Automatic Data Augmentation from Massive Web Images for Deep Visual Recognition

YALONG BAI, Harbin Institute of Technology, China
KUIYUAN YANG, DeepMotion, China
TAO MEI, JD AI Research, China
WEI-YING MA, Bytedance, China
TIEJUN ZHAO, Harbin Institute of Technology, China

Large-scale image datasets and deep convolutional neural networks (DCNNs) are the two primary driving forces for the rapid progress in generic object recognition tasks in recent years. While lots of network architectures have been continuously designed to pursue lower error rates, few efforts are devoted to enlarging existing datasets due to high labeling costs and unfair comparison issues. In this article, we aim to achieve lower error rates by augmenting existing datasets in an automatic manner. Our method leverages both the web and DCNN, where the web provides massive images with rich contextual information, and DCNN replaces humans to automatically label images under the guidance of web contextual information. Experiments show that our method can automatically scale up existing datasets significantly from billions of web pages with high accuracy. The performance on object recognition tasks and transfer learning tasks have been significantly improved by using the automatically augmented datasets, which demonstrates that more supervisory information has been automatically gathered from the web. Both the dataset and models trained on the dataset have been made publicly available.

CCS Concepts: • **Information systems** → *Web mining*; • **Computing methodologies** → *Image representations*; *Object recognition*;

Additional Key Words and Phrases: Dataset construction, deep convolutional neural network, dataset augmentation

ACM Reference format:
Yalong Bai, Kuiyuan Yang, Tao Mei, Wei-ying Ma, and Tiejun Zhao. 2018. Automatic Data Augmentation from Massive Web Images for Deep Visual Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 3, Article 69 (July 2018), 20 pages.
https://doi.org/10.1145/3204941

This work was done when the first author was an intern at Microsoft Research Asia.

Authors' addresses: Y. Bai, Harbin Institue of Technology, Harbin, 150090, China; email: ylbai@outlook.com; K. Yang, DeepMotion, No. 9 North 4th Ring West Road, Beijing, 100190, China; email: kuiyuanyang@deepmotion.ai; T. Mei, JD AI Research, No. 8 Beichen West Street, Beijing, 100105, China; email: tmei@jd.com; W.-Y. Ma, Bytedance, No. 43 Beisanhuan West Road, Beijing, 100080, China; email: weiyingma@bytedance.com; T. Zhao, Harbin Institue of Technology, Harbin, 150090, China; email: tjzhao@hit.edu.cn.

https://doi.org/10.1145/3204941

1 INTRODUCTION

Generic object recognition is a fundamental problem in multimedia and computer vision and has achieved steady progress with efforts from both large-scale dataset construction and sophisticated model design. Though the goal is to minimize expected errors on previously unseen images, only empirical errors can be directly optimized on a set of labeled images with respect to a function space defined by a model. According to statistical learning theory, the gap between expected error and empirical error is determined by the sample size and model capacity. The gap becomes smaller with increasing sample size, and model design tries to minimize the expected error by defining a function space to minimize the empirical error and control the model capacity. Starting from the success of AlexNet [18] on the ILSVRC-2012 dataset [4, 27], years of effort have been devoted to model designing, and a series of improved deep convolutional neural networks (DCNNs) such as ZFNet [45], VGGNet [29], GoogLeNet [32], and ResNet [10] are proposed. There are also many efforts to create new datasets for new recognition tasks [16, 22, 38, 41, 47]. However, there is little effort to increase an existing dataset to make the empirical error closer to the expected error, mainly for two reasons: one is the labeling cost scales linearly with the size of the dataset, the other is that using more human labeling to achieve better results is usually considered to be unfair comparison. In this article, we attempt to automatically augment[1] an existing dataset from the web with a pre-trained DCNN on the existing dataset.

The web hosts massive images with rich contextual information and the volume keeps growing fast, which makes many applications possible such as image search engines [46] and semantic graph building [11]. The web is also the basic source of many datasets, which are scraped from search engines without further human labeling [2, 17, 31, 35, 42, 44]. An image on a web page often comes with rich contextual information edited by web authors. For examplealt text can convey the essential visual information and can be used to replace the associated image in a pure text-based browser, page title describes what is the whole web page is about, and surrounding text around the image that are related to the image content in some manner. Nevertheless, contextual information is not purposely edited to annotate image content; it is often quite noisy. The noisy web information is often used as a weakly supervised dataset for many multimedia tasks, e.g., image annotation [39], visual concept learning [6], and image retrieval [21].

DCNNs trained on large-scale datasets have achieved superior performance, which inspires us to investigate the possibility to use DCNN replace humans to do the laborious labeling task. In our early study, we found that DCNN trained on ImageNet performs much worse on web images, due to that both images and categories are not following the same distribution as the training set, and results in many false positives for each category. The problem can be alleviated by setting high thresholds for the prediction score; however, in this way, the collected images can provide limited additional information to improve the pre-trained DCNN since the DCNN is already quite confident on these images.

DCNN extracts image's visual information while the web provides an image's contextual information, which is complementary and can jointly provide additional information to an existing dataset. The noise of contextual information can be removed by the DCNN using visual information, while rich contextual information helps to achieve high prediction accuracy, even with a lower threshold for the prediction score of a DCNN. Together, we can augment an existing dataset in a scalable, accurate, and informative way. Specifically, we automatically augment ILSVRC-2012 with an additional 12.5 million images from the web. By training the same DCNN on the augmented

[1]This is different with the common practice of data augmentation for DCNN training, which randomly crops training samples from an image to avoid overfitting and achieve translation/scale invariance.

dataset without human-labeled images, significant performance gains are observed, which demonstrates a well-trained DCNN can further improve itself by self-labeling more images from the web. Another encouraging experimental result is that we can boost the performance of ResNet-50 on the ILSVRC-2012 validation set from 74.55% to 77.35%, even by using our augmented dataset, which is labeled by the lower performance AlexNet. We release the dataset and models[2] to facilitate the research on learning-based object recognition and transfer learning tasks.

The rest of this article proceeds as follows: After an overview of related work in Section 2, automatic dataset augmentation is introduced in Section 3. We evaluate the quality of augmented datasets in Section 4, and conclude with a discussion in Section 5.

2 RELATED WORK

Dataset is the basic input for statistical learning algorithms to train models, and significant efforts have been made to construct datasets for various recognition tasks. In this section, we discuss related efforts according to the degree of labor cost during constructing datasets.

2.1 No Human Labeling

Some datasets are directly collected from image search engines or social networks without human labeling. TinyImage [35] contains 80 million 32×32 low resolution images, collected from image search engines by using nouns in WordNet as queries. YFCC100M [33] is another large database of approximately 100 million images associated with metadata collected from Flickr. Krause et al. [15] only use web images to fine-tune DCNN pre-trained on ILSVRC-2012 for fine-grained classification and get even higher accuracies than using fine-grained benchmark datasets, which is expected since existing fine-grained benchmark datasets are quite small. Phong et al. [36] collect 3.14 million web images from Bing and Flickr for the same 1,000 categories of ILSVRC-2012.

Massouh et al. [24] proposed a framework to collect images from the web and use a visual and natural language concept expansion strategy to improve the visual variability of a constructed dataset. Li et al. [20] also constructed a dataset by directly querying images from Flickr and Google Images Search. However, DCNN trained on all of these automatically constructed datasets perform much worse than human-labeled datasets when testing on ILSVRC-2012, which reflects the noisy and highly biased nature of web images.

Recently, Sun et al. [31] constructed a large-scale dataset from a search engine, the dataset has 300 million images and is labeled with 18,291 categories; however, this dataset is still noise in labels: approximately 20% of the labels in the dataset are noisy.

2.2 Fully Human Labeling

Each image is manually labeled by one or multiple annotators to ensure high accuracy. Due to the high labeling cost, datasets constructed by fully labeling are often with small size. Some typical datasets are Caltech101/256 [7, 8], Pascal VOC [5], and several for fine-grained object recognition [14, 23, 37]. These datasets are widely used for shallow model learning, while not large enough to train a DCNN from scratch. Though challenging, million scale datasets have been constructed, such as ImageNet [4] for object recognition and Places [47] for scene recognition. With ImageNet, DCNN first proves its success and improves most object recognition tasks by the learned feature representations [18]. However, the high labeling cost limits both the number of images that can be labeled for each category, and the number of categories that can be labeled.

