Social Image Tags as a Source of Word Embeddings: A Task-oriented Evaluation

Mika Hasegawa, Tetsunori Kobayashi, Yoshihiko Hayashi
School of Science and Engineering, Waseda University
Waseda-machi 27, Shinjuku, Tokyo 1690042, Japan
mika@pcl.cs.waseda.ac.jp, koba@waseda.jp, yshk.hayashi@aoni.waseda.jp

Abstract

Distributional hypothesis has been playing a central role in statistical NLP. Recently, however, its limitation in incorporating perceptual and empirical knowledge is noted, eliciting a field of perceptually grounded computational semantics. Typical sources of features in such a research are image datasets, where images are accompanied by linguistic tags and/or descriptions. Mainstream approaches employ machine learning techniques to integrate/combine visual features with linguistic features. In contrast to or supplementing these approaches, this study assesses the effectiveness of social image tags in generating word embeddings, and argues that these generated representations exhibit somewhat different and favorable behaviors from corpus-originated representations. More specifically, we generated word embeddings by using image tags obtained from a large social image dataset YFCC100M, which collects Flickr images and the associated tags. We evaluated the efficacy of generated word embeddings with standard semantic similarity/relatedness tasks, which showed that comparable performances with corpus-originated word embeddings were attained. These results further suggest that the generated embeddings could be effective in discriminating synonyms and antonyms, which has been an issue in distributional hypothesis-based approaches. In summary, social image tags can be utilized as yet another source of visually enforced features, provided the amount of available tags is large enough.

Keywords: word embeddings, image tags, social media, semantic similarity, synonyms, antonyms.

1. Introduction

Virtually, all the methods for generating distributional/distributed word representations (Baroni et al., 2014) rely on the notion of distributional hypothesis (Firth, 1957). These approaches enable word representations to properly capture the distributional hypothesis by measuring the commonality of the linguistic contexts of word occurrences. Although these approaches are proven effective in various semantic tasks, they are limited in terms of the incorporation of perceptual and empirical knowledge: perceptually or empirically obvious objects have not been necessarily well verbalized in a corpus of written texts (Bruni et al., 2014). Yet another issue with the distributional hypothesis-based methods is that they often run into trouble when discriminating synonyms from antonyms or more vaguely related words (Hill et al., 2015).

Recently, motivated by these issues, several research works that try to incorporate human perceptual/empirical knowledge into linguistically derived representations have emerged. Most typically, such approaches combine/integrate visual features achieved from visual resources with linguistic features (word embeddings) by applying machine learning/deep learning techniques. To enable this line of research, a visual resource in which an image is accompanied by linguistic descriptions is generally required.

Although these methods compensated/improved purely linguistic representations, the source of visual features cannot be limited to image data. That is, if a content of any modality is described with a substantial amount of linguistic tags and/or descriptions, the linguistic co-occurrence observed around the content can be utilized as a source of semantic features. Once resting on this notion, the so-called social media can be exploited as an attractive resource.

When using a social media service, the user assigns tags to her/his contents so that they may be easily searched and located by other users. Sometimes, this process is referred to as folksonomy, as the tags are not constrained by pre-defined controlled keywords and/or ontology terms. Despite the nature that users can freely choose tags, it is exemplified that the vocabulary of tags in a social media service has converged and become stabilized over time (Halpin et al., 2007). Moreover, if a target social media is popular enough and maintains a huge amount of content, the set of tags can be considered as a type of corpus where a similar set of tags would be assigned to similar content. These facts validate the use of social tags as a source of semantic features. Furthermore, the media type of social media content is not necessarily limited to images, admitting the possibility of incorporating other types of modality.

In the present work, we utilize the YFCC100M (Thomee et al., 2016) dataset[1], which is a social media-originated dataset. We generate word embeddings by statistically processing the co-occurrences of linguistic tags. The empirical results of semantic similarity/relatedness tasks may allow us to conclude that social image tags can be utilized as yet another source of visually enforced features, provided the amount of available image tags is large enough.

2. Related work

The present research is inspired by the work on multimodal semantic representations (section 2.1). As most of the work in this direction deals with image features, image datasets (section 2.2) as a source of visual features are of crucial importance.

