Distilling CLIP-ResNet50 into a Smaller Vision-Language Model

Benjamin Yoon, Jessica Liang, Edward Ho, Riya Agrawal

Abstract

In this project, we study the distillation of the CLIP (Contrastive Language-Image Pretraining) model. The teacher model is a pretrained ResNet-50, while the student model is an untrained ResNet-34, initialized with random weights and lacking the learned feature representations required for effective image recognition tasks. We propose a novel loss function for CLIP model distillation that incorporates synergetic and redundant information, inspired by Partial Information Decomposition (PID). This loss function enhances the student model's ability to capture both unique and shared information across modalities, improving performance on image-to-text (I2T) and text-to-image (T2I) retrieval tasks. We compare the performance of our student model under four different loss configurations: a simple baseline model using mean square error, a strong baseline model utilizing Contrastive Loss, an Extension 1 model that integrates synergetic information from images and text, and an Extension 2 model that further incorporates redundant information. Experimental results demonstrate that the Extension 2 model achieves the best performance, highlighting the efficacy of our proposed approach.

1 Introduction

In recent years, the field of Natural Language Processing (NLP) has seen significant growth at the intersection of language and vision. Models that jointly process text and images have enabled a wide range of multimodal applications such as image captioning, visual question answering, and text-based image retrieval. Among these models, CLIP (Contrastive Language-Image Pretraining) has emerged as a highly influential approach, learning a joint embedding space in which semantically related textual and visual representations are closely aligned. However, the large teacher models that often underpin these systems can be computationally expensive and slow to deploy, motivating the use of model distillation to produce smaller, more efficient *student* variants.

This task is closely related to NLP, as language provides a rich source of semantic structure that can ground visual understanding. By learning how to encode textual information—through tokenization, contextual embeddings, and other NLP techniques—the model leverages language as a guiding force to align images and text. As a result, language-based supervision helps ensure that the student model learns more robust and semantically meaningful representations even with fewer parameters, ultimately enhancing the efficiency and applicability of multimodal models in real-world scenarios.

For example, consider an image of "a dog on the grass" and text, "pepper the aussie pup" in the CLIP model in Fig. 1 (Radford et al., 2021). The teacher model processes the inputs to generate high-dimensional embeddings and a similarity score reflecting a strong image-text alignment. The student model, being smaller, processes the same inputs, and through distillation, it learns to produce embeddings and similarity scores similar to the teacher's output. In this project, the teacher model is a pretrained ResNet50 (OpenAI, 2021), and **the student model is an untrained ResNet-34, initialized with random weights**, lacks the learned feature representations necessary for effective image recognition tasks. The input to the teacher model and the student model are pairs of images and corresponding text descriptions from datasets such as MSCOCO. A successfully distilled student model is capable of:

- Generating image and text embeddings similar to those of the teacher model.
- Computing similarity scores between images and texts, enabling tasks such as image-text retrieval.

- Performing zero-shot classification with performance close to the teacher model but at a lower computational cost.

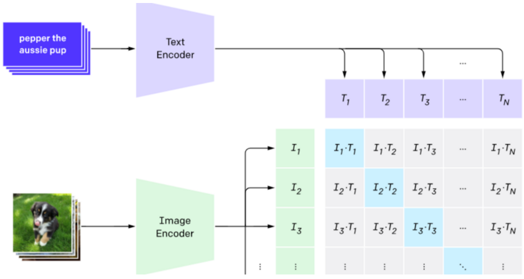


Figure 1: Illustrative Example of the CLIP (Radford et al., 2021).

We selected this task because it sits at the nexus of computational efficiency and multimodal understanding. By distilling a large-scale CLIP model into a more compact one, we maintain strong performance and semantic alignment at a fraction of the computational cost. This hands-on project allows us to deepen our understanding of cutting-edge multimodal NLP techniques, gain practical experience with knowledge distillation strategies, and explore methods for inducing richer representational structures—ultimately making advanced multimodal AI more accessible and scalable.

