ConTagNet: Exploiting User Context for Image Tag Recommendation

Yogesh Singh Rawat
School of Computing
National University of Singapore
yogesh@comp.nus.edu.sg

Mohan S Kankanhalli
School of Computing
National University of Singapore
mohan@comp.nus.edu.sg

ABSTRACT

In recent years, deep convolutional neural networks have shown great success in single-label image classification. However, images usually have multiple labels associated with them which may correspond to different objects or actions present in the image. In addition, a user assigns tags to a photo not merely based on the visual content but also the context in which the photo has been captured. Inspired by this, we propose a deep neural network which can predict multiple tags for an image based on the content as well as the context in which the image is captured. The proposed model can be trained end-to-end and solves a multi-label classification problem. We evaluate the model on a dataset of 1,965,232 images which is drawn from the YFCC100M dataset provided by the organizers of Yahoo-Flickr Grand Challenge. We observe a significant improvement in the prediction accuracy after integrating user-context and the proposed model performs very well in the Grand Challenge.

CCS Concepts

•**Information systems** → *Multimedia information systems;*

Keywords

Convolutional Neural Networks; Multi-label Image Tagging; User Context

1. INTRODUCTION

The model presented in this paper is our solution to the ACM Multimedia 2016 Grand Challenge on user tag and caption prediction presented by Yahoo-Flickr. We mainly focus on the user tag prediction task in this work and its extension to caption prediction task is discussed as a future work. We solved the user tag prediction problem as a multi-label classification and proposed a convolutional deep neural network which integrates both the image pixel values and context in a unified framework for predicting a list of user tags.

Social media users like to assign tags to the captured images before sharing with others. One of the biggest advantages of adding tags to photographs is that it makes them searchable and easily discoverable by other users. To help users annotate their photos, researchers have developed visual recognition algorithms which can recommend tags based on the visual content of the image. Deep convolutional neural networks (CNNs) have shown a great progress in visual recognition tasks in recent years [15, 21, 24]. The existing methods are mainly focused on datasets, such as ImageNet [4], that were manually annotated for single-label classification. Also, the different categories are defined by researchers and not by users, which means they are not necessarily relevant for tag recommendation.

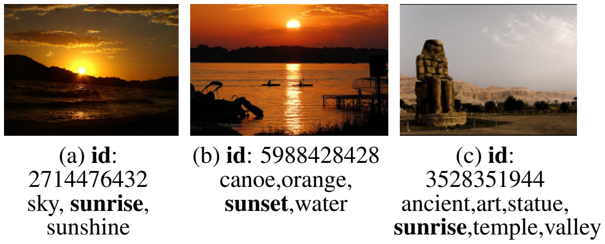

(a) **id**: 2714476432 sky, **sunrise**, sunshine
(b) **id**: 5988428428 canoe,orange, **sunset**,water
(c) **id**: 3528351944 ancient,art,statue, **sunrise**,temple,valley

Figure 1: Sample images from YFCC100M dataset [25] along with photo identifier and associated tags assigned by the users. Each photo has been assigned with multiple tags and it can be observed that the visual content of the images is not sufficient to predict the tags. Images in (a) and (b) have similar visual content, however, (a) is captured at sunrise and (b) has been captured at sunset. In addition, image (c) has been assigned 'sunrise' as a tag, which is difficult to predict using visual content of the image.

An image can have multiple annotations based on its visual content, interactions between objects present and many other factors. Motivated by this, researchers have shown interest in multi-label image classification task where an image can be annotated with multiple tags [18, 10, 29, 16, 27]. A basic approach to solving multi-label classification problem is to train independent single-label classifiers for each of the label [10]. However, when trained independently, the dependency knowledge between the various labels corresponding to an image is lost. The co-occurrence dependency between labels has also been utilized by earlier works to improve the annotation performance [31, 27].

