Distilling the knowledge in a neural network

Background

Caruana and his collaborators [1]. In their important paper they demonstrate convincingly that the knowledge acquired by a large ensemble of models can be transferred to a single small model.

Idea

1. Distilling knowledge in an ensemble of models into a single model, more suitable for deployment.
2. Introduce a new type of ensemble composed ofone or more full models and many specialist models(trained in parallel) which learn to distinguish fine-grained classes that the full models confuse.

Math statement

Assume that we have a large teacher network with parameters θ_T and a smaller student network with parameters θ_S. Let x denote an input sample and y its corresponding label. The teacher network output is denoted as $p_T(y|x;\theta_T)$, and the student network output is denoted as $p_S(y|x;\theta_S)$.

The soft target distribution is generated by applying a temperature parameter T to the teacher network output probabilities via the softmax function:

$$softmax(z_i/T) = exp(z_i/T)/sum(exp(z_j/T)) for i in 1, 2, ..., n$$

where z_i is the logit output of the teacher network for class i, and q_i is the corresponding soft target probability.

The student network is trained to minimize the difference between its output probabilities and the soft target probabilities generated by the teacher network. The loss function is defined as follows:

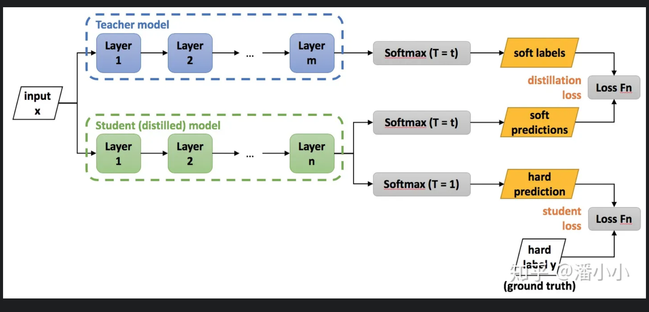
$$L = alpha * T^2 * sum(q_i/T * log(q_i/T)) + crossentropy(p_S, y)$$

where:
$alpha$ is a hyperparameter that controls the relative weight of the two terms in the loss function. (>1)
T is the temperature parameter used to generate the soft target distribution.
q_i is the soft target probability for class i generated by the teacher network.
p_S is the output probability of the student network.
y is the true label of the input sample.

Method

The goal is to transfer the knowledge learned by the teacher network to the student network in a compressed form. This is achieved by using a **soft target distribution** instead of the hard targets typically used in supervised learning.

In the soft target approach, the output probabilities of the teacher network are smoothed by applying a temperature parameter to the softmax function. This results in a more informative and less rigid target distribution.



Temperature selection

1. At lower temperatures, distillation pays much less attention to matching logits that are much more negative than the average
2. The very negative logits may convey useful information about the knowledge acquired by the cumbersome model. Needs high temperature.

[1] C. Bucilu˘a, R. Caruana, and A. Niculescu–Mizil. Model compression. In Proceedings of the 12th ACMSIGKDD International Conference on Knowledge Discovery andDataMining, KDD '06, pages 535—541, New York, NY, USA, 2006. ACM.

https://zhuanlan.zhihu.com/p/102038521