arXiv:1605.05054v1 [cs.CV] 17 May 2016

HARRISON: A Benchmark on HAshtag Recommendation for Real-world Images in SOcial Networks

Minseok Park
pms0209@kaist.ac.kr

Hanxiang Li
hanxiangli@kaist.ac.kr

Junmo Kim
junmo@ee.kaist.ac.kr

School of Electrical Engineering
KAIST
South Korea

Abstract

Simple, short, and compact hashtags cover a wide range of information on social networks. Although many works in the field of natural language processing (NLP) have demonstrated the importance of hashtag recommendation, hashtag recommendation for images has barely been studied. In this paper, we introduce the HARRISON dataset, a benchmark on hashtag recommendation for real world images in social networks. The HARRISON dataset is a realistic dataset, composed of 57,383 photos from Instagram and an average of 4.5 associated hashtags for each photo. To evaluate our dataset, we design a baseline framework consisting of visual feature extractor based on convolutional neural network (CNN) and multi-label classifier based on neural network. Based on this framework, two single feature-based models, object-based and scene-based model, and an integrated model of them are evaluated on the HARRISON dataset. Our dataset shows that hashtag recommendation task requires a wide and contextual understanding of the situation conveyed in the image. As far as we know, this work is the first vision-only attempt at hashtag recommendation for real world images in social networks. We expect this benchmark to accelerate the advancement of hashtag recommendation.

1 Introduction

A hashtag is defined as any word attached to the prefix character '#' that is used in online social network services (SNS) such as Facebook, Twitter, and Instagram. With the growth of online social networks, hashtags are commonly used to summarize the content of a user's post and attract the attention of followers. In Instagram, for example, simple hashtags such as *#dog* and *#beach* describe simple objects or locations in a photo. Emotional hashtags such as *#happy* express a user's feelings, abstract hashtags such as *#fashion* and *#spring* categorize topics, and inferential hashtags such as *#colourful* and *#busy* represent situational or contextual information. There are even advertising hashtags such as *#likeforlike*, which are not related to the photo's content. Figure 1 demonstrates examples of posted images

Dataset is available at https://github.com/minstone/HARRISON-Dataset

#food #foodporn #foods #foodpics #foodie #foodgasm #instafood #foodpic #yummy #yum #amazing #instagood #photooftheday #hot #lunch #breakfast #fresh #tasty #delish #delicious #eating #eat #hungry #makeansampaislim #tagsforlikes #like4like

#fashionable #fashion #fashionblog #fashionista #fashionpost #blogger #fashionblogger #beautiful #matching #girl #gorgeous #goals #beauty #photoofteday #instapic #style #stylish #streetstyle #outfit #ootd #inspo #webstagram #lookdodia

#relax #friends #afternoon #goodtimes #goodvibes

#doggy #dog #cockerspaniel #bella #instagood #love #ariel #instaday
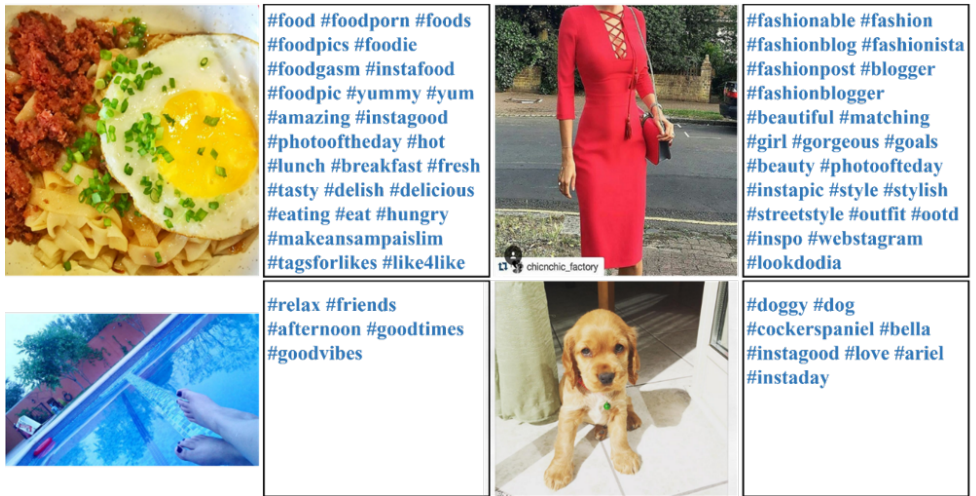
Figure 1: Examples of Instagram photos and their hashtags

and hashtags on Instagram. Considering their wide variety, the recommendation of proper hashtags is a highly interesting and useful task in the age of social media.

