Knowledge Distillation on Zero-Shot Model for Landmark Recognition

Kiet Chu
kietchu@umass.edu

Hung Pham
hdpham@umass.edu

Huy Cao
hcao@umass.edu

Abstract

Knowledge distillation has shown promise for transferring capabilities from larger models to more efficient ones, yet its effects on tasks like landmark recognition, both closed-domain classification and open-domain recognition, remain unexplored. In this project, we first train a large model for zero-shot classification in landmark recognition and subsequently employ knowledge distillation to transfer its capabilities to a smaller student model. Our approach involves distilling information separately for image and text encoders using logit matching, feature matching, and a teacher-assistant strategy. By independently refining each encoder and then integrating them, we aim to equip the student model with effective recognition capabilities for both landmark classification on a trained dataset and zero-shot recognition on an unseen dataset. Preliminary results suggest that this targeted distillation approach allows the student model to achieve competitive performance with improved efficiency, highlighting the potential of knowledge distillation for domain-specific applications.

1. Introduction

Landmark recognition is a challenging domain in computer vision, involving both closed-domain classification, where the objective is to correctly identify known landmarks from a predefined set, and open-domain recognition, where the model is expected to generalize and identify unseen landmarks. As the need for efficient and scalable models grows, especially for real-world applications, there is increasing interest in transferring knowledge from large, pre-trained models to more compact architectures while preserving predictive performance.

Traditional approaches to landmark recognition leverage large models capable of handling intricate features and semantic nuances. However, these models are computationally expensive and difficult to deploy in resource-constrained environments. Knowledge distillation offers a compelling solution by transferring the learned representations of a large teacher model to a smaller student model, thus achieving a balance between performance and efficiency. Yet, the specific impact of knowledge distillation on tasks like closed-domain and open-domain landmark recognition remains undiscovered.

In this work, we employ CLIP (Contrastive Language-Image Pre-Training) [4] as our baseline teacher model. We fine-tune CLIP for landmark recognition , achieving a model well-suited for extracting meaningful visual and textual representations. Our KD approach focuses on transferring CLIP's learned capabilities to a smaller, student model, optimizing it separately for the image and text encoders. Our main strategies are as follows:

- Logit Matching[2]: Aligning the student model's output logits with those of the CLIP teacher model.
- Feature Matching[2]: Minimizing the difference between feature representations produced by the teacher and student models.
- Teacher-Assistant Strategy[3]: Using an intermediate model (teacher assistant) to gradually transfer knowledge from CLIP to the student model for smoother and more effective distillation.

By distilling information separately for image and text encoders and later integrating them, we aim to equip the student model with efficient recognition capabilities. The distilled student model is evaluated on two main tasks: landmark classification on a closed-domain dataset and zero-shot recognition on an open-domain dataset.

We use Google landmark recognition dataset, which includes diverse scenes and landmarks for both training and evaluation. The performance of our student model is measured using metrics like accuracy for classification, allowing us to compare its efficiency and effectiveness against the CLIP baseline.

2. Related work

2.1. Zero-shot learning

A zero-shot model [6] focuses on making predictions or performing tasks on data it has never seen or been explicitly trained on. The term "zero-shot" refers to the model's ability to generalize to new tasks or categories without requiring any task-specific training examples. This is achieved through learning general representations or relationships that allow it to infer information about unseen data. Zero-

shot learning often relies on auxiliary information such as natural language descriptions, semantic embeddings, or attributes of unseen classes. In this paper, we will utilize CLIP [4] (Contrastive Language-Image Pretraining) which is a zero-shot model that can understand images based on textual descriptions to make predictions on our Google Landmark dataset. [5]

2.2. CLIP as a zero-shot image classification model

CLIP (Contrastive Language-Image Pre-Training) [4] is a model designed for zero-shot image classification that operates by learning a joint embedding space for visual and textual data. The model is trained on a large corpus of image-text pairs using a contrastive learning framework. This involves a dual encoder architecture, where separate neural networks are employed to process images and text, respectively. During training, the model optimizes a contrastive loss function that encourages alignment between the embeddings of matching image-text pairs while pushing apart the embeddings of non-matching pairs. As a result, CLIP can compute the similarity between an image embedding and the embeddings of class labels represented as text. This allows the model to classify images into categories that were not part of its training data, demonstrating its capability for zero-shot classification. The efficiency of CLIP lies in its ability to generalize across various tasks without the need for task-specific fine-tuning.

