Heterogeneous Architecture Knowledge Distillation

Musa Aftab, Tauseef Mohsin, Areeb Zahid, Jasir Hussain Khan

Abstract

Knowledge Distillation has emerged as an effective approach to transferring information from large, high-capacity teacher models to smaller, lightweight student models. This process enables the development of compact models suitable for deployment in resource-constrained environments without significant loss in performance. While much progress has been made in KD, challenges persist when the teacher and student architectures are heterogeneous, as differences in representational capacity and inductive biases hinder efficient knowledge transfer. Cross-architecture KD has gained attention as a means of addressing this gap, aiming to bridge the disparity between diverse model architectures such as convolutional neural networks and transformers. By aligning feature representations and optimizing the transfer process, we seek to maximize the utility of the teacher's knowledge while maintaining the efficiency of the student model. This paper explores the complexities of cross-architecture KD, discussing the challenges it poses and the opportunities it presents for advancing model compression and transfer learning.

1 Introduction

Knowledge Distillation (KD) has emerged as a pivotal technique in deep learning, enabling the transfer of knowledge from large, high-capacity models to smaller, computationally efficient models. This paradigm has gained increasing relevance in applications where deploying large models is impractical due to hardware constraints or energy limitations. Among recent advancements, the Contrastive Language-Image Pretraining (CLIP) model, which leverages a Vision Transformer (ViT) backbone, has demonstrated exceptional performance across a wide range of vision-language tasks. However, deploying such large-scale models on resource-constrained devices remains a significant challenge. Existing efforts in KD primarily focus on teacher-student models with similar architectures, such as transferring knowledge between convolutional neural networks (CNNs). However, the inherent representational differences between architectures like ViT and CNNs pose unique challenges for effective knowledge transfer. While CLIP's ViT excels at capturing global, contextual features, smaller CNN models are designed for localized, hierarchical feature extraction. This architectural disparity complicates the distillation process, highlighting a pressing need for novel approaches to cross-architecture KD. In this project, we aim to train a compact CNN model by distilling knowledge from a pretrained CLIP ViT model. Our objectives include exploring and evaluating the student model's performance and the change in the model's inherent ability of encoding semantic meaning. To achieve this, we propose a methodology that aligns features across different architectural paradigms while preserving semantic knowledge critical for generalization. By addressing the challenges of cross-architecture KD, this work seeks to advance the applicability of large-scale pretrained models in real-world scenarios where computational resources are limited. The anticipated outcome is a lightweight CNN capable of leveraging the rich semantic understanding of CLIP's ViT, thus balancing efficiency and performance effectively.

CS-6304 Advanced Topics in Machine Learning
Template from NeurIPS 2024

2 Related Work

Knowledge distillation (KD) has been widely employed to compress neural networks, enhancing their efficiency while maintaining performance. Traditional KD methods, such as those introduced by Hinton et al. [Hinton et al., 2015], focus on logits-based approaches where a student model mimics the soft outputs of a larger teacher model. These methods have proven effective for homogeneous architectures, where the teacher and student share similar structures, such as ResNet variants. However, they struggle to generalize in heterogeneous settings where the architectures of the teacher and student differ significantly [Hao et al., 2023, Liu et al., 2022].

In heterogeneous architecture knowledge distillation (HAKD), feature-based methods often encounter challenges due to discrepancies in feature representation across architectures. For example, convolutional neural networks (CNNs) encode spatially localized patterns, while transformers utilize self-attention mechanisms to capture global dependencies. These differences necessitate specialized strategies for effective feature alignment [Wu et al., 2024]. Existing solutions include One-for-All Knowledge Distillation (-KD), which introduces a projection mechanism to align features into a shared logits space. This approach reduces architecture-specific information, enabling more effective intermediate-layer distillation. Additionally, -KD integrates adaptive target enhancement to mitigate the influence of irrelevant information in the teacher's predictions, achieving notable performance gains on datasets such as CIFAR-100 and ImageNet-1K [Hao et al., 2023].