Hashtag recommendation for Twitter text posts has been actively studied in the field of natural language processing (NLP). Previous works focused on the content similarity of tweets [11, 22] and unsupervised topic modeling with Latent Dirichlet Allocation (LDA) [4, 10, 16]. Many other approaches [6, 9, 12] have also been studied and deep learning approaches [21] have recently been adopted. Although the importance of hashtag recommendation has been proved by many works in NLP task, hashtag recommendation for images has barely been studied in the field of image understanding. Recently, the study of [3] is presented for the hashtag recommendation systems, but they made use of image data with additional metadata of user information, such as gender and age. As far as we know, our work is the first vision-only attempt to recommend hashtags for real world images in social networks.

In the field of computer vision, studies on image understanding and visual analysis have exploded, with various works attempting to challenge such topics as object classification [5, 17, 18], object detection [14], scene classification [23], action recognition [19], image captioning [8], and even visual question answering [1]. In particularly, image annotation task [13, 20] shares similarity with hashtag recommendation task in regard to the diversity of labels. The labels of annotations mostly consist of surface information such as objects in image and locations of image, but hashtags include the inferential words, which require contextual understanding of images, and trendy words in social networks as well as surface information. For these reasons, hashtag recommendation for images is very challenging and is a worthy topic of research in the field of image understanding.

In this paper, we introduce the novel benchmark for image hashtag recommendation, titled HARRISON, or the **HA**shtag **R**ecommendation for **R**eal world **I**mages in **SO**cial **N**etworks. The HARRISON dataset is a realistic dataset, which provides real posted images with their associated hashtags in online social network, Instagram. We also construct the baseline models based on convolutional neural networks (CNN), and then experiment and evaluate the performance of our baselines on the HARRISON dataset. Many previous
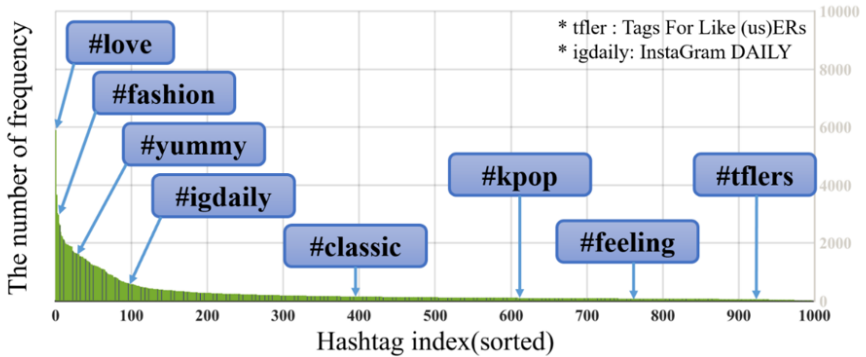
Figure 2: Histogram of hashtags in the HARRISON dataset

works have been accelerated by the release of benchmark datasets. For example, annual challenges, called ILSVRC [15], ensure that object classification task surpasses human level performance. The HARRISON dataset is expected to accelerate the research and development of hashtag recommendation task.

2 Constructing the HARRISON Dataset

2.1 Collecting Images and Hashtags

It is important for hashtag recommendation systems to suggest hashtags that are frequently used in the real world. This means that we need to collect Instagram photos based on the most popular hashtags. A hashtag ranking website[1] is consulted and 50 hashtags are manually selected out of top 100 based on the meaning of hashtags.

