Image Tag Recommendation by Exploiting Social-Network Metadata

Amin Salehi, Avinash Reddy Kaitha

Abstract—With the unprecedented surge of images shared on social media platforms, users are overloaded with many choices which can result in poor decision making. Social tagging alleviates this problem by enabling users to categorize and also retrieve images by using social tags. However, the success of social tagging depends on the participation of users. To facilitate the process of tagging images for users in order to increase their participation, many image tag recommender systems have been proposed. These systems face one or both of the following challenges. First, they do not capture the most representative features reflecting the visual content of images. Second, they do not efficiently exploit social-network metadata particularly social interactions between users. Recently, deep learning has achieved great success in image-related tasks (e.g., image classification) thanks to its capability to capture salient features. Moreover, according to social influence theory, users interacting with each other can change one another's attitude and behavior. Motivated by this social science finding, we also demonstrate that social interactions between users in social media result in their similar tagging behavior. In this project, we propose a deep learning framework which exploits the visual content of images and social interactions to recommend the most relevant tags to users. Through experimental results, we aim to demonstrate the efficacy of our proposed framework.

Index Terms—Image Tag Recommendation, Social Tags, Image Tagging, Social Interactions.

1 Introduction

THe number of images shared on social media platforms is enormously growing in recent years. As a result, users are facing with the challenge of finding the relevant images from a huge number of available ones. Fortunately, many social media platforms enable users to tag their images. User-generated tags provide insight to the image categories important to users. Therefore, users can utilize such categories to find the relevant images. Moreover, image labeling, a time consuming and costly task in image retrieval, can be replaced by social tagging.

The success of social tagging depends on the participation of users. However, some users might not tag many images because (1) they might not come up with relevant tags for such images, and (2) they might find this process tedious. Many image tag recommender systems have been developed in order to help users to assign relevant tags to images. Therefore, these systems help the searchability of images to increase, as the number of tagged images grows.

Although a large body of work has been proposed to tackle image tag recommendation problem, the majority of them merely focuses on the visual content of images in order to recommend tags to users. Recently, many efforts [18], [23], [24] have demonstrated that deep learning performs very well in the context of extracting the most relevant features for images. However, the majority of tag image recommender systems do not exploit deep learning to extract the most representative features of images. Therefore, deep learning can be used to enhance image tag recommender systems to suggest more relevant tags for users' images according to the correspondence between the tags and the visual content of images.

Social interactions between users are one of the most important sources of information which can be found in social media because they result in social influence between users. According to social influence theory [1], [2], social influence can be defined as a change in an individual's attitudes and behavior resulting from social interactions with other individuals. Social interactions have been used for a broad range of applications such as detecting online communities [15], suggesting items to users [16], and online marketing [17]. Moreover, the tagging behavior of users can be impacted by social influence. That means users interacting with each other tend to use the same tags for similar images. Therefore, social interactions have the potential to improve the quality of image tag recommendation.

In this project, we aim to present a deep learning framework to recommend tags for users' images according to the visual content of images and social interactions between users. To capture the most salient features of images, our framework utilizes the deep learning ideas borrowed from the well-know architectures proposed for image classification. Moreover, it also exploits the social interactions between users in order to improve the quality of image tag recommendation by relying on social influence theory.

The rest of the paper is organized as follows. In Section 2, we first review the related work. In Section 3, we formally define the problem of image tag recommendation using social interactions. Our dataset is presented in Section 4. In section 5, we present our framework as well as some analysis demonstrating the usefulness of using social interactions for image tag recommendation. In Section 6, we present a summary of our evaluation methodology. We conclude our proposed framework in section 7. In section 8, we review some of the failed attempted in this project. We finally present the progress of our project in Section 8.

2 Related Work

A large body of work has been devoted to image tag recommendation [3], [4], [5], [6], [7], [8], [9], [20], [21], [22], [25]. For instance, Sun *et.al* [8] utilize the co-occurrence of social tags in order to detect concepts; a concept is defined as a set of highly co-occurred tags. Next, they recommend tags to users based on

TABLE 1
Notations used in the paper.

