
Deep Convolutional Ranking for Multilabel Image Annotation

Yunchao Gong
UNC Chapel Hill
yunchao@cs.unc.edu

Yangqing Jia
Google Research
jiayq@google.com

Thomas K. Leung
Google Research
leungtk@google.com

Alexander Toshev
Google Research
toshev@google.com

Sergey Ioffe
Google Research
sioffe@google.com

Abstract

Multilabel image annotation is one of the most important challenges in computer vision with many real-world applications. While existing work usually use conventional visual features for multilabel annotation, features based on Deep Neural Networks have shown potential to significantly boost performance. In this work, we propose to leverage the advantage of such features and analyze key components that lead to better performances. Specifically, we show that a significant performance gain could be obtained by combining convolutional architectures with approximate top- k ranking objectives, as they naturally fit the multilabel tagging problem. Our experiments on the NUS-WIDE dataset outperforms the conventional visual features by about 10%, obtaining the best reported performance in the literature.

1 Introduction

Multilabel image annotation [25, 14] is an important and challenging problem in computer vision. Most existing work focus on single-label classification problems [6, 21], where each image is assumed to have only one class label. However, this is not necessarily true for real world applications, as an image may be associated with multiple semantic tags (Figure 1). As a practical example, images from Flickr are usually accompanied by several tags to describe the content of the image, such as objects, activities, and scene descriptions. Images on the Internet, in general, are usually associated with sentences or descriptions, instead of a single class label, which may be deemed as a type of multitagging. Therefore, it is a practical and important problem to accurately assign multiple labels to one image.

Single-label image classification has been extensively studied in the vision community, the most recent advances reported on the large-scale ImageNet database [6]. Most existing work focus on designing visual features for improving recognition accuracy. For example, sparse coding [33, 36], Fisher vectors [28], and VLAD [18] have been proposed to reduce the quantization error of “bag of words”-type features. Spatial pyramid matching [21] has been developed to encode spatial information for recognition. Very recently, deep convolutional neural networks (CNN) have demonstrated promising results for single-label image classification [20]. Such algorithms have all focused on learning a better feature representation for one-vs-rest classification problems, and it is not yet clear how to best train an architecture for multilabel annotation problems.

In this work, we are interested in leveraging the highly expressive convolutional network for the problem of multilabel image annotation. We employed a similar network structure to [20], which contains several convolutional and dense connected layers as the basic architecture. We studied



Figure 1: Sample images from the NUS-WIDE dataset, where each image is annotated with several tags.

and compared several other popular multilabel losses, such as the ranking loss [19] that optimizes the area under ROC curve (AUC), and the cross-entropy loss used in Tagprop [14]. Specifically, we propose to use the top- k ranking loss, inspired by [34], for embedding to train the network. Using the largest publicly available multilabel dataset NUS-WIDE [4], we observe a significant performance boost over conventional features, reporting the best retrieval performance on the benchmark dataset in the literature.

1.1 Previous Work

In this section we first review related works on multilabel image annotation and then briefly discuss works on deep convolutional networks.

Modeling Internet images and their corresponding textural information (e.g., sentences, tags) have been of great interest in the vision community [2, 10, 12, 15, 30, 29, 34]. In this work, we focus on the image annotation problem and summarize several important lines of related research. Early work in this area was mostly devoted to annotation models inspired by machine translation techniques [1, 8]. The work by Barnard et al. [1, 8] applied machine translation methods to parse natural images and tried to establish a relationship between image regions and words.

More recently, image annotation has been formulated as a classification problem. Early works focused on generative model based tagging [1, 3, 26], centred upon the idea of learning a parametric model to perform predictions. However, because image annotation is a highly nonlinear problem, a parametric model might not be sufficient to capture the complex distribution of the data. Several recent works on image tagging have mostly focused on nonparametric nearest-neighbor methods, which offer higher expressive power. The work by Makadia et al. [25], which proposed a simple nearest-neighbor-based tag transfer approach, achieved significant improvement over previous model-based methods. Recent improvements on the nonparametric approach include TagProp [14], which learns a discriminative metric for nearest neighbors to improve tagging.

