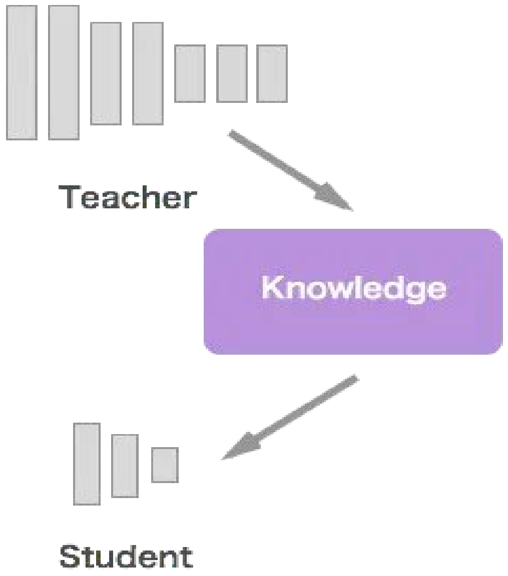
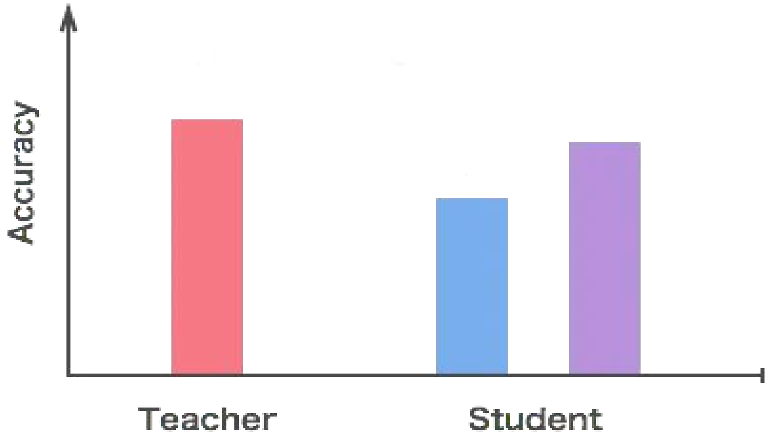
Knowledge Distillation on Neural networks

- Poornima Devi Krishnasamy Karthikeyan -





Contents :

- Introduction
- Knowledge Distillation (KD)
  - Types of Knowledge Distillation
  - Knowledge Distillation Schemas
- Methodology
  - Terms
  - Hybrid approach
  - Loss function
- Dataset information
- Implementation :
  - Teacher model – Resnet5O
  - Student model – Resnet18
    - Hyperparameter tuning
  - Student model without KD
- Results

Introduction:

Problem:

- Large deep learning models ☾ high accuracy and cost. Real-world deployment ☾ smaller, faster models.

