Deep Model Compression: Distilling Knowledge from Noisy Teachers

Introduction:

Recently, many methods have been proposed for deep model compression, but almost all of them focus on reducing storage complexity. The compressed models have to be decompressed at runtime, which means, issues of deployability on mobile devices continue to remain. This paper extends the teacher-student framework for deep model compression to address the runtime and train time complexity. The authors propose a simple methodology to include a noise-based regularizer while training the student from the teacher. Implementing this simulates learning from multiple-teachers for better model compression, and it provides a healthy improvement in the performance of the student network.

Proposed Methodology

Student Learning using Logit Regression

The authors use the method proposed by Ba and Caruana as a baseline for comparison. Ba and Caruana proposed a method to train the student directly on the log probability values z, also called logits, which is the output of the layer before softmax activation. The student network is trained in a regression setting using the logits, with its training data given by:

$$\{(x^{(1)}, z^{(1)}), ..., (x^{(i)}, z^{(i)}), ..., (x^{(n)}, z^{(n)})\}$$

The L2 loss minimized during the training is given by:

$$L(x, z, \theta) = \frac{1}{2T} \sum_i \left\| g(x^{(i)}; \theta) - z^{(i)} \right\|$$

where T is the mini-batch size, $x^{(i)}$ is the i^{th} training sample in the mini-batch, $z^{(i)}$ is the corresponding logit output of the pretrained teacher for $x^{(i)}$, θ is the set of student model parameters, $g(x^{(i)}; \theta)$ is the student model's logit output for $x^{(i)}$. The authors build on this approach to develop their idea of noisy teachers.

Noisy Teachers: Student Learning using Logit Perturbation

The authors propose a methodology to simulate the effect of multiple teachers, by injecting noise and perturbing the logit outputs of a teacher. The perturbed outputs, not only, simulate a multiple-teacher setting, but also results in noise in the loss layer, thus producing the effect of a regularizer. This new noisy teacher thus acts as a target-cum-regularizer and helps the student to learn better and produce results closer to the teacher network.

Mathematical Formulation

Let ξ be a vector of Gaussian noise with mean $\mu = 0$ and standard deviation σ. Dimension of ξ is equal to the number of classes/logits in the teacher network. If $z^{(i)}$ is the logit layer output of the teacher model for $x^{(i)}$, then modify $z^{(i)}$ as follows:

$$z'^{(i)} = (\mathbf{1} + \xi) \cdot z^{(i)}$$

where $\mathbf{1}$ is a vector of ones and $i \in \mathbb{R}^n$ where n is the number of classes. The loss function then becomes:

$$L(x, z', \theta) = \frac{1}{2T} \sum_i \left\| g(x^{(i)}; \theta) - z'^{(i)} \right\|$$

Equivalence to Noise-Based Regularization

Bishop showed that adding an L2 regularization term in the loss function is equivalent to adding Gaussian noise in the input data. The regularised loss function is given as:

$$L(x', \theta, z) = L(x, \theta, z) + R(\theta)$$

where x' corresponds to x with Gaussian noise and $R(\theta)$ is the L2 regularizer.

In the paper, authors perturb the target output z with noise, instead of the input data x. But below is the proof that perturbing the target output z, the logit values of the teacher, is equivalent to adding a noise-based regularization term to the loss function:

$$z'^{(i)} = (\mathbf{1} + \xi) \cdot z^{(i)} = z^{(i)} + \xi \cdot z^{(i)}$$

Hence, the modified L2 loss can be written as:

$$\begin{aligned} L(x, \theta, z') &= \|(z^{(i)} - g(x^{(i)}, \theta)) - \xi.z^{(i)}\|_2^2 \\ &= \|z^{(i)} - g(x^i, \theta)\|_2^2 + \|\xi.z^{(i)}\|_2^2 \\ &\qquad + 2\|z^{(i)} - g(x^{(i)}, \theta)\|_2 * \|\xi.z\|_2 \\ &= L(x, \theta, z) + E_R \end{aligned}$$

where E_R is the new regularizer that is based on the noise ξ.