2.3. Knowledge Distillation

Originally proposed in a paper of Bucila, Caruana, and Niculescu-Mizil (2006) [1] and popularized by a paper of Hinton, Vinyls, and Dean (2015) [2]. Knowledge distillation is a technique used to compress large neural networks by transferring knowledge from a large "teacher" model to a smaller "student" model. The goal is to train the student model to mimic the performance of the teacher while significantly reducing the model size and computational requirements. This process involves using the teacher's output, either as hard labels (final predictions) or soft labels (probability distributions), to guide the student's training. This technique enables lightweight and efficient neural networks that can be deployed on resource-constrained devices such as mobile phones, IoT devices, or embedded systems, where the full-scale model would be impractical. Since then, knowledge distillation has been used as a reliable technique and widely adopted in a variety of deep learning models and tasks.

2.4. Google Landmarks Dataset

Weyand et al. [5] introduced the Google Landmarks Dataset v2 as a large-scale benchmark for landmark recognition and retrieval. The Google Landmarks Dataset v2 consists of over 5 million images representing more than 200,000 distinct landmarks. This dataset has been associated with two Kaggle challenges focused on these tasks. However, neither challenge addressed the complexities of large-scale models and zero-shot classification, which remain open areas for exploration. Due to hardware constraints, we employ a subset of 180,000 images covering over 3000 landmarks, ensuring our study remains computationally feasible while maintaining diversity within the subset for effective model training and evaluation.

3. Method

Describe the methods you intend to apply to solve the given problem.

3.1. Synthetic Caption Training

We create synthetic captions for each landmark in the follow format:
A photo of $\{$*name*$\}$*, which is a* $\{$*type*$\}$ *located in* $\{$*location*$\}$.
In this format, $\{$*name*$\}$ denotes the name of the landmark, $\{$*type*$\}$ denotes the category type of the landmark (e.g. castle, museum, park) and $\{$*location*$\}$ denotes either the city or country where the landmark is located. This structure helps the model understand landmark characteristics even when the name itself is unfamiliar.

3.2. Logit Matching and Feature Matching

We plan to implement these methods based on the foundational work of Hinton, Vinyals, and Dean in the paper Distilling the Knowledge in a Neural Network [2].

3.2.1 Logit Matching

The Logit Matching Loss using KL divergence can be expressed as:

$$\mathcal{L}_{\text{KLD}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \hat{p}_{i,j} \log \frac{\hat{p}_{i,j}}{p_{i,j}}$$

where:
- N is the number of samples,
- C is the number of classes,
- $\hat{p}_{i,j}$ is the target probability for the j-th class of the i-th sample,
- $p_{i,j}$ is the predicted probability for the j-th class of the i-th sample, calculated as $p_{i,j} = \text{softmax}(\mathbf{z}_i)_j$.

3.2.2 Feature Matching

The Feature Matching Loss using Mean Squared Error (MSE) can be expressed as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{f}_i - \hat{\mathbf{f}}_i \right\|^2$$

where:
- N is the number of samples,
- $\mathbf{f}_i$ is the feature vector from the model for the i-th sample,
- $\hat{\mathbf{f}}_i$ is the target feature vector for the i-th sample,
- $\|\cdot\|^2$ represents the squared Euclidean norm.

3.3. Knowledge Distillation Using Teacher Assistants (TAKD)

This method uses the same loss as the Logit Matching loss, which is the KL Divergence with the softened outputs of student model y_s = softmax($logits_s/\tau$) and the teacher model y_t = softmax($logits_t/\tau$)

$$\mathcal{L}_{\text{KD}} = \tau^2 \text{KL}(y_s, y_t)$$

where τ is a hyperparameter controlling the temperature scaling factor.
The teacher assistant and the student network are then trained under the following loss function:

$$\mathcal{L}_{\text{student}} = (1-\lambda)\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{KD}}$$

where λ is a second hyperparameter controlling the trade-off between the two losses.