2 Literature Review

The CLIP Knowledge Distillation (KD) in (Yang et al., 2024), focuses on distilling smaller CLIP models, with supervision provided by a larger teacher CLIP model. Several distillation strategies are introduced, including relation-based, feature-based, gradient-oriented, and contrastive methods, to assess the effectiveness of CLIP-KD. Results indicate that simple feature mimicry with Mean Squared Error loss performs remarkably well. Furthermore, interactive contrastive learning between teacher and student encoders contributes to significant performance improvements. The success of CLIP-KD is attributed to maximizing feature similarity between teacher and student models. This unified methodology is applied to distill various student models trained on CC3M+12M datasets. CLIP-KD consistently enhances the performance of student CLIP models on zero-shot ImageNet classification and cross-modal retrieval benchmarks.

The TinyCLIP in (Wu et al., 2023), is an innovative cross-modal distillation approach designed for large-scale language-image pre-trained models. TinyCLIP leverages two primary techniques: affinity mimicking and weight inheritance. Affinity mimicking enables student models to replicate the teacher models' approach to cross-modal feature alignment, capturing interactions within a visual-linguistic affinity space. Weight inheritance, on the other hand, transfers pre-trained weights from teacher models to student models, significantly enhancing distillation efficiency. Additionally, TinyCLIP incorporates a multi-stage progressive distillation strategy to preserve informative weights during intense compression. Extensive experiments validate TinyCLIP's effectiveness, showing that it can reduce the size of the CLIP ViT-B/32 model by 50% while maintaining similar zero-shot performance. With weight inheritance, the distillation process achieves training speeds 1.4 to 7.8 times faster than training from scratch.

The work in (Li et al., 2023) explores the distillation of visual representations from large teacher vision-language models into compact student models, using a small- or mid-sized dataset. A primary focus of this study is on open-vocabulary out-of-distribution (OOD) generalization, a challenging area that has received little attention in previous model distillation research. Two guiding principles are proposed to enhance the OOD generalization of student models from both visual and language perspectives: (1) by closely replicating the teacher model's visual representation space and strengthening alignment with the teacher's vision-language relationships; (2) by enriching the language representations of the teacher with detailed semantic attributes, enabling finer discrimination across labels. Several metrics are introduced, and extensive experiments are conducted to assess these techniques. The findings demonstrate considerable improvements in zero-shot and few-shot performance of student models for open-vocabulary OOD classification, underscoring the effectiveness of the proposed methods.

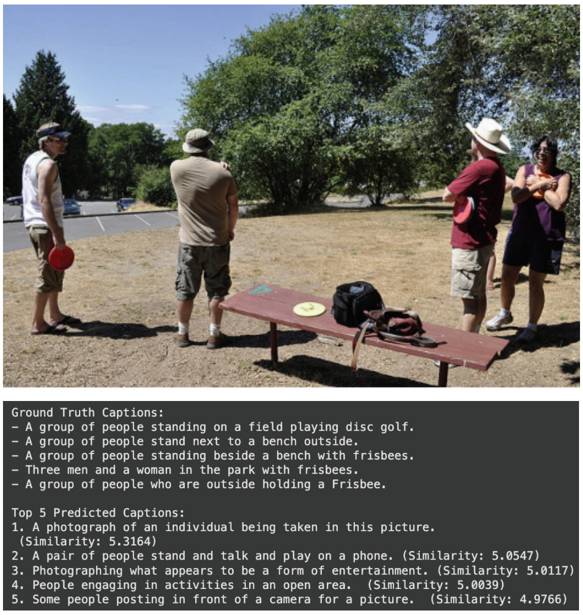
3 Experimental Design

3.1 Data

For this project, we utilize the Microsoft COCO 2014 dataset (MSCOCO) (Lin et al., 2014), a widely used benchmark dataset for vision-language tasks. MSCOCO contains annotated images with captions and segmentation data, which makes it ideal for tasks such as image-to-text (I2T) and text-to-image (T2I) retrieval. The data splits are shown in Table 1, providing an overview of the dataset size for training and validation splits. In Fig. 2, we provide an example of image and caption from MSCOCO.

