Interactive Multi-Label CNN Learning with Partial Labels

Dat Huynh
Northeastern University
huynh.dat@husky.neu.edu

Ehsan Elhamifar
Northeastern University
eelhami@ccs.neu.edu

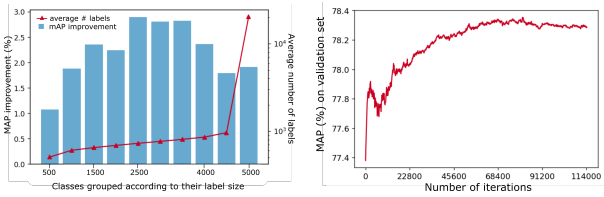

Figure 1: Left: Improvement of mAP score (%) of our method with respect to Logistic regression for groups of classes with least annotations (left bar) to most annotations (right bar) in Open Images. The red curve shows the number of available training images in each group. Our method significantly improves over Logistic especially for the middle groups that have sufficient annotations. Right: mAP score of our method on the validation set during training. We observe that the performance curve fluctuates at the early stage of the training since the model struggles to refine noisy label graph and learn the data manifold. However, after 20k iterations, the learning curve monotonically improves until convergence. Here, each iteration corresponds to a gradient descent step on a minibatch.

1. Additional Results on Open Images

Figure 1 shows the mAP improvement of our method (IMCL) over the logistic regression when dividing labels into 10 groups (instead of 5, discussed in the main paper). Notice that our method improves the mAP score over the logistic regression for all the 10 groups. This shows the effectiveness of using the collaborative similarity learner in conjunction with the logistic model, which regularizes the network and takes advantage of semantic similarities across images and labels. Also, the right plot in Figure 1 shows the improvement of the mAP score on the validation set during training. We observe that the performance fluctuates at the early stage of the training as the model tries to refine the noisy label graph and learn the data manifold. However, after 20k iterations, the learning curve monotonically improves until convergence.

2. Detailed Results on CUB Dataset

Table 1 shows the exact performance of different methods as a function of the percentage of missing attributes on the CUB dataset (in the main paper, we showed bar charts).

Missing attributes	Logistic	CNN-RNN	Fast0Tag	Curriculum Labeling	IMCL $a = 1$	IMCL $a = 10$
90%	23.8	23.3	23.3	23.8	25.3	**25.7**
80%	25.2	25.0	25.0	25.7	**26.4**	**26.4**
60%	26.4	27.3	27.3	27.4	27.3	**27.9**
40%	26.6	26.9	27.2	**27.8**	27.7	**27.8**
20%	26.6	27.3	27.6	**27.8**	27.5	27.6
0%	27.9	27.2	27.9	27.9	28.6	**28.7**

Table 1: mAP scores (%) as a function of the percentage of missing attributes in the CUB dataset.

CNN-RNN	LSEP	Wsabie	Fast0Tag	Latent Noise (relevant)	Latent Noise (visual)	Curriculum Labeling	IMCL
-0.6	0.1	0.1	0.0	0.9	1.0	1.5	1.9

Table 2: Improvement of mAP score (%) of different methods over the logistic regression on the MS-COCO dataset.

3. Detailed Results on MS-COCO Dataset

Table 2 shows the exact mAP improvement of different methods over the logistic regression on the MS-COCO dataset (in the main paper, we showed bar charts).

4. Interactive Learning Framework

In this section, we present the derivations of our proposed Joint Nonnegative OMP algorithm for finding semantically similar images, in the similarity learning step in our framework, presented in the main paper. We then discuss the computational complexity of our algorithm.

4.1. Joint Nonnegative OMP Derivation

Consider our objective function with the classifier parameters, $\{\boldsymbol{\theta}_j\}_{j=1}^{|\mathcal{C}|}, \boldsymbol{w}$, written as

$$\min_{\boldsymbol{w},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{|\mathcal{C}|}} \sum_i \mathcal{L}_{\mathrm{c}}^{(i)}\Big(\boldsymbol{w}, \{\boldsymbol{\theta}_j\}_{j=1}^{|\mathcal{C}|}\Big) + \mathcal{L}_{\mathrm{s}}^{(i)}\Big(\boldsymbol{w}, \{\boldsymbol{\theta}_j\}_{j=1}^{|\mathcal{C}|}\Big), \tag{1}$$

where the cross entropy loss for image i is defined as

$$\mathcal{L}_{\mathrm{c}}^{(i)} \triangleq -\sum_{j\in\Omega_i} y_{j,i}^{o} \log(p_{j,i}) + (1 - y_{j,i}^{o}) \log(1 - p_{j,i}), \tag{2}$$

and the prediction smoothness loss for image i is defined as

$$\begin{aligned}\mathcal{L}_{\text{s}}^{(i)} \triangleq \min_{\{c_{i',i},\bar{c}_{i',i}\}} & \lambda_y \Big\| \boldsymbol{y}_i - \tanh\Big(\sum_{i'=1}^{N} c_{i',i}\boldsymbol{A}\boldsymbol{y}_{i'}\Big)\Big\|_2^2 \\ & + \lambda_f \|\boldsymbol{f}_i - \sum_{i'=1}^{N} \bar{c}_{i',i}\boldsymbol{f}_{i'}\|_2^2 \\ \text{s.t.} \ \sum_j \mathrm{I}\Big(\big\|\big[c_{i',i},\ \bar{c}_{i',i}\big]\big\|\Big) \leq k,\ & c_{i',i}, \bar{c}_{i',i} \geq 0,\ \forall i'. \end{aligned} \tag{3}$$