Convolutional neural networks (CNNs) [20, 22, 23, 17, 5] are a special type of neural network that utilizes specific network structures, such as convolutions and spatial pooling, and have exhibited good generalization power in image-related applications. Combined with recent techniques such as Dropout and fast parallel training, CNN models have outperformed existing handcrafted features. Krizhevsky et al. [20] reported record-breaking results on ILSVRC 2012 that contains 1000 visual-object categories. However, this study was mostly concerned with single-label image classification, and the images in the dataset only contain one prominent object class. At a finer scale, several methods focus on improving specific network designs. Notably, Zeiler et al. [37] investigated different pooling methods for training CNNs, and several different regularization methods, such as Dropout [16], DropConnect [32], and Maxout [13] have been proposed to improve the robustness and representation power of the networks. In addition, Earlier studies [7] have shown that CNN features are suitable as a general feature for various tasks under the conventional classification schemes, and our work focuses on how to directly train a deep network from raw pixels, using multilabel ranking loss, to address the multilabel annotation problem.

2 Multilabel Deep Convolutional Ranking Net

In our approach for multilabel image annotation, we adopted the architecture proposed in [20] as our basic framework and mainly focused on training the network with loss functions tailored for multi-label prediction tasks.

2.1 Network Architecture

The basic architecture of the network we use is similar to the one used in [20]. We use five convolutional layers and three densely connected layers. Before feeding the images to the convolutional layers, each image is resized to 256×256 . Next, 220×220 patches are extracted from the whole image, at the center and the four corners to provide an augmentation of the dataset. Convolution filter sizes are set to squares of size 11, 9, and 5 respectively for the different convolutional layers; and max pooling layers are used in some of the convolutional layers to introduce invariance. Each densely connected layer has output sizes of 4096. Dropout layers follow each of the densely connected layers with a dropout ratio of 0.6. For all the layers, we used rectified linear units (RELU) as our nonlinear activation function.

The optimization of the whole network is achieved by asynchronous stochastic gradient descent with a momentum term with weight 0.9, with mini-batch size of 32. The global learning rate for the whole network is set to 0.002 at the beginning, and a staircase weight decay is applied after a few epochs. The same optimization parameters and procedure are applied to all the different methods. For our dataset with 150,000 training images, it usually takes one day to obtain a good model by training on a cluster. Unlike previous work that usually used ImageNet to pre-train the network, we train the whole network directly from the training images from the NUS-WIDE dataset for a fair comparison with conventional vision baselines.

2.2 Multilabel Ranking Losses

We mainly focused on loss layer, which specifies how the network training penalizes the deviation between the predicted and true labels, and investigated several different multilabel loss functions for training the network. The first loss function was inspired by Tagprop [14], for which we minimized the multilabel softmax regression loss. The second loss was a simple modification of a pairwise-ranking loss [19], which takes multiple labels into account. The third loss function was a multilabel variant of the WARP loss [34], which uses a sampling trick to optimize top- k annotation accuracy.

For notations, assume that we have a set of images \mathbf{x} and that we denote the convolutional network by $f(\cdot)$ where the convolutional layers and dense connected layers filter the images. The output of $f(\cdot)$ is a scoring function of the data point \mathbf{x} , that produces a vector of activations. We assume there are n image training data and c tags.

