Generating Diverse Image Datasets with Limited Labeling

Niluthpol Chowdhury Mithun, Rameswar Panda, Amit K. Roy-Chowdhury
University of California, Riverside, CA 92521, USA
nmithun@ece.ucr.edu, rpand002@ucr.edu, amitrc@ece.ucr.edu

ABSTRACT

Image datasets play a pivotal role in advancing multimedia and image analysis research. However, most of these datasets are created by extensive human effort and extremely expensive to scale up. There is high chance that we may have no instances for some required concepts in these data-sets or the available instances do not cover the diversity of real-world scenarios. In this regard, several approaches for learning from web images and refining them have been proposed, but these approaches either include significant redundant instances in the dataset or fail to guarantee a diverse enough set to train a robust classifier. In this work, we propose a semi-supervised sparse coding framework to collect a diverse set of images with minimal human effort, which can be used to both create a dataset from scratch or enrich an existing dataset with diverse examples. To evaluate our method, we constructed an image dataset with our framework, which is named as DivNet. Experiments on this dataset demonstrate that our method not only reduces manual effort, but also the created dataset has excellent accuracy, diversity and cross-dataset generalization ability.

Keywords

Image dataset construction, Sparse coding, Active learning.

1. INTRODUCTION

The efficiency of several visual recognition tasks depends upon the ability to identify suitable training examples to learn initial models. The majority of the success in this regard has been achieved by models trained on large-scale hand-labeled image datasets (e.g., SUN [34], ImageNet[24]). Although, these datasets cover large numbers of categories, expanding them to new categories or providing new examples to an existing category, is extremely costly and labor-intensive [20]. Moreover, there exist various types of bias in the popular image datasets and hence, they do not demonstrate satisfactory cross-dataset generalization (training on a dataset, testing on a different dataset) capability [19, 30]. Future multimedia and image analysis research requires examining even a greater number of visual categories and adapt to higher intra-class variation present within a category[7]. Complexity of the models will increase over time to cope with this. Hence, creating high-quality image dataset and continuously updating existing datasets with new diverse examples is becoming more important over time. Complete human labeling based solution is unlikely to keep pace with this growing need.

To address the issues stated above and inspired by streams of images available on the web, there has been lot of recent interest in learning directly from noisy web data [20, 28, 18, 4, 14, 15, 5] or automatically curating web images for creating a dataset [33, 23, 31, 6, 1]. These methods reduce labeling cost and show promising results in recognition tasks by having a noise-aware model. However, these methods usually assume that most of the returned images from the web will be relevant to the query. It may work well for simple categories, (e.g. bike, car) but it is unlikely to be true for the complex categories, (e.g. birthday party, concert), especially when looking beyond the top few matches. Furthermore, most of these approaches consider images from a single search engine (e.g. Google [20, 14], Flickr [4]), or generate synthetic data from a clean set [28]. This may bring up issues of bias, low accuracy and lack of diversity.

A few semi-supervised approaches have been developed that minimize human effort for dataset creation using an active learning framework [35, 9, 7]. However, there is no guarantee that the images collected represent a diverse enough set to train a robust classifier. Similar to automatic approaches, most of the semi-automatic approaches primarily aim at collecting as many relevant images as possible. Hence, in spite of causing serious wastage of space, the dataset loses quality and training with these images may not provide expected performance gain. Moreover, these approaches [7, 35] may select many samples for human labeling that have significant information overlap [12].

Motivated by the above, the main goal of this work is to develop a method for construction of high-quality image datasets with limited budget (e.g. labeling, storage etc). The images in each category of the dataset should be relevant and diverse. The second goal, is to provide an online framework, that is capable of collecting more discriminative images continuously as new data becomes available, which is suitable for enriching existing image datasets. In order to achieve these goals, *we propose a novel sparse coding framework with human in the loop.* Our method builds upon several machine learning tools, e.g., active learning [25, 17, 2], sparse coding [11, 32, 36] and deep learning [22, 21].