[1] http://yfcc100m.appspot.com/

Dataset	# of images	Annotation type	Who annotated?
ImageNet	14M	WordNet synsets	Crowdworkers
ESP-Game	350K	tags	ESP-Game participants
MS COCO	120K	categories, captions	Crowdworkers
YFCC100M	100M	tags	Flickr contributors

Table 1: Representative image datasets.

2.1. Multimodal semantic representation

Theoretically supported by the concept of grounded cognition (Barsalou, 2008) and technically endorsed by the progress of machine learning techniques, work on distributed word representation (word embeddings) has extended its research scope to *multimodal semantic representation* in which perceptual information, such as visual features, is combined with or integrated into corpus-derived linguistic embeddings (Silberer and Lapata, 2014; Bruni et al., 2014; Kiela and Bottou, 2014; Kiela et al., 2016). Mainstream approaches employ deep learning techniques to integrate/combine visual features with linguistic features (Lazaridou et al., 2015; Kodirov et al., 2017; Hasegawa et al., 2017). The achieved results in standard semantic similarity/relatedness tasks are generally promising, suggesting that corpus-derived word embeddings can be successfully enhanced by visual features.

2.2. Source of image/visual features

As far as a method for inducing multimodal semantic representation relies on image features, the role of the source image dataset is crucial. Image datasets can be classified in terms of the different types of collected images, linguistic annotations, and originators (who tagged images). Table 1 contrasts representative image datasets with YFCC100M, which is the central ingredient of the present work.

ImageNet (Krizhevsky et al., 2012) has been playing a leading role in improving visual object recognition techniques. The ESP-Game (von Ahn and Dabbish, 2004) dataset is often employed in the work on multimodal semantic representations. These two datasets are contrastive in a sense: ImageNet images clearly portray a focused object, whereas ESP-Game images often depict more natural scenes, showing multiple objects and the relations among them. This means that the ESP-Game images are noisier in terms of visual object recognition (Kiela et al., 2016). MS COCO (Lin et al., 2014), however, has been heavily employed in caption generation research.

YFCC100M (Thomee et al., 2016) collects images and the associated metadata from a social media service Flickr, which is a Web-based service for sharing visual contents. This dataset is different from others in that the tags attached to a posted image is given by the contributor. This nature makes a difference when it is utilized as a source of semantic features, as discussed in the rest of this paper. In Flickr, each image is annotated with a variety of tags, including the name of a depicted object, the place where the picture is taken, and the emotional feeling expressed by the contributor. The amount of data made possible by the popularity of Flickr is also a crucial factor; Thus meaningful co-occurrence statistics can be collectively obtained from this huge dataset.

3. Generating word embeddings from social media data

As described earlier, we aim to construct word semantic representations (word embeddings) by exploiting a social media service as a source of visually enforced semantic features. More specifically, we generate word embeddings, first by constructing a tag co-occurrence matrix, and then converting the raw counts to more effective quantities, and finally applying a dimensionality reduction technique to the co-occurrence matrix. It should be emphasized here that we only employ textual tags, meaning that we have never applied any visual feature extraction to the maintained images. This process assigns each tag word a dense and low-dimensional vector, which can be utilized as a word embedding vector.

The rationale behind this approach is that visual co-occurrences of objects could be naturally captured by the co-occurrence of image tags. We further suppose that the intention of a contributor who wants to disseminate her/his photo to a broader audience may be reflected in the attached tags. Therefore, the tags attached to an image can be considered as a proxy to the image that may partake social implications.

Constructing a tag word co-occurrence matrix: We constitute a tag word co-occurrence matrix M, where $M_{i,j}$ counts the number of times that tag word w_i and tag word w_j are attached to the same image. The shape of the matrix M is $N \times N$ if the number of word types equals to N.

Transforming the matrix: As the raw counts do not properly dictate the strength of co-occurrence, we transform the co-occurrence matrix by computing positive pairwise mutual information (PPMI) (Church and Hanks, 1990), which is formulated as follows.

$$M_{i,j} = \max\left(0, \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}\right) \quad (1)$$

As this formulation suggests, PPMI alleviates the influence of high frequency tags, allowing us to properly measure the strength of tag co-occurrence.