Another significant contribution in this domain is the ViT-to-CNN distillation framework, which employs partially cross-attention (PCA) and group-wise linear (GL) projectors. These mechanisms align the feature spaces of a transformer teacher and a CNN student by mapping the CNN features into the transformer's attention and feature spaces. This alignment enables the student to learn complementary global and local features, overcoming the structural discrepancies between the architectures [Liu et al., 2022].

Further advances include the Low-Frequency Component-based Contrastive Knowledge Distillation (LFCC) framework, which focuses on aligning low-frequency components of intermediate features to facilitate effective distillation across architectures. By extracting semantic information from these low-frequency components, LFCC minimizes noise and achieves alignment in a compact space. Additionally, LFCC employs sample-level contrastive learning, leveraging intra-sample similarities and inter-sample disparities to enhance the student model's performance. This method has demonstrated significant improvements on challenging benchmarks, outperforming other state-of-the-art techniques [Wu et al., 2024].

Relational Knowledge Distillation (RKD) shifts focus from direct feature alignment to relational knowledge transfer, capturing relationships between data points in the feature space. This method has shown promise in scenarios where direct feature-level alignment is infeasible due to substantial architectural differences [Park et al., 2019]. Similarly, Contrastive Relational Distillation (CRD) adapts contrastive learning to maximize mutual information between teacher and student features, making it highly effective for heterogeneous architectures [Tian et al., 2022].

Another noteworthy approach is the Instance Relationship Graph (IRG), which explores relationships among instance features, their transformations, and their relative distributions in the feature space. IRG has proven effective in bridging architectural gaps by emphasizing higher-order relationships rather than simple pointwise alignments [Liu et al., 2019].

Despite these advancements, challenges persist in fully bridging the representational gap between heterogeneous architectures. The aforementioned methods highlight the importance of innovative alignment strategies, robust feature extraction, and relational knowledge representation in addressing these challenges. By building on these approaches, our work aims to further enhance the effectiveness and generalization of knowledge distillation across diverse architectures.

3 Methodology

This project aims to investigate and analyze cross-architecture knowledge distillation (KD). The goal is to systematically evaluate the transfer of knowledge between diverse model architectures (CNNs and ViTs) using feature alignment. We aim to see if cross-architecture knowledge distillation allows for a model to learn past its inherent biases. i.e CNN becoming more shape (global feature) biased. The following subsections detail our approach, techniques, implementation specifics, and the conceptual rationale behind our choices.

3.1 Description of Approach

The primary objective of this study is to explore the transfer of knowledge across and within architectural paradigms. Specifically, we investigate:

- Knowledge distillation from CLIP's Vision Transformer (ViT) to Convolutional Neural Networks (CNNs).
- Knowledge distillation within the same paradigm (CNN-to-CNN).

3.2 Techniques and Procedures

The knowledge transfer process involves:

- Aligning student and teacher feature respresentations using two projectors:
  1. **Partially Cross-Attention (PCA) Projector:** Aligning the teachers global features with students corresponding features.
  2. **Group-Wise Linear (GL) Projector:** Aligning pixel-level feature maps between the teacher and student models.
- Evaluating the effects of distillation through metrics such as classification accuracy, bias robustness, feature similarity (CKA), and interpretability visualizations (e.g., GradCAM).

3.3 Conceptual Rationale

The choice of cross-architecture KD is motivated by the complementary strengths of ViTs (e.g., global attention) and CNNs (e.g., local attention). We employ a feature alignment approach with the PCA and GL projectors. By combining the projectors, the distillation process helps ensure that the student learns the internal features of the teacher instead of just mimicking the teacher's output.

