Feature Learning-based Knowledge Distillation to train Teacher and Student Networks

Akshat Maithani (21UCC014)
Akshay Anand(21UCC015)
Patel Het Manojkumar (21UCC125)
Supervisor: Dr Upendra Pratap Singh

Feature Learning-based Knowledge Distillation to train Teacher and Student Networks

Akshat Maithani
(21UCC014)
Akshay Anand(21UCC015)
Patel Het Manojkumar (21UCC125)
Supervisor: Dr Upendra Pratap Singh

The LNM Institute Of Information Technology Jaipur

Feature Learning-based Knowledge Distillation to train Teacher and Student Networks

Akshat Maithani (21UCC014)
Akshay Anand(21UCC015)
Patel Het Manojkumar (21UCC125)
Supervisor: Dr Upendra Pratap Singh

Introduction

Feature Learning-based Knowledge Distillation to train Teacher and Student Networks

Deep learning models are often too large for resource-limited devices. Knowledge distillation addresses this by training a smaller model(student) with insights from a larger model(teacher). [1]

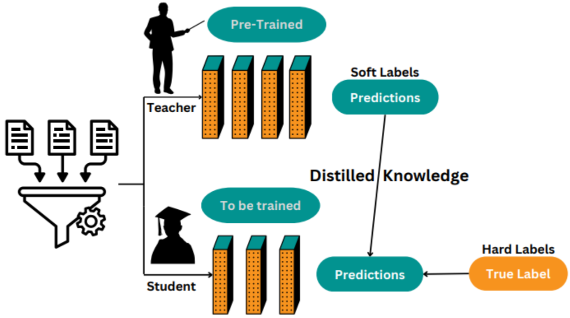


Figure: Model Distillation Architecture [2]

Motivation

Why Knowledge Distillation?

- Enhance the efficiency and deployment of models.
- Facilitate integration with other technologies.
- To enable the execution of models in environments with limited computational resources.

Mathematical Modeling

Formal Definition of KD

Neural networks use a softmax function to generate logits (zi) (output before softmax) to class probabilities.

$$\boxed{\sigma(z_i, T) = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}}$$

Here i,j=0,1,2,..,C-1 where C is the number of classes. T is temperature which is normally set to 1.[2]

Mathematical Modeling

Distillation Loss

$$\boxed{L_{\text{distillation}} = \alpha \cdot L_{\text{soft}} + (1 - \alpha) \cdot L_{\text{hard}}}$$

where $L_{\text{distillation}}$ represents the total loss function, which is a combination of two components:

- L_{hard}: Categorical cross-entropy loss computed between the true labels (y_{true}) and the student's predictions (y_{pred}).
- L_{soft}: Categorical cross-entropy loss computed between the softened outputs of the teacher model (teacher preds) and the softened predictions of the student model (y_{pred}).

The parameter $\alpha \in [0, 1]$ is a weight that balances the contributions of L_{hard} and L_{soft}.

Mathematical Modeling

Distillation Loss with Nuclear Norm

The distillation loss can be written as:

$$\boxed{L_{\text{distillation}} = \alpha \cdot L_{\text{soft}} + (1-\alpha) \cdot L_{\text{hard}} + \beta \cdot \|W\|_*}$$

Here,

- L_{soft} is the soft loss,
- L_{hard} is the hard loss,
- $\|W\|_*$ is the nuclear norm of the first layer's weights of the student model,
- α is the weight parameter balancing soft and hard losses,
- β is the regularization weight (set α for simplicity).

Mathematical Modeling

Significance Of Nuclear Norm

- **Promotes Simplicity:** Encourages simpler representations by focusing on the most important features of data.
- **Noise Reduction:** Helps suppress irrelevant details or noise, highlighting meaningful patterns.
- **Improves Generalization:** Prevents overfitting by regularizing the model's weights, making it better at handling new, unseen data.

Akshat Maithani (21UCC014)
Akshay Anand(21UCC015)
Patel Het Manojkumar (21UCC125)
Supervisor: Dr Upendra Pratap Singh

- **1. Precision**
  - **Definition**: Precision measures the proportion of true positive predictions among all positive predictions. It indicates how many of the predicted positive cases were actually correct.
  - **Mathematical Formula**:

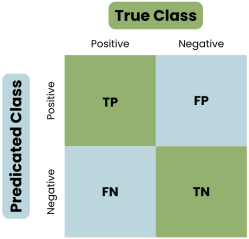
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **2. Recall (Sensitivity)**
  - **Definition**: Recall measures the proportion of true positive cases that were correctly identified by the model.
  - **Mathematical Formula**:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **3. Accuracy**
  - **Definition**: Accuracy is the ratio of correctly predicted observations to the total observations.
  - **Mathematical Formula**:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Number of Samples}}$$



- **4. F1 Score**
  - **Definition**: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall.
  - **Mathematical Formula**:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **5. Inference Time**
  - **Definition**: Inference time is the time it takes for the trained model to make a prediction on new data.