[2]The dataset and models can be found at https://auto-da.github.io/.

2.3 Partially Human Labeling

To alleviate human-labeling cost and use the limited budget in more effective ways, there are several active learning-based approaches are proposed to label images that are considered as informative for a model. Collins et al. [3] propose a method to do image labeling and model training iteratively. In their work, some randomly selected images are first labeled as seed training set to train an initial model, then the model is applied to a set of unlabeled images; at last, human annotators are further asked to label a subset of images of which the model is mostly uncertain. The process is iterated until the classification accuracy converges or the budget is run out. Krause et al. [15] present a similar scheme for fine-grained object recognition by using DCNN. Since informative images are selected based on some specific model, human involvement is always required for newly designed models.

To decouple human labeling from model training, Tong et al. [40] propose to train DCNN for clothing classification with both a clean dataset manually labeled by annotators and millions of images with noisy labels provided by sellers from online shopping websites. Though noisy, the accuracy of images from online shopping websites (~62% [40]) is much higher than general web images (~10% [35]). Sukhbaatar et al. [30] try to train DCNN with 0.3M clean ILSVRC-2012 training images and 0.9M noisy web images, and show marginal improvement with a noise layer to model noise, but still with much higher error rate than DCNN directly trained on 1.2M ILSVRC-2012 training images.

Different from the work of Li et al. [19], which aims to learn robust image classifiers by considering the noisy textual information accompanied with web images, in this article, we try to automatically scale up an existing image data in an automatically way. Moreover, both the high diversity and high accuracy should be ensured for the constructed dataset. Since the accuracy of web images is relatively low, the number of web images needs to be orders of magnitude larger than existing datasets to contain enough relevant images. Thus, we aim to use as many web images as possible; till July 30, 2017, we have used 186.4 million web images as candidate images to augment several labeled image datasets. These augmented image datasets achieve high performance on object recognition tasks than human-labeled datasets with significantly more training images. To the best of our knowledge, this is the first work that uses DCNN to label web images and demonstrates a well-trained DCNN can automatically improve itself by surfing the web.

3 AUTOMATIC DATASET AUGMENTATION

Starting from a human-labeled image dataset $\mathcal{D}$, we are targeting at augmenting it to a much larger dataset $\mathcal{D} \cup \mathcal{E}$, where $\mathcal{E}$ is automatically labeled from web images by a DCNN trained on $\mathcal{D}$. Labeling images is an intelligent process, which requires sufficient intelligence and knowledge. In this section, we will first investigate two separated labeling methods by DCNN and the web, respectively, then present our method, which labels image by the web and DCNN jointly. Without special mention, AlexNet designed by Krizhevsky et al. [18] will be used as the basic DCNN in this article, considering it is with a relatively low computational cost for large-scale experiments.

3.1 Labeling By DCNN

DCNNs have achieved remarkable prediction accuracy on validation set and testing set of ILSVRC-2012 [27] by end-to-end learning on the training set, which inspires us to use DCNN to replace humans to do image-labeling tasks. We defined the "confidence score" of a given image I relevant to category c as the probability for c by DCNN. Given a DCNN trained on the labeled dataset $\mathcal{D}$, which maps an image I to a set of confidence scores $f_c(I)$ for each pre-defined category $c \in \{1, \ldots, C\}$, it is intuitive to use it for image labeling. A new image I can be labeled as an instance of a category

TOMM1403-69 ACMJATS Trim: 6.75 X 10 in July 4, 2018 16:28

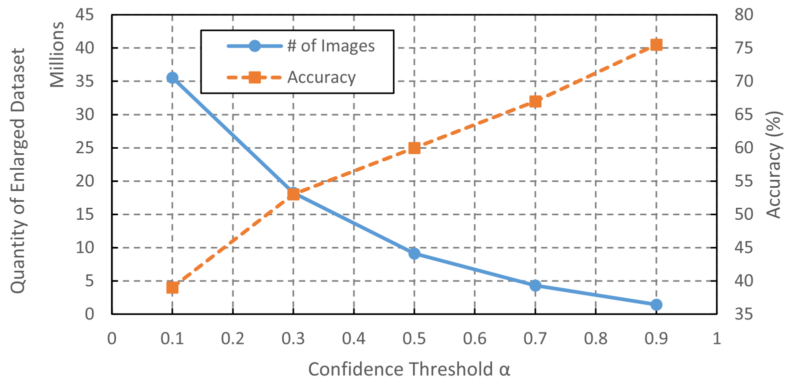

Fig. 1. The distributions of quantity and accuracy of dataset $\mathcal{E}_V$ across confidence threshold α.

c if I has a confidence score of c exceeds some predefined threshold α, i.e.,

$$f_c(I) \geq \alpha. \tag{1}$$

To avoid ambiguity, images with multiple labels that exceed the threshold are ignored. Then an augmented dataset $\mathcal{E}_V$ can be labeled by applying the DCNN on a large unlabeled image set $\mathcal{U}$, i.e.,

$$\mathcal{E}_V = \{\langle I, c\rangle : f_c(I) \geq \alpha, I \in \mathcal{U}, c \in \{1, \ldots, C\}\}. \tag{2}$$

The labeling process is fully automatic, which only requires feedforward calculation on an unlabeled image set. We investigate this method by using the DCNN learned from the ILSVRC-2012 training set to label an unlabeled candidate image set randomly collected from the web. By analyzing the labeling results, we find several properties of labeling by DCNN.

Low Accuracy. Figure 1 shows the quantity and accuracy of automatically labeled dataset $\mathcal{E}_V$ by setting different thresholds α, where accuracy is estimated by manually inspecting randomly sampled images (10 images per category) from 100 categories in the constructed dataset. As expected, a higher threshold will result in a smaller dataset with higher accuracy. However, even with the relatively high threshold 0.9, the achieved accuracy 75.5% is still much lower than the accuracy 99.7% achieved by human labeler on ImageNet [4]. Figure 2 shows some incorrectly labeled false positive images, where most noises are out of the 1,000 categories used for training, but visually similar to these categories in some aspects. The result also shows that the DCNN is still hard to generalize to a testing set with many out-of-class images.

Less Informative. Higher accuracy can be obtained by keeping increasing the threshold. However, this will cause two problems. One is the number of images that can be collected will be reduced for a fixed unlabeled dataset, and the unlabeled dataset needs to grow larger to collect enough images. The other problem is even worse, images labeled by high confidence scores are iconic samples and with high similarity with images in the existing training set, as shown in the third row of Figure 3. These images can bring little new supervisory information to the existing training set.

3.2 Labeling by the Web

The web hosts trillions of images with rich metadata, which provides a "free" way to label images since labels are already in the metadata provided by web users. Image search engines directly leverage these metadata to index massive web images and make them retrievable. Though image search engines provide a convenient way to collect web images by searching words or word phrases that describe a category, they are with several limitations for dataset construction because

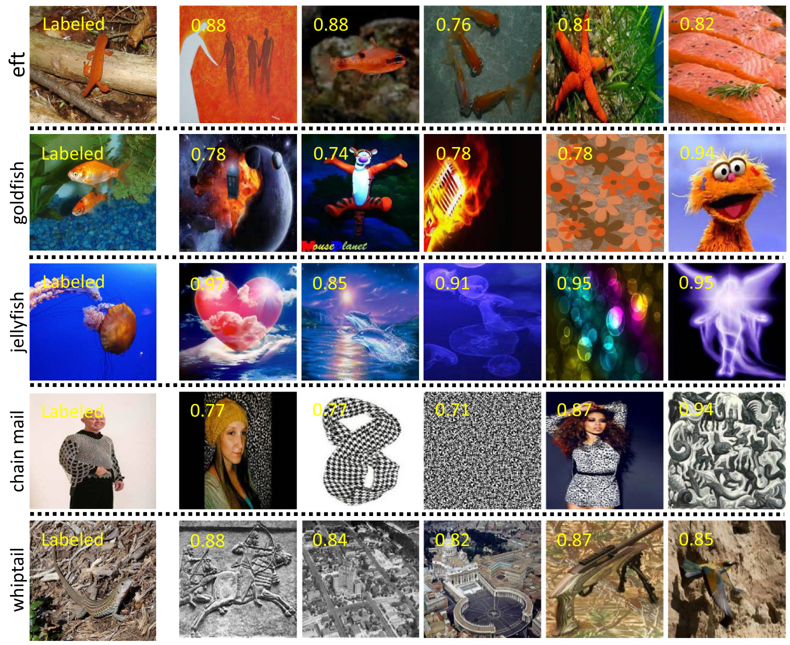

Fig. 2. Noisy images that predicted one of the categories with high confidence by DCNN. The first column in this figure shows an example image from the labeled dataset for each category. The other columns show noisy images from unlabeled dataset with high-confidence DCNN predictions for the categories in a different row, respectively. The confidence scores are shown on each noisy image.

