Abstract

In this work, we propose a new algorithm for boosting Deep Convolutional Neural Networks (BoostCNN) to combine the merits of boosting and these networks. To learn this new model, we propose a novel algorithm to incorporate boosting weights into the deep learning architecture based on least square objective function. We also show that it is possible to use networks of different structures within the proposed boosting framework and BoostCNN is able to select the best network structure in each iteration. This not only results in superior performance but also reduces the required manual effort for finding the right network structure. Experiments show that the proposed method is able to achieve state-of-the-art performance on several fine-grained classification tasks such as bird, car, and aircraft classification.

1 Introduction

Deep convolutional neural networks (CNNs) have recently produced outstanding results in learning image representations for several vision tasks including image classification [15, 24, 41] and object detection [12, 16, 31]. These neural networks usually consist of several components including convolutional, fully connected and pooling layers. By stacking several of these layers, deep CNNs are capable of learning complex features that are highly invariant and discriminant [24, 37, 38, 39, 44]. Krizhevsky et al. [24] proposed an eight layer deep network architecture that produced state-of-the-art performance on the ImageNet Challenge [32]. Given the success of this network, it has been applied widely to many other problems such as video classification [20], face recognition [40] and action recognition [28]. However, the optimal image representation for each computer vision task is unique and finding the optimal deep CNN structure for extracting that representation is a challenging
problem. There are some general guidelines, inspired by biological vision systems, for designing these deep networks such as putting convolutional layers early in the network. These guidelines, however, do not address how to set structural parameters of the networks such as the number of layers, number of units in each layer or the size of the receptive fields in the pooling layers. As a result designing a new network can take weeks of trial-and-error.

To address this design challenge, we propose to combine boosting and deep CNNs. The idea is to leverage the ability of boosting to combine the strength of multiple weaker learners to simplify the complicated design process of deep CNNs. This characteristic of boosting has been shown to be very significant for several Computer Vision tasks. For example in the Viola and Jones face detector [42] boosting searches a very large pool of simple classifiers (thresholds on Haar wavelet responses) and trains a classifier that is able to detect a complex object such as human face. Our approach is also motivated by the successful combination of decision trees [29] and boosting. As in the case of CNNs, the design of decision trees is not straightforward, e.g. what is the optimal depth of a trees? how many branches should there be per node? what is the optimal criteria for splitting tree nodes? These are all important aspects of learning a decision tree and much research has been devoted to these questions in the 1990s. For example there are several proposals for node splitting criteria such as Gini impurity, information gain [26] and there are several tree induction algorithms such as CART [3] or C4.5 [30]. Today, boosted decision trees [9, 10, 11] have eliminated most of these problems via an optimal weighted vote over decision trees that are individually sub-optimal. Similar to the success story of boosted decision trees, we believe that combination of boosting and deep learning can significantly reduce the challenges in designing deep networks.

The idea of boosting neural networks or, more generally, working with ensembles of neural networks has been around for many years; see for example [1, 5, 6, 7, 8, 14, 19, 34, 35, 46]. All these works demonstrated advantages of using an ensemble of networks over using a single large network. These works, however, either rely on simple averaging of several networks or rely on some heuristic weighting mechanism to impose boosting weights in the training process. In addition, some of these methods do not scale to the object recognition tasks that pervade the modern computer vision literature.

In this work, we propose a new algorithm for boosting deep networks (BoostCNN) to combine the merits of boosting and deep CNNs. To learn this new model, we propose a novel algorithm to incorporate boosting weights into the deep learning architecture based on least square objective functions. Experiments show that the proposed method is able to achieve state-of-the-art results on several fine-grained classification task as such bird, car, and aircraft classification. Finally we show that it is possible to use networks of different structures within the proposed boosting framework and BoostCNN is able to find the best network in each iteration. This not only results in superior performance, but also eliminates the required manual effort for finding the right network structure.