3.4. Proposed Neural Knowledge Distillation (NKD) Loss Function

We introduced a novel Neural Knowledge Distillation (NKD) Loss function designed to facilitate more nuanced knowledge transfer between teacher and student models. The proposed loss function addresses the limitations of traditional knowledge distillation approaches by incorporating three distinct loss components: original loss, soft target loss, and distributed loss.[7]

3.4.1 Loss Function Components

The NKD Loss is mathematically formulated as:

$$L_{\text{NKD}} = L_{\text{ori}} + L_{\text{soft}} + \alpha \cdot L_{\text{distributed}} \tag{1}$$

Where:
- L_{ori} represents the original class probability loss
- L_{soft} captures the soft target knowledge transfer
- $L_{\text{distributed}}$ focuses on the probability distribution of non-target classes
- α is a hyperparameter controlling the distributed loss component's influence

3.5. Temperature-Scaled Probability Transformation

To effectively transfer "dark knowledge," we implemented temperature scaling:

$$P_{\text{scaled}} = \text{softmax}\left(\frac{\text{logits}}{\lambda T}\right) \tag{2}$$

Where:
- λT represents the temperature scaling factor
- Lower temperatures create sharper probability distributions
- Higher temperatures produce more uniform probability distributions

3.5.1 Original Loss Component

The original loss L_{ori} focuses on the correct class probability:

$$L_{\text{ori}} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{\text{student},y_i} + \epsilon) \tag{3}$$

Where:
- N is the batch size
- p_{student,y_i} is the student model's probability for the true class
- ϵ is a small constant to prevent logarithm of zero

3.5.2 Soft Target Loss Component

The soft target loss L_{soft} incorporates the confidence of the teacher model:

$$L_{\text{soft}} = -\frac{1}{N}\sum_{i=1}^{N} p_{\text{teacher},y_i}\log(p_{\text{student},y_i} + \epsilon) \tag{4}$$

3.5.3 Distributed Loss Component

The distributed loss $L_{\text{distributed}}$ captures non-target class probability distributions:

$$L_{\text{distributed}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq y_i} p_{\text{teacher},j}\log(p_{\text{student},j} + \epsilon) \tag{5}$$

3.5.4 Hyperparameter Configuration

Two critical hyperparameters were introduced:
- α: Scaling factor for distributed loss
- λT: Temperature Scaling Parameter

These hyperparameters enable fine-grained control over the knowledge transfer process, allowing precise tuning of the distillation mechanism.

3.5.5 Experimental Rationale

The proposed NKD Loss addresses key challenges in knowledge distillation:
- Captures nuanced probability distributions beyond point estimates

- Enables transfer of "dark knowledge" from complex teacher models
- Provides flexible knowledge transfer through temperature scaling
- Mitigates information loss during model compression

4. Cross-Modal Knowledge Distillation Framework

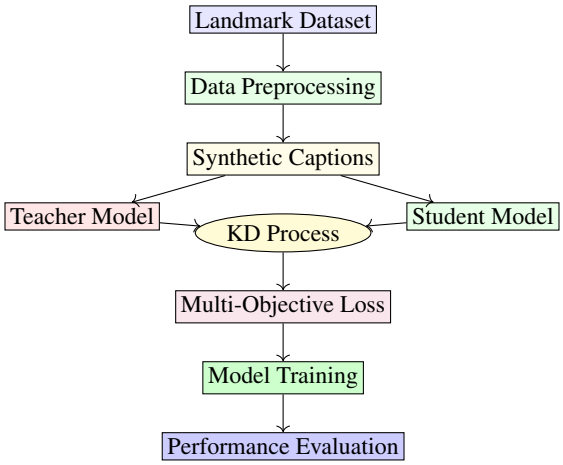

Figure 1. Knowledge Distillation Pipeline Workflow

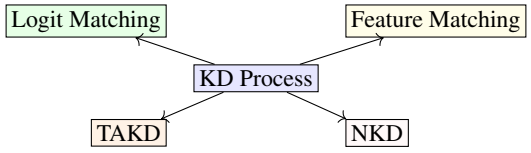

Figure 2. Components of the Knowledge Distillation process

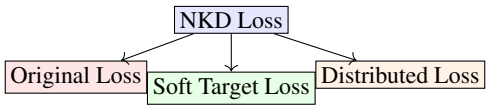

Figure 3. NKD Loss Components

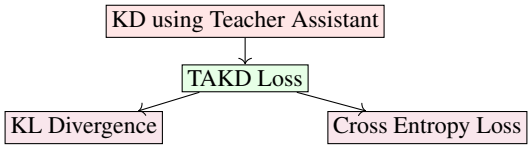

Figure 4. TAKD Loss Components

We developed a comprehensive cross-modal knowledge distillation framework designed to transfer sophisticated representations from a large teacher model (CLIP RN-101) to a computationally efficient student model (CLIP RN-18) using methods mentioned in the previous section.