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967285

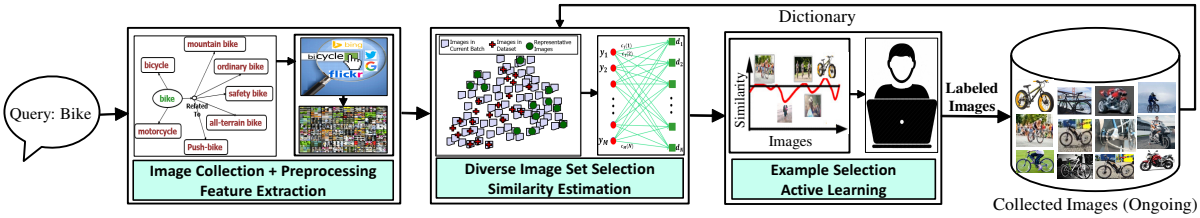


Figure 1: Brief illustration of our proposed framework for collecting images. Please see the text in Section 2 for details.

Contributions. The main contributions of this work are following. **First,** our proposed diversity aware sparse representative selection based active learning approach considers both images in current unlabeled batch and images in existing dataset to select the best subset for further processing. Hence, our approach efficiently minimizes the chance of including redundant instances in dataset and utilizes human effort in labeling most informative and discriminative instances. **Second,** the proposed framework is a generalized one which makes high-precision image dataset creation feasible with no initial cost (e.g. providing positive and/or negative seeds for categories) and limited labeling budget. Moreover, our approach can efficiently update an already created dataset, when new data becomes available. **Third,** experiments demonstrate that the dataset created by our method shows excellent cross-dataset performance, diversity and scalability. Additionally, it is worth mentioning that, even with no human labeling, our method shows high-precision in collecting images automatically.

2. PROPOSED METHODOLOGY

Overview. In our framework, collecting images of one category is independent of other categories. Hence, images for different categories can be collected in parallel. Fig. 1 summarizes our incremental image collection framework for a category 'bike'. Initially, we collect images related to the category from different web-sources. If no image of the category is available beforehand, a set of top ranked images from a reliable search engine is considered as initial dataset. During each run of incremental update, we process a small batch from the collected images. First, we employ a diversity aware sparse representative selection approach to choose a smaller set of representative images that not only best represents this batch, but also is discriminative to the images in current dataset. Then, we calculate similarity of each representative images with the images in dataset. Based on the similarity score, we employ active learning to decide whether to label an image manually or not. As collecting images for one category is independent of others, we ask only binary questions to annotator. We continuously update the dataset with new images labeled by the system.

2.1 Image Collection and Feature Extraction

Since the number of returned images from a web search based on a query is limited, we use a query expansion scheme. The expansion is done using Google Search and ConceptNet [27]. We select only synonyms and derived phrases as expanded queries, as these are highly relevant. For example, given a query 'Bike', we expand to queries such as 'Bicycle', 'Ride Bike', 'Mountain Bike' etc. The expanded queries are used to collect images from different web sources. We filter out the images having sufficiently low quality, e.g., out-of-focus or blurred, too white or black, empty or too small. A mini batch of images from the remaining collection is processed during each run of incremental update until required number of images for the category is selected. Batch-size is chosen experimentally and depends on available resources.

We apply deep CNN to extract features, as they are now the state-of-the-art image features [22, 21]. The best performance among CNN models has been achieved by very deep networks with large number of layers [26, 29], but the processing time per image in such networks can be very high. Hence, to keep dataset construction scalable, we use the architecture proposed in [21], which requires significantly less processing time. It is worth mentioning that our method does not depend on this particular choice of CNN, except for the scalablity issue mentioned above. We remove the last classification layer of the CNN [21] and treat rest of the network as a fixed feature extractor. For an image f at the input layer, we extract feature vector x_f ($x_f \in R^d$, $d = 4096$) from fc7 layer of the CNN.