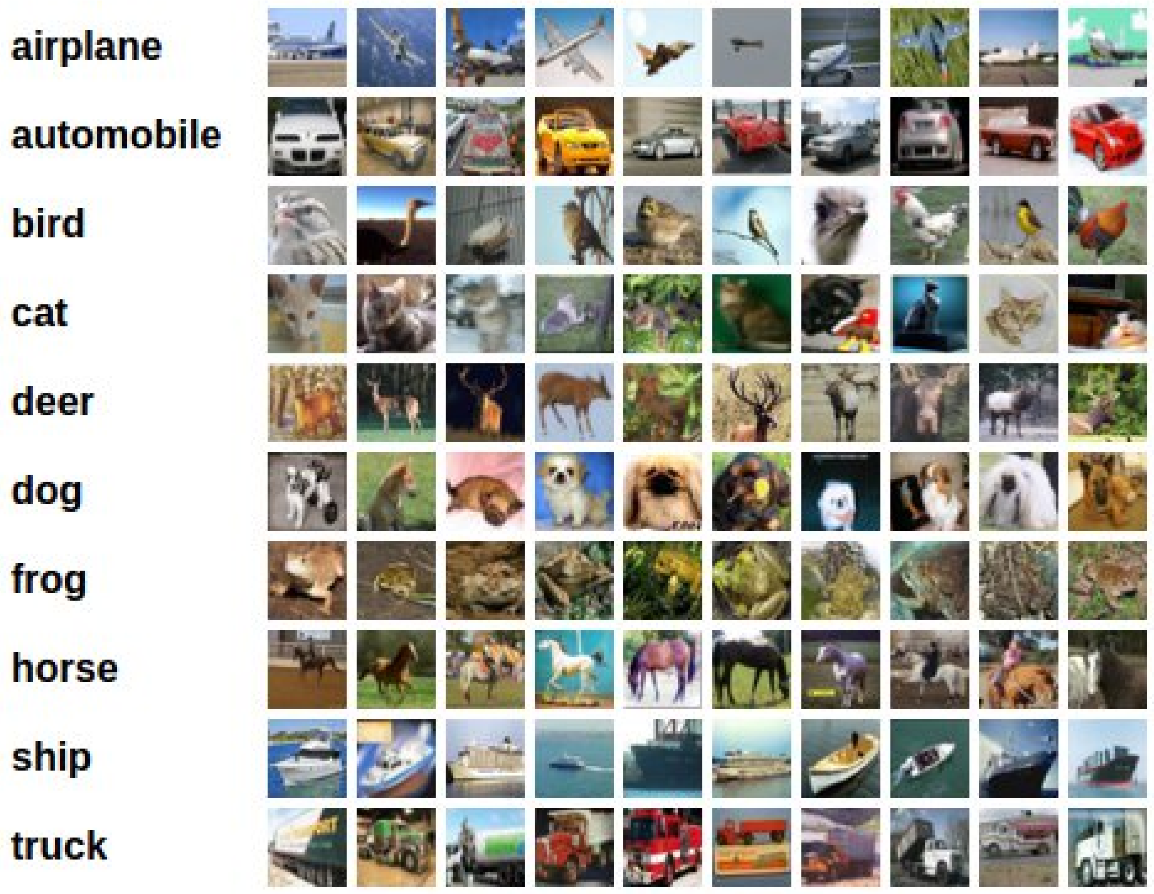
Goal :

- Knowledge distillation ☾ lighter models for resource constraint deployment without losing the performance.
  - Transferring knowledge from a large model (Teacher) to a small model (Student).