2 Multiclass boosting

We start with a brief overview of multiclass boosting. A multiclass classifier is a mapping $F: \mathcal{X} \rightarrow \{1 \ldots M\}$ that maps an example $x_i$ to its class label $z_i \in 1 \ldots M$. Since this is not a continuous mapping, a classifier $F(x)$ is commonly trained through learning a predictor
\( f : \mathcal{X} \rightarrow \mathbb{R}^d \) for some \( d \). The classifier \( F(x) \) is then implemented by
\[
F(x) = \arg \max_{k=1\ldots M} \langle y_k, f(x) \rangle,
\]
where \( y_k \) is a unit vector that represents label of the \( k^{th} \) class and \( \langle \cdot, \cdot \rangle \) is the dot product.

For example in binary classification, labels are \( y_1 = +1 \) and \( y_2 = -1 \) and (1) is equivalent to the popular \( F(x) = \text{sign}[f(x)] \) decision rule. Another example is one-vs-all multiclass classifiers. In this method, for each class \( k \), a predictor \( f_k(x) : \mathcal{X} \rightarrow \mathbb{R} \) is trained to discriminate between examples of that class versus others. In order to classify a new example \( \hat{x} \), \( f_k(\hat{x}) \) is computed for all \( k = 1\ldots M \) and the class of largest predictor is declared as the label. This procedure is equivalent to (1) by defining \( f(x) = [f_1(x) \ldots f_M(x)] \in \mathbb{R}^M \) and \( y_k = 1_k \in \mathbb{R}^M \), i.e. \( k^{th} \) element is one and the rest are zero. In general, the choice of labels are not restricted to the canonical basis in \( \mathbb{R}^M \) and it is possible to use any set of \( M \) distinct unit vectors \( y_1 \ldots y_M \in \mathbb{R}^d \). For simplicity, in the rest of this paper, we assume that \( d = M \) and \( y_k = 1_k \).

Multiclass boosting is a method that combines several multiclass predictors \( g_i : \mathcal{X} \rightarrow \mathbb{R}^d \) to form a strong committee \( f(x) \) of classifiers, i.e., \( f(x) = \sum_{i=1}^{N} \alpha_i g_i(x) \) where \( g_i \) and \( \alpha_i \) are the weak learner and coefficient selected at \( t^{th} \) boosting iteration. There are several approaches for multiclass boosting such as [4, 5, 6] and we use GD-MCBoost method of [6] in this paper. GD-MCBoost trains a boosted predictor \( f(x) \) by minimizing risk of classification
\[
\mathcal{R}[f] = E_{X,Z}\{L(z,f(x))\} \approx \frac{1}{\|D\|} \sum_{(x_i,z_i)\in D} L(z_i,f(x_i)),
\]
where \( D \) is a set of training examples and
\[
L(z_i,f(x_i)) = \sum_{j=1,j\neq z_i}^{M} e^{-\frac{1}{2}[(z_j-f(x_i))-(y_j,f(x_i))]}.
\]

The minimization is via gradient descent in function space. GD-MCBoost starts with \( f(x) = 0 \in \mathbb{R}^d \ \forall x \) and iteratively computes the directional derivative of the risk, (2), for updating \( f(x) \) along the direction of \( g(x) \)
\[
\delta \mathcal{R}[f; g] = \left. \frac{\partial \mathcal{R}[f + \varepsilon g]}{\partial \varepsilon} \right|_{\varepsilon=0} = -\frac{1}{2\|S\|} \sum_{(x_i,z_i)\in D} \sum_{j=1}^{M} g_j(x_i)w_j(x_i),
\]
where
\[
w_k(x_i) = \begin{cases} -e^{-\frac{1}{2}[(z_i-f_k(x_i))]} & k \neq z_i \\ \sum_{j=1,j\neq k}^{M} e^{-\frac{1}{2}[(z_i-f_j(x_i))]} & k = z_i \end{cases}
\]
GD-MCBoost then selects/trains a weak learner \( g^* \) that minimizes (4),
\[
g^* = \arg \min_{g \in \mathcal{G}} \delta \mathcal{R}[f; g],
\]
and compute the optimal step size along \( g^* \),
\[
\alpha^* = \arg \min_{\alpha \in \mathbb{R}} \mathcal{R}[f + \alpha g^*],
\]
using a line search. The boosted predictor \( f(x) \) is finally updated as
\[
f = f + \alpha^* g^*.
\]
3 Boosting convolutional neural networks

