Clustering and Locality Penalties for Semi-Supervised Learning

Ankan Bansal Joaquin Zepeda

1 Problem

The problem is how to leverage the large amount of unlabeled data in addition to the existing labeled data for training deep networks.

We have two sets of images, one with labels, $\mathcal{S} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{S}$, and the other without labels, $\mathcal{U} = \{\mathbf{X}_j\}_{j=S+1}^{S+U}$. Here, $\mathbf{X}_i$ is the i^{th} image and $y_i \in \mathcal{C}$ is the corresponding label and $\mathcal{C}$ is the set of classes. Using this data, we want to train an image classifier, $f(\mathbf{X}; \Theta)$, which takes an image as input and outputs the probability distribution $\mathbf{q}$ over classes $\mathcal{C}$.

2 Idea

The first idea is based on unsupervised clustering of images. We can cluster the images into $|\mathcal{C}|$ clusters. However, without any labels, the clusters do not necessarily correspond to semantic classes. We can use the labeled samples $\mathcal{S}$ to seed the clusters. This ensures that the clusters formed have a semantic meaning.

Let $H(.)$ represent the entropy of a probability distribution defined as $H(\mathbf{p}) = -\sum_{i=1}^{C} p_i \log(p_i)$. Say that, for an image $\mathbf{X}_t$, the output of the classifier is a probability distribution, $\mathbf{q}_t$, over the classes $\mathcal{C}$. We draw a batch of size T and compute the following two losses over the batch.

Mean Entropy Loss (MEL)

$$J_M = \frac{1}{T}\sum_{t=1}^{T} H(\mathbf{q}_t) \tag{1}$$

Negative Batch Entropy Loss (NBEL)

$$J_B = -H(\frac{1}{T}\sum_{t=1}^{T} \mathbf{q}_t) \tag{2}$$

The first (MEL) tries to increase the confidence of the predicted class and the second (NBEL) ensures that the predictions do not collapse to a single label and are evenly spread out. NBEL is based on the assumption that, when classes are uniformly distributed, uniform sampling from the dataset should lead to uniform sampling over the output classes.

For the supervised images, we are given a ground-truth probability distribution, $\mathbf{p}$ (This is just a one-hot probability vector in our case). The objective of a classifier, f, is to minimize the cross-entropy $E = -\sum_{i=1}^{C} p_i \log(q_i)$ for an image $\mathbf{X}$. Here, p_i and q_i are the elements of $\mathbf{p}$ and $\mathbf{q} = f(\mathbf{X}; \Theta)$ respectively. For a mini-batch of size T, the cross-entropy loss can be defined as $J_C = \frac{1}{T}\sum_{t=1}^{T} E_t$, where E_t is the cross-entropy for image $\mathbf{X}_t$.

Consider a mini-batch, $\mathcal{B} = \{(\mathbf{X}_k, y_k)\}_{k=1}^{R} \cup \{\mathbf{X}_k\}_{k=R+1}^{T}$, where $\{(\mathbf{X}_k, k_i)\}_{k=1}^{R} \in \mathcal{S}$, and $\{\mathbf{X}_k\}_{k=R+1}^{T} \in \mathcal{U}$. Our total loss is given as:

$$\mathcal{L} = J_C + \alpha J_M + \beta J_B \tag{3}$$

$$\mathcal{L} = \frac{1}{R}\sum_{t=1}^{R} E_t + \alpha\frac{1}{T}\sum_{t=1}^{T} H(\mathbf{q}_t) - \beta H(\frac{1}{T}\sum_{t=1}^{T} \mathbf{q}_t) \tag{4}$$

We find α, β, and R by cross-validation. The ratio $\frac{R}{T}$ can be thought of as fine-adjustment for α (and β?).

2.1 Locality Loss

The second idea is a structured sparsity loss. This is a way of incorporating the prior knowledge that an object occupies a small region in the image. We use the sparsity inducing norms introduced in [1, 2]. Consider a matrix, $C \in \mathbb{R}^{N \times N}$. Let us define a subset of the power-set of the support of C as

$$\mathcal{G} = \{g_i \subseteq \text{support}(C)\}_i \tag{5}$$

where $\text{support}(C) = [1 \dots N] \times [1 \dots N]$ and $\cup_{g \in \mathcal{G}} = \text{support}(C)$. The following norm, when used as a loss, behaves like an L-1 norm on the group level and therefore, induces group sparsity.

$$\Omega(C, \mathcal{G}) = \sum_{g \in \mathcal{G}} \|C_g\|_2 \tag{6}$$

where C_g is the vector formed by the values in C indexed by g and $\|.\|_2$ is the L-2 norm. This loss encourages each C_g to go to zero. However, within $C_g,\ g \in \mathcal{G}$, the L-2 norm does not promote sparsity.

We focus on the case where $\mathcal{G}$ is a set of horizontal or vertical half-spaces. We use class activation maps $C \in \mathbb{R}^{N \times N}$ [3] and define four sets of groups and the corresponding losses. The first set (left-right) contains nested vertical half-spaces, i.e. collections of columns of C and is defined as: $\mathcal{G}_1 = \{g_{1k}\}_{k=1}^{N}$ where $g_{1k} = [1 \dots k] \times [1 \dots N]$. The left-right loss is then given as: $l_1 = \Omega(C, \mathcal{G}_1) = \sum_{g \in \mathcal{G}_1} \|C_g\|_2$. This encourages the left-most columns to be zero. The right-left, top-bottom, and bottom-top losses can be defined similarly.