Next, we collect a set of Instagram photos for each selected hashtag. Through the platform's public APIs (Application Program Interfaces) in a social network analysis website[2], photos and their associated hashtags of public posts are collected. For each selected hashtag, list of recent Instagram photos tagged with it and associated hashtags for each photo is gathered. Removing repeated images from the lists, a total of 91,707 images are obtained, with an average of 15.5 associated hashtags per image. The total number of hashtags in this collected dataset is approximately 1.4 million from 228,200 kinds of words.

2.2 Organizing the Dataset

Since there are no rules for making and tagging hashtags, they can be diversely generated and freely used. For example, there are simple cognate hashtags in both the singular and plural form (e.g. *#girl*, *#girls*), hashtags in the lower and upper case (e.g. *#LOVE*, *#love*), hashtags in various forms of the same root word (e.g. *#fun*, *#funny*), sentence-like hashtags (e.g. *#iwantthissomuch*, *#kissme*), slang-inspired hashtags (e.g. *#lol*), and meaningless hashtags to gain the attention of followers (e.g. #like4like, #followforfollow). Moreover, users can repeatedly tag the same hashtag for emphasis.

---

[1] http://top-hashtags.com/instagram
[2] https://netlytic.org/index.php

Figure 3: Overview of images and hashtags in the HARRISON dataset

To construct a high quality dataset, post-processing on hashtags are performed. First, hashtags containing non-alphabetic characters, e.g. Chinese and Korean, are rejected. Next, lemmatization [8] is applied to all hashtags, which is the process of grouping the different inflect forms of a word. For example, *#walked*, *#walks*, and *#walking* have the same base form *#walk*. Based on lemmatization process, repeated hashtags associated with the same image are removed.

Through the above processes, approximately 165,000 unique hashtags are collected. Observing the frequency of these hashtags, the top 1,000 hashtags such as *#love* and *#friend* form 59% of the total hashtags and the top 5,000 hashtags account for 74.6% of the total. Since the rest of hashtags which appear less than 22 times are unused (e.g. *#mancrush-sunday*) or out of style (e.g. *#sightseeing*), we choose only the 1,000 most frequently used hastags as classes of the dataset and the others are eliminated from the dataset. Finally, we check the number of associated hashtags per image and exclude images with either no hashtag or more than 10 hashtags since photos with too many hashtags are likely to be commercial posts. In this manner, the HARRISON dataset is organized with selected images and corresponding stemmed hashtags from 1,000 classes.

2.3 Description of the HARRISON Dataset

Figure 3 shows overall distribution of the HARRISON dataset. The HARRISON dataset has a total of 57,383 images and approximately 260,000 hashtags. Each image has an average of 4.5 associated hashtags (minimum 1 and maximum 10 associated hashtags) due to the removal of infrequently used hashtags. The ground truth hashstags for each image are made up of the 1,000 most frequently used hashtags, encoded with numbers based on frequency ranking results, as shown in Figure 4. For evaluation, we randomly split the dataset into 52,383 data for training and 5,000 data for test.

illustration (318)	art (45)	architecture (181)	ootd (94)
drawing (137)	coffee (135)	hollywood (698)	son (502)
boyfriend (30)	tired (13)	vintage (130)	boy (53)
		design (98)	instagood (17)
		losangeles (397)	love (0)
		california (259)	
		yellow (4)	

Figure 4: Examples of the HARRISON dataset, consisting of images and hashtags with encoded class numbers

3 Baseline Methods

We consider the hashtag recommendation task as a multi-label classification problems. The baseline algorithm consists of two main steps, the visual feature extractor and the multi-label classifier, as shown in Figure 5. In the visual feature extractor, two kinds of visual features are extracted from the input image. Next, the extracted features are used as the inputs of a trained multi-label classifier, and the score of each class of hashtag is obtained. The scores are sorted, and the top ranked hashtags are recommended for the input images. Details of each step are presented below.