To efficiently solve for image and label similarities, we use a first-order approximation of the hyperbolic tangent function in $\mathcal{L}_{\text{s}}^{(i)}$, which is $\tanh(x) \approx x$. This is a good approximation as long as tanh is not saturated. Using this approximation, we can rewrite (3) as

$$\begin{aligned}\mathcal{L}_{\text{s}}^{(i)} \triangleq \min_{\{c_{i',i},\bar{c}_{i',i}\}} & \lambda_y \|\boldsymbol{y}_i - \sum_{i'=1}^{N} c_{i',i}\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2 \\ & + \lambda_f \|\boldsymbol{f}_i - \sum_{i'=1}^{N} \bar{c}_{i',i}\boldsymbol{f}_{i'}\|_2^2 \\ \text{s.t.} \ \sum_j \mathrm{I}\Big(\big\|\big[c_{i',i},\ \bar{c}_{i',i}\big]\big\|\Big) \leq k,\ & c_{i',i}, \bar{c}_{i',i} \geq 0,\ \forall i'. \end{aligned} \tag{4}$$

To compute the loss function $\mathcal{L}_{\text{s}}^{(i)}$, we need to solve the joint nonnegative sparse optimization in (4) over both $c_{i',i}$ and $\bar{c}_{i',i}$. We generalize [1], a greedy nonnegative sparse solver, to efficiently solve for both nonnegative similarity coefficients. More specifically, we use a generalization of OMP that starts from the empty set, at each iteration selects the best image i' that minimizes the smoothness loss the most (i.e., whose label and feature vectors best reconstruct the label and feature vectors of image i), updates residual errors and repeats to select the next best candidate, until k candidates are chosen. Let $\mathcal{N}_i$ denote the set of similar images chosen so far. We compute the the optimal similarity coefficients $\{c_{i',i}, \bar{c}_{i',i}\}$ and the residual errors for the label, $\boldsymbol{r}_y$, and feature ,$\boldsymbol{r}_f$, as

$$\begin{aligned} c_{i',i}^* &= \max\big(0, \arg\min \|\boldsymbol{y}_i - \sum_{i'\in\mathcal{N}_i} c_{i',i}\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2\big), \\ \bar{c}_{i',i}^* &= \max\big(0, \arg\min \|\boldsymbol{f}_i - \sum_{i'\in\mathcal{N}_i} \bar{c}_{i',i}\boldsymbol{f}_{i'}\|_2^2\big), \\ \boldsymbol{r}_y &= \boldsymbol{y}_i - \sum_{i'\in\mathcal{N}_i} c_{i',i}^*\boldsymbol{A}\boldsymbol{y}_{i'}, \\ \boldsymbol{r}_f &= \boldsymbol{f}_i - \sum_{i'\in\mathcal{N}_i} \bar{c}_{i',i}^*\boldsymbol{f}_{i'}. \end{aligned} \tag{5}$$

For each image i' not in the current set $\mathcal{N}_i$, we compute the loss $\mathcal{L}_{\text{s}}^{(i)}(\mathcal{N}_i \cup \{i'\})$ and then select the best i' for which we have the minimum loss function. To do so, we fix the similarity coefficients for $\mathcal{N}_i$ and compute

$$\begin{aligned}\mathcal{L}_{\text{s}}^{(i)}(\mathcal{N}_i \cup \{i'\}) = \min_{c_{i',i}} & \lambda_y \|\boldsymbol{r}_y - c_{i',i}\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2 \\ & + \lambda_f \|\boldsymbol{r}_f - \bar{c}_{i',i}\boldsymbol{f}_{i'}\|_2^2 \end{aligned} \tag{6}$$

We select the image i' that achieves the minimum loss function, i.e., $s = \arg\min_{i'} \mathcal{L}_{\text{s}}^{(i)}(\mathcal{N}_i \cup \{i'\})$. For each i', we compute the closed-form of (6) by setting the derivative with respect to $c_{i',i}$ and $\bar{c}_{i',i}$ to zero,

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{s}}^{(i)}(\mathcal{N}_i \cup \{i'\})}{\partial c_{i',i}} &= \frac{\partial \|\boldsymbol{r}_y - c_{i',i}\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2}{\partial c_{i',i}} = 0, \\ (\boldsymbol{A}\boldsymbol{y}_{i'})^T\Big(c_{i',i}^*\boldsymbol{A}\boldsymbol{y}_{i'} - \boldsymbol{r}_y\Big) &= 0, \\ \Longrightarrow c_{i',i}^* &= \frac{\langle \boldsymbol{r}_y, \boldsymbol{A}\boldsymbol{y}_{i'}\rangle}{\|\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2}. \end{aligned} \tag{7}$$