3.4 Figures and Diagrams

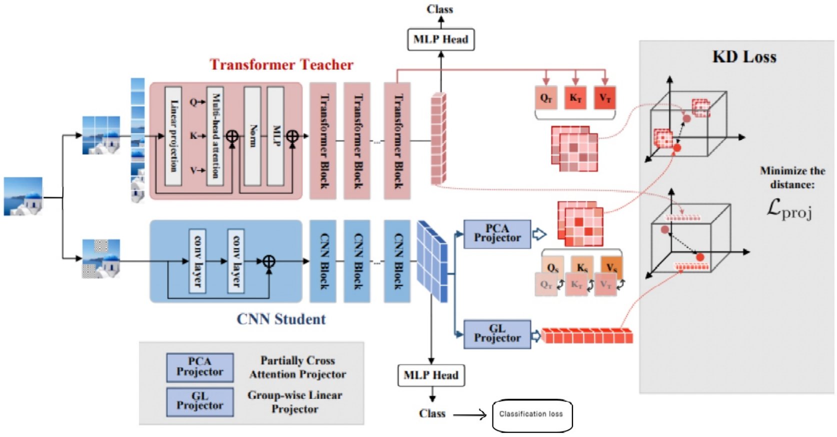

Figure 1: Knowledge Distillation Architecture Diagram

4 Experimental Design

This section outlines the experimental framework used to evaluate the proposed methodology for cross-architecture knowledge distillation (KD). The experiments aim to address key research questions related to the effectiveness and characteristics of knowledge transfer between diverse model types.

4.1 Research Questions or Hypotheses

The experiments are designed to address the following research questions:

1. How does cross-architecture KD influence the student performance (ViT B-16 Teacher to ResNet50 Student)?
2. How do KD and control students compare across various datasets and bias scenarios?
3. To what extent can KD transfer robustness and mitigate biases (i.e. shape, spatial, texture)

4.2 Experimental Setup

4.2.1 Datasets

The following datasets are used for training and evaluation:

- **CIFAR-10:** This dataset is widely used for classification and provides baseline image classification tasks.
- **STL-10:** A high-resolution image classification dataset, with occlations applied to the images to test for modeels reliance on global vs local feature.
- **FashionMNIST:** A dataset of fashion images used for training and testing, suitable for testing shape and texture biases in classification tasks. Texturized background to test for texture bias in models.

All datasets undergo standard preprocessing, including normalization and resizing to fit the input requirements of CNN and ViT models (e.g., 224x224 for ViT).

4.2.2 Teacher - Student Pairs and Control

- **ViT-B16 - ResNet-50:** Cross-Architecture Knowledge Distillation with teacher ViT-B16 and student ResNet-50.
  ViT-B16 was selected as it provides reliable global context understanding, which makes it robust to certain biases (e.g., occlusions or background textures).
  ResNet-50 was selected as it is a versatile and lightweight model, compared to the teacher. It relies on local spatial features, which help highlights the global reliance gained from the teacher.
- **Wide ResNet-101 - ResNet-50 :** Within-Architecture Knowledge with teacher Wide ResNet101 and student ResNet-50. This acts as the control baseline model.
  Wide ResNet-101 was selected as we wanted a CNN teacher model with as many or more parameters compared to the ViT we used. Wide ResNet-101 contains about 126 million parameters and ViT-B16 contains about 86 million parameters.
  ResNet-50 was selected to maintain the student for both cases to ensure, the only variable in control and treatment models was the teacher.

4.2.3 Evaluation Metrics

In order to understand the extent at which KD transfers the knowledge that mitigates biases, we make use of three metrics that qualitatively and quantitatively measure this bias mitigation. Namely:

1. **GradCAM Visualizations:** A powerful visualization tool that leverages gradients flowing back from the output layer to localize critical regions in the input. This enables an intuitive understanding of whether the knowledge-distilled (KD) student and the control student focus on the same regions as the teacher model. It is particularly useful for diagnosing model performance issues, especially when working with datasets prone to bias.
2. **Frequency Spectrum Analysis:** Investigates the frequency components of an image, distinguishing between low and high frequencies. Models with a bias toward texture tend to rely more on high-frequency components (High frequencies correspond to finer details, such as edges, textures, and small patterns within the image. These details are more localized and provide information about texture or minute differences in regions,) whereas those prioritizing global structure emphasize low-frequency components (Low frequencies capture the broader, smoother variations in an image, such as overall shapes, structures, and spatial relationships. These are features that provide global context and are critical for understanding the "big picture" of an image.) This method offers a quantitative framework

to assess whether KD students are learning robust global features, aligning their biases more closely with the teacher.