Dataset	# of Images	# of Captions
Train2014	82,783	414,113
Val2014	40,504	202,654

Table 1: Size of the MSCOCO 2014 dataset splits.



```
Ground Truth Captions:
- A group of people standing on a field playing disc golf.
- A group of people stand next to a bench outside.
- A group of people standing beside a bench with frisbees.
- Three men and a woman in the park with frisbees.
- A group of people who are outside holding a Frisbee.

Top 5 Predicted Captions:
1. A photograph of an individual being taken in this picture.
 (Similarity: 5.3164)
2. A pair of people stand and talk and play on a phone. (Similarity: 5.0547)
3. Photographing what appears to be a form of entertainment. (Similarity: 5.0117)
4. People engaging in activities in an open area.  (Similarity: 5.0039)
5. Some people posting in front of a camera for a picture.  (Similarity: 4.9766)
```

Figure 2: An Example of image and captions from MSCOCO.

The input to the CLIP model could be:

- Images: RGB images of various sizes.
- Text: Captions describing the image content.

The outputs from CLIP model could be:

- Image-to-Text (I2T) Retrieval: Given an image, retrieve the most relevant caption(s).
- Text-to-Image (T2I) Retrieval: Given a caption, retrieve the most relevant image(s).

3.2 Evaluation Metric

In this project, we evaluate the performance of the model using **Recall@K**, a standard metric for image-to-text (I2T) and text-to-image (T2I) retrieval tasks (Mikolov et al., 2013). Recall@K measures the proportion of queries (images or texts) for which the correct item appears in the top K retrieved results. For example, Recall@1 calculates the percentage of queries where the correct result is ranked first. This metric reflects the ability of the model to return relevant results in its top predictions and is widely used in vision-language retrieval benchmarks.

Task	Recall@1	Recall@5	Recall@10
I2T	0.22	0.97	1.83
T2I	0.33	1.39	2.50

Table 2: Zero-shot retrieval performance of student models (ResNet34) using simple baseline.

The metric Recall@K is formally defined as:

$$\frac{\text{Number of queries with correct result in top } K}{\text{Total number of queries}}$$

where:

- K: The number of top results considered.
- Numerator: Counts the queries where the correct item appears in the top K.
- Denominator: Total number of queries.

The Recall@K metric has been described and utilized extensively in vision-language retrieval literature (Lin et al., 2014)(Radford et al., 2021). High Recall@K values indicate that the model is effectively retrieving relevant results, making this metric suitable for evaluating both image-to-text and text-to-image retrieval tasks.

3.3 Simple baseline

The simple baseline is feature distillation, which minimizes the loss between teacher and student embeddings. It uses Mean Squared Error (MSE) to align intermediate features. To preserve the teacher's representation quality, the MSE (L2 distance) between student and teacher embeddings is minimized:

$$\ell_{\text{L2}} = \frac{1}{2}\big(\|\mathbf{z}_i^{(S)} - \mathbf{z}_i^{(T)}\|_2^2 + \|\mathbf{z}_t^{(S)} - \mathbf{z}_t^{(T)}\|_2^2\big) \quad (1)$$

This is a very straightforward way to do CLIP distillation. In Table 2, we summarize the performance of simple baseline. Observe the performance is not good.

Task	Recall@1	Recall@5	Recall@10
I2T	0.39	1.77	3.16
T2I	0.67	2.55	4.28

Table 3: Zero-shot retrieval performance of student models (ResNet34) using published (strong) baseline.

