through the text encoder to get its text embedding, denoted as E_j^T.

2. **Generate Image Embeddings:** Given a dataset of N images, pass each image through the image encoder to get a list of N image embeddings, denoted as $E_{[1...N]}^I$.

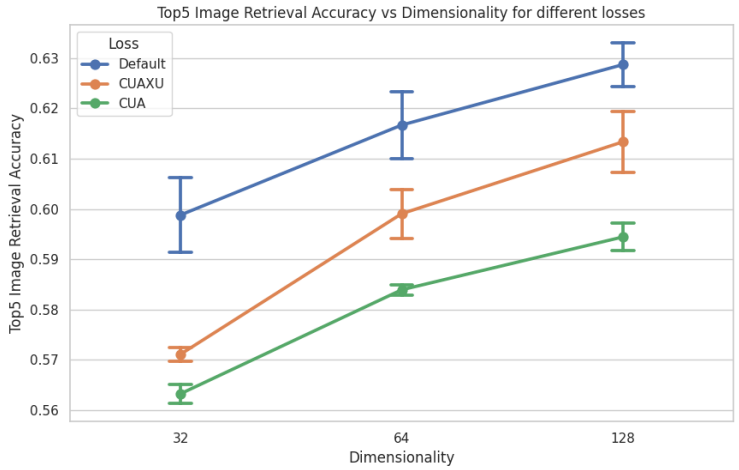
3. **Find the Most Similar Image:** Find the image that best matches the input caption by calculating the cosine similarity between E_j^T and each of $E_{[1...N]}^I$. The image with the highest similarity score is considered the best match for the caption.

The text retrieval task is similar, but in reverse: we retrieve the most relevant caption for a given input image.

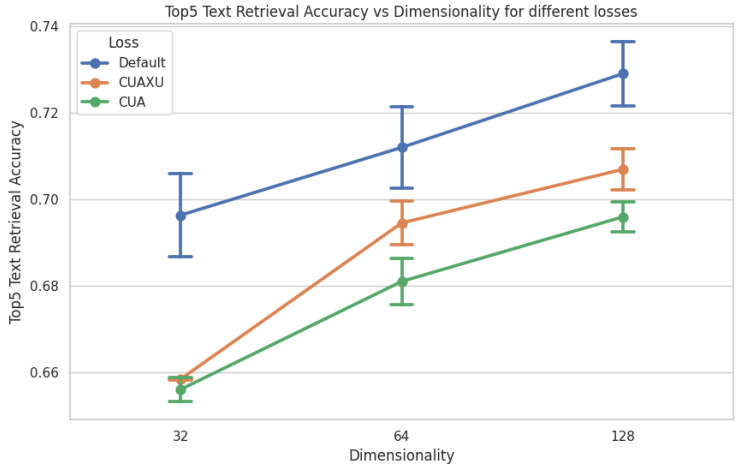
We present the results of fine-tuning and evaluating CLIP on MS-COCO in 5.4. For both the text-retrieval ($I \to T$) and image-retrieval tasks ($T \to I$), models fine-tuned with L_{CLIP}outperform the other models by a small margin. The results suggest that the performance in the image-text retrieval task is *not* well correlated with uniformity, alignment, and size of contrastive gap.

5.5 Zero-Shot Transfer

Previously, we saw that optimizing for uniformity and alignment reduces the size of the contrastive gap by encouraging the image and text embeddings to be spread more uniformly on the unit hypersphere and lie on higher dimensional



(a) Top 5 recall for Image retrieval task for different fine-tuning losses. L_{CLIP}outperforms other losses for this task.



(b) Top 5 recall for text retrieval task for different fine-tuning losses. L_{CLIP}outperforms other losses for this task.

Figure 5.4: Top 5 recall for image to text and text to image retrieval tasks on the MS-COCO validation set. Fine-tuning with L_{CLIP}gives slightly higher recall values compared to fine-tuning with L_{CUA} and L_{CUAXU}.

manifolds in $\mathbb{R}^d$. Now, we analyze the effects this has on zero-shot image classification, a common downstream task for assessing the quality of CLIP embeddings. We evaluate our fine-tuned CLIP models on the standard image-classification datasets outlined in Table 5.2.