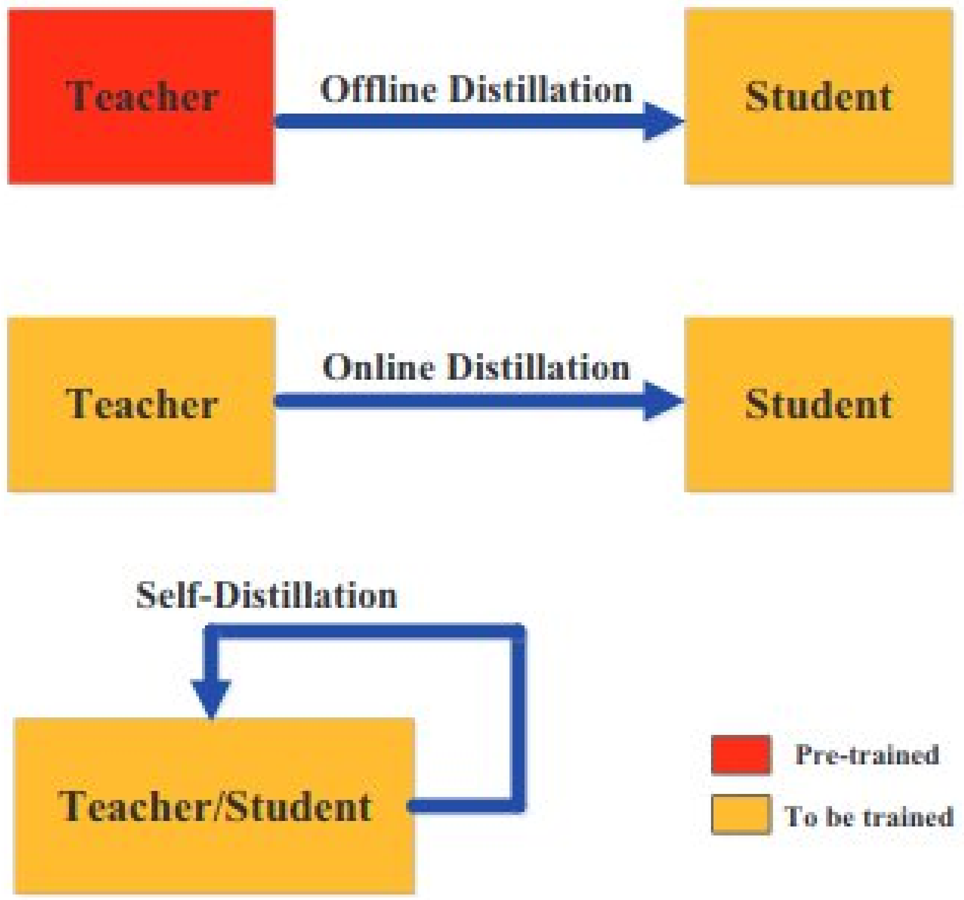
Dataset Information:

CIFAR-10 :

- 5000 training images
- 1000 validation images
- 2 x 32 x 32 images
- 10 classes