Notation	Explanation
$\mathtt{U}$	The set of users
$\mathtt{P}$	The set of communities
$\mathtt{T}$	The set of issues
n	The number of users
m	The number of images
k	The number of tags
$\mathbf{X}$	Social interaction matrix
$\mathcal{A}$	Tag assignment tensor

these concepts by using retrieval techniques. Another work by Haifeng *et.al* [7] utilizes the social tags assigned to users' favorite images as well as users' friendships to recommend tags. However, aforementioned methods do not leverage the visual content of images. Several efforts [3], [4], [5] focus on exploiting visual similarity between images, using simple similarity measures, in order to recommend tags. Moreover, Su *et.al* [20] predict image tags by fusing different models of k-NN and Linear SVM as well as image metadata (e.g., upload time). Li *et.al* [25] further learn tag relevance by aggregating votes from visually similar neighboring images. Another work by Panagopoulos *et.al* [9] incorporates not only image similarity but also friendships, geotags, user groups, and historical tags into a tensor decomposition framework to suggest tags. Shah *et.al* [21] also recommend personalized tags to a user by determining a group of users similar to him/her in their tagging behavior. They first find candidate tags from visual content, textual metadata, and tags of neighboring photos. Then, they initialize scores of the candidate tags using asymmetric tag co-occurrence probabilities and normalized scores of tags after neighbor voting, and later perform random walk to promote the tags that have many close neighbors and weaken isolated tags. Furthermore, Rawat *et.al* [6] propose a convolutional neural network which models visual content of images and the context in which images are captured separately. Then, their proposed CNN merges these two models by using a concatenation layer. Another work by Zhang *et.al* [19] captures semantic relationship among tags as well as the visual content of images by fusing the deep multimodal feature representation and cross-modal correlation mining.

3 Problem Statement

We first begin with the introduction of the notations used in the paper as summarized in Table 1. Let $\mathtt{U} = \{u_1, u_2, ..., u_n\}$ be the set of n users, $\mathtt{P} = \{p_1, p_2, ..., p_k\}$ indicate the set of m images, and $\mathtt{T} = \{t_1, t_2, ..., t_k\}$ denote the set of k tags. For matrix indexing, $\mathbf{Z}_{ij}$ is used to indicate the entry at the intersection of the i-th row and j-th column of matrix $\mathbf{Z}$. $\mathbf{G} \in {\mathbb{R}_+}^{n\times n}$ denotes the matrix of users' social interactions, where $\mathbf{G}_{ij}$ implies this is a social interaction between user u_i and user u_j. Furthermore, $\mathcal{A} \in {\mathbb{R}_+}^{n\times k\times m}$ indicates the tag assignment tensor, in which $\mathcal{A}_{ilj}$ corresponds to the number of times that user u_i has assigned tag t_l to item p_i. $\mathcal{M} \in {\mathbb{R}_+}^{m\times r\times c}$ also denotes the image tensor, where $\mathcal{M}_{j::}$ indicates a matrix representing the visual content (pixels) of image p_j.

By using the above notations, the problem of image tag recommendation using social interactions can be defined as: *Given tensors $\mathcal{A}$ and $\mathcal{M}$ as well as social interaction matrix $\mathbf{G}$, we have a multi-label classification problem in which we aim to determine the most relevant tags from $\mathtt{T}$ for a given image p_j of user u_i by using the visual content of image p_j (i.e., matrix $\mathcal{M}_{j::}$) and the social interactions of user u_i (i.e., $\mathbf{G}_{i:}$) .*

TABLE 2
The statistics of the crawled dataset.

Attribute	Count
The number of users	368
The number of edges	17,918
The number of images	36,800
The number of distinct Tags	2648
The number of all Tags	403,654

4 Data Description

Flickr is one the most well-known image sharing platform, in which users not only share images but they also are able to tag them. Moreover, users can follow each other in this platform. These properties make Flickr one of the best sources of information for our project. Therefore, we used Flickr API to crawl images, tags, and social interactions (i.e., followship) of users. However, there is a major challenge when it comes to crawling the social interactions in Filckr. This platform only provides the users that a particular user follows; not his/her followers. This property results in crawling a sparse network of followships between users. To have a denser social network, when we are crawling, we crawl the user who has the most number of connections to the existing crawled users. Furthermore, to have a more range of images, we started with 10 different users (i.e., seeds) who has uploaded different kinds of images. The statistics of our crawled dataset is shown in table 2.