To test the zero-shot transfer capabilities of our CLIP models, we adopt the evaluation strategy of Goel *et al.* [11], which is also recommended by Radford

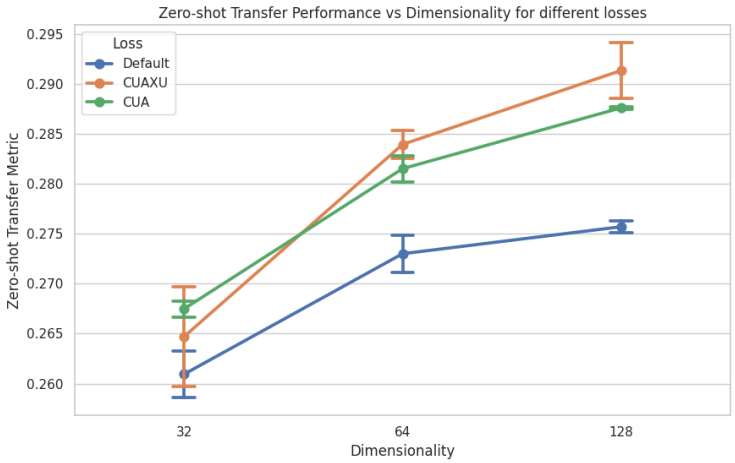


Figure 5.5: Average zero-shot transfer performance for fine-tuned CLIP on the different losses. We plot the average metric value of all the datasets as shown in Table 5.2. CLIP losses with uniformity and alignment terms added consistently get better zero-shot performance than default fine-tuned CLIP on the same dimensionality.

et al. [28]: We generate prompts using the class names to form sentences like "a photo of a {class name}", "A sketch of a {class name}", etc. We then pass these sentences through the text encoder to get prompt embeddings. We average all the prompt embeddings to get a class embedding for each class. Finally, to classify an image, we pass the input image through the image encoder and obtain its image embedding. We then determine the predicted class by finding the class embedding closest to the image embedding using cosine similarity.

Figure 5.5 shows the average zero-shot metric values across all the datasets for each of the losses and dimensionalities. CLIP losses with added alignment and uniformity terms consistently outperform the default CLIP loss, and XU-niform adds additional benefit. Further, the improvement in zero-shot transfer performance is more significant in higher dimensionalities of CLIP.

This is a similar trend to that in Figure 5.2, where the difference in the size of the contrastive gap for the different losses was more significant in higher dimensionalities. Thus, representational spaces with smaller contrastive gap (as learned by models fine-tuned with L_{CUA} and L_{CUAXU}) appear to correlate

Dataset	Classes	Test size	Evaluation metric
CIFAR-10	10	10,000	Accuracy
CIFAR-100	100	10,000	Accuracy
SUN397	397	19,850	Accuracy
Pascal VOC 2007	20	4,952	11-Point mAP
Oxford-IIIT Pets	37	3,669	Mean Per Class
Caltech-101	102	6,085	Mean Per Class
ImageNet	1000	50,000	Accuracy
ImageNet-V2	1000	10,000	Accuracy
ImageNet-Sketch	1000	50,000	Accuracy
ImageNet-A	200	7,500	Accuracy
ImageNet-R	200	30,000	Accuracy
ImageNet-O	200	2,000	Accuracy
ObjectNet	113	50,000	Accuracy

Table 5.2: Datasets evaluated on to test zero-shot image-classification performance of the different CLIP losses.

with higher performance for the zero-shot transfer task.

5.6 Multi-modal Arithmetic

A high quality multi-modal representational space should maintain consistent structural relationships between the different modalities it learns, such as images and text. This means that similar concepts should be represented in a structurally coherent way across both modalities. To evaluate such consistency between CLIP's image and text embeddings, we use SIMAT (Semantic IMage Transformation), from Couairon *et al.* [8]. SIMAT assesses how well the structural relationships are preserved between the two modalities by transforming an image representation using *text delta vectors*. These vectors capture the change from one text embedding to another, and are used to adjust the image embedding accordingly. Specifically, SIMAT computes a new image representation with the formula $E^I_{\text{target}} = E^I_{\text{input}} + \lambda \cdot (E^T_{\text{target}} - E^T_{\text{input}})$, where $(E^T_{\text{target}} - E^T_{\text{input}})$ is the text delta vector and λ is a hyperparameter that controls the strength of this transformation. In our experiments, we set λ to 1. We use this approach to retrieve the closest image to the transformed embedding. By examining the retrieved image, we gain insight into how well

the image and text embeddings align in the representational space.