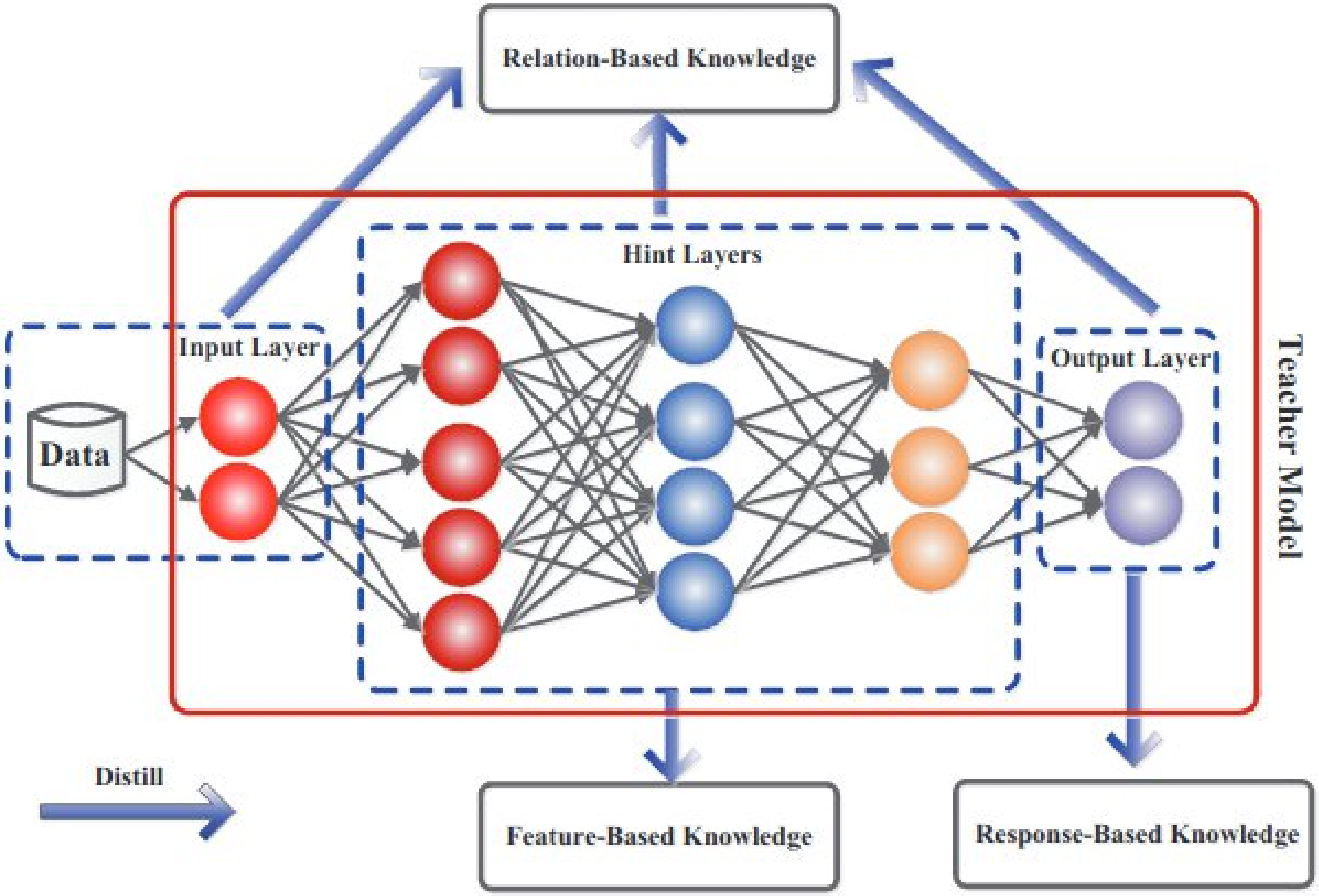


Knowledge distillation schemes:



Offline distillation method is used for the further implementation of KD

Types of Knowledge Distillation:



Methodology : Terms

Softmax :

- raw logits ☾ probability distribution [0.95, 0.02, 0.01, 0.01, 0.01] ☾ Not informative beyond the top prediction.

$$P_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

Temperature (T) :

- Softens the probability output from softmax [0.85, 0.05, 0.04, 0.03, 0.03]

$$\frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

  - Higher T → softer, more informative distribution

Methodology : Terms

Soft Labels :

- Teacher's output probabilities [0.4, 0.3, 0.2, 0.1] ☾ richer info than just the correct class.

Hard Labels :

- One-hot true labels [1, 0, 0, 0]) ☾ standard classification loss

Alpha (α):

- Controls the balance between:
    - Distillation loss (soft labels)
    - Classification loss (hard labels)

Methodology - Knowledge Distillation (KD):

Model architecture:

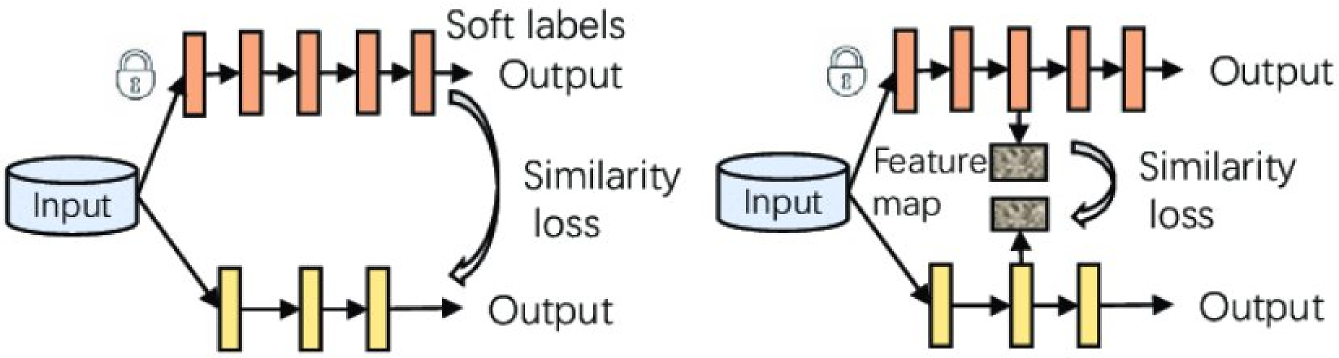
- Teacher : large pre-trained model (ResNet50) + fine tuned to CIFAR 10
- Student : (ResNet18)