Combining boosting and convolutional neural networks is possible by using convolutional neural networks (CNN) as weak learners in GD-MCBoost algorithm. In this case, the weak learner \( g(x) \in \mathbb{R}^M \) is a deep network, e.g. Alex-Net [24] without the last soft-max loss layer. Using CNNs as weak learners (base-learners) requires training CNNs to minimize (4) in each iteration of boosting. However, the learning algorithms for CNNs are typically based on minimizing the error rate, e.g. log-likelihood or soft-max and these objective functions are independent of the boosting weights and can be very different from (4). A possible solution is to replace the soft-max loss layer with a layer to directly optimize (4) in the back-propagation algorithm. However, this is not practical and diverges quickly as (4) is unbounded, e.g. scaling \( g(x) \) can make it infinite.

In order to address this issue, we first note that \( \delta R[f; g] \) of (4) is equivalent to

\[
\delta R[f; g] = -\frac{1}{2||D||} \sum_{(x_i, z_i) \in D} \langle g(x_i)w(x_i) \rangle,
\]

where \( \langle \cdot, \cdot \rangle \) is the Euclidean dot product. This shows (4) is a summation of dot products between the network output \( g(x_i) \) and boosting weights \( w(x_i) \). Therefore (4) measures the similarity between those vectors and thus the optimal network output, \( g^*(x_i) \), has to be aligned with the boosting weights, i.e.

\[
g^*(x_i) = \beta w(x_i),
\]

where \( \beta > 0 \). Note that the exact value of \( \beta \) is not crucial during the training of the network because \( g^*(x) \) will be scaled appropriately by the optimal \( \alpha \) in (7) afterwards. Therefore without loss of generality we can assume that \( \beta = 1 \) and the optimal network output has to replicate boosting weights. This is equivalent to train a network \( g(x) = [g_1(x) \ldots g_M(x)] \in \mathbb{R}^M \) to minimize the square error loss

\[
L_{se}(w, g) = \sum_{x_i \in D} \sum_{j=1}^M (g_j(x_i) - w_j(x_i))^2.
\]

Using \( L_{se} \) loss function for learning a CNN, the back-propagated derivatives are

\[
-\frac{\partial L_{se}}{\partial g_k(x_i)} = 2(w_j(x_i) - g_k(x_i)).
\]

The proposed algorithm (BoostCNN) is summarized in Algorithm 1. It starts by initializing \( f(x) = 0 \in \mathbb{R}^M \). In each iteration, it first computes the boosting weights, \( w(x) \in \mathbb{R}^M \) according to (5) and trains a network \( g^*(x) \) to minimize the square error between the network output and boosting weights using (11). Once the network is trained, BoostCNN finds the boosting coefficient by minimizing the boosting loss, (7), and adds the network to the ensemble according to \( f(x) = f(x) + \nu \alpha^* g^*(x) \) where \( \nu \in (0,1] \) is the shrinkage parameter which has been shown to act as a regularizer for boosting algorithms [11].

3.1 Implementation Details:

We implemented the proposed algorithm using Caffe library [18]. For training networks to minimize the square loss, (11), we replaced the soft-max loss layer of the popular networks
such as Alex-Net [24] with the Euclidean loss layer. In each iteration of boosting we need to compute the weights, according to (5), however, computing $f(x)$ for each example requires running all of the networks in the current ensemble and can be very expensive. To avoid this problem, we note that if $w_t(x_i)$ are the weights at iteration $t$ then using (5) and (8)

$$w_{k+1}^t(x_i) = \begin{cases} -e^{-\frac{1}{2}^\nu \alpha^*(g^*_t(x_i) - g^*_k(x_i))} w_k^t(x_i) & k \neq z_i \\ -\sum_{j=1, j \neq k}^M w_{j+1}^t(x_i) & k = z_i \end{cases}$$

(13)

where $g^*(x)$ and $\alpha^*$ are the network and the coefficient learned at iteration $t$. Similarly finding an accurate $\alpha^*$ by a line search in (7) can be computationally expensive and instead we used binary search to solve $\frac{\partial R[f + \alpha g^*]}{\partial \alpha} = 0$. Finally, note that for training a deep network in each boosting iteration it is possible to initialize it by random parameters or by parameters of the network learned in the previous iteration. According to our experiment the later is more effective and we will further discuss this issue in section 5.