4.1. Model Initialization

The models were initialized as follows:

- **Teacher Model:** The Teacher Model was initialized with pre-trained CLIP ResNet-101 (RN-101) weights. This pre-trained embedding space provides a starting point for aligning visual and textual modalities, enabling faster convergence of the student model during the distillation process. Although the pre-trained teacher provides a foundational embedding space, it is retrained as part of the process to better suit the specific task requirements, making high-quality supervision signals secondary to efficient initialization.
- **Student Model:** The Student Model utilized a hybrid architecture, with the text encoder initialized using pre-trained CLIP RN-101 weights and the image encoder initialized using ResNet18 weights. ResNet18 was selected as the image encoder for its computational efficiency and simplicity, which makes it well-suited for scenarios with limited computational resources while still maintaining competitive performance.

4.2. Training Configuration

By freezing the text encoder weights, both the Teacher Model and the Student Model preserved the general-purpose semantic understanding encoded in CLIP RN-101, reducing the risk of overfitting to task-specific data. The fine-tuning of the ResNet18 image encoder enabled the student model to learn domain-specific visual representations while maintaining alignment with the teacher model's embedding space. This initialization strategy balanced the need for computational efficiency with the preservation of robust generalization capabilities, ensuring the effectiveness of the student model in both supervised and zero-shot settings.

4.2.1 TAKD Training

For the Knowledge Distillation using Teacher Assistant method, we introduced an intermediate-sized CLIP model referred to as the Teacher-Assistant (TA). This TA helps bridge the gap between the teacher CLIP model and the student CLIP model, facilitating better knowledge transfer. In this paper, we will use CLIP RN-50 as the TA model to help with the knowledge distillation process. This architecture choice is considered to have the reasonable intermediate size between the teacher (CLIP RN-101) and the student (CLIP RN-18), and lies closer to the student in terms of accuracy. The TA model is first trained by the teacher and then used to train the student, making the knowledge transfer more efficient and successful, and for the student to achieve better accuracy.

4.3. Distillation Objectives

The knowledge transfer process incorporated a multi-objective loss function:

1. **Knowledge Distillation Loss (KD Loss):** Minimized divergence between teacher and student output probabilities using temperature-scaled soft targets.
2. **Feature Alignment Loss:** Utilized cosine similarity to synchronize intermediate feature representations between models.
3. **Cross-Entropy Loss (CE Loss):** Ensured student model predictions were aligned with ground-truth labels, enhanced with label smoothing techniques.
4. **Contrastive Loss:** Integrated an advanced contrastive mechanism to preserve intra-class feature similarities while maintaining inter-class feature separability.

This comprehensive approach enabled effective knowledge transfer, significantly improving the student model's performance while maintaining computational efficiency.

5. Results

5.1. Closed-Domain Landmark Classification

Table 1 is the validation accuracy of two baseline CLIP models (RN-101 and RN-18) after being trained on the Google Landmarks dataset. Given the extensive class count of the dataset of more than 3000, the models demonstrate promising performance. The result also shows the potential of having the RN-101 model as the teacher and the RN-18 model as the student for the knowledge distillation task.

Model	Top-1	Top-5	Top-10
CLIP RN-101	36.74	48.01	52.64
CLIP RN-18	27.02	45.49	53.24

Table 1. Validation Accuracy (%) on Closed-Domain Landmarks Recognition

Table 2 presents the validation accuracy of the student models (CLIP RN-18) on the Google Landmarks dataset after the knowledge distillation was performed. The results still show that the student models perform similar to the baseline model before implementing knowledge distillation process.

KD Method	Top-1	Top-5	Top-10
Logit Matching	26.63	44.60	52.38
Feature Matching	26.66	44.67	52.22
TAKD	25.89	41.32	47.58
NKD	20.74	30.23	32.54

Table 2. Validation Accuracy (%) for the Distilled RN-18 Model on Closed-Domain Landmarks Recognition.

5.2. Open-Domain Landmark Classification

Table 3 presents the accuracy of the baseline trained CLIP models on a dataset consisting of 2000 unseen landmarks. Although the accuracy is lower than its closed-domain performance, it still shows some potential results. The training data's class imbalance likely contributed to overfitting, causing the models to focus on dominant class characteristics and thereby reducing their ability to generalize effectively to new, unseen landmarks.

Model	Top-1	Top-5	Top-10
CLIP RN-101	4.04	11.77	17.08
CLIP RN-18	2.83	9.57	14.80

Table 3. Accuracy (%) of the baseline models on Open-Domain Landmarks Recognition

Table 4 presents the validation accuracy of the student models (CLIP RN-18) on an open-domain dataset with 2000 unseen classes (landmarks) after the knowledge distillation using teacher assistant was performed. Similar to the closed-domain classification, knowledge distillation process doesn't show any improvement in accuracy.

