Enhancing Multimodal Embeddings with Word Semantic Relations for Image Search Applications*†

Mejorando representaciones de baja dimensionalidad con relaciones semánticas de palabras para aplicaciones de búsqueda de imágenes

Marco A. Gutiérrez
Robolab, University of Extremadura
Avda. de la Universidad, Cáceres, Spain
marcog@unex.es

Resumen: La generación de leyendas para imágenes juega un papel esencial en las aplicaciones de búsqueda de imágenes ya que nos permiten generar automáticamente descripciones de imágenes. Sin embargo a veces las palabras en estas leyendas generadas no son exactas y además pueden encontrarse abiertas a criticas subjetivas. También cuando buscan una imagen, los usuarios puede que no usen exactamente las mismas palabras que las existentes en esas leyendas sino otras con cierta similitud semántica. Por lo tanto presentamos un trabajo en el que expandimos el ámbito de nuestras leyendas generadas a partir de imágenes comparando la relación semántica entre la consulta y las palabras en la leyenda. En este trabajo usamos un pipleine codificador-decodificador que unifica representaciones de baja dimensionalidad de modelos imagen-texto con modelos de lenguage multimodales neuronales para generar descripciones de imágenes. Luego extendemos la semántica de estas descripciones utilizando vectores de palabras entrenados sobre grandes conjuntos de palabras para representar eficientemente su similitud semántica. Finalmente mostramos que haciendo uso de estas relaciones semánticas entre palabras somos capaces de encontrar conceptos mostrados en las imágenes que no estaban directamente escritos en las descripciones generadas incialmente.
Palabras clave: Representaciones de baja dimensionalidad, relaciones semánticas, Redes neuronales convolutivas, Vectores de palabras, Búsqueda de imágenes

Abstract: Image caption generation play a key role in image search applications as they allow us to automatically generate language based description of pictures. However sometimes the words on these generated captions might not be accurate and the result is open to criticism of subjectivity. Also, when searching for an image, users might not use the exact same words as the ones in generated captions but others with a semantic similarity. Therefore we present a work were we expand the scope of our image generated captions by looking at the semantic relation between the query and the words in the captions. We use an encoder-decoder pipeline that unifies joint image-text embedding models with multimodal neural language models to generate image captions. Then we extend the semantics of those captions making use of word vectors trained over large word datasets in order to effectively represent word semantic similarity. We finally show that by making use of these word semantic relations we are able to find concepts shown in the image that were not directly written in the initially generated captions.
Keywords: Embedding, Semantics, Convolutional Neural Networks, Word Vectors, Image search

1 Introduction

Words can have multiple degrees of similarity (Mikolov, Yih, and Zweig, 2013). On top of that different users might query in different ways when looking for the same thing. Also systems might label images and generate descriptions in different manners that can even

* Thanks to Dr. Rafael E. Banchs from the Institute for Infocomm Research for his advice and support on this work.
† The author conducted this work as part of his A*STAR Research Attachment Programme (ARAP) at the Human Language Technology Department of Institute for Infocomm Research, Singapore.

be subjectively considered proper or not. Expanding the semantic scope of these image descriptions and the users queries can benefit the search of images producing a wider and more accurate range of results.

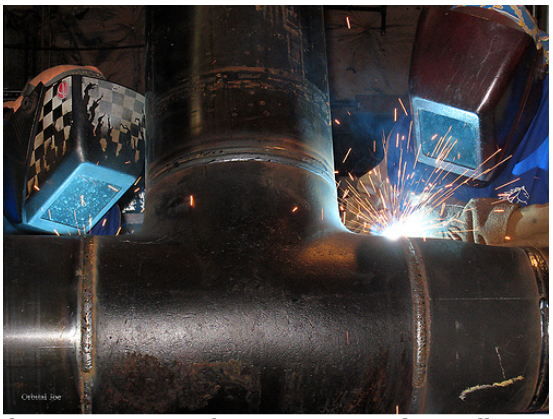


Figure 1: Top result of our system when looking for a picture with *smoke*. Note that the word *smoke* does not appear in the generated captions but there is still smoke on the picture.

Recent works like (He et al., 2015), (Vinyals et al., 2014) or (Xu et al., 2015) prove the big advances that have been done automatically generating captions for images. However these captions are usually short and, even though they could provide accurate descriptions, they do not contain all the information that is showed in the picture. Same objects can be described with different words. Therefore people can differ on how they would call something in an image. In order to extend the information contained on these generated sentences semantic relations between words can be exploited.

There are several techniques that provide semantic similarity between words (Christoph, 2016). Some approaches exploit manually created ontologies or taxonomies like WordNet (Fellbaum, 1998) or Freebase (Bollacker et al., 2008). These ontologies are manually created and maintained, sometimes being very costly. In consequence, only a few domains have a suitable ontology, limiting the applicability of similarity measures based on one of them. Dense vector representation approaches exploit the statistics over large text corpora by representing words as high dimensional sparse word count vectors. We use the skip-gram negative sampling approach (Mikolov et al., 2013b). These models are trained using windows extracted from a natural language corpus (i.e. an unordered set of words which occur nearby in a text sequence in the corpus). This allows us to easily retreain the system with new word scopes to cover new semantic areas. The final model is trained to predict, given a single word from the vocabulary, those words that will likely occur nearby in a text.