Dimensionality reduction: As the number of tag types easily increases with the size of a dataset, the matrix M would generally be sparse. We thus apply singular value decomposition (SVD), and reduce the matrix from $N \times N$ to $d \times d$, where $d \ll N$.

$$M = U \cdot \Sigma \cdot V^T \quad (2)$$

$$\approx U^{(d)} \cdot \Sigma^{(d)} \cdot V^{T^{(d)}} \quad (3)$$

This gives us the word embedding vector for word w_t as $v_{w_t} = U_{w_t}^{(d)} \cdot \sqrt{\Sigma^{(d)}}$. As frequently argued, dense and low-dimensional representations may yield the benefit of data reduction as well as the effect of data abstraction.

4. Experimental settings

We evaluate the efficacy of tag-originated word embeddings (henceforth, tag embeddings) in standard semantic similarity/relatedness tasks. This section describes the experimental settings, and the following section discusses the results.

4.1. Source of social image tags

We used the Yahoo Flickr Creative Commons 100M (YFCC100M) dataset (Thomee et al., 2016). This dataset collects almost 100 M images and approximately 700 K movies posted on Flickr. Each image is described by a set of metadata, including image titles, user-generated tags, and machine-generated tags. We constructed word semantic representations by feeding only the user-generated tags (tags that are given by the contributors of contents) to the process described in the previous section. The number of user-generated tags amounts to approximately 69M, among which 68.5M tags are assigned to images, and the rest 420K are assigned to movies. In average, approximately seven tags are assigned to each instance of the content.

Figures 1 (a) and (b) show examples of the images and tags. Figure 1 (a) portrays a `cat`, which is further detailed by the hyponyms "kitten" and "kitty," as well as the hypernym "pet." Figure 1 (b) artistically shows a scenery for which abstract words like "calm" and "quiet" are attached. Moreover tag words like "summer" (time) or "favignana" (place) are being assigned, which would not be annotated even by state-of-the-art computer vision techniques.

In the experiments, a word co-occurrence matrix was constructed for the selected 20,943 words that were used for describing more than 1,500 images. We excluded multi-word tags and numbers, and the remaining words were converted to lowercase. The total number of tag instances counts at a value of 10 M. In the dimensionality reduction by SVM, the dimensionality d of word embeddings is set to 300.

Figure 1: Examples of YFCC100M images and tags.

4.2. Evaluation tasks and the datasets

We evaluated the efficacy of constructed word embeddings with word similarity/relatedness tasks in which the predicted scores were compared against the gold data given in the following test datasets. The Spearman's rank correlation coefficient was employed as the performance measure of the experiment that uses one of the datasets.

- **YP130** (Yang and Powers, 2006): This dataset that maintains 130 verb pairs was built for the evaluation of verb similarities.
- **WordSim353** (Finkelstein et al., 2002): This dataset contains 353 word pairs for which semantic relatedness scores are assigned. Note that semantic similarity that essentially measures the degree of synonymy can be considered as a subclass of semantic relatedness.
- **SimLex999** (Hill et al., 2015): SimLex999 provides word similarity (rather than relatedness or association) judgments for 999 word pairs. Note that the parts of speech of compared words are always the same.
- **USF Assoc** (Nelson et al., 2004): This dataset, University of South Florida Free Association Norms (abbreviated as USF Assoc), collects the free association scores for 5,019 stimulus words. In the experiments, we used the pairs of words included in the SimLex999 dataset. Needless to say, free association relations include a wider range of semantic relationships.
- **MEN** (Bruni et al., 2014): This dataset presents semantic relatedness scores for 3,000 word pairs. This dataset was specially made to evaluate multimodal representations. The parts of speech of compared words are not necessarily the same. The words are biased to concrete concepts, as they are chosen from the tags in the ESP-Game and Flickr data.
- **SemSim / VisSim** (Silberer et al., 2016): This is a dataset of 7,576 word pairs, each of which is annotated using not only semantic similarities (SemSim) but also visual similarities (VisSim); therefore, the user can compare the performances of her/his model in predicting different types of similarities.