How I trained my student model :

Hybrid Approach = response based KD + feature based KD

- soft outputs ☾ teacher
- Align intermediate features ☾ representation alignment
- Harc

Response based KD



Feature based KD

Methodology – Loss function

Loss Components:

- Distillation Loss: KL divergence loss
- Classification Loss: Cross entropy loss
- Feature Loss: MSE

Total Loss:

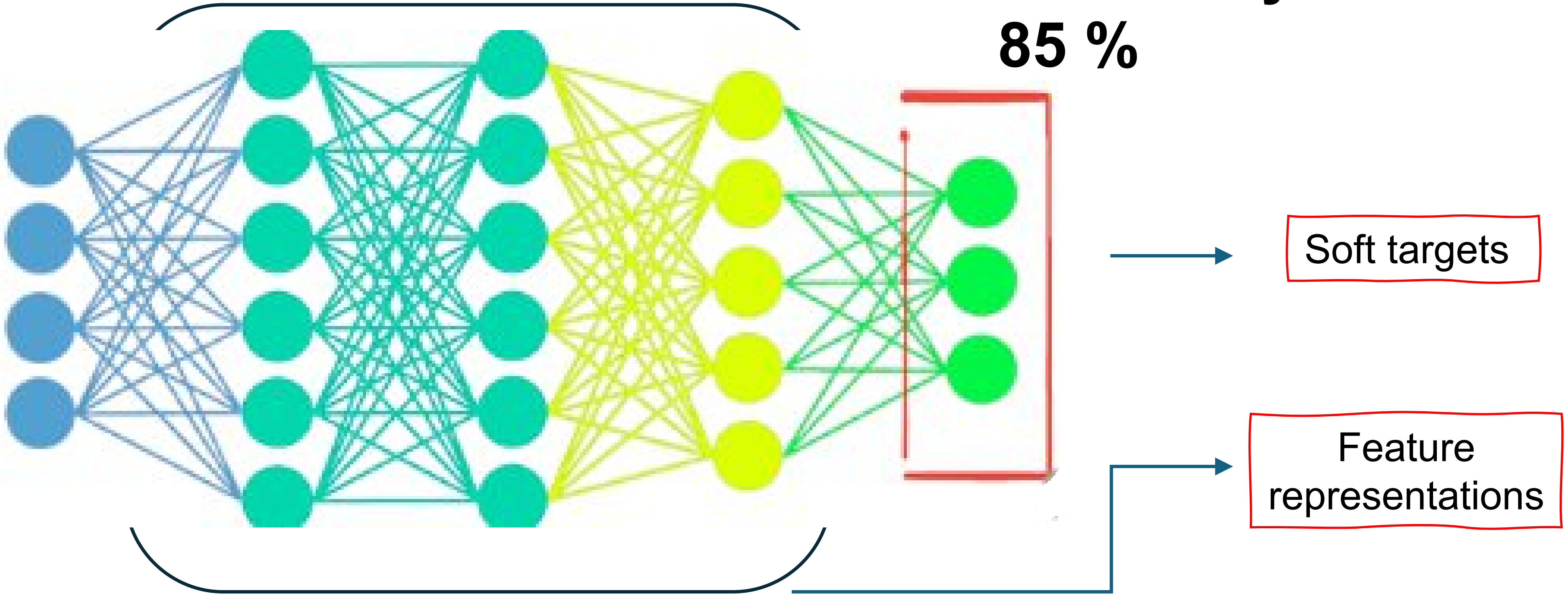
$\mathcal{L}$ = α × Distillation Loss + (1 − α) × Classification Loss + 0.1 × Feature loss.