Figure 5.6 further explains the methodology of SIMAT. SIMAT performs image retrieval, guided by an input image and a query describing a text transformation. The text transformation is mapped to a delta vector, which, when added to the input image embedding, results in a transformed embedding. SIMAT retrieves the closest image to the tranformed embedding from its image database.

Finally, the evaluation module checks whether the retrieved image corresponds correctly to the text-transformed caption. This validation is performed using OSCAR (Li *et al.* [19]), an image-text matching oracle that assigns a probability for a given image-caption pair, indicating how likely they are to be associated with each other. For the SIMAT evaluation, the retrieved image is considered correct if the OSCAR probability is greater than 0.5. Figure 5.7 illustrates the concept by showing examples of *expected* images to be retrieved given an input image and a transformation query.

We present the results of SIMAT evaluation on the different loss functions in Figure 5.8. Adding uniformity and loss terms to the CLIP loss leads to increased SIMAT scores, indicating that the representational space learned with uniform and align terms is more consistent with arithmetic operations between modalities. Notably, the difference in SIMAT scores between the baseline and the modified CLIP losses becomes more significant as the dimensionality of CLIP increases. We see a similar trend in Figure 5.2, where fine-tuning with L_{CUA} and L_{CUAXU} lead to representational spaces with significantly lower contrastive gap, especially in higher dimensionalities, compared to fine-tuning with L_{CLIP}.

As an illustrative example of this improvement, Figure 5.9, shows that the CLIP model fine-tuned with L_{CUAXU} retrieves an image that is more representative of the transformed caption compared to CLIP model fine-tuned with L_{CLIP}.

From these empirical findings, we conclude that having a smaller contrastive gap is well correlated with higher performance in the multi-modal arithmetic task. Our results suggest that closing the contrastive gap by fine-

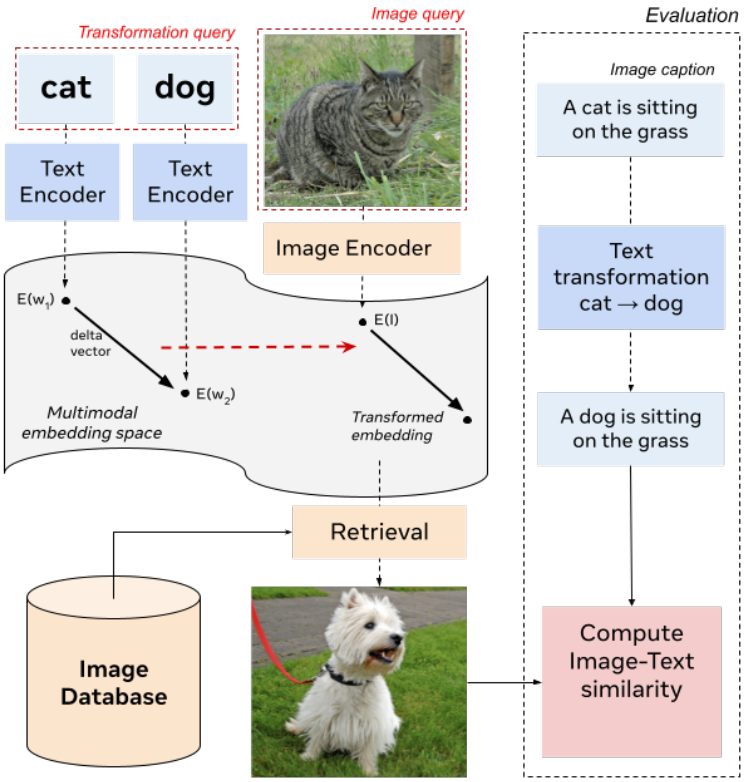


Figure 5.6: SIMAT Methodology: Image retrieval guided by text transformation query. Figure taken from Couairon *et al.* [8]

Figure 5.7: Examples of images that are expected to be retrieved given input images and text transformations. Figure taken from Couairon *et al.* [8].