In the visual feature extractor stage, high-level feature representation and the diversity of visual information are required. To learn the deep hierarchical features, the VGG-16 layers model [17] is used as the feature extractor. In addition, we extract two different visual features to obtain the diversity of visual information. Since object and scene task are well-organized on the large-scale dataset, we select object-based and scene-based features as a representative visual features. We train our feature extractor on two different datasets. One is trained on 1.2 million images of the ImageNet dataset [15] and the other is trained on 2.5 million images of the Places Database [23] with 205 scene categories. According to [23], a simple visualization of the receptive fields for CNN units of two models shows that object-based CNN and scene-based CNN differ in their internal representations. The visual features extracted from the final fully connected layer of each model have 4,096 dimensions, denoted by *VGG-Object* and *VGG-Scene*, respectively.

Next, a multi-label classifier with two fully-connected layers and a sigmoid cross entropy layer is constructed. The concatenation of the previous two visual features ($N = 4,096 \times 2$) is used as the input of the classifier, and each fully connected layer of classifier have 4,096 neurons. The output of this multi-label classifier is the confidence scores of the 1,000 hashtag classes. To recommend the K number of hashtags, we sort the scores and extract the top K classes.

We first construct single feature-based models, which use only *VGG-Object* features or only *VGG-Scene* features as the input of classifier. In these models, unused features are
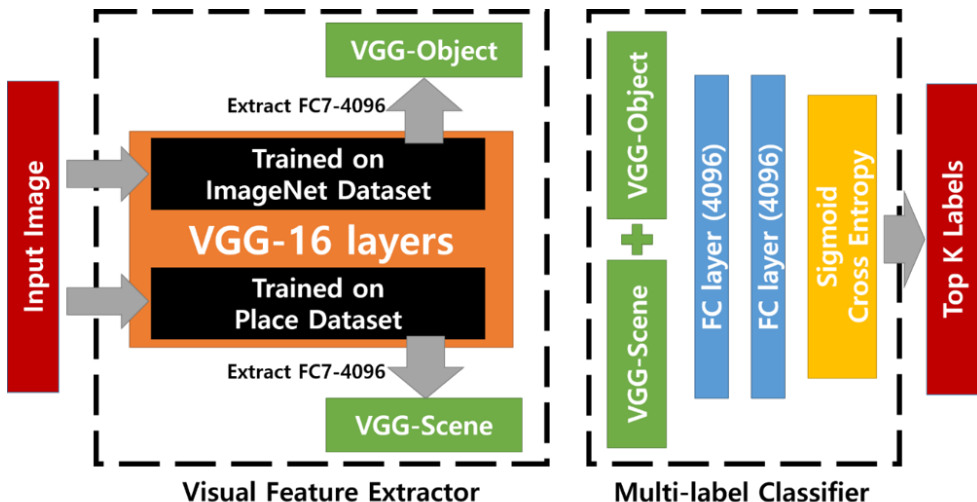
Figure 5: Overall flow of baseline methods

zero-padded in the input layer of classifier. For comparison with these single feature-based models, the integrated model, *VGG-Object + VGG-Scene*, is constructed using two visual features. In training phase, the feature extractor and the classifier are trained separately.

4 Experiments

We evaluated the baseline approaches described in Section 3 on our HARRISON dataset. The experiment was performed using the open source Caffe library [7] on a single NVIDIA GeForce Titan X. In the visual feature extractor, we use pre-trained VGG-16 layers model on ImageNet dataset [15] and Place Database [23]. Using two visual features, we trained our multi-label classifier for 100,000 iterations with 100 samples per mini-batch. The learning rate is initially set to 10^{-1}, and then is decreased by a factor of 10 every 30,000 iterations.

4.1 Evaluation Measures

In previous works on tag recommendation [4, 21] and image annotation [13, 20], two performance measures have been mainly used. The first one is precision, which shows how well hashtags are predicted, and the other one is recall how well recommendation system covers the wide range of hashtags. Only precision measure is risky since reasonable hashtags can be missed in real hashtags. In particularly, precision at 1 and recall at 10 is commonly used. The detail explanations of parameters are explained below. In this paper, we select three measures for the fair evaluation of hashtag recommendation, similar to [3], Precision and recall are primarily used, and accuracy is also evaluated to check how many mistakes this system makes.