they are optimized for human users. For example, search engines typically limit the number of images retrievable for each query (in the order of a few hundred to a thousand), and the retrieved images are often iconic, presenting a single, centered object with a simple background, which is not representative of natural conditions. Thus, we directly resort to raw images with textual metadata from the web as our source data. Specifically, four textual fields are collected for each image, including:

- *Anchor text* T^1 is the visible, clickable text in a hyperlink linked to an image, which usually gives the user relevant description about the content of the linked image.
- *Alt text* T^2 is shown when an image cannot be displayed to a reader. Thus, it can be regarded as a textual counterpart to the visual content of an image.
- *Page title* T^3 is an important field for the page to state the main content of the web page.
- *Surrounding text* T^4 consists of the text paragraphs around an image in a web page. The surrounding text is in many cases semantically related to the image content. However, since the surrounding text can also contain information that is uncorrelated to the image, this field as a contextual information source can be very noisy.

Then a data item from the web can be denoted by $\langle I, T^1, T^2, T^3, T^4 \rangle$. Figure 4 shows a web image and its four types of textual metadata, where rich information about goldfinch" is embedded in metadata for the image.

Given a web image dataset denoted by $\mathcal{W} = \{\langle I_i, T_i^1, T_i^2, T_i^3, T_i^4 \rangle\}_{i=1}^{|\mathcal{W}|}$, then labeling by the web can be directly carried out through string match. $|\mathcal{W}|$ is the number of elements of the closed set $\mathcal{W}$. Let each category c be represented by a set of word phrases from its WordNet synonyms [25]

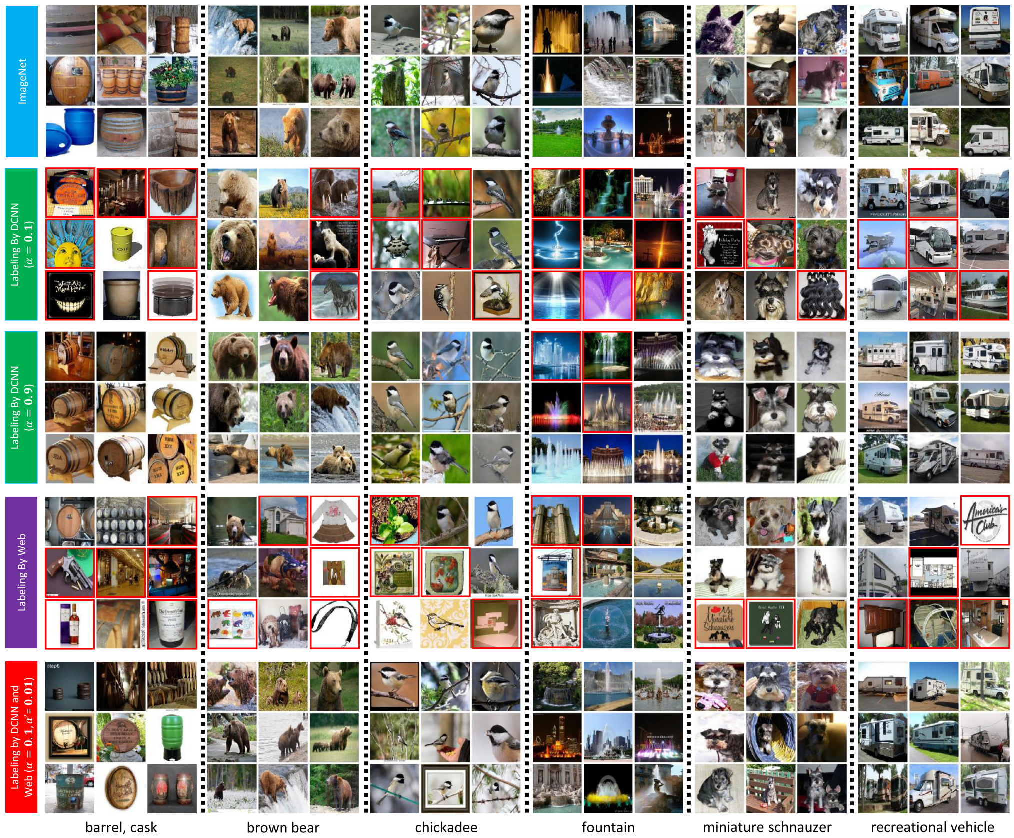

Fig. 3. Snapshots of human-labeled dataset ImageNet and four automatically constructed datasets on six randomly sampled categories in ILSVRC-2012: the first row is from the ImageNet; the second and third row are from the dataset labeled by DCNN with confidence threshold $\alpha = 0.1$ and $\alpha = 0.9$, respectively; the fourth row is from the dataset labeled by the web; the last row is from the dataset labeled jointly by DCNN and the web with confidence threshold $\alpha = 0.1, \alpha' = 0.01$. For each category,nine randomly sampled images are presented. Images marked with red boxes are noisy images.

and relevant descriptions in 12 different languages (including AR, ZH, EN, FR, DE, EL, HE, HI, IT, JA, RU, ES) from BabelNet [26], denoted by $\mathcal{S}_c = \{s_j\}_{j=1}^{|\mathcal{S}_c|}$. An image I_i is labeled as an instance of category c if at least one textual field contains at least one element in $\mathcal{S}_c$, i.e.,

$$\delta_i^c = \begin{cases} 1 & : s_j \subseteq T_i^k, \exists s_j \in \mathcal{S}_c, \exists k \in \{1,\ldots,4\} \\ 0 & : otherwise \end{cases}. \tag{3}$$

Then an augmented dataset $\mathcal{E}_T$ can be labeled by web data $\mathcal{W}$, i.e.,

$$\mathcal{E}_T = \{\langle I, c\rangle : \delta_i^c = 1, i \in \{1,\ldots,|\mathcal{W}|\}, c \in \{1,\ldots,C\}\}. \tag{4}$$

The labeling process is also fully automatic and very fast after $\mathcal{W}$ has been collected. By the method, we collect a dataset with 186.4 million images for the 1,000 categories from ILSVRC-2012 dataset. Here, we summarize several properties observed from the dataset.

Figure 5 shows the percentage of images collected by each textual field. We can find that surrounding text has the greatest contribution since most images are with surrounding texts and

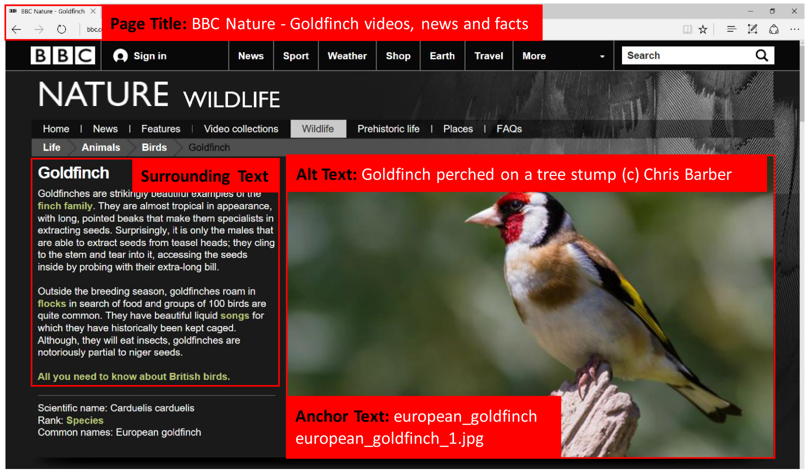

Fig. 4. Illustration of the textual metadata associated with an image in a web page. The web page used in this figure is from http://www.bbc.co.uk/nature/life/European_Goldfinch.

