Improving Word Representations via Global Visual Context

Ran Xu
Department of Computer Science
SUNY at Buffalo
rxu2@buffalo.edu

Jiasen Lu
Department of Computer Science
SUNY at Buffalo
jiasenlu@buffalo.edu

Caiming Xiong
Department of Statistics
UCLA
caimingxiong@ucla.edu

Zhi Yang
Department of Computer Science
SUNY at Buffalo
zhiyang@buffalo.edu

Jason J. Corso
Department of Electrical Engineering and Computer Science
University of Michagan
jjcorso@eecs.umich.edu

Abstract

Visually grounded semantics is a very important aspect in word representation, largely due to its potential to improve many NLP tasks such as information retrieval, text classification and analysis. We present a new distributed word learning framework which 1) learns word embeddings that better capture the visually grounded semantics by unifying local document context and global visual context, 2) jointly learns word representation, image representation and language models and 3) focus on better word similarity rather than relatedness. We apply a data set that contains 1 million image-sentence pairs for training and the evaluation on word similarity demonstrates our model outperforms linguistic model without global visual context.

1 Introduction

Distributed word representations [1, 2] have shown to be very effective and efficient to capture syntactic and semantic word relationships, and have been used in many NLP tasks such as sentiment analysis and computer vision tasks such as video to text [3]. But as Huang et al. [4] noted, most continuous word representation models are built with only local context. There are some works that incorporate global text or paragraph with local window to train word vector [4, 5] and achieved better performance in word relationships, text classification and sentiment analysis tasks. Although global word context can provide useful topical information, the learned representations still capture more **relatedness or association** between words rather than **similarity**. For example, with a skip-gram model [2] trained with 100 billion words, the cosine similarity between "cat" and "dog" is 0.76 while between "cat" and "tabby cat" is 0.63.

In this work, we propose to use global visual context to help learn better word representations that embed visually grounded semantics. Some researchers have explored multi-modality between vision and language, Frome et al. [6] learn similarity metric between output of a deep visual model and a distributional language model. Socher et al. [7] build a multi-modal representation between

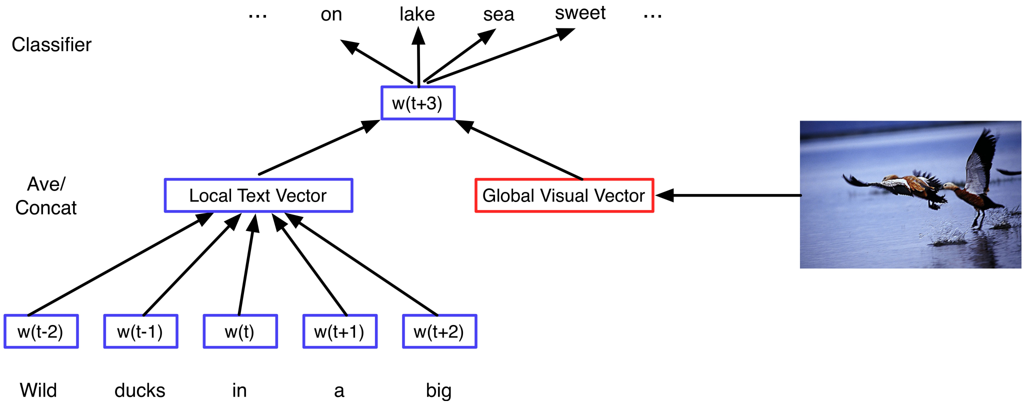

Figure 1: Framework of our model, left part is language model with input from sliding windows of word sequence; right part is our global visual context model with input from image.

compositional sentence vectors and image vector representation. More recently, Kiela and Bottou [8] construct multi-modal concept representations by concatenating a skip-gram linguistic representation vector with a deep visual representation vector. All of these methods explore the role of visual information in multi-modal transformation, sentence compositionality or extra information for better word semantics, but neither the word representation nor language model would be influenced by visual information.

WIth our framework, the word representation, the language model and the visual representation are learned together, where visual information can be used as a global context to improve the word semantics. In this way, we are able to measure how global visual information contributes (or affects) word representation. Further in the paper, we show that our approach will improve word semantics.

2 Algorithms

To train our model, we use image-sentence pair as our training data, where each image corresponds to a number of sentence descriptions. As Fig. 1 illustrates, our framework is composed of two parts, the language model on the left is inspired by Continuous Bag-of-Words model [1, 5], where each word in the vocabulary is initialized as random vector with fixed length. Right side of the figure shows our global visual context, we initialize the global visual vector with convolutional neural networks [9] output from the corresponding image. We average the word vectors from a local window of training sentences to construct the local text vector, and directly use CNN feature as our global visual vector, both of them will contribute to the classification of next word. Then, classification error will be back propagated to both local text vector, global visual vector and the model.