2.2.1 Softmax

The softmax loss has been used for multilabel annotation in Tagprop [14], and is also used in single-label image classification [20]; therefore, we adopted it in our context. The posterior probability of an image \mathbf{x}_i and class j can be expressed as

$$p_{ij} = \frac{\exp(f_j(\mathbf{x}_i))}{\sum_{k=1}^c \exp(f_k(\mathbf{x}_i))}, \quad (1)$$

where $f_j(\mathbf{x}_i)$ means the activation value for image \mathbf{x}_i and class j . We then minimized the KL-Divergence between the predictions and the ground-truth probabilities. Assuming that each image has multiple labels, and that we can form a label vector $\mathbf{y} \in R^{1 \times c}$ where $y_j = 1$ means the presence of a label and $y_j = 0$ means absence of a label for an image, we can obtain ground-truth probability by normalizing \mathbf{y} as $\mathbf{y}/\|\mathbf{y}\|_1$. If the ground truth probability for image i and class j is defined as \bar{p}_{ij} , the cost function to be minimized is

$$J = -\frac{1}{m} \sum_{i=1}^n \sum_{j=1}^c \bar{p}_{ij} \log(p_{ij}) = -\frac{1}{m} \sum_{i=1}^n \sum_{j=1}^{c_+} \frac{1}{c_+} \log(p_{ij})$$

where c_+ denotes the number of positive labels for each image. For the ease of exposition and without loss of generality, we set c_+ to be the same for all images.

2.2.2 Pairwise Ranking

The second loss function we considered was the pairwise-ranking loss [19], which directly models the annotation problem. In particular, we wanted to rank the positive labels to always have higher

scores than negative labels, which led to the following minimization problem:

$$J = \sum_{i=1}^n \sum_{j=1}^{c_+} \sum_{k=1}^{c_-} \max(0, 1 - f_j(\mathbf{x}_i) + f_k(\mathbf{x}_i)), \quad (2)$$

where c_+ is the positive labels and c_- is the negative labels. During the back-propagation, we computed the sub-gradient of this loss function. One limitation of this loss is that it optimizes the area under the ROC curve (AUC) but does not directly optimize the top- k annotation accuracy. Because for image annotation problems we were mostly interested in top- k annotations, this pairwise ranking loss did not best fit our purpose.

2.2.3 Weighted Approximate Ranking (WARP)

The third loss we considered was the weighted approximate ranking (WARP), which was first described in [34]. It specifically optimizes the top- k accuracy for annotation by using a stochastic sampling approach. Such an approach fits the stochastic optimization framework of the deep architecture very well. It minimizes

$$J = \sum_{i=1}^n \sum_{j=1}^{c_+} \sum_{k=1}^{c_-} L(r_j) \max(0, 1 - f_j(\mathbf{x}_i) + f_k(\mathbf{x}_i)). \quad (3)$$

where $L(\cdot)$ is a weighting function for different ranks, and r_j is the rank for the j th class for image i . The weighting function $L(\cdot)$ used in our work is defined as:

$$L(r) = \sum_{j=1}^r \alpha_j, \text{ with } \alpha_1 \geq \alpha_2 \geq \dots \geq 0. \quad (4)$$

We defined the α_i as equal to $1/j$, which is the same as [34]. The weights defined by $L(\cdot)$ control the top- k of the optimization. In particular, if a positive label is ranked top in the label list, then $L(\cdot)$ will assign a small weight to the loss and will not cost the loss too much. However, if a positive label is not ranked top, $L(\cdot)$ will assign a much larger weight to the loss, which pushes the positive label to the top. The last question was how to estimate the rank r_j . We followed the sampling method in [34]: for a positive label, we continued to sample negative labels until we found a violation; then we recorded the number of trials s we sampled for negative labels. The rank was estimated by the following formulation

$$r_j = \lfloor \frac{c-1}{s} \rfloor, \quad (5)$$

for c classes and s sampling trials. We computed the sub-gradient for this layer during optimization.