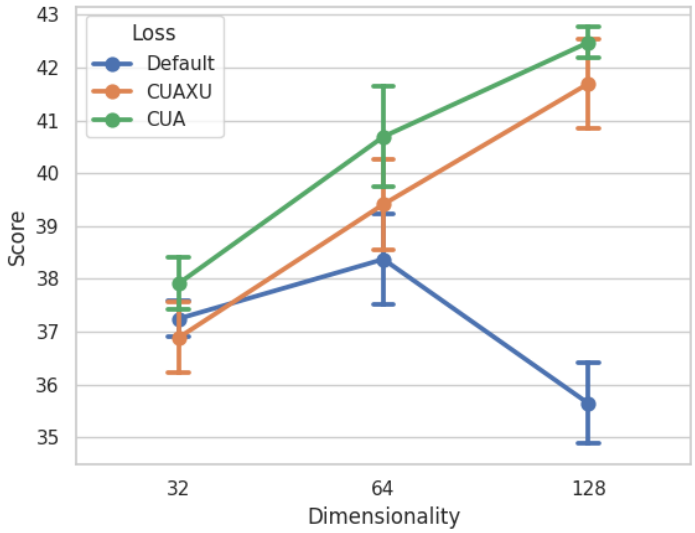


Figure 5.8: SIMAT Score vs Dimensionality plot. Higher SIMAT scores indicate more consistency in the arrangement of image and text embeddings in CLIP space. The plot shows that L_{CUA} and L_{CUAXU} have achieve higher SIMAT scores than L_{CLIP}. The difference in SIMAT scores becomes more significant as CLIP dimensionality increases from 32D to 128D.

tuning with added uniformity/alignment terms could benefit applications that rely on the geometric structure and consistent arithmetic properties in the latent space.

5.7 Chapter Conclusion

In this chapter, we studied the effects of fine-tuning on losses L_{CLIP}, L_{CUA}, and L_{CUAXU}, characterizing their impact on the size of the contrastive gap, image-text retreival accuracy on MS-COCO, zero-shot accuracy across various standard image classification datasets, and performance in multi-modal arithmetic tasks.

We conclude this chapter with the following key takeaways:

- Adding uniform and align terms to CLIP loss (as done in L_{CUA} and L_{CUAXU}) significantly reduces the size of the contrastive gap, with the reduction being more drastic compared to L_{CLIP} as dimensionality increases.

