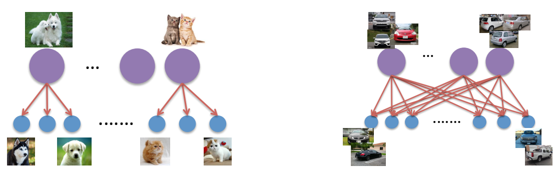
Hyper-class Augmented and Regularized Deep Learning for Fine-grained Image Classification

Saining Xie[1], Tianbao Yang[2] Xiaoyu Wang[3], Yuanqing Lin[4]
[1]University of California, San Diego. [2]University of Iowa. [3]Snapchat Research. [4]NEC Labs America, Inc.

Fine-grained image classification (FGIC) is challenging because (i) fine-grained labeled data is much more expensive to acquire (usually requiring domain expertise); (ii) there exists large intra-class and small inter-class variance. In this paper, we propose a systematic framework of learning a deep CNN that addresses the challenges from two new perspectives: (i) identifying easily annotated hyper-classes inherent in the fine-grained data and acquiring a large number of hyper-class-labeled images from readily available external sources, and formulating the problem into multi-task learning, to address the data scarcity issue. We use two common types of hyper-classes to augment our data, with one being the **super-type** hyper-classes that subsume a set of fine-grained classes, and another being named **factor-type** hyper-classes (e.g., different view-points of a car) that explain the large intra-class variance. (ii) a novel learning model by exploiting a regularization between the fine-grained recognition model and the hyper-class recognition model to mitigate the issue of large intra-class variance and improve the generalization performance. The proposed approach also closely relates to attribute-based learning, since one can consider that factor-type hyper-classes are (or can be generalized to) object attributes. We demonstrate the success of the proposed framework on two small-scale fine-grained datasets (Stanford Dogs and Stanford Cars) and on a large-scale car dataset that we collected.



(a) Super-type (b) Factor-type

Figure 1: Two types of relationships between hyper-classes and fine-grained classes.

hyper-class regularized learning. As a factor-type hyper-class can be considered as a hidden variable for generating the fine-grained class, therefore we model $\Pr(y|\mathbf{x})$ by

$$\Pr(y|\mathbf{x}) = \sum_{v=1}^{K} \Pr(y|v,\mathbf{x}) \Pr(v|\mathbf{x}) \tag{1}$$

where $\Pr(v|\mathbf{x})$ is the probability of any factor-type hyper-class v and $\Pr(y|v,\mathbf{x})$ specifies the probability of any fine-grained class given the factor-type hyper-class and the input image $\mathbf{x}$. If we let $\mathbf{h}(\mathbf{x})$ denote the high level features of $\mathbf{x}$, we model the probability $\Pr(v|\mathbf{x})$ by a softmax function

$$\Pr(v|\mathbf{x}) = (\exp(\mathbf{u}_v^\top \mathbf{h}(\mathbf{x})))/(\sum_{v'=1}^{K} \exp(\mathbf{u}_{v'}^\top \mathbf{h}(\mathbf{x}))) \tag{2}$$

where $\{\mathbf{u}_v\}$ denote the weights for the hyper-class classification model. Given the factor-type hyper-class v and the high level features $\mathbf{h}$ of $\mathbf{x}$, the probability $\Pr(y|v,\mathbf{x})$ is computed by

$$\Pr(y=c|v,\mathbf{x}) = (\exp(\mathbf{w}_{v,c}^\top \mathbf{h}(\mathbf{x})))/(\sum_{c=1}^{C} \exp(\mathbf{w}_{v,c}^\top \mathbf{h}(\mathbf{x}))) \tag{3}$$

where $\{\mathbf{w}_{v,c}\}$ denote the weights of factor-specific fine-grained recognition model. Putting together (2) and (3), we have the following predictive probability for a specific fine-grained class, and we use this equation to make the final predictions

$$\Pr(y=c|\mathbf{x}) = \sum_{v=1}^{K} \frac{\exp(\mathbf{w}_{v,c}^\top \mathbf{h}(\mathbf{x}))}{\sum_{c=1}^{C} \exp(\mathbf{w}_{v,c}^\top \mathbf{h}(\mathbf{x}))} \frac{\exp(\mathbf{u}_v^\top \mathbf{h}(\mathbf{x}))}{\sum_{v'=1}^{K} \exp(\mathbf{u}_{v'}^\top \mathbf{h}(\mathbf{x}))} \tag{4}$$