5 The Proposed Framework

In this section, we present our framework which exploits the visual content of images and social interactions in order to recommend tags. Before delving into how we model the problem, in the next subsection, we focus on why social interactions are important and how they can be used in deep learning architectures.

5.1 Exploiting Social Interactions

In this section, we first examine the impact of social influence, resulting from social interactions, on users' tagging behavior using the Flickr dataset. Next, we propose how to exploit social interactions in our framework.

5.1.1 The Impact of Social Influence on Tagging Behavior

In this section, we investigate the impact of social influence on users' tagging behavior. According to social science findings, users interacting with each other can change one another's attitude and behavior [1], [2]. However, the findings borrowed from social sciences do not necessarily hold in social media due to many factors, such as the validity and representativeness of available information [10], [11]. Therefore, we first investigate the correlation between the existence of a social interaction between two users and their similar tagging behavior. In other word, we aim to answer to the following question: *Are the tagging behavior of two users connected in social media similar to each other?*

To answer to this question, we use the following procedure. For each pair of users (u_i, u_j) who have a social interaction between

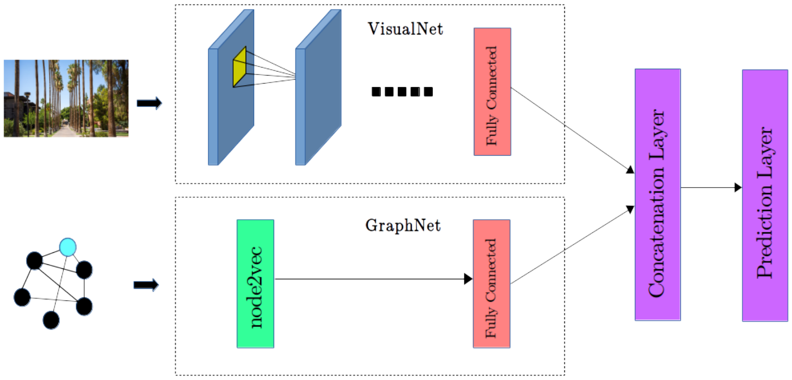

Fig. 1. The proposed framework which has three inputs: (1) an image and (2) the social-network position of the user who has uploaded the image. VisualNet is the component processing the visual content of image , and GraphNet is the component processing the user's position in the social network. Finally, the results from these two components are merged in order to make prediction of relevant tags for the image.

each other (i.e., $\mathbf{G}_{ij} > 0$), we randomly select a user u_k where users u_i and u_k do not have a social interaction between each other (i.e., $\mathbf{G}_{ik} = 0$). Next, for each tag that user u_i has assigned, we check which one of users u_j and u_k have used the tag. If user u_j has used the tag, we $s_p = 1$; otherwise $s_p = 0$. Similarly, if user u_k has used the tag, we $s_r = 1$; otherwise $s_r = 0$. Let vector S_p denote the set of all s_ps for pairs of users who are connected in social media, and let vector S_r denote the set of all s_rs for pairs of users who are not connected.

We conduct a two-sample t-test on S_p and S_r for each dataset. The null hypothesis H_0 and alternative hypothesis H_1 are defined as follows:

$$H_0 : S_p \leq S_r, \quad H_1 : S_p > S_r \tag{1}$$

The null hypothesis is rejected at significance level $a = 0.01$ with p-value of 0. However, the p-value of 0 is too good for our case to be true. We realized that p-value is highly probable to be zero when sample size is very large; our sample size is $2,504,370$. Therefore, we aim to utilize a measure not sensitive to the size of samples. The effect size [28] is such a measure which quantifies the difference between two sets of samples. The effect size is calculated as follows:

$$\frac{\widetilde{S_p} - \widetilde{S_r}}{N} \tag{2}$$

where $\widetilde{S_p}$ and $\widetilde{S_r}$ are the means of S_p and S_r, respectively. Moreover, N is as follows:

$$N = \sqrt{\frac{|S_p|v_p + |S_r|v_r}{|S_p| + |S_r| - 2}} \tag{3}$$

where v_p and v_r are the variance of S_p and S_r, respectively. Moreover, $|S_p|$ and $|S_r|$ indicate the size of S_p and S_r, respectively.

