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

Saining Xie
University of California, San Diego
s9xie@eng.ucsd.edu

Tianbao Yang
Department of Computer Science
University of Iowa
tianbao-yang@uiowa.edu

Xiaoyu Wang
Snapchat Research
xiaoyu.wang@snapchat.com

Yuanqing Lin
NEC Laboratories America, Inc.
ylin@nec-labs.com

Abstract

Deep convolutional neural networks (CNN) have seen tremendous success in large-scale generic object recognition. In comparison with generic object recognition, fine-grained image classification (FGIC) is much more 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. Most recent work exploiting deep CNN for image recognition with small training data adopts a simple strategy: pre-train a deep CNN on a large-scale external dataset (e.g., ImageNet) and fine-tune on the small-scale target data to fit the specific classification task. In this paper, beyond the fine-tuning strategy, 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 (e.g., image search engines), and formulating the problem into multi-task learning; (ii) a novel learning model by exploiting a regularization between the fine-grained recognition model and the hyper-class recognition model. 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.

1. Introduction

The goal of FGIC is to recognize objects that are both semantically and visually similar to each other. Since the seminal work of [23], deep convolution neural networks (CNN) have achieved the state-of-the-art performance on large-scale image classification [41, 33]. An important factor in the success of deep CNNs is the access of large-scale labeled training data. However, because it is often expensive to obtain a large number of labeled images in fine-grained image classification tasks, it can be difficult to train a good deep CNN on a small dataset without suffering from significant overfitting. Several works have used the ImageNet dataset (containing 1.2 million images from 1,000 classes) to pre-train a deep CNN and then directly use the resulting CNN to extract features that are then directly used in a fine-grained image recognition task at hand [1]. However, as we report below, the features learned from a generic dataset might not be well suited for a specific FGIC task. Recently, more attempts followed the strategy in [17], where network parameters are fine-tuned on the target FGIC data.

Figure 1. Illustration of large intra-class variance due to different views.

Another challenge for fine-grained image classification is that intra-class variation can be quite large due to differences in pose, view-point, etc. One example of this issue is illustrated in Figure 1 for fine-grained car recognition.

In this paper, we propose a principled framework to explicitly tackle the challenges of learning a deep CNN for FGIC.

Our first contribution is to propose a task-specific data augmentation approach to address the data scarcity issue.

[1] https://sites.google.com/site/fgcomp2013/

We augment the FGIC dataset with external data annotated by some hyper-classes, which are inherent attributes of fine-grained data. We can easily acquire a large number of hyper-class-labeled images from readily available sources, such as online search engines (either keyword-based or content-based). 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. Then we formulate the problem into multi-task deep learning, allowing the two tasks (fine-grained classification on target data and hyper-class classification on augmented data) with disjoint label set to share and learn the same feature layers.

Our second contribution in the paper is to propose a novel regularization technique in the multi-task deep learning that exploits the relationship between the fine-grained classes and the hyper-classes to provide explicit guidance to the learning process at the classifier level. When exploiting factor-type hyper-classes that explain the intra-class variance, the proposed learning model is able to mitigate the issue of large intra-class variance and improve the generalization performance. We name the proposed framework as **hyper-class augmented and regularized deep learning**.

To demonstrate the effectiveness of the proposed framework, we first perform experiments on two relatively small-scale fine-grained datasets, namely Stanford Dogs and Stanford Cars. We augment the fine-grained dogs data by using the super-type hyper-class (dog) and simultaneously learn a dog vs cat recognition model using a publicly available dataset in which each image is annotated as dog or as cat. To augment the Stanford Cars data, we exploit the factor-type hyper-classes of 3D view-points, by utilizing an online image search engine.

Our 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, 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 containing $157,023$ car images from 333 categories and perform experiments similar to those for the Stanford Cars data.

2. Related Work

FGIC has recently received a surge of interest in computer vision [40, 42, 14, 39, 19]. To train a fine-grained recognition model, one can first extract features from images and then train a multi-class classifier based on the derived feature representation. Many features can be computed from an image, ranging from traditional features such as SIFT [28] and HOG [7], to visual word features [6, 38, 36], and recently proposed deep convolutional activation features extracted from the activation of a deep CNN that is pre-trained on a large, fixed set of object recognition tasks (e.g., the 1,000 objects recognition task of ImageNet Challenge) [10]. There have been few works that directly train a (deep) neural network from the fine-grained images; one exception is Gnostic Fields [19], which can be interpreted as a kind of feed-forward neural network relying on hand-engineered features.