The total locality penalty for an image, $\mathbf{X}_t$, is just the sum:

$$J_L^t = \sum_{j=1}^{4} l_j(C) = \sum_{j=1}^{4} \sum_{g \in \mathcal{G}_j} \|C_g\|_2 \tag{7}$$

and the total locality penalty for a batch of size T is $J_L = \frac{1}{T} J_L^t$.

Adding this to the earlier loss formulation, the total loss is given as:

$$\mathcal{L} = J_C + \alpha J_M + \beta J_B + \gamma J_L \tag{8}$$

Again, for a batch, J_C is calculated only for the supervised images, and J_M, J_B, and J_L are calculated for both supervised and unsupervised images.

3 Evaluation

We use ImageNet as the training and testing dataset. We fix about 64,000 images for which we have labels available (supervised set, $\mathcal{S}$). We use an additional 200,000 images as the unsupervised set, $\mathcal{U}$. We use the ImageNet validation set for testing. It has 50,000 images spread over all 1,000 classes. We report the top-1 accuracy for all experiments. We train DenseNet network architectures for all our experiments.

3.1 Baseline

As the baseline, we train a network using $\mathcal{S}$ as the training set and only J_C as the loss function. This model achieves a top-1 accuracy of 42.04% on the ImageNet validation set.

3.2 Cross-entropy + MEL

Now, we add the J_M term to the loss function, i.e., $\mathcal{L} = J_C + \alpha J_M$ and use $\mathcal{S} \cup \mathcal{U}$ as the training set. We set $\alpha = 0.8$. Figure 1 (left) shows the performance as we vary $\frac{R}{T}$.

Next, we vary the number of unsupervised images in the training set. Figure 1 (right) shows the validation accuracy for different amount of unsupervised data. We notice that more unsupervised data is helpful for improving the classification performance.

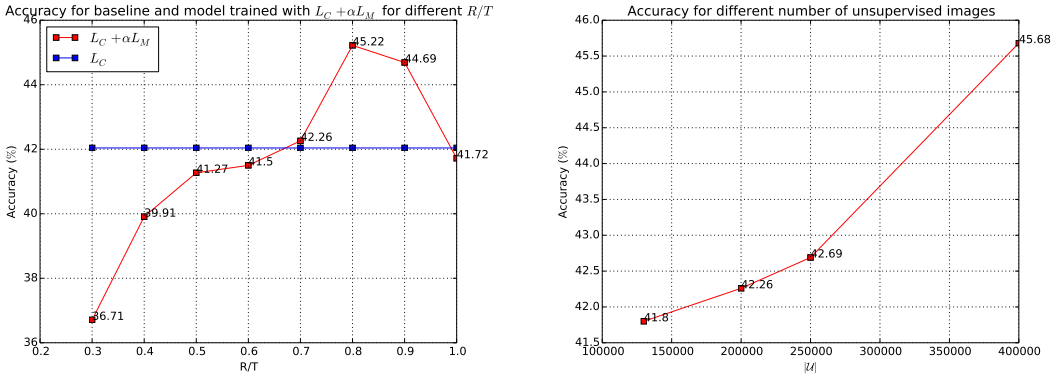

Fig. 1. (Left) Accuracy on the ImageNet validation set for models trained with the cross-entropy loss and MEL with $\mathcal{S} \cup \mathcal{U}$ (red) and only cross-entropy loss with $\mathcal{S}$ (blue). (Right) Accuracy of models trained with cross-entropy loss and MEL with $\mathcal{S} \cup \mathcal{U}$ for $\frac{R}{T} = 0.7$ for different number of unsupervised images.

3.3 Cross-entropy + MEL + NBEL

Now we add the negative batch-entropy loss to cross-entropy loss and MEL. We vary β while keeping α and $\frac{R}{T}$ fixed from the previous case. The performance (accuracy on the validation set) of these models hasn't reached the level of the previous cases. We believe that this is because of the small batch sizes that we are using. Our next steps for this are to increase the batch-size by using more GPUs or accumulating gradients over a few iterations before updating the parameters.

3.4 Cross-entropy + Loc

In this case, we take the class activation map (CAM) of the class with the highest probability and apply the locality penalty to this CAM. We have observed that CAMs are able to localize some objects, but we haven't analyzed whether this is because of the formulation of CAMs or because of our loss function. We plan to see the evolution of the activations as training progresses to see if the area covered by the activations is reducing and whether the CAMs are becoming more accurate with more training.

3.5 Transfer Learning

For a comparison with recent work on unsupervised learning, we will experiment with the transfer learning setting too. In this setting, models trained with only unsupervised losses are used as initializations for other tasks (e.g. object detection, semantic segmentation). We want to compare the unsupervised losses used by these works against ours. We also want to see if models trained with both supervised and unsupervised losses are better than models trained with only unsupervised losses for transfer learning.

References

1. Jenatton, R., Audibert, J.Y., Bach, F.: Structured variable selection with sparsity-inducing norms. Journal of Machine Learning Research **12**(Oct), 2777–2824 (2011)
2. Jenatton, R., Obozinski, G., Bach, F.: Structured sparse principal component analysis. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 366–373 (2010)
3. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)