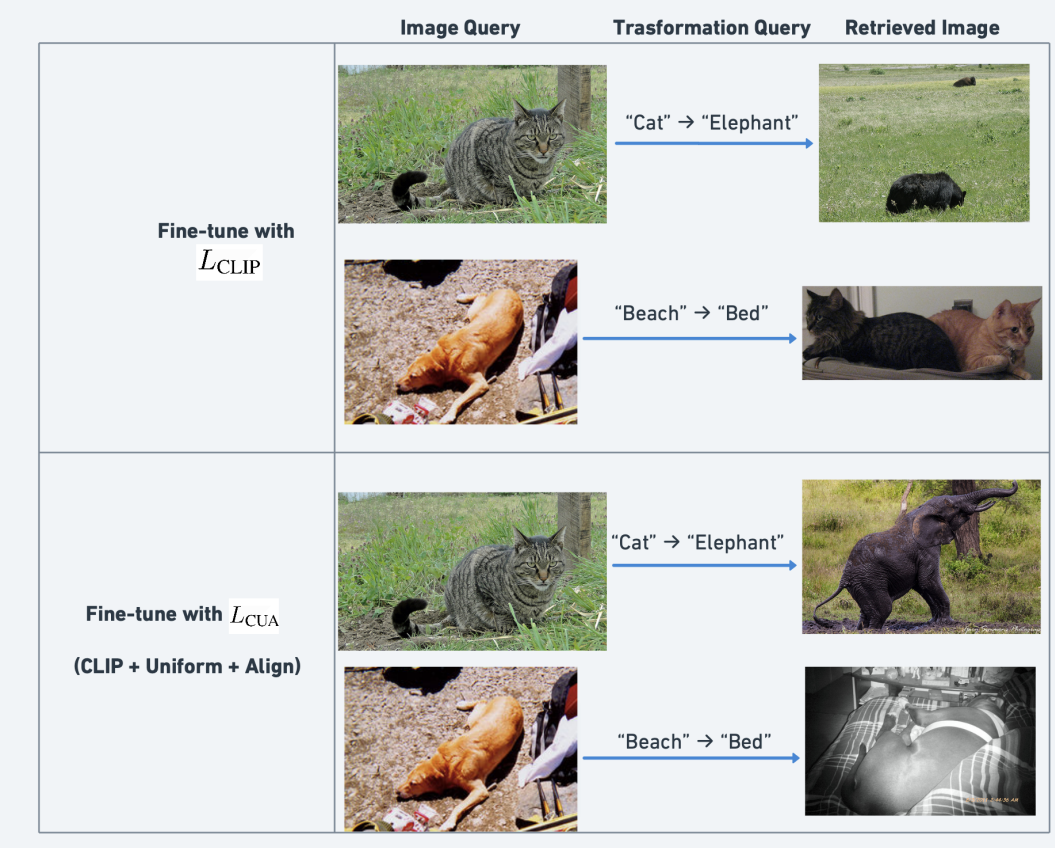


Figure 5.9: Reducing contrastive gap improves quality of images retrieved using text transformation queries. **Top 2 rows**: Arithmetic in CLIP space with default contrastive gap (with model fine-tuned with L_{CLIP}). **Bottom 2 rows:** Arithmetic in CLIP space with reduced contrastive gap (with model fine-tuned with L_{CUA})

- Despite this reduction in the contrastive gap, fine-tuning with L_{CUA} and L_{CUAXU} does not lead to improved image-text retrieval performance on the MS-COCO dataset. This indicates that factors beyond the contrastive gap, uniformity, and alignment may play a critical role in determining the quality of embeddings for this particular task.

- Fine-tuning with L_{CUA} and L_{CUAXU} and thus minimizing the contrastive gap enhances zero-shot image classification accuracy across a wide range of datasets. This improvement in zero-shot transfer over $L_{\text{CLIP}}$is more significant in higher CLIP dimensionalities. We noticed that this trend is also true for the size of the contrastive gap as dimensionality increases, suggesting that the smaller contrastive gap is correlated to higher zero-shot transfer performance.

- Minimizing the contrastive gap improves multi-modal arithmetic performance. We observed similar trends in the measures of contrastive gap and SIMAT scores, which suggest that high performance in the multi-modal arithmetic task is correlated with smaller contrastive gap, and that minimizing the gap is beneficial for tasks requiring complex reasoning across modalities.

Chapter 6

Conclusions and Future Work

Multi-modal contrastive learning is a rapidly emerging approach in the field of vision-language models. Models like CLIP and its many variants are being used in a wide variety of applications ranging from simple image-text retrieval all the way to studying representations formed from various stimuli in the brain. Therefore, the contrastive gap is an emergent phenomenon that affects these applications, often in ways that are not trivial to analyze.

In this thesis, we studied the representations learned by multi-modal contrastive learning algorithms and analyzed the contrastive gap phenomenon. We showed that eliminating all reasons commonly thought to cause the gap does *not* close it. Thus, this is *not a modality gap*: We showed that the gap is an inherent property that arises from the CLIP loss itself, and is not due to properties of the data (such as mismatched pairs), the network (i.e. cone effect), nor the loss landscape (CLIP loss getting stuck in local minima), as prior work has suggested. We instead proposed the term *contrastive gap* to describe this phenomenon. By studying the behaviour of the contrastive loss in 3D, we deduced that the most important factor behind the gap is low uniformity of the embeddings in the unit hypersphere. We additionally demonstrated that the contrastive gap is symptomatic of the representations lying on a lower dimensional manifold in the latent space.

Motivated by the idea that the gap stems from low uniformity, we added adding explicit uniformity and alignment terms to the CLIP loss. We showed that directly optimizing for uniformity and alignment in the latent space sig-

nificantly reduces the gap. This supports our claim that low uniformity in the representational space causes the contrastive gap.

We further showed that closing the gap by simply fine-tuning CLIP with added uniformity and alignment terms improved zero-shot image classification and multi-modal arithmetic performance, which suggests that a smaller contrastive gap may lead to higher performance for these tasks.

In this work we explored the contrastive gap in the context of limited data (fine-tuning CLIP on MS COCO). In the future we would like to expand our scope and include larger datasets. Training on larger datasets could lead to more insights into the extent to which the contrastive gap closes by optimizing uniformity and alignment at scale. Another interesting direction for future work is to analyze the contrastive gap and downstream task performance using all five captions per image in MS COCO, rather than just the first caption as we did in our study. Since each of the five captions typically describes different aspects of the same image, this approach could provide insights into how well the model's representation space aligns varied textual descriptions of the same visual content.