Our model is built upon the recent success of deep CNN for visual recognition. In the notable work by Krizhevsky et al. [23], the authors developed a large deep CNN with 5 convolutional layers and 3 fully connected layers.The similar deep learning architecture has achieved state-of-the-art performance on other visual tasks, including face recognition [34], object detection [30, 17], and human pose estimation [35]. Novel variants [27, 26] have been proposed and achieved new state-of-the-art performance on tiny-image datasets like CIFAR-10 and CIFAR-100 [22]. In ImageNet 2014 Challenge, [31] and [33] designed very deep CNN architectures and achieved impressive results. The development in generic image object recognition brings benefits to the FGIC community: Training those deep models requires significant time and resources, but one can easily utilize the trained models by fine-tuning these models on the new dataset and get great performance boost. The problem of having scarce labeled data is alleviated because the pre-trained deep networks can generalize extremely well and provides good layer initializations in the fine-tuning process. Our implementation is based on Krizhevsky's widely used CNN model, but the proposed learning framework can apply to any deep learning architecture.

Large variation of view-point, pose, appearance etc. for visual recognition has long been recognized and studied in computer vision. Out of several proposed methods, part-based methods have gained significant recent attention to tackle view-point and pose variation, including DPM model [15] and poselets [2]. Part-based methods have also been applied to FGIC with the key concept of pose-normalization [42, 14], which localizes object parts and establishes their correspondences for deriving an intermediate level representation of fine-grained images. Zhang et al. [43] proposed a method that combines part-based models and deep learning by training pose-normalized CNNs for inferring human attributes. However, these methods usually require intensive human annotations of the data and therefore are restricted to small datasets. There are some other studies exclusively devoted to addressing particular variations, e.g. view-points for car recognition [11, 24, 21]. Nevertheless, as far as we know, there has been no study on designing a deep CNN for FGIC to explicitly model large

intra-class variance.

The proposed learning framework is closely related to neural networks with multi-task learning [3]. The idea is to jointly train multiple related tasks by allowing them share the same feature layers of the neural network. Multi-task learning has been incorporated into deep learning in several applications. For example, Collobert and Weston [5] proposed a multi-task deep learning method for natural language processing (NLP), which trains jointly multiple NLP prediction tasks, e.g., predicting part-of-speech tags, chunks, named entity tags, semantic roles, etc. Seltzer and Droppo [29] applied multi-task learning to deep neural networks for improving phoneme recognition. It has been observed that multitask learning can improve the generalization of the shared tasks. However, in contrast to traditional multi-task learning for deep neural networks that aims to transfer knowledge only by sharing the lower level features in a "blind" way, we explicitly use the classifier learned for hyper-class recognition to regularize the classifier for the fine-grained recognition, enabling the sharing of weights among fine-grained classes while maintaining discriminative power.

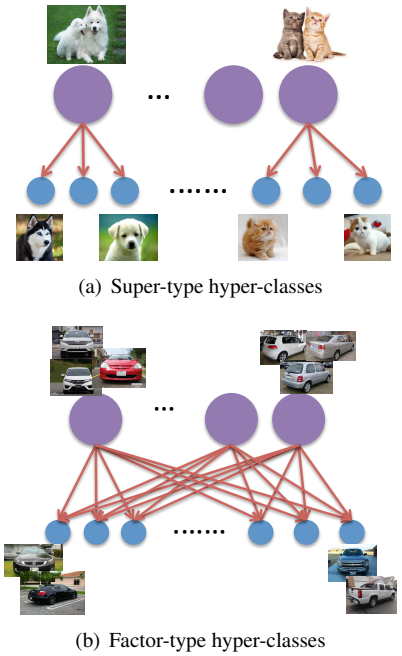
It should be noted that the proposed approach also closely relates to attribute-based learning [13, 25, 37], since one can consider that factor-type hyper-classes are (or can be generalized to) object attributes. To the best of our knowledge, our work is the first to exploit attribute-based learning and information sharing in a unified deep learning framework.

