Literature Review on Latent-alignment aware knowledge distillation on CLIP

Muhammad Faraz Ahmad[1], Amna Iftikhar[1], and Areeb Khalid[1]

[1]LUMS

February 5, 2026

Contents

1 Introduction

This document contains the key points of all the papers we've found relevant for our project *Latent-alignment aware knowledge distillation on CLIP*.

2 Research Papers:

2.1 CLIP-KD: An Empirical Study of CLIP Model Distillation [Link]

- Gives a generalized framework for distilling CLIP into smaller student models. Their framework, unlike TinyCLIP, does not require student models to have the same architecture as that of the teacher model.
- Essentially, proposes the following six distillation strategies:
    1. **Contrastive Relational Distillation:** Minimize the KL-divergence between the probability distributions of the student and the teacher.
    2. **Feature Distillation (FD):** Minimize the mean-squared error (MSE) between the image/text embeddings of the student and the teacher.

3. **Masked Feature Distillation:** Essentially FD, with the only difference being that the student receives a masked image as an input and tries to minimize MSE between its embedding on the masked image and the teacher's embedding on the full original image.
4. **Gradient Distillation (GD):** Tries to match the student's and the teacher's gradients.
5. **Interactive Contrastive Learning (ICL):** Uses the idea of matching the image embedding of one model with the text embedding of another (and vice versa). It takes student image embedding as an anchor and contrasts it against the teacher's text embeddings (and vice versa).
6. **Augmented Feature Distillation (AFD):** This uses a fusion encoder to mix student and teacher embeddings before calculating the loss.

- The simplest method, FD, works the best.
- **Relevance:** We aimed to use MSE between features as well, and this paper proves that it works very well. It also gives different strategies we could incorporate into our method.
- **Date:** 7 May 2024

2.2 Is CLIP ideal? No. Can we fix it? Yes! [Link]

- A problem with CLIP is that its latent space fails at handling **complex visual-textual interactions**. CLIP suffers when the semantics get complicated.
- Four conditions that must be satisfied for CLIP to precisely understand images and texts:
  1. Represent basic descriptions and image content
  2. Attribute binding
  3. Spatial locations and relationships
  4. Represent negation (CLIP can't represent "Not X" correctly)
- The authors prove that CLIP's geometry, which represents images and texts as n-dimensional unit vectors on a unit hypersphere and compares them via cosine-similarity, is incapable of handling complex semantics (can't satisfy any two out of the four conditions at the same time).
- **Solution:** Don't use only `<EOS>` (for text) and `<CLS>` (for images). Instead, compute cosine similarity between every text token and every image patch. This way, we have a 2D map of scores, which preserves the topology of the image. Finally, feed this map into a lightweight CNN (2 conv. layers with hidden dimension of 128) to get the final scores.
- **Relevance:** Will pose a challenge, as a student model inherits such discrepancies from the teacher. So, any student model will also fail in such cases.
- **Date:** 10 March 2025.

2.3 The Double-Ellipsoid Geometry of CLIP [Link]

- Argues that CLIP's primary embeddings (pre-normalized) form two linearly separable, tilted ellipsoid shells for each modality: image and text.
- Also notes that:

  - **Non-origin centering:** the ellipsoids are shifted away from the origin.
  - **Thin-shell phenomenon:** most of the embedding mass concentrates within a narrow range.

- Authors argue that this geometry emerges naturally out of the loss function to handle uncertainty. They introduce *Conformity*, a metric to measure how common or rare a concept is. It's defined as the expected value of the cosine similarity of a vector v_j with all other vectors v_k in a set S.

- Paper claims that common concepts (higher conformity, vague), e.g., a human or a dog, get embedded close to the mean embedding of their modality. On the other hand, rare concepts (low conformity, less vague), e.g., "A vintage yellow refrigerator surrounded by wood cabinetry," get embedded away from the mean embedding of the ellipsoid.

- **Semantic Blur:** Authors also say that shifting ellipsoids away from the center allows the model to control the sharpness of embeddings. By keeping common concepts near the center of a shifted ellipsoid, the model ensures there is high cosine similarity between a concept and its similar concepts (e.g., two images of dogs). This reduces false negatives.

- **Relevance:** As per this paper, CLIP's embedding space contains two tilted ellipsoids shifted away from the origin. While this helps the original CLIP, a student CLIP may not be able to model this complex geometry. To deal with this, the next paper of W-CLIP provides a solution for whitening the teacher CLIP before distillation.

- **Date:** 24 May 2025

2.4 Whitened CLIP as a Likelihood Surrogate of Images and Captions [Link]

- This paper proposes a novel transformation of CLIP's latent space (the embeddings) via an invertible matrix W. This transformation leads to each feature in the whitened space having zero mean, unit variance, and no correlation with all other features.

- They assess if the distribution of the whitened space approximates the standard normal distribution using two statistical tests: Anderson-Darling test and D'Agostino-Pearson test. The tests confirm this.

- **Methodology:** Every raw embedding x is converted into a whitened embedding y using $y = W(x - \mu)$, where W is computed with a data-driven process using the covariance of a representative dataset (like MS-COCO in the author's case). As a result of this process, the double ellipsoid geometry of CLIP transforms into an Isotropic Hypersphere.

- The main focus of the paper, however, was to compute likelihood estimates for images, which is not a trivial task. They were the first to study CLIP's embedding space probabilistically. After transformation, the log-likelihood of images can be estimated easily using the norm of the whitened embedding.

- **Relevance:** For our purposes, the likelihood estimation is not that important; the transformation of the embedding space is important so as to help the student model learn better.

- **Date:** 11 May 2025

2.5 CLIP-SDMG [Link]

- **Date:** 28 July 2025

3 Other Relevant Papers:

3.1 A Sentence Speaks a Thousand Images (RISE paper) [Link]

- **Date:** 21 Sep 2023