For each image, *Precision@K* is defined as the portion of hashtags in the top K ranked hashtags which match with the ground truth hashtags. *Recall@K* is defined as the portion of hashtags in the ground truth hashtags which match with the top K ranked hashtags.

Predicted Hashtag	GroundTruth Hashtag
nike	adidas
adidas	nike
shoe	onlineshop
fashion	sepatumurah
sneaker	ootd
	fashion
	makeup

Predicted Hashtag	GroundTruth Hashtag
hair	hairstyle
red	haircolor
love	redhair
hairstyle	red
hairstylist	Inspiration
	fashionblog
	fashionblogger

Predicted Hashtag	GroundTruth Hashtag
yummy	cooking
instafood	food
food	yummy
home	cleaneating
green	

Predicted Hashtag	GroundTruth Hashtag
tree	spring
spring	cloud
nature	blossom
school	tree
pink	cherryblossom
	pretty

Figure 6: Examples of successful hashtag recommendation results with the baseline model *VGG-Object + VGG-Scene*. Matched hashtags between the prediction and ground truth are indicated by red colour.

Table 1: Evaluation results of the baseline methods with average precision@1, average recall@5, and average accuracy@5.

Baseline Methods	Precision@1	Recall@5	Accuracy@5
VGG-Object	28.30 %	20.83 %	50.70 %
VGG-Scene	25.34 %	18.66 %	46.30 %
VGG-Object + VGG-Scene	30.16 %	21.38 %	52.52 %

Accuracy@K is defined as 1 if at least one match between the top K ranked hashtags and the ground truth hashtags exists. Equations of three measures are shown as follows:

$$Precision@K = \frac{|Result(K) \cap GT|}{|Result(K)|}$$

$$Recall@K = \frac{|Result(K) \cap GT|}{|GT|}$$

$$Accuracy@K = \begin{cases} 1 & \text{if } Result(K) \cap GT \neq \emptyset \\ 0 & \text{if } Result(K) \cap GT = \emptyset \end{cases}$$

where $Result(K)$ corresponds to a set of the top K hashtags in predicted results and GT corresponds to a set of the ground truth hashtags. In our experiments, we set K to 1 for precision and 5 for recall and accuracy considering the average number of associated hashtags per image in the HARRISON dataset.

4.2 Results

We evaluated the baseline results on *Precision@1*, *Recall@5*, and *Accuracy@5* by averaging over all images in the test set. Table 1 shows the evaluation results for our baseline
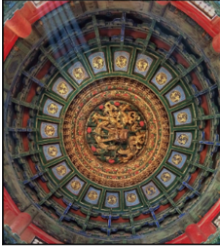
	Predicted Hashtag	GroundTruth Hashtag			Predicted Hashtag	GroundTruth Hashtag
	truth	architecture			school	sneaker
	music	interiordesign			home	streetwear
	goodtimes	style			tired	kick
	makeup	art			family	fashion
	beauty	food			travel	style
		cake				shoe
						new
	Predicted Hashtag	GroundTruth Hashtag			Predicted Hashtag	GroundTruth Hashtag
	old	colorful			truth	kobe
	red	park			beach	lakers
	explore	traveling			love	nba
	love				friday	nike
	adventure				repost	

[722] Los Angeles Lakers +4½ Points -115 for the Game (Buy +½) Risking 2,300.00 T 2,000.00

Figure 7: Examples of failure cases for the baseline model.

	Predicted Hashtag	GroundTruth Hashtag			Predicted Hashtag	GroundTruth Hashtag
	truth	depressed			food	work
	love	quote			yummy	tired
	quote	love			foodporn	
	life	loveher			yum	
	boyfriend	lovehim			lunch	
		loveyou				

"I get way too sensitive when I get attached to someone. I can detect the slightest change in the tone of their voice, and suddenly I'm spending all day trying to figure out what I did wrong."
— me. (via itcuddles)

Figure 8: Examples of challenging cases and their hashtag recommendation results by the baseline model.

methods on the HARRISON dataset. Compared with [4], all three measures in our results were distributed higher, since the number of hashtag classes in the HARRISON dataset is 10 times smaller while the average number of associated hashtags per image is about 2 times greater.