2.2 Diverse Representative Set Selection

The goal of this step is to find an optimal subset of the current batch of images. In particular, we are trying to represent the current batch of images by selecting only a few representative images, which are also dissimilar to the images in current dataset. Therefore, our natural goal is to establish a image level sparsity which can be induced by performing l_1 regularization on rows of the sparse coefficient matrix. By introducing the row sparsity regularizer, the problem can now be succinctly formulated as

$$\begin{aligned} &\min \ \ ||X - XZ||_F^2 \\ &\text{s.t. } ||Z||_{2,1} \leq \tau, \ \ ||D^T XZ||_F^2 \leq \kappa \end{aligned} \tag{1}$$

Here, $X \in \mathbb{R}^{B \times N}$ is the feature matrix for all images in current batch, where $X = \{x_i \in \mathbb{R}^B, i = 1, \cdots, N\}$. Each x_i represents the feature descriptor of an image in current batch in B-dimensional feature space. N denotes the number of images in the batch. $Z \in \mathbb{R}^{N \times N}$ is the sparse coefficient matrix and $\|Z\|_{2,1} \triangleq \sum_{i=1}^{N} \|z_i\|_2$ is the row sparsity regularizer, i.e., sum of l_2 norms of the rows of Z. τ and κ are trade-off parameters. $D \in \mathbb{R}^{B \times M}$ is the feature matrix of current dataset, where $D = \{d_j \in \mathbb{R}^B, j = 1, \cdots, M\}$. M denotes the number of images of the category in current dataset.

The objective function in Eq. 1 is intuitive: the first constraint i.e., $l_{2,1}$ regularizer is to induce row level sparsity in representative selection [8, 11], whereas the second constraint tries to select images that are less correlated with images in current dataset. Minimization of Eq. 1 leads to a sparse solution for Z in terms of rows, *i.e.*, the sparse coefficient matrix Z contains few nonzero rows which constitute the representative set. Optimization of Eq. 1 attempts to obtain a sparse set of images non-redundant with previously selected images.

Optimization. Here, we briefly describe the strategy to solve the convex optimization problem in Eq. 1. Using Lagrange multipliers, optimization problem in Eq. 1 can be

written as

$$\min \frac{1}{2}\|X - XZ\|_F^2 + \lambda\|Z\|_{2,1} + \frac{\alpha}{2}||D^T XZ||_F^2 \quad (2)$$

where λ and α are trade-off parameters associated with sparsity and diversity regularization. We implement the algorithm using an Alternating Direction Method of Multipliers (ADMM) optimization framework [3]. We refer the reader to check [3] for more details on the ADMM. We choose columns of X corresponding to the nonzero rows of final Z and denote the feature matrix of the representative set as Y. Here, $Y \in \mathbb{R}^{B\times K}$ is the feature matrix for all images in representative set, where $Y = \{y_i \in \mathbb{R}^B, i = 1, \cdots, K\}$. y_i represents the feature descriptor of ith representative sample in B-dimensional feature space. K denotes the total number of images in selected representative set.

2.3 Active Learning for Image Labeling

After we have a diverse representative set Y, the next goal is to estimate the similarity of each image in Y to the images in the dataset D. Based on the similarity score, active learning module will determine images to be labeled.

Our goal here is to find how likely a sample image belongs to a particular class, given only some examples of the same class. Specifically, given a sample y_i, we compute its sparse representation c_i based on dictionary D.

Then, we select a sample y_i as relevant based on how well the nonzero entries in the estimate c_i are associated with the columns of D. Given the above stated goals, the optimization problem can be written as,

$$\min \ ||Y - DC||_F^2 \ \text{ s.t.}||c_i||_1 \leq s \quad (3)$$

Here, $C \in \mathbb{R}^{M\times K}$ is the sparse coefficient matrix, where $C = \{c_i \in \mathbb{R}^M, i = 1, \cdots, K\}$. s is a tradeoff parameter.

In Eq. 3, the constraint, i.e., l_1 regularizer is to induce element wise sparsity in a column. The objective is logical as any new sample of a category will approximately lie in the linear span of some samples in dataset associated with the same category. We require the coefficient matrix C to be sparse by solving the optimization program in Eq. 3. We use similar ADMM procedure stated in Sec. 2.2, to solve this optimization problem.

After calculating C, the similarity score for each image in representative set is calculated as follows,

$$\zeta_i = 1 - \frac{\|y_i - Dc_i\|_2}{\|y_i\|_2} \quad (4)$$

Here, ζ_i is similarity score of ith sample. $\|y_i - Dc_i\|_2$ indicates the residual between y_i and Dc_i, which is reconstruction of y_i using samples of same category from dataset.

