1. Introduction

Cross Entropy Hard vs. Soft Class Labels:
\[ L_{CE} = - \sum_{c=1}^{C} \log (\pi(c|x, \theta)) \]

Table: The label information and the Model predictions

- Category: Audi, BMW, Carrot
- Label: Hard Label 1, 0, 0, Model-A 0.6, 0.39, 0.01

CE+Hard: Loss_A = Loss_B  CE+Soft: Loss_A < Loss_B

Drawbacks of Hard Label based Cross Entropy:
- Considering no correlation between classes.
- Prone to overfitting.

Solution: Knowledge Distillation

2. Methodology

Knowledge Distillation by On-the-Fly Native Ensemble

Figure 2: Overview of online distillation training of ResNet-110 by the proposed On-the-Fly Native Ensemble (ONE). With ONE, we reconfigure the network by adding \( m \) auxiliary branches. Each branch with shared layers makes an individual model, and their ensemble is used to build the teacher model.

Multi-Branch Design:
In auxiliary branches with the same configuration, each serving as an independent efficient classification model.

Gate Network:
A gate which learns to ensemble all \((m + 1)\) branches to build a stronger teacher:
\[ z_e = \sum_{i=0}^{m} g_i \cdot z_i \]

On-the-Fly Knowledge Distillation:
Compute soft probability distributions at a temperature of \( T \) for branches and the ONE teacher as:
\[ \tilde{\pi}_c(x, \theta_e) = \frac{\exp(x'_c/T)}{\sum_{c=1}^{C} \exp(x'_c/T)} \]
\[ \tilde{\pi}_c(x, \theta_e) = \frac{\exp(x'_c/T)}{\sum_{c=1}^{C} \exp(x'_c/T)} \]

Distill knowledge from the teacher to each branch:
\[ \mathcal{L}_{kl} = \sum_{i=0}^{m} \mathcal{L}_{klc} = \mathcal{L}_{kl} + T^2 \cdot \mathcal{L}_{kl} \]

Overall Loss Function:
\[ \mathcal{L} = \sum_{i=0}^{m} \mathcal{L}_{klc} + \mathcal{L}_{kl} + T^2 \cdot \mathcal{L}_{kl} \]

3. Experiments

ImageNet test

Knowledge Distillation and Ensemble Comparisons

4. Further Analysis

ONE vs. Model Ensemble (ME)

(1) Model variance: average prediction differences between every two models/branches.
(2) Mean model generalisation capability.

- ONE leads to higher correlations due to the learning constraint from the distillation loss;
- ONE yields superior mean model generalisation capability with lower error rate 26.61 vs 31.07 by ME.

5. Reference