More formally, given the training text we can obtain a vocabulary and construct a vocabulary matrix $W \in \mathbf{R}^{m \times d}$, where m is vocabulary size and d is dimension of word vector. For images, we extract CNN feature vectors $V \in \mathbf{R}^{n \times l}$ as global visual context, where n is total number of images and l is dimension of image vector. For a sequence of training words $w_1, w_2, w_3, ..., w_T$, the objective function is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k}, ..., w_{t+k}, v_t^{(a)}), \quad a = 1, 2, ...n \tag{1}$$

where T is training words size, $2k+1$ is local text window size, $v_t^{(a)}$ is visual representation of image a that corresponds to w_t. The probability can be formulated by softmax:

$$p(w_t|w_{t-k}, ..., w_{t+k}, v_t^{(a)}) = \frac{\exp(y_{w_t})}{\sum_i \exp(y_i)} \tag{2}$$

Table 1: WordSim Spearman correlation between linguistic baseline and our model with global visual context. Results with different window sizes and two training methods are reported.

Window/Method	Linguistic (NS)	Global Visual Context (NS)	Linguistic (HS)	Global Visual Context (HS)
8	0.338	**0.362**	0.324	**0.330**
10	0.331	**0.344**	**0.386**	0.377
12	0.314	**0.353**	0.293	**0.336**

$$\begin{aligned} y &= M_t h(w_{t-k}, .., w_{t+k}, v_t; W, V) \\ &= M_t([ave(w_{t-k}, .., w_{t+k}), \alpha v_t^{(a)}]) \end{aligned} \tag{3}$$

$M \in \mathbf{R}^{m \times (d+l)}$ is softmax parameter matrix and M_t is the vector corresponds to word w_t, and h is a function that combine local words representation and global image representation, y is prediction of softmax. In our implementation, we average the word vectors and then concatenate with image vector with tuning parameter α. Similar as [1, 2], we use Hierarchical Softmax and Negative Sampling to speedup the training. Both word vectors and image vectors are trained with stochastic gradient descent and back propagation.

3 Experiments

Our experiments focus on measuring word semantic similarity, especially the relationship between word representation and its visually grounded meaning.

Dataset Our corpus and corresponding images are from SBU Captioned Photo Dataset [10], where 1 million well captioned pictures are collected from Flicker. In practical, we totally collected 929499 images because some of photos are not available from Flicker, each image is described by one sentence. The whole corpus contains 13.4 million words and thus we build a vocabulary of 298469 words.

Implementation Details To train the model, we initialize each word in the vocabulary as 100-dimensional random vector. The image vectors are initialized by computing convolutional neural networks [9], we choose the 7th layer output after ReLU and get the 4096-dimensional vector. In training, we test both hierarchical softmax and negative sampling to speed up, and also test different window sizes, word vector sizes and global visual contribution factor α.

Evaluate Semantics Similarity We use WordSim353 Similarity test set [11] and SimLex-999 [12] to measure the words semantics similarity.

For WordSim 353 data set, our vocabulary covers 95$\%$ of the annotated ground truth so we only evaluate on the 192 concept pairs. We evaluate the Spearman correlation between system output, i.e., cosine similarity between a pair of words, and human rated ground truth.

Table. 1 shows the Spearman correlation score of linguistic baseline and our model with global visual context, three different word window sizes and 2 different training methods are compared. The scores confirm significant improvement with the help of visual context. We also tested different visual contribution parameter α and find it generally performs well with value close to 0.1.

To evaluate whether higher scores really capture better visually grounded semantics, we compare the cosine similarity between linguistic model and our model, and find top 5 word pairs when our similarity scores are higher than linguistic model, and top 5 pairs that are lower. Table. 2 reports results from two methods (NS in first 4 columns and HS in last 4 columns), it is clear that our word vector trained with images capture better semantics. For example, the similarity is higher when dealing word pairs with strong visual information, such as “planet”-“star” and “vodka”-“gin”; on the contrary, we also find word vectors with global visual context tend to return lower similarity when the word pair has more abstract relationship but limited visually grounded meaning, such as “psychology”-“discipline” and “morality”-“marriage”, or the word pair captures more relatedness rather than similarity such as “cell”-“phone”. In summary, the qualitative results are consistent with correlation scores and indicate the effectiveness of our method.