3. **Cross-Model Feature Alignment:** Centered Kernel Alignment (CKA) is a quantitative metric designed to measure the similarity of representations learned by different models. By comparing the internal features of the teacher and the KD student, CKA provides a direct measure of cross-model feature alignment. A higher CKA score indicates closer alignment between the student's features and the teacher's, making it a critical tool for evaluating the success of knowledge distillation.

Mathematical Definitions provided at the end.

4.3 Plan for Analysis

The analysis plan involves:

- Comparing student model performance (accuracy) against baselines and teacher models.
- Evaluating representational similarity using similarity scores across layers.
- Analyzing robustness under adversarial conditions to assess the transfer of robustness traits.
- Interpreting GradCAM visualizations to understand changes in attention focus post-distillation.

These analyses aim to provide comprehensive insights into the effectiveness and implications of cross-architecture KD.

5 Results and Findings

This section outlines the experimental findings and provides an analysis aimed at addressing the research questions. A detailed presentation of the results from our knowledge distillation approaches, along with the evaluations of the models, is included.

5.1 Results

The table below summarizes the accuracy of student models alongside their respective teachers trained using control and treatment configurations.

article booktabs

Table 1: Accuracy Comparison of Teacher and Student Models Across Different Architectures and Datasets

Data Set	**Architecture**	**Model**	**Accuracy (%)**
CIFAR-10	ViT-B16 → ResNet50	Teacher	91.68
		Student	78.17
	ResNet101 → ResNet50	Teacher	79.91
		Student	79.78
STL-10	ViT-B16 → ResNet50	Teacher	95.16
		Student	36.825
	ResNet101 → ResNet50	Teacher	90.00
		Student	26.65
FMNIST	ViT-B16 → ResNet50	Teacher	83.31
		Student	32.49
	ResNet101 → ResNet50	Teacher	76.21
		Student	45.86

The results highlight several key findings, which coincide with our treatments to the datasets to induce certain biased conditions (i.e. shape, spatial, and texture).

The changes to the datasets were as follows:

- **CIFAR-10**: No changes - used as control data set.
- **STL-10**: Occlusions added - tests local vs global biases
- **FMNIST**: Texture changes - tests high frequency/texture bias

We notice that the greatest improvement from the control to treatment was seen with the STL-10 dataset, which can be attributed to improved global spatial awareness of Visual Transformers being distilled to the CNN student, helping it deal with occlusions better.

5.2 Evaluation Metrics Results

We now move onto the actual evaluations performed on the teacher-student models. Note that we had a control model trained independently which we compared our student model's performances with.

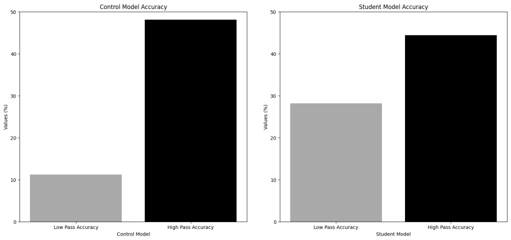

Figure 2: Student vs. Control Frequency Spectrum Analysis

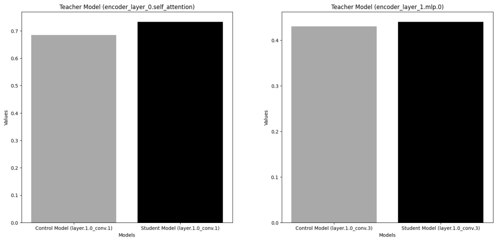

Figure 3: CKA Scores for 2 Initial Layers (Student and Control) compared with Teacher