We further observe that a user assigns tags to a captured image based on not only the image content but also his or her current context, such as time and location. In addition, context can also help in improving the visual recognition task. For example, sunset and sunrise photographs may have similar visual content (Figure 1), however, with available time context, it becomes simpler to identify the correct annotation. In addition, an appropriate tag for an image does not always have to be about the image visual content, it can also be related to the current user context (Figure 1c). Therefore, apart from image content, user context also plays an important role in the recommendation of image tags. Inspired by this, we propose to explore the role of context in multiple-label user-tag prediction using a deep neural network.

In this work, we propose a unified CNN-NN framework to integrate context with image content for multi-label tag prediction. The framework of the proposed model is shown in Figure 2. In comparison with state-of-the-art multi-label tag prediction methods, the

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2964284.2984068

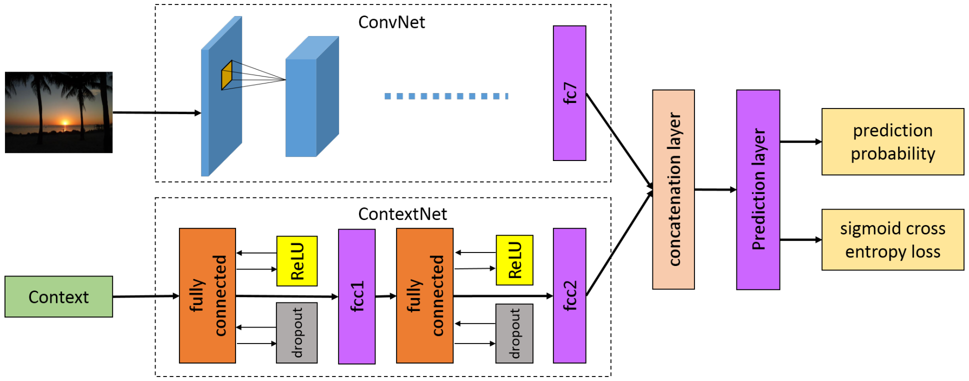


Figure 2: Overview of the proposed model. An input image is passed through a CNN (ConvNet) component of the network which has been adopted from AlexNet [15]. The associated context of the image is processed through a parallel NN (ContextNet). The output of ConvNet (fc7 layer) and ContextNet (fcc2 layer) are merged together and passed to the prediction layer for multi-label prediction. The proposed model can be trained end-to-end using gradient descent optimization.

proposed framework has the following advantages. First of all, it incorporates the context information associated with the image for tag prediction. Secondly, it integrates the user context with image content in a unified way and the complete model can be trained end-to-end. And finally, the model is trained on a large set of user defined tags (1540) which is much bigger as compared to other publicly available datasets for multi-label classification [3, 5, 17]. We evaluate the proposed framework on a dataset exclusively drawn from YFCC100M provided by Yahoo-Flickr [25]. The experimental evaluation shows that integrating context significantly improves the tag prediction accuracy.

The rest of the paper is organized as follows. First of all, we will discuss the related work in section 2. In section 3 we will discuss the proposed method followed by experimental results in section 4. Finally, we will conclude this paper with a conclusion and future research direction in section 5.

2. RELATED WORK

In past few years, image classification has made a significant progress due to development in deep convolutional neural networks [15] and availability of large-scale labeled datasets such as ImageNet [4]. Apart from single-label classification multi-label classification is also making progress because of available datasets, such as NUS-WIDE [3], VOC PASCAL [5] and Microsoft COCO [17]. Some of the attempts to solve multi-label classification task using deep learning include works which treat each label independently and ignore any correlation between corresponding labels [8, 28]. Gong et al. [8] investigated several ranking-based loss functions for CNN and proposed weighted approximated-ranking loss for multi-label annotation. Wei et al. [28] proposed Hypothesis-CNN-Pooling (HCP) where outputs from multiple hypotheses are aggregated using max pooling.