Literature Survey

Classification

Knowledge Distillation can be subdivided into:

1. **Offline Distillation** is the process of training a smaller student model using a pre-trained larger teacher model's outputs as targets.[2]
2. **Online Distillation** is a training technique where both the teacher and student models are updated simultaneously, allowing for real-time learning and improvement of the student model's performance.[2]
3. **Self-Distillation** involves a model learning from its own predictions, using them to generate soft labels for its data, and then refining itself based on these labels, resulting in enhanced accuracy and robustness.[2]

Literature Survey

Research Prospects

- Image Classification: For incomplete and ambiguous images knowledge distillation is proposed to increase efficiency for complex image classification.[2]
- NLP: Future advancements in KD can lead to better model compression, deployment, and efficient training.[2]
- Object Detection: Future advancements can lead to better efficieny, speed, adaptability, and generalization.[3]

Datasets Used

Akshat Maithani (21UCC014)
Akshay Anand(21UCC015)
Patel Het Manojkumar (21UCC125)
Supervisor: Dr Upendra Pratap Singh

Dataset	Classes	Images	Image Size	Source	Application Focus
UC Merced	21	2,100	256 × 256	USGS National Map	Land-use pattern analysis
EuroSAT	10	27,000+	64 × 64	Sentinel-2	Land cover classification
AID	30	∼10,000	600 × 600	Google Earth	Land-use classification
NWPU-RESISC45	45	31,500	256 × 256	Google Earth	Scene classification

Table: Datasets Used

Data Preprocessing

- **Image Resizing:** All images are resized to 32×32 pixels to:
  - Reduce memory usage for handling large datasets.
  - Speed up computations during model training and inference.
  - Ensure consistent dimensions for batch processing in neural networks.
- **Normalization:** Pixel values are normalized to a [0, 1] range for faster and more stable model convergence.

Results

Model Distillation On UC Merced Dataset

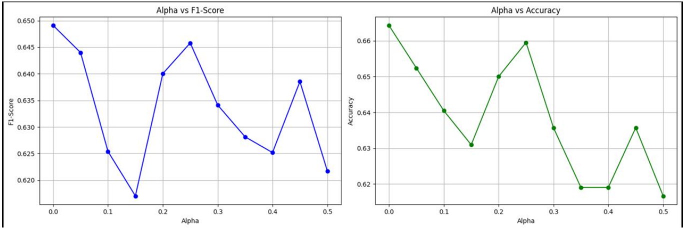


Figure: Plots of Alpha vs F1-Score and Alpha vs Accuracy. The x-axis represents the alpha values, and the y-axis represents the F1-score and Accuracy.

The accuracy of the teacher model is 63.57%. The plot illustrates that both α and $1 - \alpha$ play significant roles in the student model's performance. For certain values of α, the F1-score and accuracy are higher, while for others, they are lower. This highlights the balanced contribution of both α and $1 - \alpha$.

Results

Model Distillation On UC Merced Dataset Considering Nuclear Norm

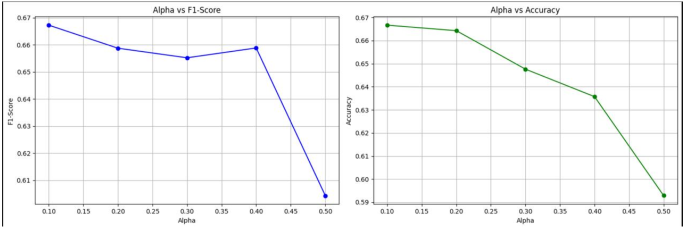


Figure: Plots of Alpha vs F1-Score and Alpha vs Accuracy. The x-axis represents the alpha values, and the y-axis represents the F1-score and Accuracy.

The accuracy of the teacher model is 65.48%. The plot illustrates that both α and $1 - \alpha$ play significant roles in the student model's performance. For certain values of α, the F1-score and accuracy are higher, while for others, they are lower. This highlights the balanced contribution of both α and $1 - \alpha$.

Results and Analysis

Performance Metrics from Model Distillation

5 Alpha UC Merced	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.68	0.65	0.65	20.66
Student	0.71	0.69	0.69	5.81

11 Alpha UC Merced	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.65	0.64	0.63	41.13
Student	0.65	0.65	0.64	4.26

Model	Alpha Values	Accuracy	F1 Score
11 Alpha UC Merced	0	0.66	0.65
	0.05	0.65	0.65
	0.1	0.64	0.63
	0.15	0.63	0.62
	0.2	0.65	0.64
	0.25	0.66	0.65
	0.3	0.64	0.63
	0.35	0.62	0.63
	0.4	0.62	0.63
	0.45	0.64	0.65
	0.5	0.62	0.62
5 Alpha UC Merced	0.1	0.67	0.67
	0.2	0.66	0.66
	0.3	0.65	0.66
	0.4	0.64	0.66
	0.5	0.59	0.6

Figure: Tabular Representation of Alpha Values vs Performance Metrics. Metrics include Accuracy, F1-Score, Precision, Recall, and Inference Time.

Detailed plots and tables for other datasets are available in the project report for further reference.

A. Alkhulaifi, F. Alsahli, and I. Ahmad, "Knowledge distillation in deep learning and its applications," *PeerJ Computer Science*, vol. 7, p. e474, 2021.

J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus,

S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

L. Wang, Y. Chen, X. Wang, R. Wang, H. Chen, and Y. Zhu, "Research on remote sensing image classification based on transfer learning and data augmentation," in *International Conference on Knowledge Science, Engineering and Management*, pp. 99–111, Springer, 2023.

X. Liu, K. H. Ghazali, F. Han, and I. I. Mohamed, "Review of cnn in aerial image processing," *The Imaging Science Journal*, vol. 71, no. 1, pp. 1–13, 2023.


THANK YOU