Similarly, we obtain the optimal value for $\bar{c}_{i',i}$ as

$$\bar{c}_{i',i}^* = \frac{\langle \boldsymbol{f}_{i'}, \boldsymbol{r}_f\rangle}{\|\boldsymbol{f}_{i'}\|_2^2}. \tag{8}$$

Substituting (7) and (8) into (6), we can compute the optimal loss function for any given i', $\mathcal{L}_{\text{s}}^{(i)}(\mathcal{N}_i \cup \{i'\})$ as

$$\begin{aligned} & \lambda_y \|\boldsymbol{r}_y - \frac{\langle \boldsymbol{r}_c, \boldsymbol{A}\boldsymbol{y}_{i'}\rangle}{\|\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2}(\boldsymbol{A}\boldsymbol{y}_{i'})\|_2^2 + \lambda_f \|\boldsymbol{r}_f - \frac{\langle \boldsymbol{f}_{i'}, \boldsymbol{r}_f\rangle}{\|f_{i'}\|_2^2}\boldsymbol{f}_{i'}\|_2^2 \\ &= \lambda_y\Big(\frac{\langle \boldsymbol{r}_y, \boldsymbol{A}\boldsymbol{y}_{i'}\rangle^2 \|\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2}{\|\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^4} - 2\frac{\langle \boldsymbol{r}_y, \boldsymbol{A}\boldsymbol{y}_{i'}\rangle^2}{\|\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2} + \|\boldsymbol{r}_c\|_2^2\Big) \\ &\quad + \lambda_f\Big(\frac{\langle \boldsymbol{r}_f, \boldsymbol{f}_{i'}\rangle^2 \|\boldsymbol{f}_{i'}\|_2^2}{\|\boldsymbol{f}_{i'}\|_2^4} - 2\frac{\langle \boldsymbol{r}_f, \boldsymbol{f}_{i'}\rangle^2}{\|\boldsymbol{f}_{i'}\|_2^2} + \|\boldsymbol{r}_f\|_2^2\Big) \\ &= -\lambda_y \frac{\langle \boldsymbol{r}_y, \boldsymbol{A}\boldsymbol{y}_{i'}\rangle^2}{\|\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2} - \lambda_f \frac{\langle \boldsymbol{r}_f, \boldsymbol{f}_{i'}\rangle^2}{\|\boldsymbol{f}_{i'}\|_2^2} + \text{constant}. \end{aligned} \tag{9}$$

Thus, we select the best next sample in Algorithm 2 via

$$s = \arg\max_{i'} \lambda_y \frac{\langle \boldsymbol{r}_y, \boldsymbol{A}\boldsymbol{y}_{i'}\rangle^2}{\|\boldsymbol{A}\boldsymbol{y}_{i'}\|_2^2} + \lambda_f \frac{\langle \boldsymbol{r}_f, \boldsymbol{f}_{i'}\rangle^2}{\|\boldsymbol{f}_{i'}\|_2^2}. \tag{10}$$

4.2. Speeding up Training

The line 4 of Algorithm 2 in the main paper requires comparing the residual vectors with label vector and feature vector of every image. For large datasets, such as Open Images or MS-COCO, where there is a lot of redundancy in images, we could significantly reduce the computational time of this step by using a subset of images, obtained using random sampling or subset selection techniques. In our experiments, we construct the smaller dictionary by randomly selecting, for each label in the training set, 10 images that contain that label. Thus, the dictionary for Open Images

and MS-COCO have the sizes of $50,000$ and $10,000$, respectively (we used the whole CUB dataset since there are only $11,000$ images in the dataset). This simple strategy speeds up the training while improving the state-of-the-art results, as shown in the paper.

4.2.1 Complexity Analysis

For the Joint Nonnegative OMP in Algorithm 2 of the main paper, the dominant complexity cost of the algorithm is to find similarities (line 4) and solve the non-negative least squares (line 9 and 10). Let N_S be the size of the dictionary used to find similar images and let l and f be, respectively, the dimension of $\boldsymbol{y}_i$ and $\boldsymbol{f}_i$. When the search for similar images is performed on the dictionary of size N_S, the complexity is $\mathcal{O}(N_S(l+f))$. On the other hand, solving non-negative least squares has the complexity of $\mathcal{O}(k^3 + k^2(l+f))$. Finding all k similar images thus takes $\mathcal{O}(N_S k(l+f)+k^4+k^3(l+f))$. With $N_S \ll N$, as is in our experiments on the Open Images and MS-COCO datasets, the complexity of the Joint Nonnegative OMP for one image would be $\mathcal{O}(1)$ in the size of the dataset. Thus, finding similarities for all images has $\mathcal{O}(N)$ complexity. Note that (1) is optimized in a mini-batch fashion, where we iteratively construct images similarity for a small batch of data. Thus, we have low memory complexity, $\mathcal{O}(B)$ with B being the size of the minibatch, in terms of the size of the training data in each iteration.

References

[1] T. H. Lin and H. T. Kung, "Stable and efficient representation learning with nonnegativity constraints," *International Conference on Machine learning*, 2014. 2