4 Analysis

The proposed BoostCNN algorithm has a couple of interesting properties. First note that there are two objective functions in the algorithm 1) the boosting loss function that is used to compute the boosting weights, (5), and coefficients (7), and 2) the square loss function (11) which is used to train deep networks. We can optionally change these objective functions, e.g. use logistic loss as boosting objective or use weighted error rate as network training objective as long as minimizing the network objective function is consistent with minimizing directional gradient of the boosting objective function (4).

Next we provide some intuition about the boosting weights and their effects on training CNN learners. According to (5), the boosting weights in BoostCNN algorithm are $M$-dimensional. These weights encode two types of information. First, the norm of vector $w(x) \in \mathbb{R}^M$ is proportional to how well example $x$ is classified by the current ensemble of weak learners. If $x$ is correctly classified then $f_k(x_i)$ will be larger than $f_k(x_i) \forall k \neq z_i$, the terms in the exponents of (5) will be small and thus $w(x)$ will have a small norm. On the other hand, if $x$ is mis-classified, some of the exponent terms in (5) will be positive and $w(x)$ will have larger norm. Second, the $k^{th}$ components of vector $w(x_i) \in \mathbb{R}^M$, encodes the importance of $k^{th}$ class in classification of example $x_i$. For an incorrect class label $k \neq z_i$, if $f_k(x_i) > f_z(x_i)$ then $w_k(x)$ will be large. In addition $w_k(x)$ will increase exponentially by
increasing $f_k(x_i) > f_z(x_i)$. Therefore weights of incorrect classes will get magnified exponentially. Similarly, if $f_k(x_i) < f_z(x_i)$, $w_k(x_i)$ will be a small value. These two weighting mechanisms will modulate the network output $g(x)$ in (4) and help CNN learning procedure to focus on more difficult examples and more confusing classes.

Another interesting property of BoostCNN method is that, for any network $\tilde{g}(x)$ for which (4) is non-zero, adding $\tilde{g}(x)$ (or $-\tilde{g}(x)$) can help the gradient descent procedure to further minimize the classification risk of the boosted classifier, (2) and improve the performance. In fact if $g'(x)$ is a network with uniform random output, i.e.

$$Prob(k = \arg \max_{k=1...M} g'(x)) = \frac{1}{M},$$

then according to (4) and (5)

$$E\{\delta R[f; g']\} = -\frac{1}{2|D|} \sum_{x_i \in D} \sum_{j=1}^{M} E\{g_j(x_i)\} w_j(x_i) = -\frac{1}{2M|D|} \sum_{x_i \in D} \sum_{j=1}^{M} w_j(x_i) = 0.$$ 

Therefore, BoostCNN can utilize any network whose output is slightly better than random for boosting.

Finally, BoostCNN is not limited to a single type of base-leaner and its pool of weak learners can include networks with different structures. In this case, at each boosting iterations, we train these networks independently to approximate boosting weights using (11), and the network that leads to most reduction in the boosted classification risk (2) will be added to the ensemble. This is very significant as it can reduce the trial-and-error that is required for finding the right CNN structure for a specific task. We will discuss more about this in section 5.1.

### 4.1 Previous works

The proposed framework is similar to boosting methods proposed by Schwenk et al. [34, 35, 36] such as Diabolo classifier, multi-column deep network of [1, 5] and averaging several learning networks e.g. [24, 37].

Comparing with Diabolo classifier, the proposed BoostCNN has different boosting objective as well as different network learning objective, e.g. in Diabolo the network is trained to minimize the weighted error rate while in BoostCNN the network is trained to replicate the boosting weights. Comparing with multi-column CNN, first note that multi-column CNN trains a group of CNNs simultaneously to learn a linear combination of these network as the final predictor. This, however, will increase the complexity of the learning process as it will exponentially increase number of local minimas in the optimization problem, e.g. any permutation of columns of a local optima is also a local optima. Comparing this method with the propose method, in BoostCNN 1) networks are trained sequentially and 2) each network is trained on the mistakes of the previous networks. This sequential network learning simplifies the optimization problem and avoids the local minima problem of multi-columns networks. Finally, BoostCNN is better than averaging several independently trained networks because BoostCNN optimizes coefficients of the linear combination and trains new networks on more difficult examples and classes.
In this section we illustrate properties of the proposed BoostCNN algorithm and compare its performance with other methods on several image classification tasks. For implementation, we used the open source Caffe library [18].