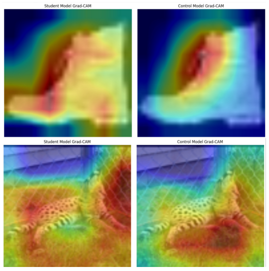

Figure 4: GradCAM Saliency Maps for Student vs. Control Model

6 Discussion

The Discussion section reflects on the implications of the findings, addressing their significance, limitations, and avenues for future exploration.

6.1 Evaluation Metric Analysis

1. **GradCAM Saliency Maps:** The student model demonstrates a greater focus on global features compared to the control model, as visualized through GradCAM. Although it isn't perfect, the student model covers considerably more ground on the saliency map than the control model.
2. **Fourier Spectrum Analysis:** The student model places increased emphasis on global (low-frequency) features (28% for the Student Model as compared to 11% for the Control Model,) reflecting an alignment with the teacher model's focus.
3. **CKA Score:** The student model exhibits higher CKA scores in its initial layers compared to an independent model (0.73 for Student and 0.68 for Independent) indicating the degree of feature alignment of the CNN student trained by the ViT teacher. However, these scores decrease significantly in deeper layers due to incomplete training, attributed to high computational requirements.

Overall, these results highlight that the student model effectively balances attention to both global and local features, rather than favoring one exclusively. The evaluation metrics confirm that the desired knowledge, aimed at mitigating or enhancing specific biases, can be successfully distilled into a smaller model.

6.2 Strengths and Contributions

This project makes the following significant contributions:

- Demonstrates that biases inherent in different architectures, such as global attention in ViTs and local feature extraction in CNNs, can be successfully distilled across architectures.
- Introduces a robust evaluation framework for comparing knowledge distillation methods across architecture types.
- Provides insights into the robustness of models trained through distillation, showing that trained students exhibit improved resilience to adversarial noise, and biases.

6.3 Limitations

While the findings are promising, several limitations warrant discussion:

- **Computational Constraints:** The experiments required substantial computational resources, which restricted the exploration of larger datasets or architectures beyond ViT and CNN variants.
- **Bias Adaptation:** Although effectively transfers biases, the extent of adaptation may depend on the complexity of the source biases. For example, certain ViT-specific biases, such as global context aggregation, were not fully leveraged by CNN students.
- **Generalization Scope:** The results are based on specific datasets, and it remains unclear whether these findings generalize to other domains, such as medical imaging.

6.4 Future Work

Building on this study, future research could explore the following directions:

- Extend the analysis to more diverse teacher-student configurations, such as distilling knowledge from hybrid architectures or lightweight models designed for edge devices.
- Investigate additional techniques for adapting cross-architecture biases, such as Contrastive Learning or Logit Matching.
- Evaluate the robustness of trained students in real-world scenarios involving domain shifts, adversarial attacks, and noisy labels.

6.5 Broader Implications

The findings of this study have several broader implications:

- The ability to distill biases across architectures opens new opportunities for model compression and efficiency, particularly for deploying ViT-like capabilities on resource-constrained devices.

By addressing these aspects, this study not only advances understanding in cross-architecture knowledge distillation but also lays the groundwork for impactful applications and future innovation.

7 Conclusion

This project explored the domain of cross-architecture knowledge distillation. It was an analytic view on whether a teacher can transfer its inherent biases to the student model, improving the student ability to view global features. The cross-architecture KD was implemented between ViT-B16 (teacher) and ResNet-50 (student) using PCA and GL projectors. Within-architecture KD was used between two CNNs (Wide ResNet-101 and ResNet-50) to create a control to help in analysis. The treatment model showed reduced bias towards local features by inheriting global attention capabilities from the ViT teacher. It should improvements in robustness compared to our control baseline.