As a minor noite, the approximate objective we optimize is a looser upper bound compared to the original WARP loss proposed in [34]. To see this, notice that in the original paper, it is assumed that the probability of sampling a violator is $p = \frac{r}{\#Y-1}$ with a positive example (x, y) with rank r , where $\#Y$ is the number of labels. Thus, there are r labels with higher scores than y . This is true only if all these r labels are negative. However, in our case, since there may be other positive labels having higher scores than y due to the multi-label nature of the problem, we effectively have $p \leq \frac{r}{\#Y-1}$.

3 Visual Feature based Image Annotation Baselines

We used a set of 9 different visual features and combined them to serve as our baseline features. Although such a set of features might not have been the best possible ones we could obtain, they already serve as a very strong visual representation, and the computation of such features is non-trivial. On top of these features, we ran two simple but powerful classifiers (kNN and SVM) for image annotation. We also experimented with Tagprop [14], but found it cannot easily scale to a large training set because of the $O(n^2)$ time complexity. After using a small training set to train the Tagprop model, we found the performance to be unsatisfactory and therefore do not compare it here.

3.1 Visual Features

GIST [27]: We resized each image to 200×200 and used three different scales [8, 8, 4] to filter each RGB channel, resulting in 960-dimensional (320×3) GIST feature vectors.

SIFT: We used two different sampling methods and three different local descriptors to extract texture features, which gave us a total of 6 different features. We used dense sampling and a Harris corner detector as our patch-sampling methods. For local descriptors, we extracted SIFT [24], CSIFT [31], and RGBSIFT [31], and formed a codebook of size 1000 using kmeans clustering; then built a two-level spatial pyramid [21] that resulted in a 5000-dimensional vector for each image. We will refer to these six features as D-SIFT, D-CSIFT, D-RGBSIFT, H-SIFT, H-CSIFT, and H-RGBSIFT.

HOG: To represent texture information at a larger scale, we used 2×2 overlapping HOG as described in [35]. We quantized the HOG features to a codebook of size 1000 and used the same spatial pyramid scheme as above, which resulted in 5000-dimensional feature vectors.

Color: We used a joint RGB color histogram of 8 bins per dimension, for a 512-dimensional feature.

The same set of features were used in [11], and achieved state-of-the-art performance for image retrieval and annotation. The combination of this set of features has a total dimensionality of 36,472, which makes learning very expensive. We followed [11] to perform simple dimensionality reductions to reduce computation. In particular, we performed a kernel PCA (KPCA) separately on each feature to reduce the dimensionality to 500. Then we concatenated all of the feature vectors to form a 4500-dimensional global image feature vector and performed different learning algorithms on it.

3.2 Visual feature + k NN

The simplest baseline that remains very powerful involves directly applying a weighted k NN on the visual feature vectors. k NN is a very strong baseline for image annotation, as suggested by Makadia et al. [25], mainly because multilabel image annotation is a highly nonlinear problem and handling the heavily tailed label distribution is usually very difficult. By contrast, k NN is a highly nonlinear and adaptive algorithm that better handles rare tags. For each test image, we found its k nearest neighbors in the training set and computed the posterior probability $p(c|i)$ as

$$p(c|i) = \sum_{j=1}^k \frac{1}{k} \exp\left(-\frac{\|x_i - x_j\|_2^2}{\sigma}\right) y_{jk}, \quad (6)$$

where y_{ik} indexes the labels of training data, $y_{ik} = 1$ when there is one label for this image, and $y_{ik} = 0$ when there is no label for this image. σ is the bandwidth that needs to be tuned. After obtaining the prediction probabilities for each image, we sorted the scores and annotated each testing image with the top- k tags.

3.3 Visual feature + SVM

Another way to perform image annotation is to treat each tag separately and to train c different one-vs-all classifiers. We trained a linear SVM [9] for each tag and used the output of the c different SVMs to rank the tags. Because we had already performed nonlinear mapping to the data during the KPCA stage, we found a linear SVM to be sufficient. Thus we assigned top- k tags to one image, based on the ranking of the output scores of the SVMs.