5. Experimental results

5.1. Major results

Table 2 compares the major experimental results (in Spearman's correlations), where the YFCC column shows the results with the tag embeddings that were generated from the tag co-occurrence matrix which records 10 M tag instances. Wiki or GNews displays the results with corpus-derived word embeddings. By applying the Word2Vec Skip-Gram model, we derived 300-dimensional word embeddings both for Wikipedia 2009 dump[2] (Wiki) and GoogleNews[3] (GNews). Notice that the dimensionalities are equalized with those of tag embeddings.

As shown in the table, the tag embeddings achieved the highest correlation of 0.81 in the MEN relatedness task, demonstrating that social image tags are good sources of visually enforced features for concrete concepts. Furthermore, the tag embeddings achieved an acceptable result

[2] `http://mattmahoney.net/dc/textdata`

[3] `https://github.com/mmihaltz/word2vec-GoogleNews-vectors`

of 0.45 in the SimLex999 similarity task, which is worse than GNews-originated embeddings, but better than Wiki-originated embeddings. This could be good news, as a similarity task is generally considered to be more difficult when compared to a relatedness task.

The table presents that the degradations in USF Assoc score compared to that of SimLex999 are evident in all the embedding types. However, the difference in YFCC (tag embedding) is larger than the other two types. This may be due to the fact that the tags attached to an image are tightly associated with the image, whereas linguistic contexts, or context windows, are more generous to include weakly associated words.

A surprise result is a correlation of 0.47 achieved by the tag embeddings in the YP130 verb similarity task. It could be unfortunately unreliable, as the coverage is as low as 16% (shown in the second column). This insists that verbs are not frequently assigned as a social image tag.

In summary, the tag embeddings could achieve comparable performances with corpus-originated embeddings in a variety of similarity/relatedness tasks.

Dataset	# of pairs	YFCC	Wiki	GNews
YP130	130 (16%)	0.47	0.35	0.24
WS353	353 (66%)	0.65	0.74	0.70
SimLex999	999 (54%)	**0.45**	0.39	0.49
USF Assoc	999 (54%)	0.34	0.38	0.44
MEN	3000 (96%)	**0.81**	0.74	0.77
SemSim	7576 (62%)	0.62	0.63	0.72
VisSim	7576 (62%)	0.49	0.50	0.55

Table 2: Results in semantic similarity/relatedness tasks (in Spearman's correlation).

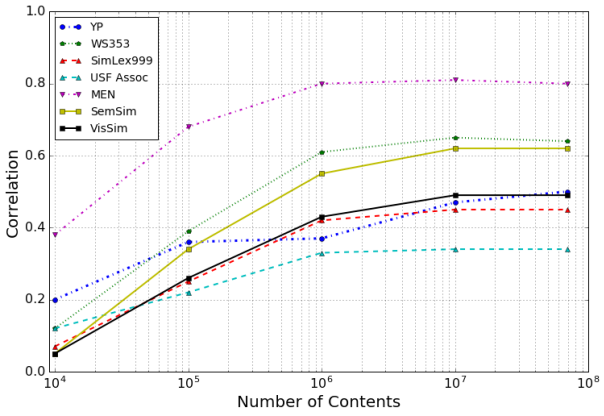

Figure 2: Relationship between the number of contents and the accuracy of semantic relatedness estimation.

The amount of data versus performances: It is often desired to know the necessary/sufficient amount of data to achieve a reasonable performance. Figure 2 displays the saturation of correlation coefficients with the increase in the amount data. As the graph shows, the performance of all datasets does not significantly improve when increased to more than 10 M contents[4], showing a limit to the effective number of tags.

5.2. Do social image tags make a difference?

An expectation to multimodal semantic representations is to address issues inherent to the purely linguistic distributional hypothesis. This expectation also applies to the tag embeddings proposed in the present work. To assess whether this could be attained, we conducted a small experiment by using WordNet semantic relationships. More specifically, for each of the selected 1,928 words that have tag embeddings, we retrieved k-nearest words in WordNet, and investigated the ranks of their antonyms, synonyms, hypernyms, and hyponyms.