4 Experimental Results

4.1 Published Baseline

The published baseline is our strong baseline using a contrastive loss to align student and teacher embeddings (Yang et al., 2024). The contrastive loss ensures the alignment of image and text embeddings in a joint feature space. Given normalized embeddings of student image features $\mathbf{z}_i^{(S)}$ and text features $\mathbf{z}_t^{(S)}$, the logits are computed as:

$$\begin{aligned}\ell_{\text{contrastive}} \quad &= \quad \frac{1}{2}\big(\ell_{\text{cross-entropy}}(\mathbf{z}_i^{(S)}\mathbf{z}_t^{(S)\top}, \mathbf{y}) \\ &\qquad +\ell_{\text{cross-entropy}}(\mathbf{z}_t^{(S)}\mathbf{z}_i^{(S)\top}, \mathbf{y})\big)\end{aligned} \tag{2}$$

where $\mathbf{y}$ represents the ground-truth labels. In Table 3, we summarize the performance for the strong baseline. Our implementation of the published baseline didn't reach the same level of accuracy as the original paper in (Yang et al., 2024). In (Yang et al., 2024), they used much more complicated student models such as ViT-B/16 and they are pretrained models. The number of parameters in ResNet34 is around 21.8M, but it's around 86M in ViT-B/16.

4.2 Extensions

4.2.1 Extension 1

We propose a new loss function consists of contrastive loss, KL-divergence, L_2 distance, and synergetic information. The contrastive loss and L_2 distance were used in (Yang et al., 2024). KL divergence has been widely used in generative AI such as variational autoencoder (Kingma and Welling, 2019) or diffusion models (Ho et al., 2020). For the first time ever, we propose to apply synergetic information to loss function. Synergetic information is from Partial Information Decomposition (PID). PID (Williams and Beer, 2010) is an important development in information theory, and has been very useful for interpretable machine learning.

KL-Divergence: The KL-divergence term aligns the student model's predicted distributions with those of the teacher. For logits $\mathbf{p}_i^{(S)}$ and $\mathbf{p}_i^{(T)}$ (student and teacher, respectively), the loss is:

$$\ell_{\text{KL}} = \frac{1}{2}\big(D_{\text{KL}}(\mathbf{p}_i^{(S)} \| \mathbf{p}_i^{(T)}) + D_{\text{KL}}(\mathbf{p}_t^{(S)} \| \mathbf{p}_t^{(T)})\big) \tag{3}$$

where D_{KL} denotes the Kullback-Leibler divergence.

Synergy and Redundancy Reward: For image and text (V, T) and the ground truth label Y, their mutual information can be decomposed using PID (Williams and Beer, 2010)

$$\begin{aligned}I(V,T;Y) \quad &= \quad U(V;Y|T) + U(T;Y|V) \\ &\qquad +R(V,T;Y) + S(V,T;Y)\end{aligned}$$

where $U(V;Y|T)$ and $U(T;Y|V)$ denote the unique information; $R(V,T;Y)$ is the redundant information; and $S(V,T;Y)$ is the synergetic information. Their relations are represented in Fig. 3 (Williams and Beer, 2010).

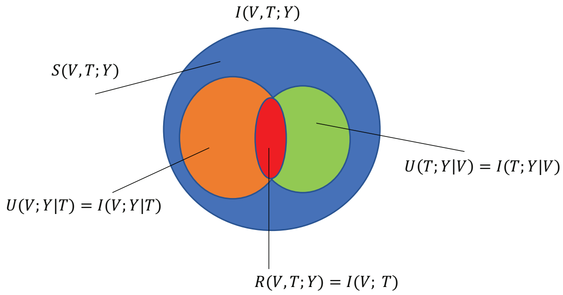


Figure 3: An illustration of PID (Williams and Beer, 2010). $U(V;Y|T)$ and $U(T;Y|V)$ are unique information; $R(V,T;Y)$ is the redundant information; and $S(V,T;Y)$ is the synergetic information.

Synergetic information quantifies the information contributed by the joint representation of image V and text T that cannot be inferred from either modality alone. Maximizing synergetic information encourages the student model to capture interactions between image and text, improving cross-modal understanding and zero-shot generalization. In our design, V is the concatenation of images from teacher and student, and T is the concatenation of texts from teacher and student.