KD Method	Top-1	Top-5	Top-10
Feature Matching	2.73	9.62	14.56
TAKD	2.87	8.53	13.52

Table 4. Accuracy (%) of the Distilled Student Models on Open-Domain Landmarks Recognition.

5.3. Discussion and Limitations

The results highlight significant challenges in fine-grained landmark classification with large-scale datasets. On the closed-domain dataset, CLIP RN-101 outperformed RN-18, showcasing the benefits of a deeper architecture. However, knowledge distillation methods such as Feature Matching and TAKD failed to significantly improve the performance of RN-18, indicating the difficulty of transferring knowledge effectively in this context. On the open-domain dataset, which included unseen landmarks, all models exhibited low accuracy, reflecting poor generalization to new classes.

The knowledge distillation process also demonstrated notable shortcomings, particularly in transferring knowledge from the larger RN-101 teacher model to the smaller RN-18 student model. One key issue is the substantial architectural gap between the teacher and student models. The RN-18 model's limited capacity was insufficient to encapsulate the robust knowledge distilled from the RN-101, which is better equipped to capture fine-grained distinctions and complex image-text relationships. Additionally, traditional CLIP models, such as RN-50, RN-101, and RN-152,

rely on larger and more expressive architectures to learn robust embeddings, making it particularly challenging for the restricted RN-18 to match their performance. These limitations suggest that the student model was unable to retain the nuanced relationships necessary for effective classification, further hindered by the inherent imbalances and noise within the dataset.

The dataset itself is characterized by significant skewness in the distribution of landmark IDs, where some landmarks are heavily overrepresented while others are severely underrepresented. This imbalance introduces several challenges. During training, the model may overfit on overrepresented classes, leading to biased predictions and reduced performance on minority classes. Furthermore, evaluation metrics such as accuracy can be disproportionately influenced by the dominant classes, masking deficiencies in handling underrepresented ones. Skewness also impacts the model's ability to generalize, particularly in zero-shot settings where a balanced understanding of all classes is critical. Addressing these challenges requires strategic interventions, such as data augmentation to generate synthetic samples for minority classes, class-weighted loss functions to reduce training bias, and advanced evaluation metrics like macro-averaged precision and recall to better capture performance across all classes. These efforts are essential to ensure that the models are robust, fair, and capable of handling real-world datasets with inherent imbalances.

The large number of classes also increased complexity, making it harder for the models to distinguish between fine-grained differences. Addressing these challenges will require improved data quality, balanced sampling, and advanced techniques to enhance both knowledge transfer and generalization.

6. Conclusion and Future Work

This study explored the challenges of using knowledge distillation to transfer capabilities from a large teacher model, such as CLIP RN-101, to a smaller student model, like CLIP RN-18, for landmark recognition. While the approach showed promise, significant difficulties arose in effectively transferring knowledge between the teacher and student. The main issue was the large difference in capacity between the models, as the smaller RN-18 struggled to learn the detailed relationships and fine-grained features captured by the teacher. Additionally, imbalances and noise in the training dataset made it harder for the student model to generalize well, resulting in lower performance in both closed-domain and zero-shot recognition tasks. These results highlight the need to find a better balance between model efficiency and the ability to handle complex data.

Future work should focus on improving both the knowledge distillation methods and the quality of the training data. Refining how information is transferred between teacher and student models, such as using better feature alignment or contrastive learning techniques, could help reduce the performance gap. Addressing dataset challenges through methods like data augmentation, generating more examples for underrepresented classes, or using class-weighted loss functions could lead to fairer and more reliable training. Combining knowledge distillation with other techniques, such as model compression, might also allow the student model to retain more of the teacher's strengths. These steps can help create smaller, efficient models that still perform well, making them practical for real-world use in resource-limited settings while maintaining their ability to recognize both known and unseen landmarks.

References

[1] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Knowledge Discovery and Data Mining*, 2006. 2

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 1, 2

[3] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2

[5] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*, 2020. 2

[6] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly, 2020. 1

[7] Zhendong Yang, Zhe Li, Yuan Gong, Tianke Zhang, Shanshan Lao, Chun Yuan, and Yu Li. Rethinking knowledge distillation via cross-entropy, 2022. 3