For an instance $y_i \in Y$, if the corresponding similarity score ζ_i is greater than a threshold δ, we assume that our system is highly confident about this instance. Hence, we label the instance y_i using the label of the query and add to the dataset. Number of instances obtained without any human supervision is not fixed and depends on the value of δ, which we set sufficiently large so that irrelevant instances are less likely to be added to the dataset. We remove instances, which have a similarity score ζ_i smaller than a threshold γ (say $\gamma = 0.3$), as we believe there is high chance of these examples to be irrelevant. Among the remaining samples, we choose the instances with lowest similarity score first for human labeling, as these examples have greater chance to increase diversity in our dataset. We sequentially choose maximum b instances like this and request for human annotation. Here, b is our labeling budget per iteration.

3. RESULTS AND ANALYSIS

Experimental Setup. For feature extraction, in all of the experiments, we use Alexnet-CNN, which is trained on ImageNet dataset, under the network architecture of [21]. For all the results shown in this paper, we have used the features extracted from the pre-trained CNN[21] and trained SVM classifiers. We set all the Lagrangian multipliers $\lambda = \lambda_0/\mu$, where $\mu > 1$ and λ_0 is computed from the input data [11]. For fair comparison, in all the experiments, except dataset enrichment, we considered no image related to the concept word is available beforehand. We take advantage of the high-precision of few top returned images for the query from Google search by utilizing them as initial dataset.

3.1 Constructed Dataset by our Framework

We constructed a dataset using the proposed framework to verify our approach, which we name as DivNet. We are continuously adding more categories in this dataset. Now, it contains images for 550 categories, averaging 1K images per category. The dataset is publicly available to download in http://www.ee.ucr.edu/~amitrc/datasets.php. We crawled images from Bing, Google and Flickr. The categories are mainly chosen from ILSVRC2016 object detection and scene classification challenge [24].

The ratio of human labeling used, compared to the total number of images in the dataset is 11.7%. The average accuracy of the labels in DivNet is estimated by manually inspecting 5.5K images (10 random images per concept) from the entire dataset. The average accuracy has been found to be 97.2%, which is slightly lower than 99.7%, reported in ImageNet. However, for collecting same number of images for any category, the manual labeling is about 9 times lower in our case on average. Fig.2 shows human labeling statistics of 35 categories from our dataset.

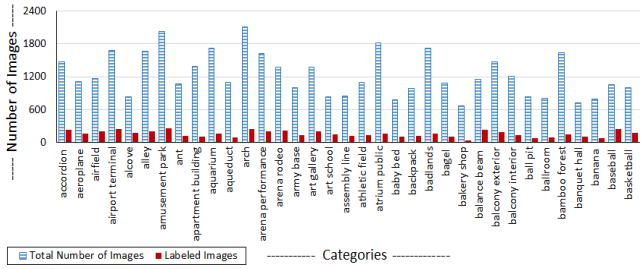
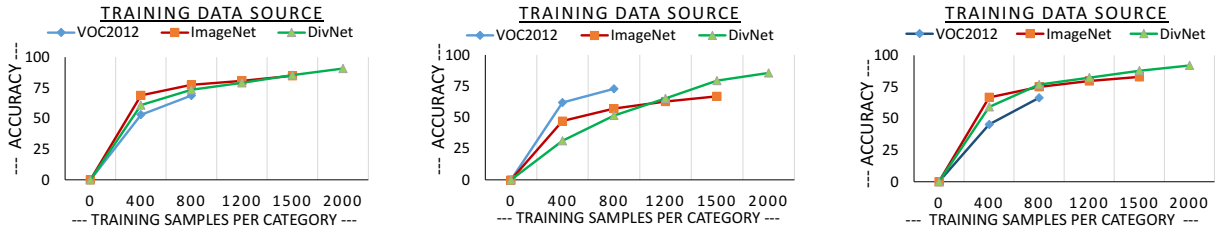


Figure 2: Number of labeled images compared to total number of selected images for 35 categories in DivNet.

