Weakly Supervised Representation Learning with Coarse Labels Supplementary

Yuanhong Xu[1] Qi Qian[2*] Hao Li[1] Rong Jin[2] Juhua Hu[3]
[1] Alibaba Group, Hangzhou, China
[2] Alibaba Group, Bellevue, WA, 98004, USA
[3] School of Engineering and Technology
University of Washington, Tacoma, WA, 98402, USA
{yuanhong.xuyh, qi.qian, lihao.lh, jinrong.jr}@alibaba-inc.com, juhuah@uw.edu

1. Theoretical Analysis

1.1. Proof of Lemma 1

Proof. To simplify the proof, we assume that each target class contains z examples as $zF = n$. Then, we define

$$\Pr\{y_i^F|f(\mathbf{x}_i), W^I\} = \frac{\exp(f(\mathbf{x}_i)^\top \bar{\mathbf{w}}_{y_i}^I)}{\sum_s^F \exp(f(\mathbf{x}_i)^\top \bar{\mathbf{w}}_s^I)}$$

where $\bar{\mathbf{w}}_s^I = \frac{1}{z}\sum_{y_j^F=s} \mathbf{w}_j^I$ averaging over the parameters from the same target class. According to the Jensen's inequality, we have

$$\exp(f(\mathbf{x}_i)^\top \bar{\mathbf{w}}_s^I) \leq \frac{1}{z}\sum_{y_j^F=s} \exp(f(\mathbf{x}_i)^\top \mathbf{w}_j^I)$$

Therefore, we have

$$\begin{aligned}
&\Pr\{y_i^F|f(\mathbf{x}_i), W^I\} \\
&\geq z\exp(f(\mathbf{x}_i)^\top \bar{\mathbf{w}}_{y_i}^I - f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^I)\Pr\{y_i^I|\mathbf{x}_i, W^I\} \\
&\geq z\alpha\exp(f(\mathbf{x}_i)^\top \bar{\mathbf{w}}_{y_i}^I - f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^I) \qquad (1)
\end{aligned}$$

where $\exp(f(\mathbf{x}_i)^\top \bar{\mathbf{w}}_{y_i}^I - f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^I)$ measures the distance from an individual example to other examples from the same target class. It cannot be bounded well by only solving the problem of instance classification.

1.2. Proof of Theorem 1

First, by optimizing the proposed classification problem, we assume

$$\forall i, \quad \Pr\{y_i^I|f(\mathbf{x}_i), W^I\} = \frac{\exp(f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^I)}{\sum_j^n \exp(f(\mathbf{x}_i)^\top \mathbf{w}_j^I)} \geq \alpha$$

and

$$\forall i, \quad \Pr\{y_i^C|f(\mathbf{x}_i), W^C\} = \frac{\exp(f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^C)}{\sum_j^C \exp(f(\mathbf{x}_i)^\top \mathbf{w}_j^C)} \geq \beta$$

By assuming the residual is lower bounded by constants a and b as

$$\forall i, \sum_{j,j\neq y_i^I}^n \exp(f(\mathbf{x}_i)^\top \mathbf{w}_j^I) \geq a; \sum_{j,j\neq y_i^C}^C \exp(f(\mathbf{x}_i)^\top \mathbf{w}_j^C) \geq b$$

we have

$$\exp(f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^I) \geq \frac{\alpha}{1-\alpha}(\sum_{j,j\neq y_i^I}^n \exp(f(\mathbf{x}_i)^\top \mathbf{w}_j^I)) \geq \frac{a\alpha}{1-\alpha}$$

and

$$\exp(f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^C) \geq \frac{\beta}{1-\beta}(\sum_{j,j\neq y_i^C}^C \exp(f(\mathbf{x}_i)^\top \mathbf{w}_j^C)) \geq \frac{b\beta}{1-\beta}$$

which leads to

$$\forall i, \quad f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^I \geq \log(\frac{a\alpha}{1-\alpha}); \quad f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^C \geq \log(\frac{b\beta}{1-\beta})$$

To guarantee the performance on the target classification problem, we have to bound $\exp(f(\mathbf{x}_i)^\top \bar{\mathbf{w}}_{y_i}^I - f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^I)$ as illustrated in Eqn. 1. Now, it can be bounded with the help from solving the coarse-class classification problem. Specifically, the instance similarity can be bounded as

$$\begin{aligned}
&f(\mathbf{x}_i)^\top \mathbf{w}_j^I - f(\mathbf{x}_i)^\top \mathbf{w}_{y_i}^I \\
&= f(\mathbf{x}_i)^\top f(\mathbf{x}_j) + f(\mathbf{x}_i)^\top (\mathbf{w}_j^I - f(\mathbf{x}_j)) \\
&- f(\mathbf{x}_i)^\top f(\mathbf{x}_i) + f(\mathbf{x}_i)^\top (f(\mathbf{x}_i)^\top - \mathbf{w}_{y_i}^I)
\end{aligned}$$