To address the above-mentioned limitation, researchers have proposed joint modeling of image and label embedding. Weston et al. [30] learn a joint representation of images and annotations that optimize top-of-the-list ranking. Similarly, Frome et al. [6] proposed a model for visual-semantic embedding to identify visual objects using both pixel values as well as image semantics. In [9], Gong et al. employ canonical correlation analysis (CCA) to map visual features and labels to the same latent space and incorporate a third view to capture high-level image semantics. These methods effectively capture the semantic relation between tags and image content, however, they ignore the co-occurrence relation between multiple tags.

The label co-occurrence dependency has also been studied by researchers for multi-label classification task [7, 20, 11, 31, 16]. In [7] Ghamrawi et al. proposed conditional random field (CRF) classification models for multi-label classification that parameterize label co-occurrences. Read et al. [20] utilize classifier chains method to model the label correlations for multi-label classification. Guo et al. [11] proposed a graphical model to represent dependencies between multiple labels. In a more recent study [27], Wang et al. proposed a deep convolutional neural network combined with the recurrent neural network (CNN-RNN) to model the dependencies between labels for multi-label classification.

Apart from tag prediction, captioning of an image has also made significant progress in recent years [1, 19, 2, 26]. In one of the recent work [14] Justin et al. proposed a model to localize as well as describe regions of an image. These existing methods for tag and caption prediction make use of the content of the image while ignoring the context in which the image was captured. Although, the works in [23] and [13] have explored the role of context in tag prediction, the methods are focused on single label prediction [23] and used relatively smaller dataset (NUS-WIDE) with lab curated small set of labels [13]. In this work, we explore the role of context for user tag prediction on a much bigger real-world dataset (YFCC100M) with a large number of user tags. We propose the use of a unified network for joint modeling of context along with image content and corresponding labels and demonstrate that context plays an important role in how a user assigns tags to their images.

3. PROPOSED METHOD

3.1 Model

We propose a novel context-aware CNN-NN framework for multi-label classification problem. The network consists of two components, CNN part, and NN part. The CNN part extracts visual information from image pixels and the NN part models the user-context. We adopted AlexNet [15] for convolutional part of the network and integrated the fully-connected (fc7) layer with the ContextNet for performing multi-label classification. An overview of the proposed model is shown in Figure 2.

The YFCC100M dataset [25] provides associated meta-data for each of the images which include time of capture and the geo-location where the image was captured. We utilize this meta information to represent user-context which indicates where and when the corresponding image was captured. We form a 6-dimensional feature vector using geo-coordinates (latitude with a range of -90.0 to 90.0 and longitude with a range of -180 to 180) and time (time in minutes with a range of 0 to 1439, weekday with a range 0-6, month with a range 1-12 and day with a range 1-31). The training data was preprocessed to ensure zero mean and unit variance before feeding into the network. The same scaling was performed on the test and validation set before predicting image tags.

The standardized context descriptor is fed into a 2-layered neural network with fully connected layers. We employ rectified linear units

(ReLU) ($f(x) = max(0, x)$) as activation functions for these layers and each of the layer comprises of 512 neurons. The output of this network (fcc2) is integrated with the last layer of the convolutional network (fc7) to perform multi-label classification. The output of the convolutional network (fc7) and context network (fcc2) are concatenated together and the label scores are computed using a fully connected layer. Finally, softmax normalization is performed on the scores to predict the label probability.

3.2 Training

The first 7 layers of the CNN module [15] in our network are pre-trained on ImageNet 2012 classification challenge dataset [4] using Caffe deep learning framework [12]. We learn the parameters for rest of the layers (while fine-tuning the pre-trained CNN parameters) by optimizing the sigmoid cross-entropy loss for the final layer of the network on a subset of YFCC100M [25] training dataset. A label corresponding to an image i is represented as a 1540-dimensional vector p_i, where each element of p_i indicates the presence or absence of the corresponding tag.

$$p_i = (x_1, x_2, ..., x_{1540}), x_i \in [0, 1] \tag{1}$$

With sigmoid cross-entropy loss, the network is trained by optimizing the following loss objective,

$$L_e = -\frac{1}{N}\sum_{i=1}^{N}[p_i log(\hat{p}_i) + (1 - p_i)log(1 - \hat{p}_i)] \tag{2}$$

where, $\hat{p}_i$ is obtained by applying sigmoid function ($f(x) = \frac{1}{1+e^{-x}}$) to each output (label scores) of the final layer of the network.