Our design purpose is to minimize ℓ_{KL} and ℓ_{L2} and to maximize synergetic information $S(V,T;Y)$, so the total loss function combining the above metrics can be represented as:

$$\ell_{total} = \ell_{\text{contrastive}} + \alpha\ell_{\text{KL}} + \beta\ell_{\text{L2}} - \gamma S(V,T;Y) \tag{4}$$

In our design, we used $\alpha = 1.0$, $\beta = 0.3$, $\gamma = 0.7$.

4.2.2 Extension 2

In Extension 2, we introduce the redundant information into the loss function. In CLIP model distillation, maximizing redundant information may seem counterintuitive, as redundancy is often associated with inefficiency. However, in this context, redundancy refers to the reinforcement of critical features across multiple modalities, enhancing the model's robustness and generalization capabilities. By ensuring that both visual and textual components capture overlapping semantic information, the student model becomes more resilient to variations or noise in either modality. This redundancy allows the model to maintain performance even when one modality is compromised, as the other can compensate by providing similar information. Additionally, redundant representations facilitate better alignment between modalities, crucial for tasks requiring integrated vision and language understanding. Therefore, maximizing redundant information during CLIP distillation leads to a more robust, adaptable, and semantically aligned student model, capable of performing effectively across diverse and challenging scenarios. Our loss function in Extension 2 is

$$\begin{aligned}\ell_{total} \quad &= \quad \ell_{\text{contrastive}} + \alpha\ell_{\text{KL}} + \beta\ell_{\text{L2}} \\ &\quad -\gamma S(V,T;Y) - \epsilon R(V,T;Y) \quad (5)\end{aligned}$$

In our design, we used $\alpha = 1.0$, $\beta = 0.3$, $\gamma = 0.7$, and $\epsilon = 0.5$.

4.2.3 Performance of Two Extensions

The performance of the two extensions were evaluated on zero-shot image-to-text and text-to-image retrieval tasks based on the validation dataset only. The results are presented in Table 4. For comparison, we also include the performance of teacher model (trained ResNet50 CLIP model by OpenAI). Extension 1 used the loss function in (4); and Extension 2 used the loss function in (5).

These results indicate that the inclusion of KL-divergence, L_2 distance, synergetic, and redundant information in the training process enhances the student model's ability to generalize to unseen data, providing robust performance on zero-shot tasks. The inclusion of synergy and redundancy maximization prioritizes the joint predictive power of image and text embeddings, so the student model can better generalize to unseen data, enhancing retrieval and classification performance.

Task	Recall@1	Recall@5	Recall@10
Teacher Model (ResNet50)			
I2T	15.27	30.73	39.05
T2I	11.68	25.52	33.50
Student Model (Simple Baseline)			
I2T	0.22	0.97	1.83
T2I	0.33	1.39	2.50
Student Model (Strong Baseline)			
I2T	0.39	1.77	3.16
T2I	0.67	2.55	4.28
Student Model (Extension 1)			
I2T	1.04	3.96	6.66
T2I	1.00	3.80	6.38
Student Model (Extension 2)			
I2T	1.38	4.86	7.92
T2I	1.33	4.75	7.81

Table 4: Zero-shot retrieval performance comparison (in %) between the Teacher model (ResNet50) and Student models (ResNet34) under different training settings.

4.3 Hyperparameter Tuning

In our exploration of the Extension 1 model, we conducted hyperparameter tuning to identify which components of the extension had the most significant impact on model performance. The parameters considered included the learning rate, temperature, α, β, and γ. Specifically:

- Learning rate: Training loop weight update speed.
- Temperature: Scaling of logits in contrastive loss.
- α: Weight of the KL-divergence term.
- β: Weight of the L_2 distance term.
- γ: Weight of the synergetic information term.