4 Experiments

4.1 Dataset

We performed experiments on the largest publicly available multilabel dataset, NUS-WIDE [4]. This dataset contains 269,648 images downloaded from Flickr that have been manually annotated, with several tags (2-5 on average) per image. After ignoring the small subset of the images that are not annotated by any tag, we had a total of 209,347 images for training and testing. We used a subset of 150,000 images for training and used the rest of the images for testing. The tag dictionary for the images contains 81 different tags. Some sample images and annotations are shown in Figure 1.

method / metric	per-class recall	per-class precision	overall recall	overall precision	N+
Upper bound	97.00	44.87	82.76	66.49	100.00
Visual Feature + kNN	19.33	32.59	53.44	42.93	91.36
Visual Feature + SVM	18.79	21.51	35.87	28.82	82.72
CNN + Softmax	31.22	31.68	59.52	47.82	98.76
CNN + Ranking	26.83	31.93	58.00	46.59	95.06
CNN + WARP	35.60	31.65	60.49	48.59	96.29

Table 1: Image annotation results on NUS-WIDE with $k = 3$ annotated tags per image. See text in section 5.4 for the definition of “Upper bound”.

method / metric	per-class recall	per-class precision	overall recall	overall precision	N+
Upper bound	99.57	28.83	96.40	46.22	100.00
Visual Feature + kNN	32.14	22.56	66.98	32.29	95.06
Visual Feature + SVM	34.19	18.79	47.15	22.73	96.30
CNN + Softmax	48.24	21.98	74.04	35.69	98.76
CNN + Ranking	42.48	22.74	72.78	35.08	97.53
CNN + WARP	52.03	22.31	75.00	36.16	100.00

Table 2: Image annotation results on NUS-WIDE with $k = 5$ annotated tags per image. See text in section 5.4 for the definition of “Upper bound”.

4.2 Evaluation Protocols

We followed previous research [25] in our use of the following protocols to evaluate different methods. For each image, we assigned k (e.g., $k = 3, 5$) highest-ranked tags to the image and compared the assigned tags to the ground-truth tags. We computed the recall and precision for each tag separately, and then report the mean-per-class recall and mean-per-class precision:

$$\text{per-class recall} = \frac{1}{c} \sum_{i=1}^c \frac{N_i^c}{N_i^g}, \quad \text{per-class precision} = \frac{1}{c} \sum_{i=1}^c \frac{N_i^c}{N_i^p} \quad (7)$$

where c is the number of tags, N_i^c is the number of correctly annotated image for tag i , N_i^g is the number of ground-truth tags for tag i , and N_i^p is the number of predictions for tag i . The above evaluations are biased toward infrequent tags, because making them correct would have a very significant impact on the final accuracy. Therefore we also report the overall recall and overall precision:

$$\text{overall recall} = \frac{\sum_{i=1}^c N_i^c}{\sum_{i=1}^c N_i^g}, \quad \text{overall precision} = \frac{\sum_{i=1}^c N_i^c}{\sum_{i=1}^c N_i^p}. \quad (8)$$

For the above two metrics, the frequent classes will be dominant and have a larger impact on final performance. Finally, we also report the percentage of recalled tag words out of all tag words as N+. We believe that evaluating all of these metrics makes the evaluation unbiased.

4.3 Baseline Parameters

In our preliminary evaluation, we optimized the parameters for the visual-feature-based baseline systems. For visual-feature dimensionality reduction, we followed the suggestions in Gong et al. [11] to reduce the dimensionality of each feature to 500 and then concatenated the PCA-reduced vectors into a 4500-dimensional global image descriptor, which worked as well as the original feature. For kNN, we set the bandwidth σ to 1 and k to 50, having found that these settings work best. For SVM, we set the regularization parameter to $C = 2$, which works best for this dataset.