Another promising direction would be to investigate the impact of reducing the contrastive gap on *generation* tasks, such as generating images from texts or vice versa using CLIP embeddings. Reducing the gap between modalities could potentially improve the process of translating one modality (like text) into the other (like images), leading to better generation results.

In our experiments, we added uniformity and alignment losses to the CLIP loss to investigate their effects on the representational space. An insightful study would be to experiment with different weightings of these uniformity and alignment terms. This would help determine the relative importance of each of these terms in reducing the contrastive gap and improving task performance. Such an empirical study could lead to the design of more effective loss functions and provide a deeper understanding of the multi-modal contrastive latent space.

6.1 Broader Impacts

The work we have presented here is quite theoretical so the broader impacts are less clear. However, any model that learns from text and images has the ability to incorporate or enhance biases that exist in the training data. For example, if some captions are harmful, creating a better representational space for them may also be harmful. The images included in MS COCO also represent a biased sample of what occurs in the real world. For example, scenes from certain countries are underrepresented. This will impact any model trained on this data and could impact the utility of the model in certain deployment scenarios.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, *Flamingo: A visual language model for few-shot learning*, 2022. arXiv: 2204.14198 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2204.14198.

[2] J. Bai, C. Liu, F. Ni, *et al.*, *Lat: Latent translation with cycle-consistency for video-text retrieval*, 2023. arXiv: 2207.04858 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2207.04858.

[3] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, *A clip-hitchhiker's guide to long video retrieval*, 2022. arXiv: 2205.08508 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2205.08508.

[4] R. Balestriero, M. Ibrahim, V. Sobal, *et al.*, *A cookbook of self-supervised learning*, 2023. arXiv: 2304.12210 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2304.12210.

[5] M. Barraco, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara, "The unreasonable effectiveness of CLIP features for image captioning: An experimental analysis," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 4661–4669, ISBN: 978-1-66548-739-9. DOI: 10.1109/CVPRW56347.2022.00512. [Online]. Available: https://ieeexplore.ieee.org/document/9857436/ (visited on 08/20/2024).

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, *A Simple Framework for Contrastive Learning of Visual Representations*, en, arXiv:2002.05709 [cs, stat], Jun. 2020. [Online]. Available: http://arxiv.org/abs/2002.05709 (visited on 03/27/2024).

[7] J. Cho, S. Yoon, A. Kale, F. Dernoncourt, T. Bui, and M. Bansal, *Fine-grained image captioning with clip reward*, 2023. arXiv: 2205.13115 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2205.13115.

[8] G. Couairon, M. Cord, M. Douze, and H. Schwenk, *Embedding arithmetic of multimodal queries for image retrieval*, 2022. arXiv: 2112.03162 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2112.03162.

[9] Z.-Y. Dou, A. Kamath, Z. Gan, *et al.*, *Coarse-to-fine vision-language pre-training with fusion in the backbone*, 2022. arXiv: 2206.07643 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2206.07643.

[10] N. Giakoumoglou and T. Stathaki, *A review on discriminative self-supervised learning methods*, 2024. arXiv: 2405.04969 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2405.04969.

[11] S. Goel, H. Bansal, S. Bhatia, R. A. Rossi, V. Vinay, and A. Grover, *CyCLIP: Cyclic Contrastive Language-Image Pretraining*, en, Oct. 2022. [Online]. Available: http://arxiv.org/abs/2205.14459 (visited on 03/28/2024).

[12] S. M. Holland, "Principal components analysis (pca)," *Department of Geology, University of Georgia, Athens, GA*, vol. 30602, p. 2501, 2008.

[13] M. Al-Jaff, "Messing With The Gap: On The Modality Gap Phenomenon In Multimodal Contrastive Representation Learning," Ph.D. dissertation, Uppsala University, Nov. 2023.

[14] C. Jia, Y. Yang, Y. Xia, *et al.*, *Scaling up visual and vision-language representation learning with noisy text supervision*, 2021. arXiv: 2102.05918 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2102.05918.

[15] Q. Jiang, C. Chen, H. Zhao, *et al.*, *Understanding and constructing latent modality structures in multi-modal representation learning*, 2023. arXiv: 2303.05952 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2303.05952.

[16] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 2021, pp. 5583–5594. [Online]. Available: https://proceedings.mlr.press/v139/kim21k.html.

[17] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, *Masked vision and language modeling for multi-modal representation learning*, 2023. arXiv: 2208.02131 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2208.02131.