We introduce the following regularization between $\{\mathbf{w}_{v,c}\}$ and $\{\mathbf{u}_v\}$,

$$R(\{\mathbf{w}_{v,c}\},\{\mathbf{u}_v\}) = \frac{\beta}{2} \sum_{v=1}^{K} \sum_{c=1}^{C} \|\mathbf{w}_{v,c} - \mathbf{u}_v\|_2^2 \tag{5}$$

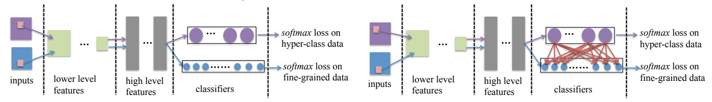
The regularization is responsible for *transferring the knowledge* to the per-viewpoint category classifier and thus helps mode the intra-class variance in the fine-grained task. The only difference for super-type hyper-class regularized deep learning is on $\Pr(y|v,\mathbf{x})$, which can be simply modeled by

$$\Pr(y=c|v_c,\mathbf{x}) = (\exp(\mathbf{w}_{v_c,c}^\top \mathbf{h}(\mathbf{x})))/(\sum_{c=1}^{C} \exp(\mathbf{w}_{v_c,c}^\top \mathbf{h}(\mathbf{x})))$$

since the super-type hyper-class v_c is implicitly indicated by the fine-grained label c. The regularization then becomes

$$R(\{\mathbf{w}_{v_c,c}\},\{\mathbf{u}_v\}) = \frac{\beta}{2} \sum_{c=1}^{C} \|\mathbf{w}_{v_c,c} - \mathbf{u}_{v_c}\|_2^2 \tag{6}$$

Figure 2: **Network Structures**



(a) HA-CNN (b) HAR-CNN

Experimental Results

Table 1: Accuracy on Stanford-Cars [2] dataset.

Method	Accuracy(%)
LLC	69.5
ELLF	73.9
ImageNet-Feat-LR	54.1
CNN	68.6
HA-CNN-Random	69.8
FT-CNN	83.1
HA-CNN (ours)	76.7
HAR-CNN (ours)	**80.8**
FT-HA-CNN (ours)	83.5
FT-HAR-CNN (ours)	**86.3**

Table 2: Accuracy on Stanford-Dogs dataset.

Method	Acc
UGA	**57.0**
Gnostic Fields	47.7
CNN	42.3
HA-CNN (ours)	48.3
HAR-CNN (ours)	**49.4**

Table 3: Accuracy on Large-scale Cars dataset.

Method	Acc
ImageNet-Feat-LR	42.8
CNN	81.6
HA-CNN (ours)	82.4
HAR-CNN (ours)	**83.6**

Our network structure and experiment settings are built upon Alex-Net [3]. experimental results demonstrate that, when training on small-scale dataset from scratch and without any fine-tuning, the proposed approach enables us to train a model that yields reasonably good performance. When integrated in the ImageNet fine-tuning process [1], our approach significantly outperforms the current state-of-the-art on Stanford Cars dataset. To further explore if the proposed framework is still useful when training on large-scale FGIC, we collect a large dataset and perform experiments similar to those for the Stanford Cars data.

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.

[2] Michael Stark Jonathan Krause, Jia Deng and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. *The Second Workshop on Fine-Grained Visual Categorization*, 2013.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pages 1097–1105, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.