4.4 Results

We first report results with respect to the metrics introduced above. In particular, we vary the number k of predicted keywords for each image and mainly consider $k = 3$ and $k = 5$. Before doing so, however, we must define an upper bound for our evaluation. In the dataset, each image had different numbers of ground-truth tags, which made it hard for us to precisely compute an upper

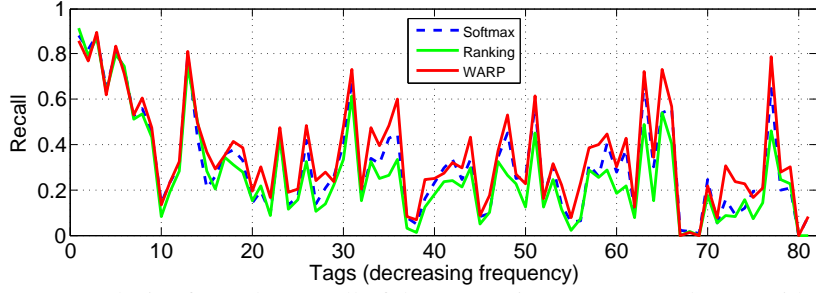


Figure 2: Analysis of per-class recall of the 81 tags in NUS-WIDE dataset with $k = 3$.

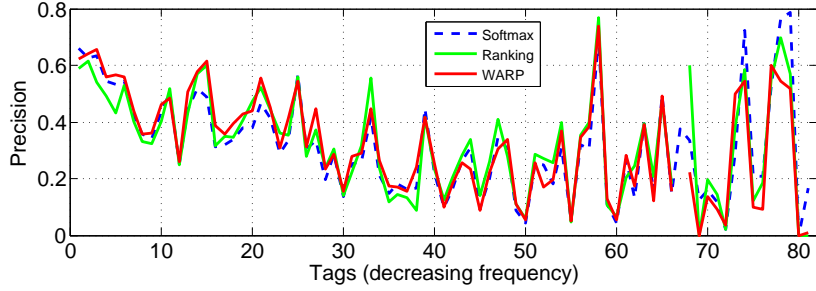


Figure 3: Analysis of per-class precision of the 81 tags in NUS-WIDE dataset with $k = 3$.

bound for performance with different k . For each image, when the number of ground-truth tags was larger than k , we randomly chose k ground-truth tags and assigned them to that image; when the number of ground-truth tags was smaller than k , we assigned all ground-truth tags to that image and randomly chose other tags for that image. We believe this baseline represents the best possible performance when the ground truth is known. The results for assigning 3 keywords per image are reported in Table 1. The results indicate that the deep network achieves a substantial improvement over existing visual-feature-based annotation methods. The CNN+Softmax method outperforms the VisualFeature+SVM baseline by about 10%. Comparing the same CNN network with different loss functions, results show that softmax already gives a very powerful baseline. Although using the pairwise ranking loss does not improve softmax, by using the weighted approximated-ranking loss (WARP) we were able to achieve a substantial improvement over softmax. This is probably because pairwise-ranking is not directly optimizing the top- k accuracy, and because WARP pushes classes that are not ranked top heavier, which boosts the performance of rare tags. From these results, we can see that all loss functions achieved comparable overall-recall and overall-precision, but that WARP loss achieved significantly better per-class recall and per-class precision. Results for $k = 5$, which are given in Table 2, show similar trends to $k = 3$.

We also provide a more detailed analysis of per-class recall and per-class precision. The recall for each tag appears in Figure 2, and the precision for each tag in Figure 3. The results for different tags are sorted by the frequency of each tag, in descending order. From these results, we see that the accuracy for frequent tags greater than for infrequent tags. Different losses performed comparably to each other for frequent classes, and WARP worked better than other loss functions for infrequent classes. Finally, we show some image annotation examples in Figure 4. Even though some of the predicted tags for these do not match the ground truth, they are still very meaningful.