- **Experimental Insights**
  - **Cifar-10:** Showed performance gains in a standard classification.
  - **STL-10:** KD student showed better robustness under spatial occlusions compared to the control student.
  - **FMNIST:** Control student performed better due to control model being more texture biased, thus performing better on the texturized dataset.
- **Evaluation Metrics**
  - **GradCAM:** Showed a more global-biased performance for the Student Model.
  - **Frequency Spectrum Analysis:** Showed lower frequencies (global features) were given more attention by the student model.
  - **CKA:** Higher CKA scores showed the alignment of ViT teacher with the CNN student as compared to the Independent model, quantifying the degree of alignment.

The key takeaway was cross-architecture KD is a viable method to transfer strengths and inherent biases (e.g., global and local feature learning) between ViTs and CNNs as well as the use of PCA and GL in aligning teacher-student features, effective knowledge transfer.
The experiments required significant computation resources, which limited the exploration of additional dataset and architectural configurations. Further, as training was conducted for few epochs (10 vs 200), the difference between the control and treatment were not as drastic as they can be. The results obtained are based on specific dataset and the generalization of these findings to other domain remains unclear

Definitions and Formulae

1. Centered Kernel Alignment (CKA)

Centered Kernel Alignment (CKA) is a metric used to measure the similarity between representations learned by two models. It is based on comparing the Gram matrices of the representations. Let the representations from the two models be denoted as $\mathbf{A} \in \mathbb{R}^{n \times d_1}$ and $\mathbf{B} \in \mathbb{R}^{n \times d_2}$, where n is the number of samples and d_1, d_2 are the respective feature dimensions.

Gram Matrices

The Gram matrix for a representation $\mathbf{X}$ is computed as:

$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ captures the pairwise similarity between the samples in $\mathbf{X}$.

Centering the Gram Matrices

To ensure alignment is not influenced by global offsets, the Gram matrices are centered. The centered Gram matrix $\mathbf{K}_c$ is computed as:

$$\mathbf{K}_c = \mathbf{K} - \frac{1}{n}\mathbf{1}_n\mathbf{K} - \frac{1}{n}\mathbf{K}\mathbf{1}_n + \frac{1}{n^2}\mathbf{1}_n\mathbf{K}\mathbf{1}_n$$

where $\mathbf{1}_n$ is an $n \times n$ matrix of ones.

CKA Similarity Metric

The similarity between the two representations is then calculated using the Hilbert-Schmidt Independence Criterion (HSIC). The CKA similarity is given by:

$$\text{CKA}(\mathbf{A}, \mathbf{B}) = \frac{\text{HSIC}(\mathbf{A}, \mathbf{B})}{\sqrt{\text{HSIC}(\mathbf{A}, \mathbf{A}) \cdot \text{HSIC}(\mathbf{B}, \mathbf{B})}}$$

where HSIC is defined as:

$$\text{HSIC}(\mathbf{A}, \mathbf{B}) = \text{Tr}(\mathbf{K}_c\mathbf{L}_c)$$

Here, $\mathbf{L}_c$ is the centered Gram matrix for $\mathbf{B}$, and $\text{Tr}(\cdot)$ denotes the trace of a matrix.

Interpretation

The value of $\text{CKA}(\mathbf{A}, \mathbf{B})$ lies in the range $[0, 1]$, where a higher value indicates greater similarity between the representations learned by the two models. This metric is particularly useful for evaluating feature alignment in tasks such as knowledge distillation.

2. Frequency Spectrum Analysis

Frequency Spectrum Analysis is a technique used to study the frequency components of an image. It is particularly useful for evaluating whether models are biased toward high-frequency (texture) or low-frequency (global structure) features.