Table 2: Visually grounded semantics comparison between linguistic model and our model. We select top 5 best and top 5 worst scoring pairs of our global visual context model with respect to the linguistic model.

Negative Sampling				Hierarchical Softmax			
word1	word2	GT	Δ_{CosSim}	word1	word2	GT	Δ_{CosSim}
car	automobile	8.94	0.085	vodka	gin	8.46	0.166
planet	star	8.45	0.074	money	dollar	8.42	0.137
hospital	infrastructure	4.63	0.072	media	trading	3.88	0.128
jaguar	cat	7.42	0.061	cup	artifact	2.92	0.117
dollar	buck	9.22	0.058	gem	jewel	8.96	0.108
psychology	discipline	5.58	-0.165	cell	phone	7.81	-0.289
consumer	confidence	4.13	-0.160	focus	life	4.06	-0.207
president	medal	3	-0.156	morality	marriage	3.69	-0.153
physics	chemistry	7.35	-0.128	man	women	8.3	-0.151
magician	wizard	9.02	-0.105	tiger	animal	7	-0.142

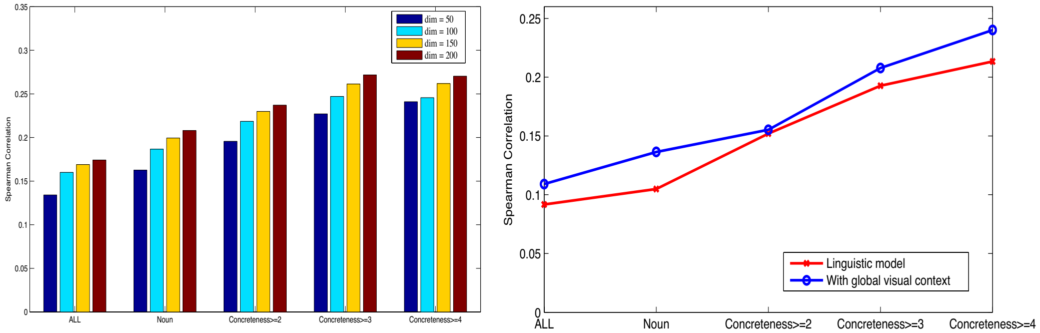

Figure 2: Figure of SimLex-999 Data set

Furthermore, we evaluate our model in a larger data set SimLex-999 [12] where 999 pairs of words are selected and rated, this data set focuses on word similarity, besides, POS tagging of words and "concreteness" of words are also annotated. Each word pair is labeled with adjective, noun and verb; and also with concreteness scores ranging from 1 to 4. Fig. 2 (a) shows Spearman correlation score of our method with Negtive Sampling over different word vector dimensions, the scores are reported for 970 SimLex words in our vocabulary, only noun word pairs (666 pairs), all word pairs that has concreteness larger than $1, 2$ and 3. The results show that our model performs better with noun, and with word pairs with higher concreteness. Fig. 2 (b) compares our model with baseline linguistic model with hierarchical softmax, it demonstrates that with global visual context, our model captures better semantic relationship than linguistic model.

4 Discussion

To conclude, in this paper, we propose to use global visual context to help train better word vector, and to the best of our knowledge, it is the first attempt to train distributed word representation directly using images. The experiments show our approach is able to capture better visually grounded semantics and outperforms pure linguistic model quantitatively and qualitatively.

Future works include 1) use both global visual information and global text information together and a skip-gram like model, 2) explore how visual context can help both semantics and syntax of words, and 3) explore the influence of image vector trained after this approach, and experiment with image/text retrieval.

References

[1] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR. (2013)

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. (2013)

[3] Sun, C., Nevatia, R.: Semantic aware video transcription using random forest classifiers. In: ECCV. (2014)

[4] Huang, E.H., Socher, R., Manning, C.D., Ng., A.Y.: Improving word representations via global context and multiple word prototypes. In: ACL. (2012)

[5] Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML. (2014)

[6] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: NIPS. (2013)

[7] Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. In: Transactions of the Association for Computational Linguistics. (2013)

[8] Kiela, D., Bottou, L.: Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In: EMNLP. (2014)

[9] Donahua, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: arXiv:1310.1531. (2013)

[10] Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: NIPS. (2011)

[11] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: WWW. (2011)

[12] Hill, F., Reichart, R., Korhonen, A.: Simlex-999: Evaluating semantic models with (genuine) similarity estimation. In: arXiv:1408.3456v1. (2014)