3.2 Cross-Dataset Generalization

To evaluate the generalization ability of our constructed dataset, we compare the dataset with two popular hand-labeled image datasets, e.g., VOC2012 [13] and ImageNet[24]. We select all twenty categories from VOC2012 dataset [13] for this experiment. We collect images from all three datasets for these categories and train classifiers. The result for different training and testing data combinations is shown in Fig. 3. Training and testing on the same dataset provides the best performance most of the times for a fixed number of samples. However, training with DivNet shows the best generalization among datasets, as the average cross dataset performance drop (e.g. training and testing on VOC, compared to training on ours and testing on VOC) is minimum. Initially with few training samples, the performance of DivNet is lower as it has very few manually labeled samples. However, as we continuously select more diverse images to be labeled by our system, the performance improves at a higher



(a) Test on ImageNet (b) Test on VOC2012 (c) Test on DivNet

Figure 3: Cross dataset performance of classifier trained on different datasets with different number of samples per category.

rate. We can compare the performance of the datasets at the point of 1500 training samples (since state of the art ImageNet has on average 1500 images per category) and see the generalization ability of DivNet is better. The comparison at the same number of training samples shows the cross-dataset generalization ability of DivNet. Moreover, DivNet can achieve even better generalization performance because of its ability to scale up with limited labeling effort.

3.3 Diversity

In order to illustrate the diversity of images in our collected dataset, we follow [10, 7], which computes the average image of each category and measure lossless JPG file size which reflects the amount of information in an image. A diverse image set should result in a blurrier average image, and the JPG file size of average image should be smaller. We resize all images to 256X256, and create average images for each category from all images of the category. Fig. 4 shows the average images and the corresponding JPG image size comparison of four categories: person, dog, monitor and airplane. The average image of DivNet is blurrier and hard to recognize out the object, while the average image of Caltech-256 [16] is more structured and sharper. DivNet has slightly smaller JPG file size than ImageNet, but significantly smaller than Caltech-256. This phenomenon is common for almost all of the categories. For randomly selected 10 categories, the average lossless JPG size has been found to be 2.3 KB in DivNet, 2.5 KB in ImageNet and 3.9 KB in Caltech-256.

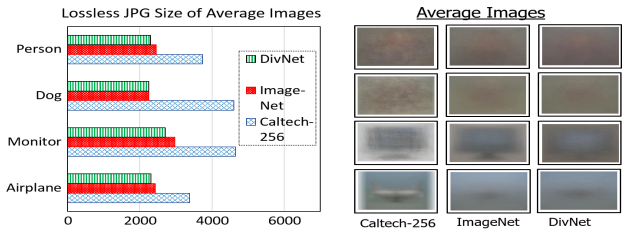
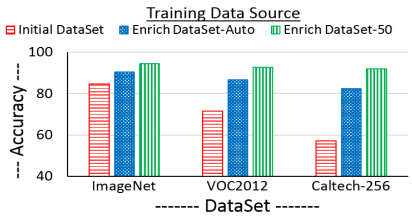


Figure 4: Average images of DivNet, ImageNet and Caltech-256 for four categories. Left chart shows the lossless JPG file sizes of average images in Bytes.

3.4 Dataset Enrichment

To evaluate the performance of our approach in dataset enrichment, we enrich ImageNet, VOC2012 and Caltech-256 by our method. For this experiment, we pick eight categories that are common in these datasets: airplane, bike, bird, car, dog, horse, monitor and person. For each category, we start our dataset construction method with images from a particular dataset as initial dataset and collect 1000 more images using our framework automatically with no labeling (Enrich Dataset-Auto) and also with a labeling budget of 50 (Enrich Dataset-50). We train image classifier on these categories with initial dataset images and also with enriched dataset images. The result in Fig. 5 shows that the performance of classifier improves after enriching dataset with our framework. Hence, our method is suitable for extending existing image collections with discriminative examples.

Figure 5: Change in image classifier performance after enriching datasets with our framework.



3.5 Scalability

Different from static dataset construction, our method can be used to dynamically update datasets. One can collect images based on desired dataset size and labeling budget. Such property makes sense as one user may be interested in collecting more image for a category, compared to others. It is also likely that a user may want to spend more time labeling images from some particular category than other categories. We investigate the scalability in labeling by collecting fixed number of images with different labeling budget. The accuracy of the classifier per category increases by 3.2% on average initially, as we increase labeling budget by 25. However, the performance improvement usually saturates after labeling around 100-200 examples(the actual number varies by category).