Fourier Transform

The frequency components of an image are obtained using the Discrete Fourier Transform (DFT). For a 2D image $\mathbf{I} \in \mathbb{R}^{h \times w}$, the DFT is defined as:

$$\mathcal{F}(\mathbf{I})(u, v) = \sum_{x=0}^{h-1}\sum_{y=0}^{w-1} \mathbf{I}(x, y) \cdot e^{-2\pi i\left(\frac{ux}{h} + \frac{vy}{w}\right)}$$

where (u, v) are the frequency coordinates, and i is the imaginary unit.

The result of the DFT is a complex-valued matrix representing the amplitude and phase of each frequency component:

$$\mathcal{F}(\mathbf{I}) = \mathbf{M}e^{i\phi}$$

where $\mathbf{M} = |\mathcal{F}(\mathbf{I})|$ is the magnitude spectrum, and ϕ is the phase.

Filtering Frequency Components

To isolate specific frequency ranges, we use masks applied to the magnitude spectrum: - **Low-pass filter**: Retains low frequencies (global structure) within a radius r from the center of the frequency domain:

$$\mathbf{M}_{\text{low}}(u, v) = \begin{cases} \mathbf{M}(u, v) & \text{if } \sqrt{(u - c_u)^2 + (v - c_v)^2} \leq r \\ 0 & \text{otherwise} \end{cases}$$

- **High-pass filter**: Retains high frequencies (texture) outside a radius r from the center:

$$\mathbf{M}_{\text{high}}(u, v) = \begin{cases} \mathbf{M}(u, v) & \text{if } \sqrt{(u - c_u)^2 + (v - c_v)^2} > r \\ 0 & \text{otherwise} \end{cases}$$

where (c_u, c_v) is the center of the frequency domain.

Reconstruction of Filtered Images

The filtered image is reconstructed by applying the Inverse Discrete Fourier Transform (IDFT) to the modified spectrum:

$$\mathbf{I}_{\text{filtered}}(x,y) = \sum_{u=0}^{h-1}\sum_{v=0}^{w-1} \mathbf{M}_{\text{filtered}}(u,v)e^{i\phi} \cdot e^{2\pi i\left(\frac{ux}{h}+\frac{vy}{w}\right)}$$

Quantitative Evaluation

To evaluate model performance on different frequency components: 1. Apply the low-pass and high-pass filters to the test images. 2. Measure model accuracy on the filtered datasets:

$$\text{Accuracy}_{\text{low}} = \frac{\text{\# Correct Predictions on Low-pass Images}}{\text{Total Images}}$$

$$\text{Accuracy}_{\text{high}} = \frac{\text{\# Correct Predictions on High-pass Images}}{\text{Total Images}}$$

Interpretation

- High accuracy on low-pass images indicates the model relies on global, structural features.
- High accuracy on high-pass images suggests the model emphasizes local, texture-based features.

Loss Functions

For, CNN student trained by ViT:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda(\mathcal{L}_{\text{PCA}} + \mathcal{L}_{\text{GL}})$$

Where:

- $\mathcal{L}_{\text{cls}}$: Classification loss (Cross-Entropy Loss)
- $\mathcal{L}_{\text{PCA}}$: PCA loss (Mean Squared Error between PCA-projected student and teacher features)
- $\mathcal{L}_{\text{GL}}$: GL loss (Mean Squared Error between GL-projected student and teacher features)

For CNN to CNN KD, classification and hint loss were used.

References

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation, 2023. URL https://arxiv.org/abs/2310.19444.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting Ding, and Liang Li. Cross-architecture knowledge distillation, 2022. URL https://arxiv.org/abs/2207.05273.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation, 2019. URL https://arxiv.org/abs/1904.05068.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2022. URL https://arxiv.org/abs/1910.10699.

Hongjun Wu, Li Xiao, Xingkuo Zhang, and Yining Miao. Aligning in a compact space: Contrastive knowledge distillation between heterogeneous architectures, 2024. URL https://arxiv.org/abs/2405.18524.