Then, we can bound each term as follows. First, the distance between an example to its individual label representation (i.e., $\mathbf{w}$) can be bounded by solving the individual

*Corresponding author

classification problem as

$$f(\mathbf{x}_i)^\top(\mathbf{w}_j^I - f(\mathbf{x}_j)) \geq -\|f(\mathbf{x}_i)^\top(\mathbf{w}_j^I - f(\mathbf{x}_j))\|_2$$
$$\geq -\|f(\mathbf{x}_i)\|_2\|\mathbf{w}_j^I - f(\mathbf{x}_j)\|_2 \text{(Cauchy-Schwarz inequality)}$$
$$\geq -c\|\mathbf{w}_j^I - f(\mathbf{x}_j)\|_2$$
$$\geq -c\sqrt{2c^2 - 2\log(a\alpha/(1-\alpha))}$$

With the similar analysis, we have

$$f(\mathbf{x}_i)^\top(f(\mathbf{x}_i)^\top - \mathbf{w}_{y_i}^I) \geq -c\sqrt{2c^2 - 2\log(a\alpha/(1-\alpha))}$$

Note that examples from the same target class also share the same coarse-class label. Therefore, the distances between examples from the same target class can be bounded as

$$f(\mathbf{x}_i)^\top f(\mathbf{x}_j) - f(\mathbf{x}_i)^\top f(\mathbf{x}_i) \geq -c\|f(\mathbf{x}_j) - f(\mathbf{x}_i)\|_2$$
$$\geq -c(\|f(\mathbf{x}_j) - \mathbf{w}_{y_i}^C\|_2 + \|f(\mathbf{x}_i) - \mathbf{w}_{y_i}^C\|_2)$$
$$\geq -2c\sqrt{2c^2 - 2\log(b\beta/(1-\beta))}$$

Combining them together, we have

$$\exp(f(\mathbf{x}_i)^\top\bar{\mathbf{w}}_{y_i}^I - f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^I)$$
$$\geq \exp\Big(-\frac{2c(z-1)}{z}\big(\sqrt{2c^2 - 2\log(a\alpha/(1-\alpha))}$$
$$+\sqrt{2c^2 - 2\log(b\beta/(1-\beta))}\big)\Big)$$

Taking it back to Eqn. 1, we can observe the desired result

$$\Pr\{y_i^F|f(\mathbf{x}_i), W^I\} \geq \alpha z h(c,\alpha,\beta)$$

where

$$h(c,\alpha,\beta) = \exp\Big(-\frac{2c(z-1)}{z}$$
$$\big(\sqrt{2c^2 - 2\log(a\alpha/(1-\alpha))} + \sqrt{2c^2 - 2\log(b\beta/(1-\beta))}\big)\Big)$$

□

1.3. Proof of Theorem 2

Proof. Following the analysis in Theorem 1, we assume

$$\forall i, \Pr\{y_i^I|\mathbf{x}_i, y_i^C, W^I\} = \frac{\exp(f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^I)}{\sum_{j:j=y_i^C}\exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^I)} \geq \alpha$$

$$\forall i, \Pr\{y_i^C|\mathbf{x}_i, W^C\} = \frac{\exp(f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^C)}{\sum_j^C\exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^C)} \geq \beta$$

$$\forall i, \quad \sum_{j=y_i^C, j\neq y_i^I}\exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^I) \geq a$$

$$\forall i, \quad \sum_{j\neq y_i^C}^{C}\exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^C) \geq b$$

Compared with the analysis for Theorem 1, if we can bound $\Pr\{y_i^I|\mathbf{x}_i, W^I\}$, the performance on the target classification task can be guaranteed. First, we try to bound the similarity between the example and the individual class. Considering $\forall j, j \neq y_i^C$, we have

$$f(\mathbf{x}_i)^\top\mathbf{w}_j^I = f(\mathbf{x}_i)^\top f(\mathbf{x}_j) + f(\mathbf{x}_i)^\top(\mathbf{w}_j^I - f(\mathbf{x}_j))$$
$$\leq f(\mathbf{x}_i)^\top f(\mathbf{x}_j) + c\|\mathbf{w}_j^I - f(\mathbf{x}_j)\|_2$$
$$\leq f(\mathbf{x}_i)^\top\mathbf{w}_j^C + c(\|f(\mathbf{x}_j) - \mathbf{w}_j^C\|_2 + \|\mathbf{w}_j^I - f(\mathbf{x}_j)\|_2)$$