3. Hyper-class Augmented and Regularized Deep Learning

In this section, we present the proposed hyper-class augmented and regularized deep learning framework for tackling the challenges of FGIC. The first challenge for FGIC is that fine-grained labels are expensive to obtain, requiring intensive labor and domain expertise. Therefore the available labeled training data is usually insufficiently large to train a deep CNN without overfitting. The second challenge is large intra-class variance vs small inter-class variance. To address the first challenge, we propose a data augmentation method. The key idea is to augment the fine-grained data with a large number of auxiliary images labeled by some hyper-classes, which are inherent attributes of fine-grained data and can be much more easily annotated. To address the second challenge, we propose a novel deep CNN model that can fully utilize the hyper-class labeled augmented data.

3.1. Hyper-class Data Augmentation

Typical data augmentation approaches in visual recognition are translations (cropping), reflections, and adding random noise to the images. However, their improvement for fine-grained image classification is limited because patches from different fine-grained classes could be more similar to each other (small inter-class variance), leading to difficulties in discriminating them, after these augmentations have been applied. We propose a data augmentation approach to address the scarcity of labeled fine-grained images. Our approach is inspired by the fact that images have other inherent "attributes" besides the fine-grained classes, which can be annotated with much less effort than fine-grained classes, and therefore a large number of images annotated by these inherent attributes can be easily acquired. We will refer to these easily annotated inherent attributes as hyper-classes.



(a) Super-type hyper-classes

(b) Factor-type hyper-classes

Figure 2. Two types of relationships between hyper-classes and fine-grained classes.

The most common hyper-class is super-type class, which subsumes a set of fine-grained classes. For example, a fine-grained dog or cat image can be easily identified as a dog or cat. We can acquire a large number of dog and cat images by fast human labeling or from external sources such as image search engines. The hierarchy of super-type hyper-classes and fine-grained classes has been exploited in many previous studies. For instance, Deng et al. [8] proposed a HEX graph to capture the hierarchical and exclusive relationships between classes and defined a joint distribution of an assignment of all classes as a conditional random field. Srivastava & Salakhutdinov [32] exploited the class hierarchy for transfer learning, in which one aims to improve the classification of lower level classes (with a small number of examples) by transferring knowledge among similar lower level classes. However, in contrast to these approaches that restrict learning to the given training data (either assuming

Figure 3. An query image (left) and retrieved images by Google (top right) and by Baidu (bottom right). Baidu search engine allows us to retrieve a large number of images with clean view annotations.

the class hierarchy is known or inferring the class hierarchy from the data), our approach is based on data augmentation which enables us to utilize as many auxiliary images as possible to improve the generalization performance of the learned features.

Besides the super-type hyper-class that captures 'a kind of' relationship, we also consider another important hyper-class to capture 'has a' relationship and to explain the intra-class variances (e.g., the pose variance). In the following discussion, we focus on fine-grained car recognition as an instance. A fine-grained car image annotated by make, model and year could be photographed from many different views, yielding images from the same fine-grained class with very different visual appearances (see Figure 1). For a particular fine-grained class, images could have different views (i.e., factor-type hyper-classes). This is completely different from the class hierarchy between a super-type hyper-class and fine-grained classes, where a fine-grained class can only belong to one single super-type class, for example, a fine-grained class "Chihuahua" belongs to a super-type class "Dog", but not "Cat". However, a fine-grained class of car in the dataset (e.g. "Acura TSX 2006") could have images from different views (e.g. "Passenger-side", "Front", "Back", etc.), and thus not necessarily belongs to one single hyper-class. From a generative perspective, the fine-grained class of a car image can be generated by first generating its view (hyper-class) and then generating the fine-grained class given the view. This is also the probabilistic foundation of our model described in next subsection. Since this type of hyper-class can be considered as a hidden factor of an image, we refer to this type of hyper-class as a factor-type hyper-class. The key difference between super-type and factor-type hyper-class is that a super-type hyper-class is implicitly implied by the fine-grained class while the factor-type hyper-class is unknown for a given fine-grained class. Without annotation of the factor class of an image, there is no way to infer that from the fine-grained class. Another example of factor-type hyper-classes is different expressions (happy, angry, smile, and etc) of a human face, where each individual can have multiple photos with different expressions. Although intra-class variance has been studied previously, to the best of our knowledge, this is the first work that explicitly models the intra-class variance to improve the performance of deep CNN. Figure 2 illustrates the two types of hyper-classes. Next, we use fine-grained car recognition as an example to discuss how to obtain a large number of auxiliary images annotated by different views. An existing approach for acquiring additional car images with specific view-points is by rendering images from 3D CAD models [21]. The issue of rendered images is that they are not photo-realistic and thus they may not follow the same distribution of real images. To surmount this issue, we put forward a more effective and efficient approach by exploiting the recent advances of online content-based image search engines. Modern image search engines have the capability to retrieve visually similar images to a given query image. We investigated several image search engines, including Google, Baidu and Bing. Bing can only find exactly matched images and corresponding source pages, which is not useful for our purpose. Google and Baidu can serve our purpose to find visually similar images. We found that images retrieved by Baidu are more suited for view prediction since it always returns visually similar images, while Google image search tries to "semantically" recognize the car and return images within the same model. To demonstrate this, we show in Figure 3 the top 9 or 10 images returned by Google and Baidu for a given query image. In our experiments, we use images retrieved from Baidu as our augmented data.