We employ stochastic gradient descent for optimizing the loss objective for the network. In our experiments, we use a dropout rate of 0.5 for all the projection layers to avoid over-fitting [22]. The momentum (μ) for the update in gradient descent was set to 0.9, the factor for dropping the learning rate (γ) was set to 0.1 and the weight decay was set to 0.0005 for regularization. We employ a global learning rate (α) of 0.001 for the pre-trained layers and a learning rate of 0.01 for the new layers for a faster learning. Both the learning rates are dropped by a factor of γ after every 100K iterations with a batch size of 64 (3.2 epochs with $\sim$ 2M images).

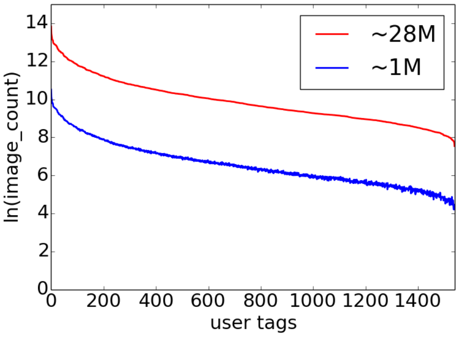
4. EXPERIMENTS

We perform our experiments on a dataset provided by the organizers of the Grand Challenge which is exclusively drawn from Yahoo-Flickr Creative Commons 100M (YFCC100M) [25]. We trained two separate models, one without context and the other integrated with context. We compared these two models for exploring the role of context in user tag prediction. We have also compared our results with other participants in the Grand Challenge. The evaluation demonstrates that the proposed context-aware model outperforms other methods in all the evaluation metrics.

4.1 Dataset

One of the major benefits of YFCC100M dataset is its volume and associated meta information available with the images along with the pixel values. For the tag prediction task, we were provided with a list 1540 tags which were extracted from YFCC100M dataset after some preprocessing by the organizers. The complete dataset was split into train and test set based on the Flickr user identifier (NSID). Only those images with at least one tag present in the master list of 1540 tags were selected for the training set. The training data contains a total of 28,157,620 images and the test set provided for evaluation consists of 46,700 images.

We divided this training dataset of around 28 million images into smaller batches of 1 million images. The order of the images was randomly shuffled before creating the batches. Out of these 29 batches we use 2 random batches with around ~2M images for training our model. The last batch of around 157K images was used as a validation set.



Dataset	Max	Min	Mean
∼28M	1017156	1881	40628
∼1M	36554	67	1445
Top frequently used tags	travel, nature, wedding, vacation, beach, people, art, architecture, summer, party		

Table 1: Distribution of the number of images for the master tag list in ∼28M dataset and a randomly sampled ∼1M dataset. The plot shows the natural log of number-of-images per user tag against the tag list sorted in decreasing order. In the table, we have the maximum, minimum and average number of photos per tag. The last row shows the top most tags present in the dataset.

fog,mist,**morning**,rural, wide+angle	sea,silhouette,sky, **sunset**,twilight
sunrise,landscape,**morning**, clouds,sky	**sunset**,sea,sky, travel,clouds
panorama,landscape,lake, nature,panoramic	**sunset,sunrise**,sea, ocean,clouds

Table 3: Images from validation set with actual and predicted tags using context-free (third row) and context-aware model (second row) along with the ground truth (first row).