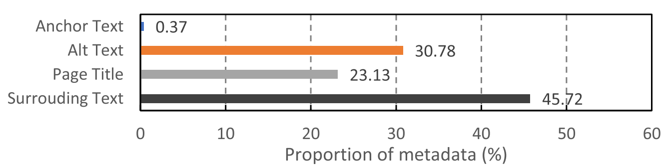

Fig. 5. The proportion of images collected according to different fields of textual metadata.

typically contain more words than other fields, while the number of images collected by anchor text is much smaller than other fields since anchor texts are typically very short and often not provided by web authors.

Besides the quantity, we also check the quality of the collected dataset. To avoid manually checking, we use the DCNN to calculate the confidence score of the labeled category of each image in $\mathcal{E}_T$, and large confidence score means a large probability of the labeled image to be correct. Figure 6 shows the distribution of confidence scores by different textual fields, where images collected by anchor text and alt text are with the larger proportion of high confidence scores, which also means these two fields are more reliable than the others. The conclusion is also consistent with experiences of using textual features for image search engines.[3]

However, as expected, images collected from the web are very noisy, where 82.8% images are with confidence scores lower than 0.05. After analyzing the noisy images, we find that the noisy images can be divided into two different types. One is that the image and its relevant textual metadata is not matching, since the poor quality of some web pages are attached with many irrelevant images. The other type of noise is introduced by ambiguities between the meaning of category and the textual metadata. A typical example is a category named "jay," which is supposed to be a bird by WordNet, lots of images about humans are collected since "jay" is often used as a human name. Though these noisy images are hard to remove by only using textual information, they are easy to remove by visual information since images of different senses of a name are typically visually distinguishable as Yao et al. [43] demonstrated.

[3]https://support.google.com/webmasters/answer/114016?hl=en.

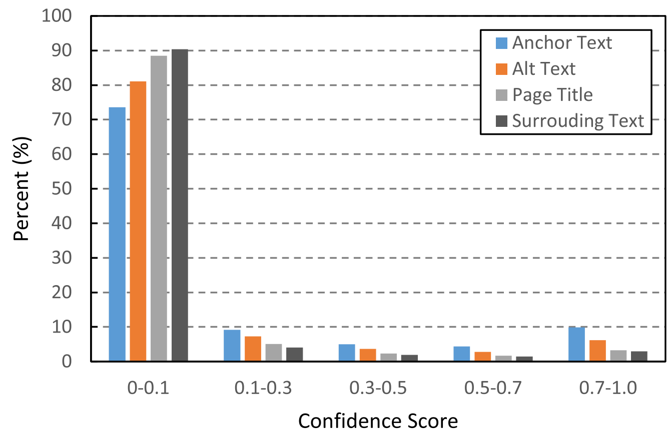

Fig. 6. The distributions of percent of images across confidence scores under different kinds of contextual information.

3.3 Labeling by Web and DCNN

Both visual labeling by DCNN and contextual labeling by the web have their limitations. For datasets labeled by DCNN, many noisy images are from categories that are out of the category set used for training DCNN, which can be easily filtered out according to semantic information. Meanwhile, for dataset labeled by the metadata of web, both of the visual irrelevant noisy images and semantic ambiguous noisy images can be easily removed by visual information, due to that the visual irrelevance images have very low confidence scores, and images of different senses of a name are typically visually distinguishable. Thus, we combine them to improve the labeling by leveraging their complementarity. We learned from the above experience that labeling by DCNN is more computational cost and tend to spend too much time on popular categories. Thus we first use the web to label a dataset $\mathcal{E}_T$ in a relatively balanced way, then use DCNN to go through the textually labeled dataset $\mathcal{E}_T$. Together, a dataset can be labeled by Web and DCNN via

$$\mathcal{E}_{VT_{web}} = \{\langle I, c\rangle : f_c(I) \geq \alpha, \langle I, c\rangle \in \mathcal{E}_T\}, \tag{5}$$

where $\mathcal{E}_{VT_{web}}$ is a filtered subset of $\mathcal{E}_T$ where lots of noisy images are filtered out by DCNN. Different from labeling by DCNN in Equation (2), the contextual labeling can filter out the majority of out-of-class noisy images, and the used $\mathcal{E}_T$ is with much higher signal-noise ratio than $\mathcal{U}$, which allows us to use lower threshold α to label more informative images. Figure 7 shows the quantity and accuracy curve concerning confidence threshold α on images labeled by the web; it is encouraging that much higher accuracy achieved even with very low confidence threshold, e.g., 94% accuracy is achieved when the threshold α is set to 0.1.

The accuracy of $\mathcal{E}_T$ is still relatively low by simply using string match, which limits us to set lower confidence threshold to absorb more diverse and informative images with keeping high accuracy. Thus, we are motivated to further decrease the noise in $\mathcal{E}_T$.

The image I_i, text T_i, metadata type t_i, and image URL domain d_i are coupled together as a single data item in our dataset, labels assigned to images by DCNN are also assigned to metadata, thus we can construct an automatically labeled textual dataset, i.e.,

$$\begin{aligned}\mathcal{T}^+ &= \{\langle T_i, t_i, d_i, y_i = c_i\rangle : \langle I_i, c_i\rangle \in \mathcal{E}_{VT_{web}}\},\\ \mathcal{T}^- &= \{\langle T_i, t_i, d_i, y_i = C+1\rangle : I_i \in \mathcal{N}_{VT_{web}}\},\end{aligned} \tag{6}$$

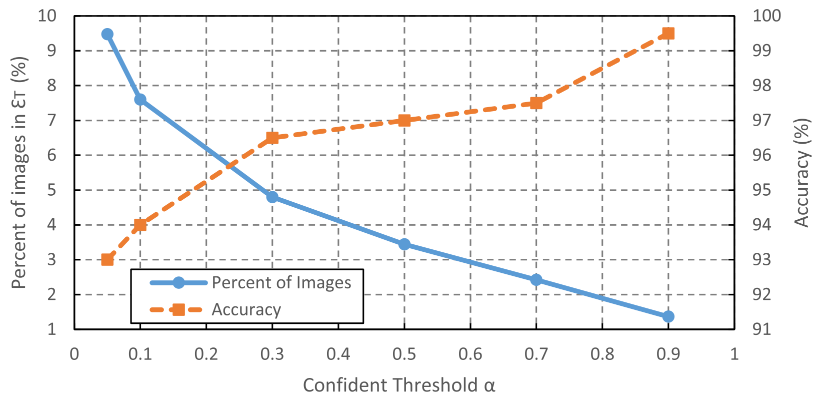

Fig. 7. The distributions of quantity and accuracy of dataset $\mathcal{E}_{VT_{web}}$ across confidence threshold α after applying visual restriction to candidate dataset $\mathcal{E}_T$.

where $\mathcal{N}_{VT_{web}} = \{\langle I, c\rangle : f_c(I) < \beta, \langle I, c\rangle \in \mathcal{E}_T, \beta \ll \alpha\}$ contains noisy images for each category by string match. Inspired by previous work on sentence classification [13], we train a two-layer fully connected network to categorize textual metadata at semantic level. The input to the network is the combination of one hot representation of metadata type t_i, image URL domain d_i, and bigrams in T_i. As Figure 6 shows, the metadata type t_i is a useful prior to the text classification task. Meanwhile, we also found that there are some special websites on which the vast majority of images are relevant to some specific categories, e.g., farnhamanglingsociety.com is a website about fishing and lots of images about tench can be found on this website. The first layer of the network generates embedding representation for inputs with weight matrix E, and the second layer classifies into categories based on the representation with weight matrix W using softmax regression,

$$
\begin{aligned}
p(y_i = c \mid T_i, t_i, d_i) &= \frac{e^{f(y_i=c|T_i,t_i,d_i)}}{\sum_{k=1}^{C+1} e^{f(y=k|T_i,t_i,d_i)}}, \\
f(y = k \mid T_i, t_i, d_i) &= \left(W_k \frac{\sum_{s_j \subseteq T_i} E \cdot s_j + E \cdot t_i + E \cdot d_i}{|T_i| + 2}\right).
\end{aligned} \tag{7}
$$