3.2. Hyper-class Regularized Learning Model

Before describing the details of our model, we first introduce some notation and terminology used throughout the paper. Let $\mathcal{D}_t = \{(\mathbf{x}_1^t, y_1^t), \ldots, (\mathbf{x}_n^t, y_n^t)\}$ be a set of training fine-grained images with $y_i^t \in \{1, \ldots, C\}$ indicating the fine-grained class label (e.g., make, model and year of a car) of image $\mathbf{x}_i^t$, and let $\mathcal{D}_a = \{(\mathbf{x}_1^a, v_1^a), \ldots, (\mathbf{x}_m^a, v_m^a)\}$ be a set of auxiliary images, where $v_i^a \in \{1, \ldots, K\}$ indicates the hyper-class label of image $\mathbf{x}_i^a$ (e.g., view-point of a car). Using v to denote a super-type hyper-class, by v_c we denote the super-type hyper-class of the fine-grained class c. In the sequel, the two terms 'classifier' and 'recognition model' are used interchangeably.

The goal is to learn a recognition model that can predict the fine-grained class label of an image. In particular, we aim to learn a prediction function given by $\Pr(y|\mathbf{x})$, i.e.,

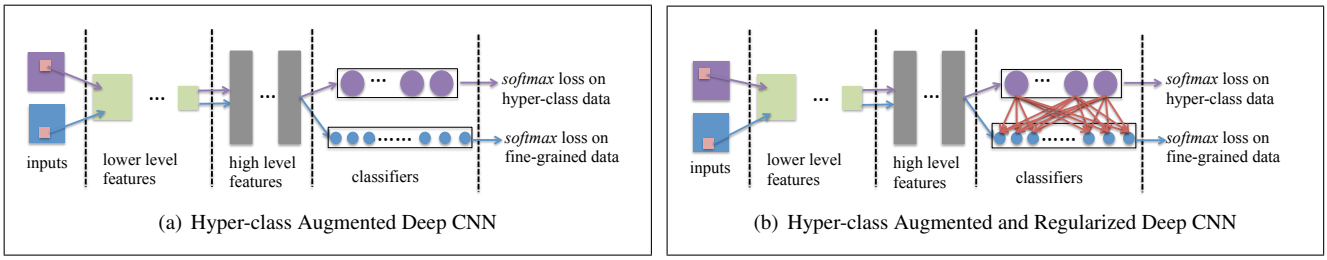


Figure 4. The network structures.

given the input image what is the probability that it belongs to a fine-grained class. Similarly, we let $\Pr(v|\mathbf{x})$ denote the hyper-class classification model. Given the fine-grained training images and the auxiliary hyper-class labeled images, a straightforward strategy is to train a multi-task deep CNN, by sharing common features and learning classifiers separately. Multi-task deep learning has been observed to improve the performance of individual tasks [3]. We reiterate that in the multi-task learning, the label sets of hyper-classes and fine-grained classes are disjoint, and we don't label the fine-grained data with hyper-class labels.

To further improve this simple strategy, we propose a novel multi-task regularized learning framework by exploiting regularization between the fine-grained classifier and the hyper-class classifier. We begin with the description of the model regularized by factor-type hyper-class.