Tuning was performed only on a random subset of the data (5000 training, 1000 validation) to reduce run-time. We tested 3 different values for learning rate and 2 for temperature, α, β and γ, respectively (48 combinations). As observed in the results presented in Table 5, the learning rate exhibits the most significant impact on improving Recall@K. Furthermore, γ also seems to exert a moderate influence compared to the other remaining parameters. This finding highlights the critical role of the training loop design and the incorporation of synergetic information in driving the model's performance.

Parameter	**Average Recall@10 Change**
Learning_rate	0.1906
γ	0.0105
Temperature	0.0064
α	0.0062
β	0.0038

Table 5: Average Recall@10 change for each parameter for extension 1 model.

4.4 Error Analysis

To evaluate the performance and limitations of our best-performing system (Extension 2), we conducted an error analysis by examining the errors made during I2T and T2I retrieval tasks. The analysis highlights the types of errors, and a comparison between the baseline and our extensions.
I2T Example Error:
Image: A photo of a person skiing down a snowy slope.
Predicted Caption: "A dog playing in the snow."
Correct Caption: "A person skiing down a snowy hill."
Error Type: Misidentification of the primary subject in the image.
T2I Example Error:
Query Text: "A bowl of fresh fruits on a wooden table."
Top Retrieved Image: An image of a bowl of soup on a kitchen counter.
Correct Image: An image of a bowl of fruits on a wooden table.
Error Type: Incorrect semantic matching of objects.

Our extensions significantly reduced errors in subject misidentification and contextual misalignment due to the integration of synergetic and redundant information. For example:
Baseline Error Example (Corrected by Extension 2):
Baseline Prediction: "A dog sitting on a porch."
Extension 2 Prediction: "A cat sitting on a windowsill looking outside."
Ground Truth: "A cat sitting on a windowsill looking outside."
Explanation: Extension 2 improved object identification by leveraging better modality alignment.

The error analysis highlights the strengths and limitations of our approach: Strengths: Improved multimodal alignment reduces major errors, particularly in subject identification and contextual relevance. Weaknesses: Challenges remain in handling fine-grained distinctions and ambiguity, which require further enhancements.

5 Conclusions

In this term project, we explored the distillation of the CLIP (Contrastive Language-Image Pretraining) model by transferring knowledge from a pretrained ResNet-50 teacher model to an untrained ResNet-34 student model. Our primary contributions include the design and implementation of novel loss functions inspired by PID, which incorporate synergetic and redundant information to improve the student model's performance on I2T and T2I retrieval tasks.

While our implementations did not achieve state-of-the-art performance, they demonstrated significant improvements over the published (strong) baseline student model trained with contrastive loss. The limitations in performance can be attributed to the use of an untrained ResNet-34 student model initialized with random weights, which lacks the learned feature representations required for effective image recognition tasks. This constraint highlights the challenge of training lightweight models from scratch without the benefit of pretraining.

Despite these limitations, the proposed loss functions showcased the potential to enhance multimodal alignment and robustness, paving the way for more efficient knowledge transfer in future work. Exploring alternative initialization strategies, incorporating pretraining for the student model, or extending the approach to larger architectures could help bridge the gap toward state-of-the-art performance.

Acknowledgements

The team would like to express our heartfelt gratitude to our TA, Artemis Panagopoulou, for her invaluable guidance and support throughout this project. Her insights and assistance were instrumental in helping us overcome challenges and achieve our goals. Without her expertise and encouragement, this project would not have been possible.

References

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Diederik P. Kingma and Max Welling. 2019. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392.

Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. 2023. Distilling large vision-language model with out-of-distribution generalizability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2492–2503.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

OpenAI. 2021. Clip: Contrastive language-image pretraining. `https://github.com/openai/CLIP`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, and Girish Sastry et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Paul L. Williams and Randall D. Beer. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint*, arXiv:1004.2515.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, and Hong Xuan et al. 2023. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21970–21980.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. 2024. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15952–15962.