The model is trained by minimizing

$$
-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C+1} 1\{y_i = k\} \log p(y_i = k \mid T_i, t_i, d_i), \tag{8}
$$

where $N = |\mathcal{E}_{VT_{web}}| + |\mathcal{N}_{VT_{web}}|$. We train this model by using stochastic gradient descent and a linear decaying learning rate. As a result, a new dataset $\mathcal{E}_{VT_{web^+}}$ labeled by our text classification model can be constructed:

$$
\begin{aligned}
\mathcal{E}_{T_{web^+}} = \{\langle I, c\rangle : & p(y = c \mid T_i, t_i, d_i) > 0.5, \\
& i \in \{1, \ldots, |\mathcal{W}|\}, c \in \{1, \ldots, C\}\}.
\end{aligned} \tag{9}
$$

The textual classification model can categorize the metadata according to the meaning of the category and the contextual information from sentences. As a result, many semantic ambiguous noisy images can be detected and filtered out. The experimental results show that the accuracy of image set $\mathcal{E}_{T_{web^+}}$ is 71.5%, which is significantly higher than $\mathcal{E}_T$ whose accuracy is only 21.3%. Naturally,

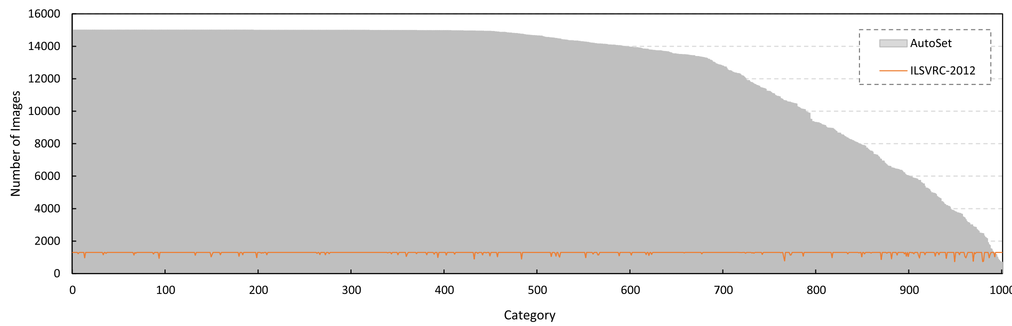

Fig. 8. The number of images per category of the ILSVRC-2012 dataset and the dataset automatically augmented from massive web images for ILSVRC-2012.

a new dataset jointly constrained by DCNN and text classification model can be constructed:

$$\mathcal{E}_{VT_{web^+}} = \{\langle I, c\rangle : f_c(I) \geq \alpha', \langle I, c\rangle \in \mathcal{E}_{T_{web^+}}\}, \tag{10}$$

where $\alpha' < \alpha$. The high-performance text classification model makes it possible to decrease the visual threshold from α to α', and to mine a more diverse and larger scale dataset without accuracy dropping, e.g., 93.8% accuracy is achieved when $\alpha' = 0.01$. Finally, we get a dataset labeled by the web and DCNN jointly,

$$\mathcal{E}_{VT} = \mathcal{E}_{VT_{web}} \bigcup \mathcal{E}_{VT_{web^+}}. \tag{11}$$

Figure 3 shows snapshots of human-labeled dataset ImageNet and four automatically constructed datasets by different methods. Compare to the dataset labeled only by DCNN or the web, the dataset constructed jointly by DCNN and the web has higher accuracy and diversity.

4 EXPERIMENTAL RESULTS

In our experiments, we augmented the ILSVRC-2012 training set ($\mathcal{D}^{1K}_{ImageNet}$) based on our proposed method. We first trained an AlexNet on $\mathcal{D}^{1K}_{ImageNet}$ that will be used for labeling and as the baseline for comparing, then use this DCNN for labeling a web-labeled dataset $\mathcal{E}_T$, which contains 186.4 million images. All of these images in $\mathcal{E}_T$ are collected from the index of Bing Image Search Engine, which crawled images from the whole web. An optimized text-matching algorithm is applied into the map-reduce framework to collect the images for $\mathcal{E}_T$ efficiently. Those images before ranking of image search engine are used to avoid bias introduced by the search engine. At last, the text classifier trained on metadata of labeled images were used to mine more informative images. For categories with more than 15,000 images, we keep 15,000 images by random sampling. Finally, there are 12.5 million images left in the augmented ILSVRC-2012. Figure 8 summarizes the statistics of the human-labeled ILSVRC-2012 dataset and our automatically labeled dataset; we can find that our method significantly increases the scale of the dataset. This automatically augmented dataset is named AutoDA and is available for download in the link as introduced in Section 1.

It is worth it to note that the main time cost of our method is computing confidence scores using DCNN. Each candidate image in $\mathcal{E}_T$ has to go through the feed-forward pass of the DCNN, and 186 million candidate images cost 186 million feed-forward passes of the DCNN in total. The cost roughly equals to training the DCNN on 1.2 million ILSVRC images for 80 epochs (i.e., each image does feed-forward and back-propagation 80 times).

In addition to quantity, quality is another import factor for a useful dataset. The work of Nizar et al. [24] has tried to evaluate the DCNN's robustness to noise. They injected noise into the

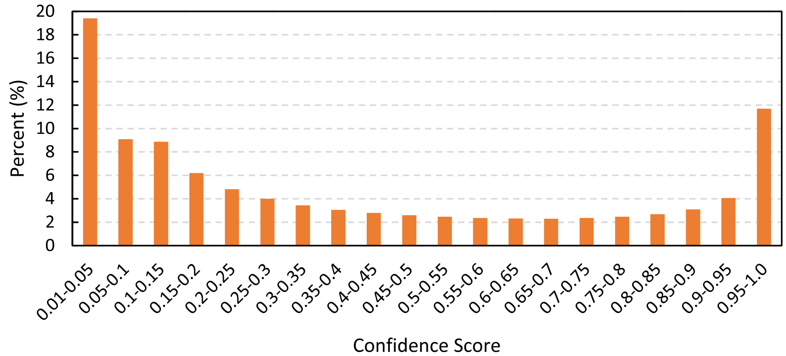

Fig. 9. The distributions of confidence score across the percent of images in AutoDA.

training dataset of two different kinds of DCNN architecture including AlexNet and GoogLeNet for ILSVRC task. The experimental results proved that a low percentage of noise (<20%) induces only a moderate reduction in classification performance. However, the model trained on a high percentage of noise (≥20%) tends to a significant performance drop. In this article, we try to collect a dataset with a high ratio between classification performance gain and dataset scale. Thus we carefully selected the hyper-parameter $\alpha = 0.1$, $\alpha' = 0.01$ for our AutoDA dataset to keep the amount of noise less than 20%, meanwhile, ensure the classification performance should be significantly improved by using as little amount of augmented images as possible. At last, we evaluated the accuracy of our finally constructed dataset AutoDA by randomly sampling 100 images per category for manual judgment. The results show that the average accuracy of AutoDA is nearly 94%.

As we know, images labeled by a higher confidence score of DCNN are usually not informative for improving the performance of DCNN further. We counted the number of images with low confidence score and images with high confidence score in Figure 9. We can find that there are nearly 28% of images whose confidence score are lower than 0.1. These images with low confidence score are usually much more difficult for DCNN training, and the image representations learned from these images have much better generalization ability. We will evaluate the quality of our augmented dataset according to the image representations learned from the augmented dataset in the following sections.

4.1 Image Classification

To quantitatively investigate AutoDA, we train the object recognition models from scratch on our augmented dataset and evaluate the trained models on the ILSVRC-2012 validation set. The test accuracy of the models on the ILSVRC-2012 validation set is used as the performance metric of the dataset quality. Although most of the categories in AutoDA have more than 10,000 images, there are still several rare categories contain fewer than 6,000 images as shown in Figure 8. Considering that an unbalanced dataset for training can lead to poor performance since the validation set is a balanced one, we balance the distribution of the augmented dataset by subsampling categories with more than 6,000 images and construct a balanced dataset $\mathcal{E}_{VT}^{1K}$ with 5.7 million of images from AutoDA.