3.2.1 Factor-type 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

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

where $\{\mathbf{u}_v\}$ denote the weights for the hyper-class classification model. Note that in all formulations we ignore the bias term since it is irrelevant to our discussion. Nevertheless it should be included in practice. 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

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

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

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

Although our model has its root in mixture models [1], however, it is worth noting that unlike most previous mixture models that treat $\Pr(v|\mathbf{x})$ as free parameters, we formulate it as a discriminative model. It is the hyper-class augmented images that allow us to learn $\{\mathbf{u}_v\}$ accurately. Then we can write down the negative log-likelihood of data in $\mathcal{D}_t$ for fine-grained recognition and that of data in $\mathcal{D}_a$ for hyper-class recognition, i.e.,

$$\begin{aligned} L(\{\mathbf{w}_{v,c}\},\{\mathbf{u}_v\}) &= -\log \Pr(\mathcal{D}) \\ &= -\sum_{i=1}^{n}\sum_{c=1}^{C} \delta(y_i^t, c) \log \Pr(y=c|\mathbf{x}_i^t) \\ &\quad -\sum_{i=1}^{m}\sum_{v=1}^{K} \delta(v_i^a, v) \log \Pr(v|\mathbf{x}_i^a) \end{aligned} \tag{5}$$

To motivate the non-trivial regularization, we note that factor-specific weights $\mathbf{w}_{v,c}$ should capture similar high-level factor-related features as the corresponding factor-type hyper-class classifier $\mathbf{u}_v$.

To this end, we introduce the following regularization between $\{\mathbf{w}_{v,c}\}$ and $\{\mathbf{u}_v\}$,

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

To understand the regularization, in our car recognition example, original fine-grained data is not capable to learn a per-viewpoint category classifier $\mathbf{w}_{v,c}$, because there is no way to infer the viewpoint hyper-class. But now we can train viewpoint classifiers $\mathbf{u}_v$ on hyper-class augmented

data, so 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 above regularization can also be interpreted by imposing a normal prior on $\mathbf{w}_{v,c}$ by

$$\Pr(\mathbf{w}_{v,c}|\mathbf{u}_v) \propto \exp\left(-\frac{\beta}{2}\|\mathbf{w}_{v,c}-\mathbf{u}_v\|_2^2\right)$$

The regularization in (6) enjoys another interesting intuition of sharing weights among the factor-type hyper-class recognition model and the fine-grained recognition model. To see this, we introduce $\mathbf{w}'_{v,c} = \mathbf{w}_{v,c} - \mathbf{u}_v$ and write the regularizer in (6) as

$$R\left(\{\mathbf{w}'_{v,c}\}\right) = \frac{\beta}{2}\sum_{v=1}^{K}\sum_{c=1}^{C}\|\mathbf{w}'_{v,c}\|_2^2$$

and $\Pr(y=c|\mathbf{x})$ is computed by

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

It can be seen that the fine-grained classifier share the same component $\mathbf{u}_v$ of the factor-type hyper-class classifier. It therefore connects the proposed model to weight sharing employed in traditional shallow multi-task learning [12, 4].

3.2.2 Super-type hyper-class regularized learning

The only difference for super-type hyper-class regularized deep learning is on $\Pr(y|v,\mathbf{x})$, which can be simply modeled by

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

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

$$R\left(\{\mathbf{w}_{v_c,c}\},\{\mathbf{u}_v\}\right) = \frac{\beta}{2}\sum_{c=1}^{C}\|\mathbf{w}_{v_c,c}-\mathbf{u}_{v_c}\|_2^2 \qquad (7)$$

It is notable that a similar regularization has been exploited in [32]. However, there is a big difference between our work and [32]. In our model, the weight $\mathbf{u}_v$ for the super-type classification is also discriminatively learned from the augmented auxiliary data.

3.3. A Unified Deep CNN

Using the hyper-class augmented data and the multi-task regularization learning technique, we reach to a unified deep CNN framework as depicted in Figure 4 (right column). We also exhibit the optimization problem:

$$\min_{\{\mathbf{w}_{v,c}\},\{\mathbf{u}_v\},\{\mathbf{w}_l\}} L(\{\mathbf{w}_{v,c}\},\{\mathbf{u}_v\}) + R(\{\mathbf{w}_{v,c}\},\{\mathbf{u}_v\}) + \sum_{v=1}^{K}r(\mathbf{u}_v) + \sum_{l=1}^{H}r(\mathbf{w}_l)$$

where $\mathbf{w}_l, l=1,\ldots,H$ denote all the weights of the CNN in determining the high level features $h(\mathbf{x})$, H denotes the number of layers before the classifier layers, and $r(\mathbf{w})$ denotes the standard Euclidean norm square regularizer with an implicit regularization parameter (or a weight decay parameter).