5 Discussion and Future Work

In this work, we proposed to use ranking to train deep convolutional neural networks for multilabel image annotation problems. We investigated several different ranking-based loss functions for training the CNN, and found that the weighted approximated-ranking loss works particularly well for multilabel annotation problems. We performed experiments on the largest publicly available multilabel image dataset NUS-WIDE, and demonstrated the effectiveness of using top- k ranking to train the network. In the future, we would like to use very large amount of noisy-labeled multilabel images from the Internet (e.g., from Flickr or image searches) to train the network.









Image	Ground truth	Predictions	Image	Ground truth	Predictions
	Boat Cloud Ocean Vehicle Water	Lake Ocean Cloud Sky Water		Grass	Cloud House Animal Grass Sky
	Cloud Flower plant Sky Valley	Grass Plant Cloud Flower Sky		Road	Grass Animal Bird Food Toy
	Beach Rock Sky Sunset Water	Rock Water Ocean Cloud Sky		Cloud Sky Snow Sunset Tree	Tree Snow Cloud Sky Water
	Animal Cloud Cow Grass	Horse Animal Cloud Grass Sky		Cloud Mountain Rock Sky	Valley Mountain Building Sky Cloud

Figure 4: Qualitative image annotation results obtained with WARP.

References

- [1] Kobus Barnard and David Forsyth. Learning the semantics of words and pictures. In *ICCV*, 2001.
- [2] Tamara Berg and David Forsyth. Animals on the web. *CVPR*, 2007.
- [3] Gustavo Carneiro, Antoni B Chan, Pedro J Moreno, and Nuno Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):394–410, 2007.
- [4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR’09)*, Santorini, Greece., July 8-10, 2009.
- [5] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc Le, Mark Mao, Marc’Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Ng. Large scale distributed deep networks. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1232–1240, 2012.
- [6] Jia Deng, W. Dong, R. Socher, Lijia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.
- [7] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.
- [8] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.
- [9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 2008.
- [10] Rob Fergus, Antonio Torralba, and Yair Weiss. Semi-supervised learning in gigantic image collections. *NIPS*, 2009.
- [11] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for internet images, tags, and their semantics. *IJCV*, 2013.
- [12] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: An procrustean approach to learning binary codes. *CVPR*, 2011.

- [13] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *ICML*, 2013.
- [14] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. *ICCV*, 2009.
- [15] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. *CVPR*, 2010.
- [16] Geoffrey Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *Arxiv*, 2012.
- [17] What is the best multi-stage architecture for object recognition? K. jarrett and k. kavukcuoglu and m. a. ranzato and y. lecun. *CVPR*, 2009.
- [18] Hervé Jégou, M. Douze, Cordelia Schmid, and Patrick Perez. Aggregating local descriptors into a compact image representation. *CVPR*, 2010.
- [19] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.
- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.
- [21] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2006.
- [22] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Handwritten digit recognition with a back-propagation network. *NIPS*, 1990.
- [23] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *ICML*, 2009.
- [24] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
- [25] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A new baseline for image annotation. In *ECCV*, 2008.
- [26] F. Monay and D. Gatica-Perez. Plsa-based image autoannotation: Constraining the latent space. In *ACM Multimedia*, 2004.
- [27] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001.
- [28] Florent Perronnin and Christopher R. Dance. Fisher kernels on visual vocabularies for image categorization. *CVPR*, 2007.
- [29] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. *CVPR*, 2007.
- [30] N. Rasiwasia, PJ Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia*, 2007.
- [31] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 2010.
- [32] Li Wan, Matt Zeiler, Sixin Zhang, Yann Lecun, and Rob Fergus. Regularization of neural networks using dropconnect. *ICML*, 2013.
- [33] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. *CVPR*, 2010.
- [34] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.
- [35] Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *CVPR*, 2010.
- [36] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. *CVPR*, 2009.
- [37] Matt Zeiler and Rob Fergus. Stochastic pooling for regularization of deep convolutional neural networks. *ICLR*, 2013.