Both of AlexNet and ResNet-50 are used for evaluating the quality of our constructed dataset. We followed the standard configuration reported in Reference [18] and [10] for AlexNet and ResNet-50 respectively. The traditional data augmentation methods such as mirror transformation, random cropping are equipped during training for all of the image recognition models in this article. For ResNet-50 training, we also used color shifting and random image resizing (the short side in the

Table 1. Single-Crop Top-1 (top-5) Accuracy of AlexNet Trained on Human-Labeled Datasets and Augmented Datasets

DCNN	#Iters	$\mathcal{D}^{1K}_{ImageNet}$	$\mathcal{E}^{1K}_{VT}$	$\mathcal{E}^{1K}_{VT} \cup \mathcal{D}^{1K}_{ImageNet}$	
				Merge	Merge (w/o dropout)
AlexNet	0.4M	56.15 (78.11)	51.99 (73.86)	56.48 (79.45)	59.90 (81.17)
	2.0M	60.36 (82.38)	56.58 (78.57)	**62.71 (83.71)**	61.72 (82.62)
ResNet-50	0.5M	74.55 (92.06)	67.25 (85.99)	75.57 (91.83)	-
	2.5M	74.44 (92.11)	70.17 (88.09)	**77.36 (93.29)**	-

Table 2. Ten-crop Top-1 (Top-5) Accuracy of AlexNet Trained on Human Labeled Datasets and Augmented Datasets

DCNN	Testing Crop	$\mathcal{D}^{1K}_{ImageNet}$	$\mathcal{E}^{1K}_{VT}$	$\mathcal{E}^{1K}_{VT} \cup \mathcal{D}^{1K}_{ImageNet}$	
				Merge	Merge (w/o dropout)
AlexNet	Central	60.36 (82.38)	56.58 (78.57)	62.71 (83.71)	61.72 (82.62)
	10-crop	63.04 (84.14)	58.40 (79.87)	**65.21 (85.50)**	64.90 (84.66)
ResNet-50	Central	74.44 (92.11)	70.17 (88.09)	77.36 (93.29)	-
	10-crop	76.12 (93.01)	71.10 (88.68)	**78.92 (94.25)**	-

range of [256, 480]) for data augmentation. Caffe toolkit is used for training and testing. To achieve closer top-1 accuracy with the reported ResNet-50 performance, we implemented a torchlike batch normalization layer for Caffe to replace Caffe's original batch normalization layer.

The experimental results in Table 1 show that the top-1 and top-5 classification accuracy on the validation set of ILSVRC-2012 with a single-crop prediction. We found that classification performance to a large extent is affected by the number of training iterations. Models training on larger training datasets need more iterations to be fully converged.

We further investigate whether the better model design and automatically labeled larger dataset can boost recognition performance together. Here, we choose ResNet-50 [10] which performs much better than AlexNet on ILSVRC-2012. Table 1 reports the results, where ResNet-50 consistently outperforms AlexNet as expected, and ResNet-50 also improves itself by using the automatically labeled data, which demonstrates that better model design and larger automatically labeled dataset can together boost the performance further.

Best performance is achieved on both AlexNet and ResNet-50 by merging the human-labeled dataset and augmented dataset. It demonstrates that well-trained DCNNs can automatically label more useful images from the web and improve themselves further. It should also be noted that the augmented dataset $\mathcal{E}^{1K}_{VT}$ is labeled by a low-performance AlexNet whose top-1 accuracy is 56.15%, but the augmented dataset can still boost a high-performance ResNet-50 from 74.55% to 77.36%. We list the classification accuracy of 10-crop testing in Table 2, the performance of ResNet-50 trained on the merged dataset is even better than the performance of ResNet-152 reported in Reference [10]. For practical applications, it means that we can apply smaller model like ResNet-50 trained on automatically augmented dataset instead of a bigger model like ResNet-152 trained on limited human-labeled dataset to save lots of computing resources.

We also evaluated the performance of DCNN without dropout layers. The experimental results in Table 1 show that the DCNN without dropout layers can converge faster, the influence of overfitting is alleviated, and better performance is achieved thanks to the large-scale augmented dataset.

Table 3. Single-Crop Top-1 Accuracy of DCNNs Trained on Augmented Datasets without Using Contextual Information from the Web

$\mathcal{D}_{ImageNet}$	$\mathcal{E}_V^{1K} \cup \mathcal{D}_{ImageNet}$		
	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
60.36	58.92	59.78	60.06

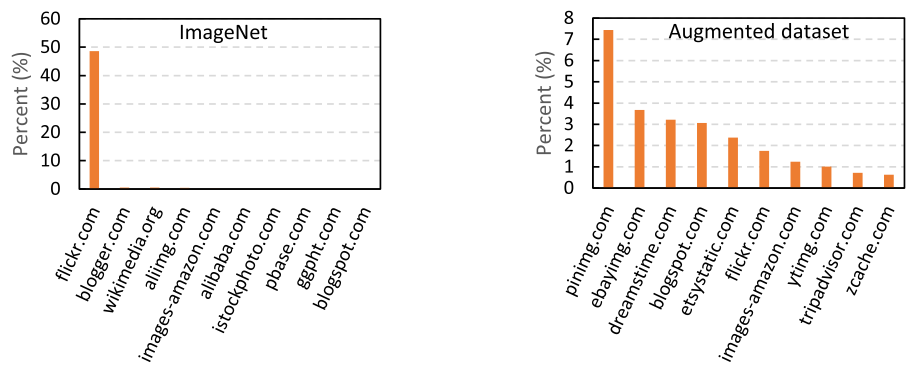

Fig. 10. The distributions of top-10 frequent domains in human-labeled datasets $\mathcal{D}_{ImageNet}$ and the automatically labeled datasets $\mathcal{E}_{VT}^{1K}$, respectively.

To investigate how the web labeling influences the quality of constructed dataset, we compare the performance of DCNNs trained on $\mathcal{E}_V^{1K}$ and $\mathcal{E}_{VT}^{1K}$. Since the accuracy of $\mathcal{E}_V^{1K}$ heavily relies on the confidence threshold α as shown in Figure 1, we try three different settings with $\alpha \in \{0, 5, 0.7, 0.9\}$ for constructing $\mathcal{E}_V^{1K}$ in this experiment. The experimental results in Table 3 show the performance of DCNN trained on $\mathcal{E}_V^{1K} \cup \mathcal{D}_{ImageNet}^{1K}$ is improved by increasing the confidence threshold since higher confidence threshold can lead to a more accurate dataset. But even with a high confidence threshold like 0.9, the overall accuracy of the $\mathcal{E}_V^{1K}$ is still relatively low as we show in Figure 1. Moreover, the visual patterns in images collected by high confidence threshold usually tend to be similar. As a result, the newly added dataset $\mathcal{E}_V^{1K}$ does not help the original dataset to achieve higher performance but hurts the performance. In general, the performance of DCNNs trained with $\mathcal{E}_V^{1K}$ is lower than the DCNN trained on $\mathcal{D}_{ImageNet}^{1K}$, which means that DCNN still cannot improve itself by self-labeling from open-ended image pool without using contextual information from the web.

4.2 Dataset Analysis

The performance by only using the automatically constructed dataset is still lower than the human-labeled dataset as shown in Tables 1 and 2.

We find that the performance gap comes from the distribution difference between the two datasets ImageNet collected about 10 years ago where visual appearance of many categories are changed over time, especially some man-made categories such as monitor and table lamp. Also, after we parse the URL domains of images in ImageNet, we find Flickr is the major source of ImageNet, while our augmented dataset is from a wider range of websites where some are even not existing during ImageNet collecting such as Pinterest.com. Figure 10 shows the difference of domain distributions of the image source of ImageNet and our augmented dataset, respectively.

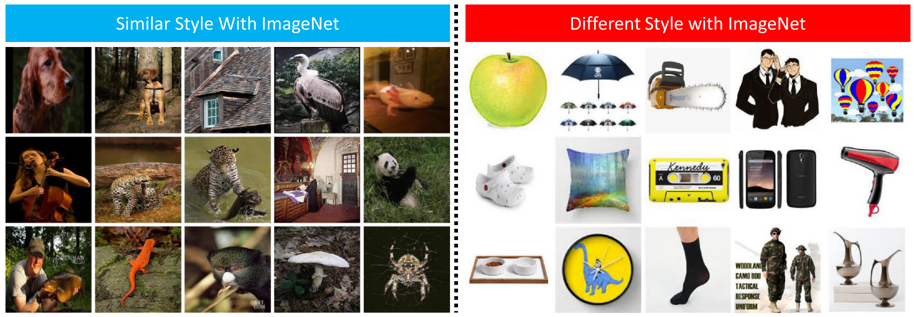

Fig. 11. Images in $\mathcal{E}_{VT}^{1K}$, which are sorted by the output of f_{critic}. Images in the left figure are all with high value, which means their image style is similar to ImageNet, while images with low f_{critic} are shown in the right figure.