Table 3 compares the mean reciprocal ranks (MRRs) of the words in each semantic relation with each embedding type. The average numbers of corresponding words in each semantic relation are as follows: 1.65 for an antonym, 3.27 for a synonym, 5.34 for a hyponym, and 3.07 for a hypernym.

Relation	# of pairs	YFCC	Wiki	GNews
antonym	798	**0.05**	0.18	0.13
synonym	3593	**0.15**	0.09	**0.16**
hyponym	1900	**0.11**	0.04	0.07
hypernym	4163	**0.06**	0.02	0.04

Table 3: MRR results for WordNet semantic relations.

The most prominent fact presented in the table is that the MRR for antonyms with YFCC embedding is far lower than that of the other two embedding types. This confirms that the proposed method could be effective in excluding antonyms from the other semantically similar/related words. Note that YFCC embedding ranked synonyms, hypernyms, and hyponyms are relatively higher than other two embedding types. This may endorse the fact that a content contributor tends to add hypernyms and/or hyponyms as tags, probably for the purpose of increasing the probability of the posted image being retrieved.

To sum up, the resulting semantic representations exhibit somewhat different and favorable behaviors from corpus-originated representations.

6. Concluding remarks

This paper proposed to exploit social image tags as a source of features for generating word embeddings, and demonstrated that the generated representations exhibit somewhat different and favorable behaviors compared to the corpus-originated representations. These results highlight that social media could be exploited as yet another source of semantic features.

This insight may open up a new way of meaning representation that optimally integrates verbal, perceptual, and social features upon a given semantic task. Other benefits potentially attained from the use of social media are dynamics and multilinguality. Social tagging would provide opportunities to capture new definitions for existing words or new

[4]Each content is associated with a set of tags.

words themselves. Tags given in multiple languages can be exploited to develop cross-lingual/multilingual semantic representations.

7. Acknowledgments

The present research was supported by JSPS KAKENHI Grant Number 17H01831 and 15K12873.

8. Bibliographical references

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *ACL*, pages 238–247.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59(August):617–645.

Bruni, E., Gatica-perez, D., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49(December):1–47.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32.

Halpin, H., Robu, V., and Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 211–220, New York, NY, USA. ACM.

Hasegawa, M., Kobayashi, T., and Hayashi, Y. (2017). Incorporating visual features into word embeddings: A bimodal autoencoder-based approach. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS2017)*, Montpellier, France.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Comput. Linguist.*, 41(4):665–695.

Kiela, D. and Bottou, L. (2014). Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. *EMNLP 2014*, pages 36–45.

Kiela, D., Clark, S., Ver, A. L., Clark, S., Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2016). Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. *EMNLP 2016*, 34(2):447–456.

Kodirov, E., Xiang, T., Gong, S., and Mary, Q. (2017). Semantic Autoencoder for Zero-Shot Learning. *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3174–3183.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.*, pages 1–9.

Lazaridou, A., Nghia, T. P., and Baroni, M. (2015). Combining Language and Vision with a Multimodal Skip-gram Model. *NAACL 2015*, pages 153–163.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. a. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behav. Res. Methods, Instruments, Comput.*, 36(3):402–407.

Silberer, C. and Lapata, M. (2014). Learning Grounded Meaning Representations with Autoencoders. *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.*, pages 721–732.

Silberer, C., Ferrari, V., and Lapata, M. (2016). Visually Grounded Meaning Representations. *IEEE Trans. Pattern Anal. Mach. Intell.*, X(X):1–1.

Thomee, B., Shamma, D. A., Friedland, G., Poland, D., and Borth, D. (2016). YFCC100M : The New Data in Multimedia Research. *Commun. ACM*.

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. *Proc. 2004 Conf. Hum. factors Comput. Syst. - CHI '04*, pages 319–326.

Yang, D. and Powers, D. M. (2006). Verb Similarity on the Taxonomy of WordNet. *Proc. GWC-06*, pages 121—-128.