[18] J. Li, D. Li, C. Xiong, and S. Hoi, *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, 2022. arXiv: 2201.12086 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2201.12086.

[19] X. Li, X. Yin, C. Li, *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," *ECCV 2020*, 2020.

[20] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, *Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning*, en, arXiv:2203.02053 [cs], Oct. 2022. [Online]. Available: http://arxiv.org/abs/2203.02053 (visited on 11/14/2023).

[21] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1.

[22] A. Luo, M. M. Henderson, M. J. Tarr, and L. Wehbe, "BrainSCUBA: Fine-grained natural language captions of visual cortex selectivity," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=mQYHXUUTkU.

[23] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, *X-clip: End-to-end multi-grained contrastive learning for video-text retrieval*, 2022. arXiv: 2207.07285 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2207.07285.

[24] R. Mokady, A. Hertz, and A. H. Bermano, *Clipcap: Clip prefix for image captioning*, 2021. arXiv: 2111.09734 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2111.09734.

[25] L. Muttenthaler, L. Linhardt, J. Dippel, *et al.*, "Improving neural network representations using human similarity judgments," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper%5C_files/paper/2023/hash/9febda1c8344cc5f2d51713964864e93-Abstract-Conference.html.

[26] C. Oh, J. So, H. Byun, *et al.*, "Geodesic multi-modal mixup for robust fine-tuning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=iAAXq60Bw1.

[27] L. Qiu, R. Zhang, Z. Guo, *et al.*, *Vt-clip: Enhancing vision-language models with visual-guided texts*, 2023. arXiv: 2112.02399 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2112.02399.

[28] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," en, 2021.

[29] S. Shen, L. H. Li, H. Tan, *et al.*, *How much can clip benefit vision-and-language tasks?* 2021. arXiv: 2107.06383 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2107.06383.

[30] A. Singh, R. Hu, V. Goswami, *et al.*, *Flava: A foundational language and vision alignment model*, 2022. arXiv: 2112.04482 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2112.04482.

[31] V. Udandarao, "Understanding and Fixing the Modality Gap in Vision-Language Models," en, 2022.

[32] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 2495–2504.

[33] T. Wang and P. Isola, "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere," en, 2020.

[34] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, *Simvlm: Simple visual language model pretraining with weak supervision*, 2022. arXiv: 2108.10904 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2108.10904.

[35] M. Welle, "UNDERSTANDING THE MODALITY GAP IN CLIP," en, 2023.

[36] T. Wolf, L. Debut, V. Sanh, *et al.*, *Huggingface's transformers: State-of-the-art natural language processing*, 2020. arXiv: 1910.03771 [cs.CL].

[37] H. Xu, G. Ghosh, P.-Y. Huang, *et al.*, "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6787–6800. DOI: 10.18653/v1/2021.emnlp-main.544. [Online]. Available: https://aclanthology.org/2021.emnlp-main.544.

[38] J. Yang, J. Duan, S. Tran, *et al.*, *Vision-language pre-training with triple contrastive learning*, 2022. arXiv: 2202.10401 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2202.10401.

[39] H. You, L. Zhou, B. Xiao, *et al.*, *Learning visual representation from modality-shared contrastive language-image pre-training*, 2022. [Online]. Available: https://arxiv.org/abs/2207.12661.

[40] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, *Coca: Contrastive captioners are image-text foundation models*, 2022. arXiv: 2205.01917 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2205.01917.

[41] R. Zhang, R. Fang, W. Zhang, *et al.*, *Tip-adapter: Training-free clip-adapter for better vision-language modeling*, 2021. arXiv: 2111.03930 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2111.03930.

[42] R. Zhang, Z. Zeng, Z. Guo, and Y. Li, *Can language understand depth?* 2022. arXiv: 2207.01077 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2207.01077.

[43] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. Langlotz, *Contrastive learning of medical visual representations from paired images and text*, 2021. [Online]. Available: https://openreview.net/forum?id=T4gXBOXoIUr.

[44] C. Zhou, F. Zhong, and C. Oztireli, *Clip-pae: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable, and controllable text-guided face manipulation*, 2023. arXiv: 2210.03919 [cs.CV].

[45] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, Jul. 2022, ISSN: 1573-1405. DOI: 10.1007/s11263-022-01653-1. [Online]. Available: http://dx.doi.org/10.1007/s11263-022-01653-1.