3.4. Training

The proposed deep learning model is trained by back-propagation using mini-batch stochastic gradient descent with settings similar to that in [23]. A key difference is that we have two sources of data and two loss functions corresponding to the two tasks. It is very important to sample both images in $\mathcal{D}_t$ and images in $\mathcal{D}_a$ in each mini-batch to compute the stochastic gradients. Using the alternative approach that alternates between training the two tasks could yield very bad solutions. This is because the two tasks may have different local optimal in different directions and the solution can be easily trapped into a bad local optimal.

4. Experiments

4.1. Model Architecture

We use exactly the same feature layers as in [23], including 5 convolutional layers and 2 fully connected layers. We refer to [23] for a detailed discussion of the architecture and training protocol. We emphasize that we do not intend to optimize the design of the feature layers but rather focus our attention on different learning strategies. The code and hyper-parameter settings are developed based on Krizhevsky's cuda-convnet. To validate our development, we first repeat their experiments on ImageNet-2012 data. Our instance of the model attains an error rate of 41.6% on the validation set. We refer to this model as Alex-net and use it to extract features for the ImageNet-Feat-LR baseline as described below.

4.2. Baselines

We compare our approach with three baselines: (i) **ImageNet-Feat-LR**, which learns a multinomial logistic regression (LR) classifier on the activation features extracted using a deep CNN pre-trained on ImageNet data; (ii) the **CNN** baseline trained directly on the given fine-grained images. (iii) the **FT-CNN** baseline in which the pre-trained network is fine-tuned on the fine-grained images. For the

ImageNet-Feat-LR baseline, we use the activation feature extracted from the fully connected layer 6 and train a logistic regression classifier with dropout. For the FT-CNN baseline, we initialize the network with the pre-trained Alex-net model and fine-tune it on our target data. For our approach, we report two results when training from scratch on the target data: one for hyper-class augmented deep CNN (**HA-CNN**) and another for hyper-class augmented and regularized deep CNN (**HAR-CNN**) for examining the effect of the two components. To show that in our multitask learning setting, HA-CNN, the hyper-class data are indeed useful, we report another baseline (**HAR-CNN-Random**), where instead of hyper-class data, irrelevant data are used as the auxiliary data. In order to show that our model is compatible with the fine-tuning strategy, we further report two results (**FT-HA-CNN**) and (**FT-HAR-CNN**) on Stanford-Car dataset, where instead of training the network from scratch, we initialize the network layers from a pre-trained model. We also report state-of-the-art results where available. Note that previous state-of-the-art methods utilized bounding box information provided in the dataset during training and testing process. In all experiments on Stanford-Cars and Stanford-Dogs datasets, for training, we mix both original and cropped images; for testing, we average the predictions of cropped and original images. Unless specified, all the results are obtained using standard 10-view testing.

4.3. Stanford Dogs

The Stanford-Dogs dataset [9] contains 20,580 images from 120 fine-grained classes. These images were taken from ImageNet for fine-grained image categorization. We use the official training/testing splitting. For the data augmentation, we use super-type hyper-class-labeled images from an external dataset – Asirra [2], which was also used in Kaggle's 2013 Dogs vs Cats contest. It contains a total of 25,000 images annotated by either a dog or a cat. 20,000 images are used as the training set in the auxiliary task to classify a dog from a cat and 5000 images are used as the validation set for parameter tuning. The regularization is only imposed on the weights for fine-grained dog recognition and the generic dog recognition. The results are shown in Table 1. Note that recently unsupervised grid alignment method [16] reports an accuracy of 57.0%. It extracts multi-channel local descriptors and utilizes unsupervised part detection and segmentation. However, our focus is to show how our framework can benefit the end-to-end deep learning framework that treat the FGIC task in a holistic recognition point of view. The relative improvement (6% for HA-CNN and 7% for HAR-CNN) compared to a CNN baseline validates our idea.

[2]http://research.microsoft.com/en-us/um/redmond/projects/asirra/

Table 1. Accuracy on Stanford-Dogs dataset. The performance of the baseline ImageNet-Feat-LR on Stanford Dogs data is not reported because the this dataset is a subset of ImageNet data.