Note that

$$(1-\beta)\exp(f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^C) \geq \beta\sum_{j\neq y_i^C}^{C}\exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^C)$$

we have

$$\forall j \neq y_i^C, \quad f(\mathbf{x}_i)^\top\mathbf{w}_j^C \leq \log(\frac{1-\beta}{\beta}) + f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^C$$

Therefore, the similarity can be further bounded as

$$f(\mathbf{x}_i)^\top\mathbf{w}_j^I \leq \log(\frac{1-\beta}{\beta}) + f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^C$$
$$+c(\|f(\mathbf{x}_j) - \mathbf{w}_j^C\|_2 + \|\mathbf{w}_j^I - f(\mathbf{x}_j)\|_2)$$
$$\leq \log(\frac{1-\beta}{\beta}) + f(\mathbf{x}_i)^\top(\mathbf{w}_{y_i}^C - \mathbf{w}_{y_i}^I + \mathbf{w}_{y_i}^I)$$
$$+c(\|f(\mathbf{x}_j) - \mathbf{w}_j^C\|_2 + \|\mathbf{w}_j^I - f(\mathbf{x}_j)\|_2)$$
$$\leq \log(\frac{1-\beta}{\beta}) + f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^I$$
$$+c(\|\mathbf{w}_{y_i}^C - f(\mathbf{x}_i)\|_2 + \|f(\mathbf{x}_i) - \mathbf{w}_{y_i}^I\|_2$$
$$+\|f(\mathbf{x}_j) - \mathbf{w}_j^C\|_2 + \|\mathbf{w}_j^I - f(\mathbf{x}_j)\|_2)$$

Note that the distance between an example and its corresponding parameters $\mathbf{w}_i$ can be bounded as in Theorem 1. Therefore, we have

$$\forall i, j : j \neq y_i^C, \exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^I) \leq \frac{1-\beta}{\beta}c'\exp(f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^I)$$

where

$$c' = \exp(2c(\sqrt{2c^2 - 2\log(a\alpha/(1-\alpha))}$$
$$+\sqrt{2c^2 - 2\log(b\beta/(1-\beta))}))$$

Then, we have

$$\Pr\{y_i^I|\mathbf{x}_i, W^I\} = \frac{\exp(f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^I)}{\sum_j^n\exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^I)}$$
$$= \frac{1}{\frac{\sum_{j:j=y_i^C}\exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^I)}{\exp(f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^I)} + \frac{\sum_{j:j\neq y_i^C}\exp(f(\mathbf{x}_i)^\top\mathbf{w}_j^I)}{\exp(f(\mathbf{x}_i)^\top\mathbf{w}_{y_i}^I)}}$$
$$\geq \frac{1}{1/\alpha + (1-\beta)c'M/\beta}$$

where M denotes the quantity of examples from different coarse classes as $M = |\{\mathbf{x}_j : y_j^C \neq y_i^C\}|$. Letting

$$\alpha' = \frac{1}{1/\alpha + (1-\beta)c''/\beta}$$

where $c'' = c'M$, we can obtain the guarantee by the similar analysis as in Theorem 1. □

2. Experiments

2.1. Synthetic Data

Besides comparing the performance on real-world data sets, we conduct an experiment on the synthetic data to illustrate the difference between patterns learned using different training labels on the same data set. The synthetic data is generated as follows. First, we randomly generate 32 big color patches and 128 small color patches as a pool of patches. Given a blank image, a big patch and a small patch are randomly sampled from the pool, and then added to the image. Finally, 512 images are obtained. Fig. 1 illustrates the process.

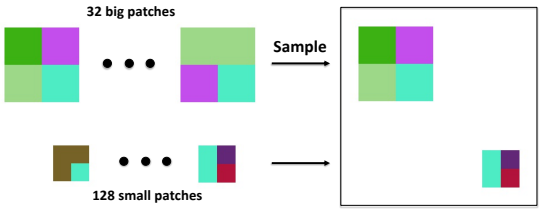

Figure 1. Data generation procedure for synthetic data.