The effect size is 0.36 which indicates that *the probability that two users connected in social network use the same tag is* 61 *percent higher than two users who are not connected.*

5.1.2 Representing Users in the Network

In the previous section, we demonstrated that social interactions between users in social media can result in similar tagging behavior. In other words, the more users are closer to each other in the social network, the more similar their tagging behaviors are. Motivated by this observation, we aim to exploit users' position in the network in order to suggest the most relevant tags. To that end, we exploit network embedding methods. These methods represent users (i.e., nodes) in a low-dimensional space of features maximizing the likelihood of preserving network neighborhoods of users [12]. Several network embedding methods [12], [13], [14] have been proposed. Among these methods, node2vec [12] is demonstrated to have a better performance. Therefore, we use node2vec in order to have a representation of users in which users close to each other in the social network having similar features.

5.2 Our Framework

In this section, we present our framework which consists of three components: (1) VisualNet and (2) GraphNet as shown in Figure 1. VisualNet extracts the visual information from the input image. We adopt AlexNet architecture [19] to construct VisualNet. GraphNet captures the social-network structural information of users who have uploaded the input image. In GraphNet, we use word2vec which capture the feature vector of any given user in the social network. After exacting the feature of a user, GraphNet utilizes a fully connected layer with ReLU activation and Dropout. Finally, we merge the results of last layers of VisualNet and GraphNet by using a concatenation layer in order to make prediction about the labels that fit an image regarding its visual content and the position of its user in social network.

5.2.1 Training

The first 7 layers of VisualNet in our framework is pre-trained on ImageNet 2012 classification challenge dataset [26]. The parameters of remaining layers are learned by minimizing the sigmoid cross-entropy loss objective applied to the final layer of our framework. In other words, we aim to minimize the following loss objective:

$$L = -\frac{1}{m}\sum_{i}^{m}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \tag{4}$$

where, $y = (y_1, y_2, ..., y_k)$, $z_j \in [0, 1]$, in which z_j determines whether a tag is assigned to image p_i or not. Moreover, $\hat{y}_i$ is obtained by applying the sigmoid function ($f(x) = \frac{1}{1+e^{-x}}$) to each output of the final layer in our framework.

To minimize loss function eq. (4), we utilize stochastic gradient descent. We run the framework for 10 epochs and also start

TABLE 3
Experimental results on the test set for different values of k.

K	Method	Accuracy@k	Precision@k	Recall@k
1	Frequent_tags	0.13	0.13	0.02
	ConTagNet	0.16	0.16	0.02
	Multi-label AlexNet	0.32	0.32	**0.04**
	Our framework	**0.41**	**0.41**	**0.04**
3	Frequent_tags	0.15	0.05	0.02
	ConTagNet	0.31	0.12	0.04
	Multi-label AlexNet	0.50	0.26	0.08
	Our framework	**0.58**	**0.32**	**0.10**
5	Frequent tags	0.15	0.03	0.02
	ConTagNet	0.39	0.10	0.05
	Multi-label AlexNet	0.59	0.22	0.11
	Our framework	**0.67**	**0.27**	**0.13**
10	Frequent_tags	0.16	0.02	0.02
	ConTagNet	0.50	0.08	0.07
	Multi-label AlexNet	0.69	0.18	0.16
	Our framework	**0.76**	**0.21**	**0.20**

with the learning rate of 0.01. The momentum for the update in gradient descent is set to 0.9. We also use dropout rate of 0.5 in order to avoid overfitting [27]. Moreover, we set the batch size to be 450.

6 Experiments

6.1 Evaluation Measures

Our framework is evaluated based on precision@k, recall@k and accuracy@k measures. To that end, we first predict top-k tags for each of the test images and compared them with the ground truth. As suggested by [6], we compute the precision@k as the proportion of top k generated tags appearing in the tags assigned to the corresponding image, recall@k as the proportion of the tags assigned to the corresponding image appearing in the top k generated tags, and accuracy as 1 if at least one of the top k generated tags is present in the user tags and 0 otherwise.