We start by illustrating properties of BoostCNN using CIFAR-10 dataset [23]. This dataset consists of 60K 32 × 32 images in 10 classes, where 50K images are for training and the rest are for testing. For training, we used CIFAR10-quick network model provided in Caffe as the base-learner and replaced its soft-max loss layer with a Euclidean loss layer. This network consists of three convolutional layers, each followed by a pooling layer and a rectified linear unit (RELU). These layers are then augmented by two fully connected layers. Using BoostCNN, we trained an ensemble of 9 networks each with 2,000 back-propagation iterations. For comparison, we trained 9 networks independently and averaged their results (bagging). We also trained a single network, using soft-max loss layer, with more than 20,000 back-propagation iterations. Figure 1 shows the error rates of these networks as a function of number of iterations. As shown in this figure, BoostCNN classifier was able to outperform both the bagged classifier and the single CNN classifier. We continued the training of the single CNN for 40K iterations, but error rate did not improve significantly. Manually changing the learning rate of the single CNN will decrease the error to around 25% which is still 3% higher than the error rate of BoostCNN. Figure 1 also shows the performance of BoostCNN where the base-learner is initialized randomly at the start of each boosting iteration, (BoostCNN with reset). As shown in this figure, the random initialization leads to inferior performance and it is better to initialize the base-learner in each boosting iteration with the last learned base-learner.

Next we evaluate performance of the BoostCNN and compare it with state-of-the-art in tasks of bird species classification, fine grained car classification and fine grained aircraft classification tasks. In these experiments, we used the Bi-linear network of [41] (B-Net) as base-learner, we initialize the network with VGG16 weights and initialized the last layer (i.e. the linear layer connecting the bilinear features to the output classes) using an externally trained linear SVM to reduce the total training time. The Linear SVM classifier is very similar to learning soft-max loss used in the network. However, we found that learning the last layer outside of the network to initialize the training leads to more stable classifiers and better results. Note that this is only possible when dealing with relatively small datasets, otherwise computing all of the features requires a significant amount of memory.

1The code is available at http://github.com/mmoghimi/BoostCNN
Table 1: Performance comparison for bird classification on CUB200 dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoostCNN</td>
<td>92.1%</td>
</tr>
<tr>
<td>Bilinear CNN (B-Net)</td>
<td>91.3%</td>
</tr>
<tr>
<td>Krause et al. [22]</td>
<td>92.6%</td>
</tr>
<tr>
<td>Chai et al. [4]</td>
<td>78.0%</td>
</tr>
<tr>
<td>Fisher Vector [3]</td>
<td>82.7%</td>
</tr>
<tr>
<td>BoostCNN (heterogeneous)</td>
<td>86.2%</td>
</tr>
<tr>
<td>Bilinear CNN (B-Net)</td>
<td>84.1%</td>
</tr>
<tr>
<td>Krause et al. [22]</td>
<td>82.0%</td>
</tr>
<tr>
<td>Pose Normalized CNN</td>
<td>75.7%</td>
</tr>
<tr>
<td>Part-based RCNN</td>
<td>73.9%</td>
</tr>
</tbody>
</table>

Table 2: Performance comparison on Cars (left) and Aircraft (right) classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoostCNN</td>
<td>88.5%</td>
</tr>
<tr>
<td>Bilinear CNN (B-Net)</td>
<td>84.1%</td>
</tr>
<tr>
<td>Fisher Vector [22]</td>
<td>80.7%</td>
</tr>
<tr>
<td>Chai et al. [4]</td>
<td>72.5%</td>
</tr>
</tbody>
</table>

Bird Species Classification: For this task we used CUB-200-2011 dataset. This data consists of 11,788 images of 200 bird species and comes with a pre-selected training and testing splits. We cropped the largest square window from the center of all images, resized them to 448 × 448, and mirrored the training images to double the training set. Table 1 compares performance of BoostCNN with other state-of-the-art results. As shown in this table, BoostCNN was able to outperform the other methods.