Table 4. Single-Crop Top-1 Accuracy of DCNNs Trained on Human-Labeled Datasets and Augmented Datasets

Domain	#Iters	Training Data	
		$\mathcal{D}_{ImageNet}$	$\mathcal{E}_{VT}$
Natural	2.0M	68.17	69.53
Artifact	2.0M	57.05	52.23
Dog	0.4M	65.80	67.56
Bird	0.4M	82.00	86.24

4.2.1 Difference between ILSVRC-2012 and AutoDA. To systematically study the distribution difference between the two datasets, we train a discriminator similar to the one used in Wasserstein Generative Adversarial Network (GAN) [1] to differentiate images in ILSVRC-12 and images in our dataset by maximizing the distance between $\mathcal{D}_{ImageNet}$ and $\mathcal{E}_{VT}^{1K}$:

$$J_{critic} = \left[\sum_{I_i \in \mathcal{D}_{ImageNet}} f_{critic}(I_i) - \sum_{I_i' \in \mathcal{E}_{VT}^{1K}} f_{critic}(I_i') \right]. \tag{12}$$

By using the trained discriminator model f_{critic}, we sorted the images in $\mathcal{E}_{VT}^{1K}$ according to the output value of f_{critic} and show the images whose styles are most different/similar with $\mathcal{D}_{ImageNet}$ in Figure 11, and found that many images that can be easily distinguished from images in ILSVRC-2012 are collected from e-commerce websites.

Considering the difference between ImageNet and our dataset are mainly on man-made categories, we split the 1,000 categories into two subsets according to WordNet ontology, one is artifact set including 522 categories, the other is natural object set including 478 categories. We compare DCNNs trained on these two subsets with DCNN trained on ImageNet, respectively. Also, we evaluate the performance on two fine-grained subsets of ILSVRC-2012, i.e., dogs (including 120 dog breeds) and birds (including 59 bird species). Since the number of categories about dog and bird is small, the recognition models of dog and bird can converge after 0.4M iters on both $\mathcal{E}_{VT}$ and $\mathcal{D}_{ImageNet}$. Table 4 summarizes the results; our dataset achieves better performance than ImageNet on natural categories since these categories have not changed much over the past decades, while

Table 5. Top-1 Accuracy of DCNNs Trained on Human-Labeled Datasets and Augmented Datasets by Using Dense Test

Training Data	Test Data	
	ILSVRC 2012 Val	WebVision Val
$\mathcal{D}^{1K}_{ImageNet}$	56.15	52.58
WebVision	47.55	57.03
$\mathcal{E}^{1K}_{VT}$	51.99	53.94
$\mathcal{D}^{1K}_{ImageNet}$ *	60.36	54.99
$\mathcal{E}^{1K}_{VT}$ *	56.58	57.98

The experimental results with mark $*$ are trained with 2.0M iterations, and the others are trained with 0.4M iterations.

our dataset achieves worse performance than ImageNet on man-made categories since many images of ImageNet are out-of-date.

4.2.2 Dataset Bias Analysis. As we know, dataset bias often leads to overfitting and poor generalization in the real world. Some previous works targeting at measuring the quality and bias of datasets, such as the work of Torralba et al. [34]. Following this work, we verify the cross-dataset generalization ability of our dataset. Cross-dataset generalization measures the performance of classifiers learned from one dataset on the other dataset. If a dataset can truly represent the real world, the model learned from this dataset can easily generalize to any other dataset in the same domain.

We compare our augmented dataset with ILSVRC-2012 and another dataset named WebVision [20]. WebVision is a dataset constructed from Flickr and Google Images Search by querying the category names in the recent period (constructed and released in 2017). The same 1,000 categories as the ILSVRC-2012 dataset are used for measuring the bias of these three datasets. We first checked the overlap between our dataset with ILSVRC and WebVision. We try to search the nearest neighbors from ILSVRC + WebVision for images in our dataset, and cosine similarity between feature vectors extracted by a pre-trained DCNN is used for measuring the similarity between images. We randomly sampled 100,000 images from our dataset as the query images; the experimental results show that there are only nearly 13.4% images in our dataset that have similar images in ILSVRC/WebVision dataset with cosine similarity larger than 0.9. It means that the overlap between our dataset and WebVision/ILSVRC-2012 is not heavy, many new/unseen images are collected in our dataset.

Table 5 shows the classification error rates. Each dataset produces a DCNN using its training set, and then evaluates the trained model on a test set from different datasets. In all of the cases, the best performance is achieved by training and testing on the same dataset. The experimental results show that our augmented dataset has better performance than ILSVRC-2012 on the validation set of WebVision. Moreover, our augmented dataset also achieves better performance than WebVision on human-labeled image dataset ILSVRC-2012. The bias between ILSVRC-2012 and WebVision may be due to two factors. One is the main body of ImageNet was collected during a limited and specific period; this can result in some classes becoming dated over time, as we mentioned in Section 4.2. The other reason is that ImageNet is collected by multiple annotators, which may involuntarily

Table 6. PASCAL VOC 2007 Object Classification Results

DCNN	Dataset	aero	bike	bird	boat	bott	bus	car	cat	chai	cow	tabl	dog	horse	mbike	person	plan	sheep	sofa	train	tv	mAP
AlexNet	ILSVRC	88.6	82.2	84.7	81.7	33.5	73.7	85.7	84.2	58.2	59.9	72.7	78.3	88.6	77.8	93.0	49.8	75.7	59.4	89.0	68.5	74.1
	Merge	91.5	83.6	88.2	83.7	37.4	76.1	86.5	87.0	58.8	67.3	72.5	83.3	89.9	81.4	93.7	51.9	77.4	62.8	90.5	68.8	**76.6**
ResNet-50	ILSVRC	98.4	93.1	94.4	92.5	57.1	85.8	91.9	94.6	68.4	83.5	83.5	93.1	93.7	88.8	95.7	62.9	87.1	76.0	96.8	82.2	86.0
	Merge	99.1	94.0	95.3	94.4	58.6	89.3	92.2	95.0	69.7	88.2	84.0	94.6	94.9	90.8	96.2	61.8	91.0	74.9	97.3	84.3	**87.3**

inject some of their views and bias on object categories. Meanwhile, the bias of WebVision is observed since the model trained on WebVision has poor performance on the validation set of ILSVRC-2012. The bias of WebVision may be due to the bias of the source of images, since search engine and Flickr have their own bias on the style of images. The search engine usually tends to popular images, while Flickr has its own styled capture bias. Overall, our dataset generalizes much better than the other two datasets; it means that our automatically constructed dataset has better ability to represent the real world.

Considering the combination of $\mathcal{E}_{VT}$ leads to a significant performance improvement as shown in Tables 1 and 2, we also try to combine the WebVison dataset with ILSVRC-2012 dataset. The experimental results show that introducing images of WebVision into ILSVRC-2012 leads to 4.5%, 5.1% performance drop for AlexNet and ResNet50, respectively, on ILSVRC-2012's validation set. We also try to merge our dataset with the WebVision dataset too, but it still results in a poor performance. It is mainly due to the bias and unbalance distributions of WebVision. Moreover, the WebVision dataset is designed for learning visual representation from noisy web data, and there are lots of noise images included in the WebVision dataset.

4.3 Evaluation of the Visual Representations

We also try to look into the power of data for visual representation learning. We evaluate the learned representations on two tasks: image classification and image retrieval.

4.3.1 Results of PASCAL VOC Object Classification. The Pascal VOC 2007 object classification task contains nearly 10,000 images of 20 classes including artifact and natural objects. The target objects in images are not centered, and, in general, the appearance of objects in PASCAL VOC 2007 is perceived to be more challenging than ILSVRC.

Following the experimental settings described in the work of Ali et al. [28], we first extracted the outputs of second last layer of AlexNet and the last pooling layer of ResNet as features for images in PASCAL VOC 2007 by CNN models learned on dataset $\mathcal{D}_{ImageNet}$ and $\mathcal{E}_{VT}$, respectively. The extracted feature vector of each image is further $L2$ normalized to unit length. Then we trained linear SVM models for all classes based on the normalized feature vectors. The results shown in Table 6 proved that the large-scale dataset augmented from massive web images is helpful to learn more powerful image representations for visual recognition task.