6.2 Baselines

We compare our framework with three baselines as follows:

- **Frequent_tags**: This baseline assigns the overall most frequent tags to every image.
- **Multi-label AlexNet**: This baseline uses the first 7 layers of AlexNet with pre-trained weights learned from ImageNet dataset. The only difference between this baseline and AlexNet is that the last layer of the baseline is the same as that of our framework in order to employ AlexNet for multi-label classification.
- **ConTagNet** [6]: This baseline has two components. The first component has the first 7 layers of Alexnet which are pre-trained with ImageNet dataset. The second component utilizes geo-location and time of capture of images as user-context metadata which is the input to two consecutive fully connected layers with ReLU and Dropout. Finally, two components are concatenated in order to predict tags for images.

6.3 Experimental Results

In this section, we compare our framework with the baselines based on three introduced measures. Table 3 shows the detailed comparison. In the table, we have reported the results for different values of k (i.e., 1, 3, 5, and 10). According to the table, we can make the following observations:

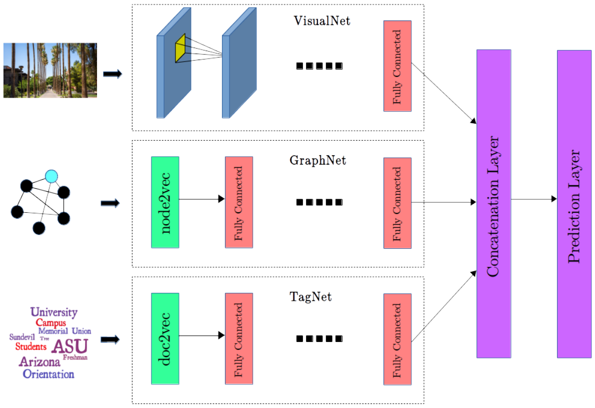

Fig. 2. The failed proposed framework which has three components.

- Our framework outperforms three baselines by a large margin using different values of k based on three introduced measures. For example, for top 5 recommended tags, our framework improves accuracy by 14% compared to the best baseline which is Multi-label AlexNet.
- The main reason that CoTagNet does not performs well is that this network utilizes geo-tags but only 5.6% of images we have crawled have geo-tags.
- The fact that all methods using AlexNet outperforms Freq_tags method by a large margin shows the importance of using deep learning in image tag recommendation task.
- Among all methods using AlexNet, only our framework achieves the best performance. This demonstrate the impact of using social interactions in improving the quality of recommendations.

7 Conclusion

In this work, we proposed a framework to recommend relevant tags for users' images. To achieve this, our framework exploits visual content of images and social interactions between users. In fact, our framework rely on social influence theory and deep learning to achieve this goal. According to social influence theory, individual interacting with each other tend to have similar behavior. Therefore, we utilized a network embedding method (i.e., node2vec) to recommend similar tags to users who have social interactions with each other. Moreover, we also exploit deep learning to capture salient visual features of images to make our tag recommendation more accurate. The experimental results demonstrate the efficacy of our framework compared to three baselines.

8 Failed Experiments

In the previous sections, we mainly mentions the successful ones. However, during our experiments, we tried some other ideas which failed. Some of these failed experiments are as follows:

- **Utilizing users' tags**: We utilized the tags that users has already assigned in a separate component namely TagNet as shown in Figure 2. As shown in the figure, we exploit doc2vec [29] in order to capture vector representations of

TABLE 4
Project progress.

Task	Subtask	Timeline	Responsibility	Progress
Data Acquisition	Phase 1	10 Mar - 31 Mar	Avinash	100%
	Phase 2	1 Apr - 15 Apr	Avinash	100%
Social Network Modeling	Modeling	10 Mar - 05 Apr	Amin	100%
	Implementation	5 Apr - 15 Apr	Amin	100%
Deep Learning	Modeling	1 Apr - 5 Apr	Amin	100%
	Implementation	5 Apr - 19 Apr	Amin and Avinash	100%
Evaluation	Metrics	19 Apr - 21 Apr	Avinash	100%
	Baseline	21 Apr - 23 Apr	Avinash	100%
Final Report	Abstract	1 Apr - 3 Apr	Amin	100%
	Introduction	1 Apr - 3 Apr	Amin	100%
	Related Work	1 Apr - 3 Apr	Avinash	100%
	Problem Statement	1 Apr - 3 Apr	Amin	100%
	Data Description	1 Apr - 3 Apr	Avinash	100%
	Proposed Framework	1 Apr - 3 Apr	Amin	100%
	Evaluation	23 Apr - 25 Apr	Amin	100%