4.2 Evaluation Metric

The proposed method is evaluated based on precision, recall and accuracy measures for the predicted tags on the validation set as well as on the test set provide by the organizers. We predict top-k tags for each of the test images and compared with the ground truth. The precision@k is computed as the proportion of top k generated tags that appear in the user tags, recall@k is computed as the proportion of the user tags that appear in the top k generated tags, and accuracy@k is computed as 1 if at least one of the top k generated tags is present in the user tags and 0 otherwise.

4.3 Results and Analysis

We evaluate both context-free and context-aware model on the test set provided by the organizers for comparison. A detailed comparison is shown in Table 4. The context-free model when trained on ∼2M images achieved an accuracy of 0.65, precision of 0.21 and a recall of 0.18 for k=5. The context-aware model, on the other hand, achieved an increased accuracy of 0.71, precision of 0.25 and a recall of 0.20.

We also evaluated the proposed model on a validation set drawn from the training data. The validation set has a total of 157620 im-

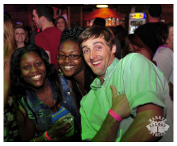
					
beach,ocean,sand, sun,vacation	club,football,soccer, stadium,wales	aircraft,airplane,airport, aviation,city,travel	beach,fort,lighthouse, seaside,travel,water	bar,club,food, fun,party	demonstration,march, protest,street
beach,sand,ocean, water,shore	football,wales,soccer, stadium,sport	airplane,plane, aircraft,jet,aviation	beach,water,travel, vacation,nature	bar,party,club, nightlife,fun	protest,black+and+white, demonstration,people,rally

Table 2: Images from validation set with actual (first row) and predicted tags (second row) using context-aware model.

Method	**Precision@5**	**Recall@5**	**Accuracy@5**
Baseline1	0.04	0.03	0.18
Baseline2	0.24	**0.20**	**0.72**
SB	0.23	0.19	0.69
P-CF	0.21	0.18	0.65
P-CA	**0.25**	**0.20**	0.71

Table 4: Comparison of evaluation results on test set provided by the organizers. Baseline1 is obtained by assigning the overall most frequent tags to every photo. Baseline2 is a convolutional neural network which is trained on around 5M images from the training dataset. SB is the second best results obtained by other participants in the Grand Challenge. P-CF (context-free) is our proposed method without context when trained for ∼2M images (approximately 7 % os the training set) and P-CA is the proposed context-aware model trained on ∼2M images.

K	**Precision@k**		**Recall@k**		**Accuracy@k**	
	P-CF	P-CA	P-CF	P-CA	P-CF	P-CA
1	0.39	0.43	0.05	0.07	0.39	0.43
3	0.32	0.35	0.12	0.16	0.60	0.65
5	0.28	0.31	0.18	0.23	0.69	0.73
10	0.22	0.23	0.27	0.33	0.79	0.83

Table 5: Evaluation results on the validation set as described in section 4.1 for different values of k obtained using proposed context-free (P-CF) and context-aware (P-CA) model.

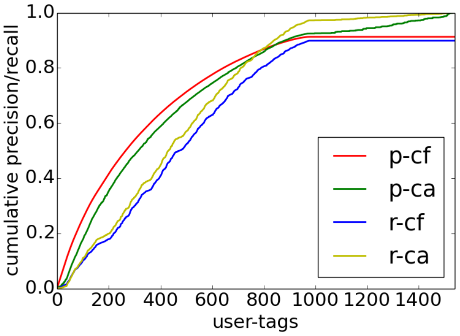


Figure 3: Plot to demonstrate that integrating context has varying effect on the prediction accuracy for different user-tags. (p-cf: cumulative precision for context-free, p-ca: cumulative precision for context-aware, r-cf: cumulative recall for context-free and, r-ca: cumulative recall for context-aware)