Car make and model classification: For this task we used the car dataset of which consists of 16,185 images of 196 different car make and models from Acura RL to Volvo XC90. This dataset comes with a pre-defined training and testing splits. The images include a variety of sizes and aspect ratios and we cropped the largest square window and resized it to 448 × 448 for pre-processing. Table 2-left compares performance of BoostCNN with other methods. In particular, BoostCNN was able to outperform its base-learner classifier B-Net and its error rate was only 0.5% higher than the state-of-the-art.

Aircraft Classification: For this task we used the FGVC-aircraft dataset which consists of 10k images of 100 different aircraft models. It also contains different sub-models of the same aircraft design, e.g. different Boeing 737s. We resized the images to 448 × 448 while ignoring the original aspect ratio for this experiment and similar to other datasets we doubled the training set by mirroring. Table 2-right compares performance of different methods. As shown in this table BoostCNN was able to outperform other state-of-the-art-methods significantly, the next best method is the Bilinear CNN (B-Net) which is used as the base-learner in this BoostCNN classifier.

5.1 Boosting heterogeneous classifiers

In the previous section we showed that BoostCNN is able to achieve state-of-the-art performance on several image classification tasks. However, each of those tasks required trial-and-error for finding the best network structure, e.g. input size. BoostCNN can eliminate the need for this trial-and-error since it is not limited to a single type of base-learner and its pool of weak learners can include networks with different structures. In this case, at each boosting
iterations, we train these networks independently to approximate the boosting weights (11), and the network that leads to most reduction in the boosted classification risk (2) will be added to the ensemble.

We used this strategy in this experiment where the pool of base-learners consists of B-Nets with 8 different input image variations, i.e. two different input sizes (448 and 224), two aspect ratios (1 and 1.25) and the choice of cropping the center patch or resizing the image to the desired size. Note that B-Nets work with any input size since the output of the bilinear layer only depends on the number of channels. Figure 2 shows details about performance of the classifier and the candidate weak-learner in each boosting iteration. In this figure, each row corresponds to performance of one network candidate and each column corresponds to one boosting iterations. The entry at $i^{th}$ row and $j^{th}$ column shows testing accuracy at iteration $i$ if $j^{th}$ network was selected in that iteration. As shown in this figure, the network with (center, 448 × 448) input structure, recipe proposed by [11], was the best base-learner in the early iterations but in the later iterations, the network with (center, 560 × 448) is more effective. Finally, Table 1 compare performance of this approach with other methods. As shown in this figure, this approach is more effective than using a single type of network as weak-learner.

### 6 Conclusion and future work

In this paper, we proposed a novel model by combining the merits of boosting and deep CNNs. We are inspired by the powerful image representation learned by deep CNNs and the ability of boosting to combine the strengths of multiple learners to improve the classification. To learn this new model, we developed an algorithm to incorporate boosting weights into the deep learning architecture. We illustrated the properties of boosted convolutional networks and demonstrated the advantages of our model via state-of-the-art results on several fine-grained classification datasets: CUB [3], Cars [21] and Aircrafts [25]. Finally we showed that BoostCNN is able to boost networks with different structures. This not only resulted in better performance but also eliminated the required manual effort for finding the right network structure. In future, we plan to apply BoostCNN to general large scale classification datasets, e.g. ImageNet, and extend our heterogeneous boosting experiments to include more
diverse set of networks with different depths and configurations.

References


the fisher vector for fine-grained classification. Pattern Recognition Letters, 49:92–98,
2014.

of aggregation algorithms. Artificial Intelligence, 163(2):139 – 162, 2005. ISSN 0004-

and Pattern Recognition (CVPR).

[16] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and
Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. arXiv


Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture

arxiv.org/abs/1503.06350.

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar,
and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for
fine-grained categorization. In Proceedings of the 2013 IEEE International Conference
on Computer Vision Workshops, ICCVW ’13, pages 554–561, Washington, DC, USA,
77.

[22] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition
without part annotations. In Proceedings of the IEEE Conference on Computer Vision


depth convolutional neural networks. In Advances in neural information processing

arxiv.org/abs/1306.5151.