Hyperparameters:

- Temperature: T
- Weight Balance: α
- Feature Weight: λ = 0.1

Implementation – Teacher model

Accuracy : 85 %



Input ☾ ResNet50 (Teacher) ☾ extracts Soft Targets + Intermediate Features information

Implementation – Teacher ☾ student

Teacher ☾ student : after projection ☾ [64,128,256,512]

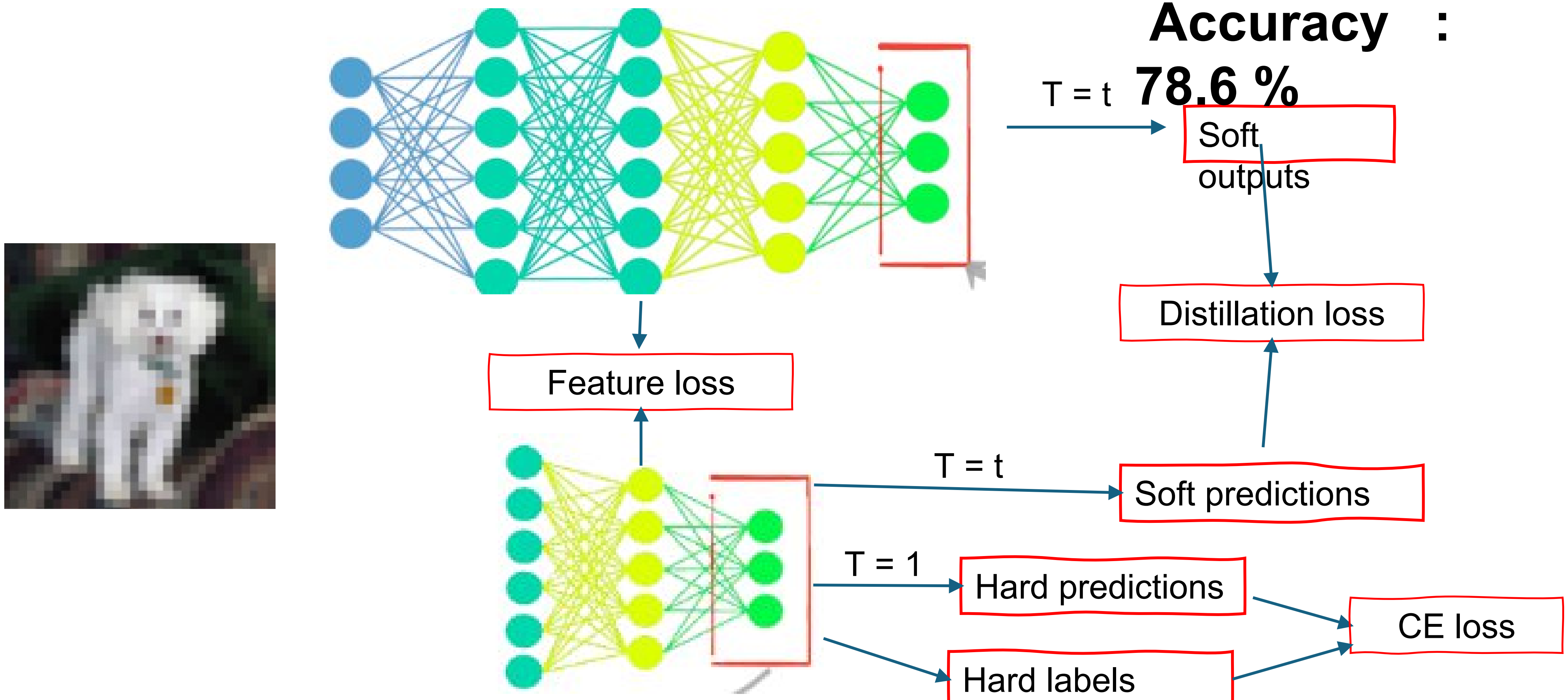
- FeatureProjector ☾ class aligns the teacher’s feature maps to the student’s feature map.
- 1x1 conv to project teachers features from in_channels (teacher channels) down to out_channels (student channels) , to match the channel dimensions.

```
# define the intermediate feature channels for both teacher and student
student_channels = [64, 128, 256, 512]
teacher_channels = [256, 512, 1024, 2048]

# create projection layers to align teacher's feature maps with student's feature maps
proj_layers = [
    FeatureProjector(in_c, out_c).to(device)
    for in_c, out_c in zip(student_channels, teacher_channels)
]
```