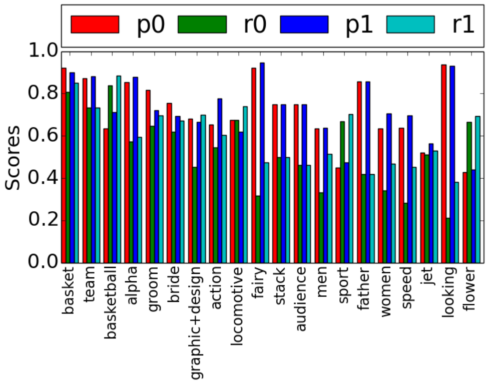


Figure 4: Precision@5 and recall@5 scores for the top predicted user-tags. (p0/r0: precision/recall without context and p1/r1: precision/recall with context.)

ages and as per the organizer's guidelines, images with at least 5 tags were selected for testing. Based on this, we have a total of 14337 images for evaluation. We computed precision@k, recall@k and accuracy@k for k=(1,3,5,10). The results are shown in table 5. We observe that the context-aware model performs better than the context-free model for all values of k. Table 2 and table 3 shows some of the sample results obtained using the proposed method. We can observe that the tags predicted using context-aware model are more meaningful as compared to the context-free model (table 3).

We computed per-tag precision/recall scores for the images in the validation set. Figure 4 shows the scores corresponding to the top-predicted tags based on their F1 score. We can observe that for most of the tags the context-aware model perform better than the context-free model. We also observe that the context-aware model does not improve the scores for all the tags equivalently. To analyze this further, we plot the cumulative prediction/recall (@5) curve for the top predicted tags arranged in decreasing order of precision. We observe that the cumulative gain in precision and recall curves for context-aware model crosses with the context-free model showing that the gain in precision and recall is not consistent across various tags. We plan to explore the context-awareness of tags in our future work.

5. CONCLUSION AND FUTURE WORK

In this work, we proposed a context-aware model which can predict multiple tags for an image which can be recommended to a user. The network can be trained end-to-end for the image content as well as the context in which the image was captured. We adopted the AlexNet convolutional model [15] for image content and combine it with another network for context to perform multi-label classification. The experimental results showed that integrating context significantly improves the tag prediction.

In the current research, we focused on exploring the role of context in image tagging and showed that the context can be an important factor in predicting how user tags their images. In future work, we will explore the role of user-context in image captioning as context also plays an important role in how user caption their images. We plan to integrate the proposed context-aware network with recurrent neural networks which are known to perform well for caption prediction tasks [14]. In addition, we also plan to explore the role of user-context in understanding the correlation between multiple tags assigned to an image by the user.

6. ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centre in Singapore Funding Initiative.

7. REFERENCES

[1] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

[2] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, page 48. ACM, 2009.

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

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

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[7] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200. ACM, 2005.

[8] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

[9] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

[10] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 309–316, Sept 2009.

[11] Y. Guo and S. Gu. Multi-label classification using conditional dependency networks. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1300, 2011.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] J. Johnson, L. Ballan, and L. Fei-Fei. Love thy neighbors: Image annotation by exploiting image metadata. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4624–4632, 2015.

[14] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] X. Li, F. Zhao, and Y. Guo. Multi-label image classification with a probabilistic label enhancement model. *Proc. Uncertainty in Artificial Intell*, 2014.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[18] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 316–329, Berlin, Heidelberg, 2008. Springer-Verlag.

[19] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

[20] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.

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

[22] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[23] Y.-C. Su, T.-H. Chiu, G.-L. Wu, C.-Y. Yeh, F. Wu, and W. Hsu. Flickr-tag prediction using multi-modal fusion and meta information. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 353–356. ACM, 2013.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[25] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, Jan. 2016.

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[27] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. *arXiv preprint arXiv:1604.04573*, 2016.

[28] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014.

[29] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, 81(1):21–35, 2010.

[30] J. Weston, S. Bengio, and N. Usunier. Wsabie: scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 2764–2770. AAAI Press, 2011.

[31] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu. Correlative multi-label multi-instance image annotation. In *2011 International Conference on Computer Vision*, pages 651–658. IEEE, 2011.