Report 2: Knowledge Distillation Progress

Karol Aleksander Cichor

April 28, 2025

1 Summary

- Trained and evaluated multiple models on TinyImageNet-200.
- Selected ConvNeXt (pretrained) as the teacher model.
- Initiated knowledge distillation experiments to lightweight students (ResNet-18 and ShuffleNetV2).
- Tuned distillation hyperparameters: temperature (T) and loss weight (α).

2 Training Setup

- **Teacher Model**: ConvNeXt, pretrained on TinyImageNet-1K.
- **Student Models**: ResNet-18, ShuffleNetV2 (both pretrained).
- **Dataset**: TinyImageNet-200.
- **Training Parameters**: batch size 128, learning rate 0.001, Adam optimizer.
- **Hardware**: 4 × NVIDIA GeForce RTX 2080 Ti GPUs.

All models were trained for 10–20 epochs depending on convergence.

3 Experimental Results

After training 10+ candidate models, the best-performing teacher was ConvNeXt (pretrained), achieving **68.63%** top-1 validation accuracy. Subsequently, knowledge distillation (KD) experiments were performed with various combinations of $T \in \{1, 2, 4, 10, 20, 100\}$ and $\alpha \in \{0, 0.1, 0.5, 0.9, 1\}$.

The distillation loss is given by:

$$L = \alpha \cdot \mathrm{CE}(\hat{y}, y) + (1 - \alpha) \cdot T^2 \cdot \mathrm{KL}\left(\log \mathrm{softmax}\left(\frac{\hat{y}}{T}\right), \mathrm{softmax}\left(\frac{\hat{y}_t}{T}\right)\right), \tag{1}$$

where $\hat{y}$ is the student's output, $\hat{y}_t$ is the teacher's output, y is the ground truth label, T is the temperature parameter, and α controls the trade-off between cross-entropy and distillation loss.

3.1 ResNet-18 Student

Table 1: Top-3 Best, Worst, and Standard Parameters for ResNet-18 Distillation

T	α	Validation Accuracy (%)
20	**0.1**	**62.84**
10	0.1	62.75
10	0.0	62.13
4	0.5	61.70
2	0.9	55.10
1	0.5	54.53
1	0.9	53.26
Baseline (no distillation): 53.63%		

3.2 ShuffleNetV2 Student

Table 2: Top-3 Best, Worst, and Standard Parameters for ShuffleNetV2 Distillation

T	α	Validation Accuracy (%)
100	**0.5**	**65.24**
20	0.5	65.09
10	0.1	64.48
4	0.5	64.42
10	0.1	60.95
20	0.0	60.19
100	0.0	58.77
Baseline (no distillation): 60.78%		

4 Observations

- Higher temperatures combined with lower α values generally yielded better performance.
- Incorrect KD settings (low T, high α) significantly harmed performance.
- Standard KD parameters ($T = 4$, $\alpha = 0.5$) still provided robust improvements.

5 Conclusions

The highest accuracies for the student models were achieved with high temperature values and low α parameters, confirming that students benefit significantly from the information contained in the teacher's output logits, beyond the hard target labels alone.