Method	Accuracy(%)
Unsupervised Grid Alignment [16]	**57.0**
Gnostic Fields [19]	47.7
CNN	42.3
HA-CNN (ours)	48.3
HAR-CNN (ours)	**49.4**

4.4. Stanford Cars

Stanford Cars dataset [19] consists of 196 classes and 16,185 images. In contrast to Stanford dogs, we exploit factor-type hyper-classes for data augmentation and regularization, because we can easily collect a huge number of view labeled images as discussed in section 3.1. To do this, we have identified eight different views: front, back, driver/passenger side, left-front, right-front, left-back and right-back with exemplar images shown in Figure 5. For each view, we randomly select seed images from the training dataset as query images and retain top 10,000 images retrieved by Baidu image search engine. In the experiments, we observe that these images have very high quality in terms of car pose labels. The results [3] are shown in Table 2. The results show that our proposed framework performs extremely well on this tasks. Note that in HA-CNN-Random baseline, we use labeled images from 50 random classes from ImageNet (to match the amount of augmented data in HA-CNN experiment) and perform multi-task training. The result shows that adding random auxiliary data during training barely increases the performance, while adding task-specific hyper-class data brings substantial gain. This conclusion well matches the observations in traditional multi-task learning research. By collecting and utilizing the properly designed hyper-class data, we are able to make the best of the discriminative power of deep learning, especially when we are training the network on small scale data from scratch.

Table 2. Accuracy on Stanford-Cars dataset.

Method	Accuracy(%)
LLC [18]	69.5
ELLF [20]	73.9
ImageNet-Feat-LR	54.1
CNN	68.6
HA-CNN-Random	69.8
HA-CNN (ours)	76.7
HAR-CNN (ours)	**80.8**

[3]Test image labels are not publicly available. An online evaluation server is available: http://ai.stanford.edu/ jkrause/cars/car_dataset.html

Figure 5. Exemplar retrieved images from different views (from left to right: front, back, side, right-front and right-back).

4.5. Hyper-class Training in Fine-tuning

Table 3. Fine-tuning Accuracy on Stanford-Cars dataset.

Method	Accuracy(%)
FT-CNN	83.1
FT-HA-CNN (ours)	83.5
FT-HAR-CNN (ours)	**86.3**

Though we already outperform the state-of-the-art by training on small FGIC data from scratch, we cannot ignore the recent success and the benefit of fine-tuning from a model pre-trained on ImageNet. Our hypothesis is that hyper-class augmentation and regularization technique should be compatible with the fine-tuning paradigm. Here we report our results on Stanford-Car dataset in Table 3. fine-tuning the ImageNet pre-trained works surprisingly well on this dataset, which already bypass the best performance we get when training from scratch. Adding hyper-class auxiliary data further increase the performance by a minor margin. This is reasonable since the problem of scarce data is mostly solved by the fine-tuning process. The hyper-class regularization, however, shows substantial impact on the learning process, when testing with only a single center view, the performance is 78.8% for FT-CNN vs 82.6% for FT-HAR-CNN. When testing in standard 10-view, our approach sets the new state-of-the-art result with 86.3% accuracy on this dataset.

4.6. Hyper-class Training at Large

Though in this paper we focus on FGIC tasks with small dataset, our another hypothesis is that our framework can still be useful when we train an FGIC task at large. To this end, we have collected a large number of car images from the Internet, which were all naturally photographed. We manually check all crawled images, retain only those that include the whole car in the images and remove images either focusing on some parts or on interior of the car. Annotations of the make, model and year are done manually with the help of meta information. Finally, we have $157,023$ training images with each labeled into one of 333 categories. Following the same procedure, we collected 7840 testing images with no overlapping with the training images. In contrast to Stanford Cars data, we do not label and use the bounding box annotations for images.

This is the largest dataset for car recognition up to date in both the number of training images and the number of categories. We use the same set of view labeled images for the view recognition task. The results of different learning strategies are shown in Table 4. Note that when enough data is available, a vanilla CNN can perform surprisingly well on this difficult task without any engineering tricks. However the experimental results demonstrate that the proposed deep learning framework is still effective even when the number of fine-grained images is large.

Table 4. Accuracy on Large-scale Cars dataset

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

From the results on all datasets, we can observe that (i) the features exacted from a pre-trained CNN on the present ImageNet data may not be suited for fine-grained classification; (ii) the proposed HAR-CNN dramatically improves the performance of fine-grained classification on small-scale datasets; (iii) exploiting the regularization between the fined-grained classes and hyper-classes further helps improve the generalization.