For the synthetic data, coarse classes are defined with big patches and target classes are defined by small patches. Consequently, there are 32 coarse classes and 128 target classes in the data set. To investigate the different patterns learned by the neural network, we train the model with the objective of "Ins", "Cos" and "Opt", respectively. After that, we visualize the spatial attention maps of different models to illustrate the patterns exploited by these models. The detailed algorithm for computing attention maps can be found in [1].

Fig. 2 shows the attention maps of these three tasks. First, we can observe that deep learning can capture the most discriminative parts for a given task. For example, it can identify the big patches for the 32-class classification task and the small patches for the 128-class classification task. Second, the learned patterns for them are totally different. When training the model with the objective of "Cos", the neural network will ignore small patches, which are essential for the 128-class classification problem. It demonstrates that the patterns learned from the conventional pipeline with coarse labels only can be inappropriate for the target task. Finally, optimizing the loss for identifying each example as "Ins" can explore all patterns in images which may introduce additional noise for the target task (i.e., 128-class).

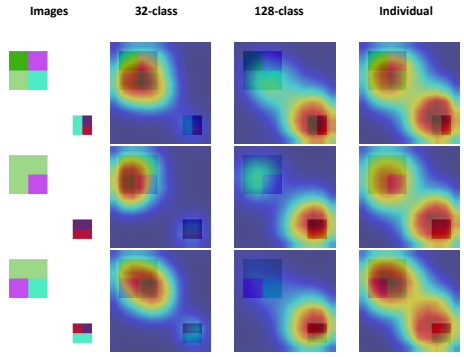

Figure 2. Illustration of different patterns learned from different tasks on the same synthetic data consisting of 512 images. According to different combinations of patches, three tasks are included: 32 coarse-class classification (i.e., 32-class with big patches), 128-class classification (i.e., 128-class with small patches) and instance-level classification (i.e., Individual with big and small patches).

2.2. Stanford Online Products

Results for coarse-class classification and retrieval are summarized in Table 1. It is evident that even when the target task is consistent with the training labels, exploring fine-grained patterns can achieve additional gain.

	Top1	Top5	R@1	R@2	R@4	R@8
Cos	80.1	97.3	76.6	84.3	89.5	93.2
$\text{CoIns}_{\text{imp}}$	80.6	97.5	77.1	84.5	89.7	93.4

Table 1. Comparison of accuracy and recall (%) for 12 coarse classes on SOP.

2.3. Ablation Study for Instance Proxy Loss

In this subsection, we conduct experiments to evaluate the effect of instance proxy loss in "CoInsP**". CIFAR-100 is adopted for the ablation study. The weight for the corresponding loss function is set as $\lambda_P = 1$.

Effect of M First, we investigate the effect of the number of epochs before the instance proxy loss is added for optimization. Since the learning rate will be first decayed at the 60-th epoch, we add the loss at $\{60, 80, 100, 120\}$ epochs and summarize the results in Table 2. The performance of "CoIns**" is included as a baseline. Apparently, including the instance proxy loss can improve the performance after sufficient training. It is consistent with our analysis that the parameters from instance classification can be applied to generate appropriate proxies for the target task when those

parameters can identify individual examples well. However, if we have the loss after another decay of learning rate at the 120-th epoch, the improvement vanishes. This phenomenon is due to the fact that the learning rate is too small to exploit the additional informative patterns effectively..

M	R@1	R@2	R@4	R@8
Baseline	60.5	71.1	79.8	86.5
60	61.4	71.5	79.7	86.3
80	61.7	71.7	80.0	86.5
100	**62.0**	**71.7**	**80.2**	**86.6**
120	60.8	70.6	79.6	86.5

Table 2. Ablation study on M before the instance proxy loss is added.

Effect of P Table 3 compares the performance by varying P. When $P = 25,000$, the number of clusters is half of that of total examples. There is no sufficient information within each cluster for CIFAR-100, since each cluster contains only about 2 instances. In contrast, a small P will lead to more aggregated clusters and can contain patterns that are related to the target task. If P is too small (i.e., the cluster size is too large), additional noise can be introduced and result in the suboptimal results as illustrated when $P = 500$. It confirms our analysis that a large P is important for the tight approximation.

P	R@1	R@2	R@4	R@8
25,000	60.5	70.8	79.6	86.6
10,000	**62.0**	**71.7**	**80.2**	**86.6**
5,000	61.2	71.2	78.8	85.4
1,000	58.9	69.0	76.5	82.3
500	57.9	67.9	76.0	82.1

Table 3. Ablation study on the number of clusters when using the instance proxy loss.

References

[1] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 3