users' tags in order to incorporate them into our framework. However, this idea does not work, and we did not observe a significant improvement. The reason behind this is that GraphNet component of our framework cap capture the knowledge that TagNet component is able to capture. In other words, each user's positions in the network can represent which tags he has used during the training of our framework. Therefore, TagNet does not improve the performance for our framework.

- **Updating the weights of pre-trained AlexNet layers**: We also tried to update the the weights of alexnet pre-trained layers used in our framework especially with lower learning rate than that of other layers. However, it did not result in a significant improvement in the performance.

9 Project Progress

In this section, we summarize the responsibility of our team members for each task in the project. Table 4 shows the details of each task in the project.

References

[1] Turner, John C. Social influence. Thomson Brooks/Cole Publishing Co, 1991.

[2] French, John RP, Bertram Raven, and D. Cartwright. The bases of social power. Classics of organization theory 7 (1959).

[3] Lindstaedt, Stefanie, et al. Automatic image annotation using visual content and folksonomies. Multimedia Tools and Applications 42.1 (2009): 97-113.

[4] Lee, Sihyoung, et al. Map-based image tag recommendation using a visual folksonomy. Pattern Recognition Letters 31.9 (2010): 976-982.

[5] Wu, Lei, et al. Learning to tag. Proceedings of the 18th international conference on World wide web. ACM, 2009.

[6] Rawat, Yogesh Singh, and Mohan S. Kankanhalli. ConTagNet: Exploiting User Context for Image Tag Recommendation. Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016.

[7] Haifeng, Guo, Su Shoubao, and Sun Zhoubao. Image tag recommendation based on friendships. Multimedia Tools and Applications (2016): 1-17.

[8] Sun, Aixin, Sourav S. Bhowmick, and Jun-An Chong. Social image tag recommendation by concept matching. Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011.

[9] Panagopoulos, Michail, and Constantine Kotropoulos. "Image tagging using tensor decomposition." Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on. IEEE, 2015.

[10] Derek Ruths and Jrgen Pfeffer. 2014. "Social media for large studies of behavior." Science 346, 6213 (2014), 10631064.

[11] Tufekci, Zeynep. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." Eighth International AAAI Conference on Weblogs and Social Media. 2014.

[12] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[13] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.

[14] Tang, Jian, et al. "Line: Large-scale information network embedding." Proceedings of the 24th International Conference on World Wide Web. ACM, 2015.

[15] Papadopoulos, Symeon, et al. "Community detection in social media." Data Mining and Knowledge Discovery 24.3 (2012): 515-554.

[16] Ma, Hao, et al. "Recommender systems with social regularization." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[17] Kozinets, Robert V. "The field behind the screen: Using netnography for marketing research in online communities." Journal of marketing research 39.1 (2002): 61-72.

[18] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[19] Zhang, Xingmeng, et al. "Image Tag Recommendation via Deep Cross-Modal Correlation Mining." China National Conference on Chinese Computational Linguistics. Springer International Publishing, 2016.

[20] Su, Yu-Chuan, et al. "Flickr-tag prediction using multi-modal fusion and meta information." Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013.

[21] Shah, Rajiv Ratn, et al. "PROMPT: Personalized User Tag Recommendation for Social Media Photos Leveraging Personal and Social Contexts." Multimedia (ISM), 2016 IEEE International Symposium on. IEEE, 2016.

[22] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

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

[24] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[25] Li, Xirong, Cees GM Snoek, and Marcel Worring. Learning social tag relevance by neighbor voting. IEEE Transactions on Multimedia 11.7 (2009): 1310-1322.

[26] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[27] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." Journal of Machine Learning Research 15.1 (2014): 1929-1958.

[28] Coe, Robert. "It's the effect size, stupid: What effect size is and why it is important." (2002).

[29] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014.