Implementation – student model tuned

Accuracy : 78.6 %



Input ☾ ResNet18 ☾ learns from soft targets + feature loss (intermediate features) + hard labels

Hyperparameter tuning :

Tuning ☾ two ways

- T , α **fixed** values.
- T , α **scheduling** through exponential decay

Hyperparameter (fixed):

- T = 0.5
- α = 0.7

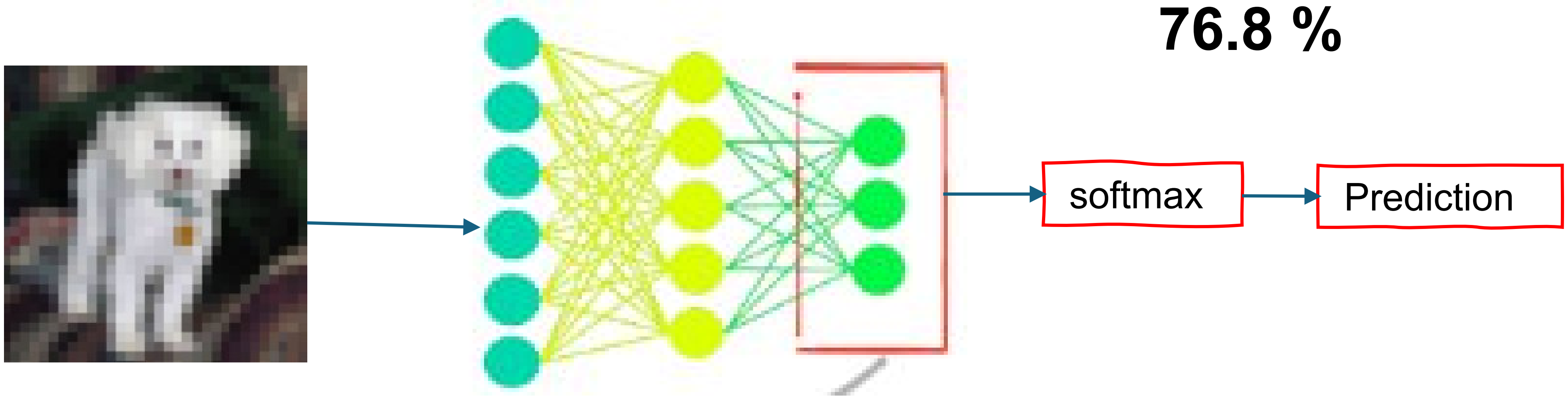
Hyperparameter (scheduling) :

- T = 0.5,0.3, 0.95
- α = 0.8, 0.5, 0.95

scheduling improved the model accuracy than fixed values

Implementation – student model without KD

Accuracy : 76.8 %



Input ☾ ResNet18 (trained from scratch) ☾ predictions (cross entropy loss)

Results :

Model	Architecture	Accuracy (%)	Parameters (M)	Latency (ms)
Teacher	ResNet50	85.11	23.5	8.72
Student distilled (tuned)	ResNet18	78.6	11.18	3.90
Student distilled	ResNet18	76.8	11.18	4.10
Student without KD	ResNet18	75.06	11.18	4.70

Thank you.