4.3.2 Results of MSR-Bing Grand Challenge. Inspired by the success of feature extractors in DCNNs learned from ILSVRC-2012, we also try to compare the generalization ability of feature extractors learned from human-labeled ILSVRC-2012 and our augmented dataset. To evaluate the quality of feature extractors more comprehensively, we test the performance of the feature extractors on an open domain image retrieval task—MSR-Bing Grand Challenge [12].

The MSR-Bing Grand Challenge task provides a training set including 11.7 million queries and 1 million images, a test set including 1,000 queries and 79,665 images. It is required to learn a ranking model based on the training set and then rank images for each query in the test set, where Normalized Discounted Cumulative Gain (*NDCG*) is used as the evaluation metric for a ranking

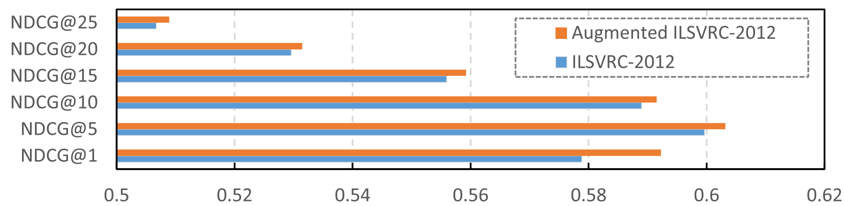

Fig. 12. The *NDCG* of CCA for image search using image representations provided by DCNNs trained on the ILSVRC-2012 training set and augmented ILSVCR-2012 training set.

list, which is defined as

$$NDCG@d = Z_d \sum_{j=1}^{d} \frac{2^{r^j} - 1}{log(1 + j)}, \tag{13}$$

where r^j = excellent = 3, good = 2, bad = 0 is the manually judged relevance for an image ranked at j with respect to the query, Z_d is a normalization factor to make the score to be 1 for an ideal case. The performance is measured by average *NDCG@d* on all queries in the test set.

We use Canonical Correlation Analysis (CCA) [9] as the basic ranking model and represent a query with bag-of-textual-words. For images, we use the outputs of the last but one fully-connected layer of a DCNN as the image representation, and two DCNNs trained on ILSVRC-2012 and augmented ILSVRC-2012 will be used. Figure 12 compares the performance of ranking results using image representations provided by the two DCNNs, where the DCNN trained on augmented ILSVRC-2012 achieves consistently better performance, which further demonstrates the generalization ability of model learned from the automatically augmented dataset.

5 CONCLUSION

In this article, we propose a method to do automatic dataset augmentation, where both the web and DCNN are used. Specifically, the web provides massive images with rich contextual information, while well-trained DCNNs are used to label these images and filter out noisy images. Meanwhile, the rich contextual information from the web ensures DCNN to achieve high labeling accuracy with relatively low confidence threshold. Together, we can augment labeled image datasets in a scalable, accurate, and informative way. Extensive experiments demonstrate that well-trained DCNNs can automatically label images from the web and further improve themselves with the automatically labeled datasets. We hope the automatically constructed large-scale datasets with rich contextual information will help further research in large neural networks.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and LãÍon Bottou. 2017. Wasserstein GAN. *arXiv:1701.07875* (2017).

[2] Yalong Bai, Kuiyuan Yang, Wei Yu, Chang Xu, Wei-Ying Ma, and Tiejun Zhao. 2015. Automatic image dataset construction from click-through logs using deep neural network. In *Proceedings of the 23rd ACM International Conference on Multimedia*. 441–450.

[3] Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei. 2008. Towards scalable dataset construction: An active learning approach. In *Proceedings of the European Conference on Computer Vision*. Springer, 86–98.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009 (CVPR'09)*. IEEE, 248–255.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision* 88, 2 (2010), 303–338.

[6] R. Ewerth, K. Ballafkir, M. Muhling, D. Seiler, and B. Freisleben. 2012. Long-term incremental web-supervised learning of visual concepts via random savannas. *IEEE Transactions on Multimedia* 14, 4 (2012), 1008–1020.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2007. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding* 106, 1 (2007), 59–70.

[8] Gregory Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 object category dataset. (2007).

[9] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16, 12 (2004), 2639–2664.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv:1512.03385* (2015).

[11] Xiaofei He, Deng Cai, Ji-Rong Wen, Wei-Ying Ma, and Hong-Jiang Zhang. 2007. Clustering and searching WWW images using link and page layout analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications* 3, 2 (May 2007), Article 10.

[12] Xian-Sheng Hua, Linjun Yang, Jingdong Wang, Jing Wang, Ming Ye, Kuansan Wang, Yong Rui, and Jin Li. 2013. Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 243–252.

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv:1607.01759*.

[14] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop on Fine-Grained Visual Categorization (FGVC'11)*, Vol. 2.

[15] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. 2015. The unreasonable effectiveness of noisy data for fine-grained recognition. *arXiv:1511.06789*.

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and others. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv:1602.07332* (2016).

[17] Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. (2009).

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[19] Wen Li, Li Niu, and Dong Xu. 2014. Exploiting privileged information from web data for image categorization. In *Proceedings of the European Conference on Computer Vision*. Springer, 437–452.

[20] Wen Li, Limin Wang, Eirikur Agustsson, and Luc Van Gool. 2017. WebVision: Visual Understanding by Learning from Web Data. Retrieved August 6, 2017 from http://www.vision.ee.ethz.ch/webvision.

[21] Z. Li and J. Tang. 2015. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Transactions on Multimedia* 17, 11 (2015), 1989–1999.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft Coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv:1306.5151* (2013).

[24] Nizar Massouh, Francesca Babiloni, Tatiana Tommasi, Jay Young, Nick Hawes, and Barbara Caputo. 2017. Learning deep visual object models from noisy web data: How to make it work. *arXiv:1702.08513* (2017).

[25] George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM* (1995).

[26] Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193 (2012), 217–250.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575

[28] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 806–813.

[29] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

[30] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2014. Training convolutional networks with noisy labels. *arXiv:1406.2080*.

[31] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, 843–852.

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[33] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM* 59, 2 (2016), 64–73.

[34] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, 1521–1528.

[35] Antonio Torralba, Rob Fergus, and William T. Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 11 (2008), 1958–1970.

[36] Phong D. Vo, Alexandru Ginsca, Hervé Le Borgne, and Adrian Popescu. 2015. On deep representation learning from noisy web images. *arXiv:1512.04785*.

[37] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The Caltech-UCSD birds-200-2011 dataset.

[38] Shuang Wang and Shuqiang Jiang. 2015. INSTRE: A new benchmark for instance-level object retrieval and recognition. *ACM Transaactions of Multimedia Computing, Communications, and Applications* 11, (Feb. 2015) 3, Article 37, 21 pages.

[39] F. Wu, Z. Wang, Z. Zhang, Y. Yang, J. Luo, W. Zhu, and Y. Zhuang. 2015. Weakly semi-supervised deep learning for multi-label image annotation. *IEEE Transactions on Big Data* 1, 3 (2015), 109–122.

[40] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2691–2699.

[41] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'15)*.

[42] Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, and Z. Tang. 2017. Exploiting web images for dataset construction: A domain robust approach. *IEEE Transactions on Multimedia* 19, 8 (2017), 1771–1784.

[43] Yazhou Yao, Jian Zhang, Fumin Shen, Wankou Yang, Pu Huang, and Zhenmin Tang. 2018. Discovering and distinguishing multiple visual senses for polysemous words.

[44] W. Yu, K. Yang, Y. Bai, H. Yao, and Y. Rui. 2015. Learning cross space mapping via DNN using large scale click-through logs. *IEEE Transactions on Multimedia* 17, 11 (2015), 2000–2007. DOI: http://dx.doi.org/10.1109/TMM.2015.2480340

[45] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer, 818–833.

[46] Lei Zhang and Yong Rui. 2013. Image search-from thousands to billions in 20 years. *ACM Transactions on Multimedia Comput. Communications, and Applications* 9, 1s (Oct. 2013), Article 36, 20 pages. DOI: http://dx.doi.org/10.1145/2490823

[47] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*. 487–495.

Received January 2018; revised April 2018; accepted April 2018