5. Conclusions

We have presented a hyper-class augmented and regularized deep learning framework for FGIC. To address the scarcity of data in FGIC, we propose a novel data augmentation approach by identifying inherent and easily annotated hyper-classes in the fine-grained data and collecting a large amount of similar images labeled by hyper-classes. The hyper-class augmented data can generalize feature learning by incorporating multi-task learning into a deep CNN. To further improve the generalization performance and deal with large intra-class variance, we have proposed a novel regularization technique that exploits the relationship between the fine-grained classes and the associated hyper-classes. We demonstrated the success of the proposed framework on both publicly available small-scale fine-grained datasets and a large self-collected car dataset.

We hope that the proposed deep joint learning and regularization framework can open up new directions of research in deep learning. For example, one could consider multi-task deep learning that incorporates regularization between different tasks. As we have mentioned, the proposed approach is closely related to attribute-based learning. Though current formulations can only use one attribute, it can be modified to handle multiple attributes by adding more tasks and using pair-wise weight regularization. We will explore this in our future work.

Acknowledgment

The authors would like to thank the reviewers for their comments and suggestions that have improved the manuscript. The work of T. Yang was supported in part by National Science Foundation (IIS-1463988). S. Xie gratefully acknowledges Hao Su for insightful discussions and the support of NVIDIA Corporation with their donation of GPUs.

References

[1] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006. 5

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1365–1372. IEEE, 2009. 2

[3] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 3, 5

[4] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 137–144. ACM, 2009. 6

[5] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008. 3

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 2

[8] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *Proceedings of the European Conference on Computer Vision (ECCV) 2014*, pages 48–64. Springer, 2014. 3

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR) 2009, IEEE Conference on*, pages 248–255. IEEE, 2009. 7

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition, 2014. 2

[11] K. Duan, L. Marchesotti, and D. J. Crandall. Attribute-based vehicle recognition using viewpoint-aware multiple instance svms. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 333–338. IEEE, 2014. 2

[12] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004. 6

[13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, pages 1778–1785. IEEE, 2009. 3

[14] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 161–168. IEEE, 2011. 2

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 2

[16] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars. Local alignments for fine-grained categorization. *International Journal of Computer Vision*, pages 1–22, 2014. 7

[17] R. Girshick, J. Donahue, T. Darrell, and J. 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. 1, 2

[18] M. S. Jonathan Krause, Jia Deng and L. Fei-Fei. Collecting a large-scale dataset of fine-grained cars. *The Second Workshop on Fine-Grained Visual Categorization*, 2013. 7

[19] C. Kanan. Fine-grained object recognition with gnostic fields. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 23–30. IEEE, 2014. 2, 7

[20] J. Krause, T. Gebru, J. Deng, L.-J. Li, and L. Fei-Fei. Learning features and parts for fine-grained recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 26–33. IEEE, 2014. 7

[21] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Computer Vision Workshops (ICCVW), 2013 IEEE*

International Conference on, pages 554–561. IEEE, 2013. 2, 4

[22] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 1(4):7, 2009. 2

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pages 1097–1105, 2012. 1, 2, 6

[24] C.-H. Kuo and R. Nevatia. Robust multi-view car detection using unsupervised sub-categorization. In *Applications of Computer Vision (WACV), 2009 Workshop on*, pages 1–8. IEEE, 2009. 2

[25] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009. 3

[26] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015. 2

[27] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 2

[28] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[29] M. L. Seltzer and J. Droppo. Multi-task learning in deep neural networks for improved phoneme recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6965–6969. IEEE, 2013. 3

[30] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[32] N. Srivastava and R. R. Salakhutdinov. Discriminative transfer learning with tree-based priors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2094–2102, 2013. 3, 6

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 1, 2

[34] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 2

[35] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014. 2

[36] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010. 2

[37] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Proceedings of the European Conference on Computer Vision (ECCV) 2010*, pages 155–168. Springer, 2010. 3

[38] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, pages 1794–1801. IEEE, 2009. 2

[39] S. Yang, L. Bo, J. Wang, and L. G. Shapiro. Unsupervised template learning for fine-grained object recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3122–3130, 2012. 2

[40] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1577–1584. IEEE, 2011. 2

[41] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV) 2014*, pages 818–833. Springer, 2014. 1

[42] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 729–736. IEEE, 2013. 2

[43] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1637–1644. IEEE, 2014. 2