As shown in Table 1, the integrated *VGG-Object + VGG-Scene* model achieved the best performance among the baseline models, with the average precision of 30.16 %, the average recall of 21.38 %, and the average accuracy of 52.52 %. Due to the gap in visual information, the results of other models which use a single visual feature were slightly lower. They also show that object-based features are closer to the information available in hashtags than scene-based features.

Figure 6 - 8 demonstrate the examples of baseline results for the HARRISON dataset in various cases. Simple non-inferential labels, such as objects and colours, are well detected by our baseline model. For example, in Figure 6, #hair, #food, #shoe, #red, #green, #pink are detected. Among these labels, fine-grained classes and non-salient objects are still difficult problems. For example, in Figure 7, #kobe is too specific and #shoe is too imperceptible to detect. Even if these problems can be solved by increasing training data, inferential hashtags which are not directly extracted from images are highly challenging problems. For example,

we should inference *#colourful* from the various colours, *#depressed* from the content of the quoted text in image, and *#tired* from the many same dishes which user maybe prepared for a long time, as shown in Figure 8.

As we mentioned above, success cases of our baseline algorithms are mostly simple non-inferential labels. These observations are supported by the performances of our baseline methods, which are relatively high in *Precision*@1 and *Accuracy*@5 and relatively low in *Recall*@5. This facts show that two visual features are outstanding for the simple non-inferential labels, but insufficient to cover the information available in hashtags. Also, our baseline models ignore the dependencies between hashtags since we considered hashtags as independent labels to use multi-label classifier. Combining with such techniques as word similarity in NLP task could be an option to improve our baseline models.

In respect of the HARRISON dataset, our result show that hashtag recommendation task is highly challenging due to the contextual understanding. Further attempt such as multiple instance detection, fine-grained classification, and even text recognition will be helpful to understand contextual information and inference the user's intention.

5 Conclusions

In this paper, we introduced the HARRISON dataset, a benchmark dataset for hashtag recommendation of real world images in social networks. To evaluate our dataset, we constructed a baseline framework with CNN-based visual feature extractor and multi-label classifier. After applying two visual features, object-based features and scene-based features, we evaluated our baseline models on three evaluation measures. Associated with this dataset, we outlined challenging issues of hashtag recommendation systems: understanding wide range of visual information, using dependencies between hashtag classes, and understanding contextual information. As far as we know, this work presents the first vision-only attempt to recommend hashtags from images. We expect this benchmark dataset to aid the development of hashtag recommendation systems.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[3] Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1731–1740, 2015.

[4] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 593–596, 2013.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

[6] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 531–538, 2008.

[7] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[8] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[9] Elham Khabiri, James Caverlee, and Krishna Y Kamath. Predicting semantic annotations on the real-time web. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 219–228, 2012.

[10] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 61–68, 2009.

[11] Tianxi Li, Yu Wu, and Yu Zhang. Twitter hash tag prediction algorithm. In *ICOMPâĂŹ11-The 2011 International Conference on Internet Computing*, 2011.

[12] Allie Mazzia and James Juett. Suggesting hashtags on twitter. *EECS 545m, Machine Learning, Computer Science and Engineering, University of Michigan*, 2009.

[13] Venkatesh N Murthy, Subhransu Maji, and R Manmatha. Automatic image annotation using deep learning representations. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 603–606, 2015.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

[16] Jieying She and Lei Chen. Tomoha: Topic model-based hashtag recommendation on twitter. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pages 371–372, 2014.

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

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

[19] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[20] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *International Joint Conference on Artificial Intelligence*, volume 11, pages 2764–2770, 2011.

[21] Jason Weston, Sumit Chopra, and Keith Adams. # tagspace: Semantic embeddings from hashtags. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2014.

[22] Eva Zangerle, Wolfgang Gassler, and Gunther Specht. Recommending #-tags in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings*, volume 730, pages 67–78, 2011.

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