Learning Efficient Object Detection Models with Knowledge Distillation

Introduction:

The paper proposes a new framework to learn compact and fast object detection networks with improved accuracy using knowledge distillation and hint learning. Although knowledge distillation has demonstrated excellent improvement for classification setups, the complexity of detection poses new challenges in the form of regression, region proposals and less voluminous labels. The paper proposes several innovations to address the above problems - a weighted cross-entropy loss to address class imbalance, a teacher bounded loss to handle the regression component and adaptation layers to better learn from intermediate teacher distributions.

Method:

The paper adopts Faster-RCNN as the object detection framework. Faster-RCNN is composed of three modules:

1. A shared feature extraction through convolutional layers
2. a region proposal network (RPN) that generates object proposals
3. a classification and regression network (RCN) that returns the detection score as well as a spatial adjustment vector for each object proposal.

Overall structure:

The paper claims to learn strong but efficient student object detectors by using the knowledge of a high capacity teacher for all the three components:

1. Use hint-based learning so that the feature representation of a student network is similar to that of the teacher network.
2. Learn stronger classification modules in both RPN and RCN using knowledge distillation framework.
3. To handle class imbalance issue, apply weighted cross entropy loss for the distillation framework.
4. Transfer the teacher's regression output as a form of upper bound (i.e., if the student's regression output is better than that of teacher, no additional loss is applied).

The overall learning objective can be written as follows:

$$L_{RCN} = \frac{1}{N}\sum_i L_{cls}^{RCN} + \lambda\frac{1}{N}\sum_j L_{reg}^{RCN}$$

$$L_{RPN} = \frac{1}{M}\sum_i L_{cls}^{RPN} + \lambda\frac{1}{M}\sum_j L_{reg}^{RPN}$$

$$L = L_{RPN} + L_{RCN} + \gamma L_{Hint}$$

Knowledge Distillation for classification with imbalanced classes:

Suppose the dataset is represented as $\{x_i, y_i\}, i = 1, 2, ..., n$ where x_i is the input image and y_i is its class label. Let t be the teacher model, with $P_t = softmax(\frac{Z_t}{T})$ its prediction and Z_t the final score output. Here, T is a temperature parameter (no temperature i.e., $T = 1$ works best for object detection problems). Similarly, one can define $P_s = softmax(\frac{Z_s}{T})$ for the student network s. The student s is trained to optimize the following loss function:

$$L_{cls} = \mu L_{hard}(P_s, y) + (1 - \mu) L_{soft}(P_s, P_t)$$

The object detection problem faces several imbalance across different categories (i.e., the background dominates). To address this, the authors adopt class-weighted cross entropy as the distillation loss:

$$L_{soft}(P_s, P_t) = -\sum w_c P_t \, log P_s$$

Knowledge Distillation for regression with teacher bounds:

Learning a good regression model is critical to ensure good object detection accuracy. Unlike distillation for discrete categories, the teacher's regression outputs can provide wrong guidance to the student model, since the real-valued regression outputs are unbounded. Thus, instead of using the teacher's regression output directly as a target, it is used as an upper bound for the student to achieve. The paper calls this loss the teacher bounded regression loss, L_b, which is used further to formulate the regression loss, L_{reg}, as follows:

$$L_b(R_s, R_t, y) = \begin{cases} \|R_s - y\|_2^2, & \text{if } \|R_s - y\|_2^2 + m > \|R_t - y\|_2^2 \\ 0, & \text{otherwise} \end{cases}$$
$$L_{reg} = L_{sL1}(R_s, y_{reg}) + \nu L_b(R_s, R_t, y_{reg}),$$

where m is a margin, y_{reg} denotes the regression ground truth label, R_s is the regression output of the student network, R_t is the prediction of teacher network and ν is a weight parameter. Here, L_{sL1} is the smooth L1 loss.

Hint Learning with feature adaptation:

The FitNets paper demonstrates that using the intermediate representation of the teacher as a hint can help the training process and improve the final performance of the student. This paper uses the L1 distance between feature vectors V and Z:

$$L_{Hint}(V, Z) = \|V - Z\|_1$$

where Z represents the intermediate layer selected as hint in the teacher network and V represent the output of the guided layer in the student network. If the number of neurons (channels, width and height) is not same between the corresponding layers in the teacher and student, an adaptation is added after the guided layer to match the size of the hint layer.