4. CONCLUSIONS

In this paper, we propose a semi-supervised framework for collecting images from web, which is suitable for dataset construction, or enriching existing datasets with discriminative examples. Our system provides flexibility that permits us to filter out irrelevant images and obtain a reliable set of diverse images based on resource and labeling budget available, so that a high-precision large-scale image classifier can be trained. The experimental results demonstrate that our approach is not only useful in reducing the manual annotation efforts, but also successful in collecting images with high precision and diversity, and robust image classifiers can be trained from these images. Future works will investigate domain-adaptation techniques and other image metadata available on the web to further improve the dataset construction framework in terms of budget and quality.

Acknowledgments: This work was partially supported by US NSF grants IIS-1316934 and CPS-1544969. We would like to thank Andrew Kwon, Cody Simons and Nishanth Babu, three current UCR undergraduate students, for helping in the construction of the DivNet dataset.We also thank NVIDIA for donating a Tesla K40 GPU.

5. REFERENCES

[1] Y. Bai, K. Yang, W. Yu, C. Xu, W.-Y. Ma, and T. Zhao. Automatic image dataset construction from click-through logs using deep neural network. In *Proc. ACM Conf. Multimedia*, pages 441–450. ACM, 2015.

[2] J. H. Bappy, S. Paul, and A. K. Roy-Chowdhury. Online adaptation for joint scene and object classification. In *ECCV*. Springer, 2016.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[4] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *ICCV*, pages 1431–1439, 2015.

[5] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, pages 1409–1416, 2013.

[6] D. S. Cheng, F. Setti, N. Zeni, R. Ferrario, and M. Cristani. Semantically-driven automatic creation of training sets for object recognition. *Computer Vision and Image Understanding*, 131:56–71, 2015.

[7] B. Collins, J. Deng, K. Li, and L. Fei-Fei. Towards scalable dataset construction: An active learning approach. In *ECCV*, pages 86–98. Springer, 2008.

[8] Y. Cong, J. Yuan, and J. Luo. Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Trans. Multimedia*, 14(1):66–75, 2012.

[9] Y. Cui, F. Zhou, Y. Lin, and S. Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. *arXiv preprint arXiv:1512.05227*, 2015.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[11] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, pages 1600–1607. IEEE, 2012.

[12] E. Elhamifar, G. Sapiro, A. Yang, and S. Sasrty. A convex optimization framework for active learning. In *ICCV*, pages 209–216, 2013.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[14] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, volume 2, pages 1816–1823. IEEE, 2005.

[15] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*, pages 522–530, 2009.

[16] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. California Institute of Technology, 2007.

[17] M. Hasan and A. K. Roy-Chowdhury. Context aware active learning of activity recognition models. In *ICCV*, pages 4543–4551, 2015.

[18] H. Izadinia, B. C. Russell, A. Farhadi, M. D. Hoffman, and A. Hertzmann. Deep classifiers from image tags in the wild. In *Proc. of the Workshop on Community Organized Multimodal Mining: Opportunities for Novel Solutions*, pages 13–18. ACM, 2015.

[19] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *ECCV*, pages 158–171. Springer, 2012.

[20] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. *arXiv preprint arXiv:1511.06789*, 2015.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[22] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1–14, 1995.

[23] L.-J. Li and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *IJCV*, 88(2):147–168, 2010.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[25] B. Settles. Active learning literature survey. *Univ. of Wisconsin, Madison*, 52(55-66):11, 2010.

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

[27] R. Speer and C. Havasi. Conceptnet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*, pages 161–176. Springer, 2013.

[28] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*. IEEE, 2015.

[30] A. Torralba, A. Efros, et al. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528. IEEE, 2011.

[31] S.-Y. Wang, W.-S. Liao, L.-C. Hsieh, Y.-Y. Chen, and W. H. Hsu. Learning by expansion: Exploiting social media for image classification with few training examples. *Neurocomputing*, 95:117–125, 2012.

[32] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[33] Y. Xia, X. Cao, F. Wen, and J. Sun. Well begun is half done: Generating high-quality seeds for automatic image dataset construction from web. In *ECCV*, pages 387–400. Springer, 2014.

[34] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010.

[35] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[36] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai. Graph regularized sparse coding for image representation. *IEEE Trans. Image Processing*, 20(5):1327–1336, 2011.