The system presented in this work weights the semantic relations between a query and image generated captions in order to improve the ranking of images to produce a result on a possible image search application. Therefore when a query is submitted to the system, nouns and adjectives from the query and from the captions are selected using the Natural Language Toolkit (NLTK) (Bird, Klein, and Loper, 2009). The neural network encoder-decoder pipeline described in (Kiros, Salakhutdinov, and Zemel, 2014) generates captions that describe a set of images. Then pre-trained word vectors helps finding semantic similarities between words on the captions and the ones selected form the query using the Skip-gram model described in (Mikolov et al., 2013a). Those similarities are calculated using the cosine distance in the vector space between the selected words in the query and the ones in the captions. Results are sorted by their calculated similarity weight, the best ones would be the ones with the highest similarity value. This process allows the expansion of the semantic domain of the words on the image generated captions being able to find things that are not explicitly noted in those sentences. Even in the case of querying for something that is not on the image dataset, the output will be more relevant than a random ordering of the images.

2 System design

Given a query our system otputs the top images that are most likely to contain what is described in the query. It accepts queries in the form of "get me a cup" or longer ones like "look for a cup on a table". As shown in Figure 2, the system contains two main em-

bedding subsystems. A multimodal encoder-decoder pipeline that generates the captions for a set of images and a word vector representation for the word semantic expansion. Words from the captions and the query are weighted on their semantic similarity and images are sorted on the average semantic value.

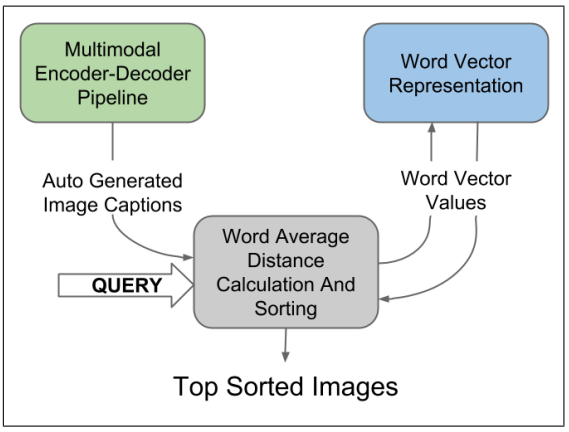


Figure 2: System architecture.

2.1 Multimodal encoder-decoder pipeline

This system is able to generate realistic image captions. The encoder is learned with a joint image-sentence embedding where sentences are encoded using long short-term memory (LSTM) recurrent neural networks (Hochreiter and Schmidhuber, 1997) Image features from the top layer of a deep convolutional network trained from the ImageNet classification task (Krizhevsky, Sutskever, and Hinton, 2012) are projected into the embedding space for the LSTM hidden states. A pairwise ranking loss is minimized in order to learn to rank images and their descriptions. For decoding the structure-content neural language model (SC-NLM) described in (Kiros, Salakhutdinov, and Zemel, 2014) is used which takes into account the content in the sentences.

2.2 Word Semantics Relationships

We decided to use neural networks for this task as they perform better than Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) for preserving linear irregularities among words and in terms of computational cost when dealing with large training datasets (Mikolov, Yih, and Zweig, 2013) (Zhila et al., 2013). We use an improved version of the Skip-gram model (Mikolov et al., 2013a) to find word representations that predict the surrounding words in a document. The version used here makes use of negative sampling (Mikolov et al., 2013b) instead of the hierarchical softmax which tries to differentiate data from noise by means of logistic regression. With this, we build a word vector space that encodes semantics relations on the words of the training data. This semantic relationships are used in our system to weight the semantic relation through the cosine distance of these words.

2.3 Word matching system

As a query comes it gets analyzed using NLTK and the nouns and adjectives are extracted. This words are the ones that will be used, since we consider them the most relevant on the query. The semantic weight of an image k is obtained by calculating the average of the cosine distance in vector space from each name or adjective from the query to each name or adjective in the top 5 generated captions of that image. Equation 1 shows the formal expression of this calculation, where n is the number of nouns in the query, m the number of nouns in the captions and d_{ij} is the cosine distance from word i from the query to word j on the captions.

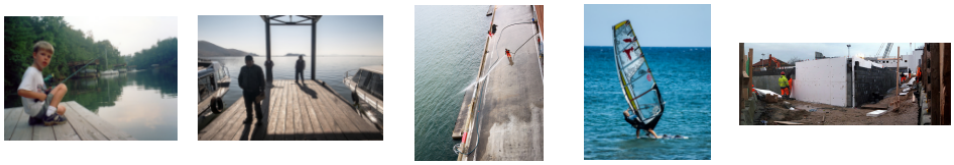
$$W_k = \frac{1}{n+m}\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} = \frac{1}{n+m}\left(d_{11} + \cdots + d_{nm}\right) \tag{1}$$

Finally when all images weights are computed for the given query they are ranked by their weight value. The ones with the highest score will be the images whose captions have a highest semantic similarity to the query.

3 Experiments

As stated by (Besser, 1990) among others, a manual interpretation of the contents of an image will always be open to criticism of subjectivity. Therefore the difficulty of quantitatively evaluate the retrieval effectiveness of our approach. We tested our system against a direct-match approach where instead of using our semantic matching system the words are just directly matched. In this approach for each of the nouns and adjectives from the query that appear on the generated captions of an image will add a value of 1 to the weight of that image otherwise 0 will be added. This will end up with

(a)



(b)

Figure 3: Top results of the query *look for a pet in a river*, being the first the top left one and last the down right one. a) These are the results using the word semantics relations. None of the generated captions specifically contained ether the word *pet* or *river*. b) These are the results for the direct matching experiment. Only five of all the captions contained the word river. The rest are not shown since they all got a score of 0.

an image-to-query similarity weight equal to the number of nouns and adjectives they share. As on our system, the images are sorted by weight and those with a higher weight will be the top result of the approach. We show for each query the top results and the generated captions of our approach versus those with the direct matching approach. Due to space limits we can only show some results, for a wider overview please refer to: http://magutierrez.com/semantics-embeddings

For the experiments the LSTM encoder and SC-NLM decoder of the pipeline described in Section 2.1 have been trained on a concatenation of training sentences from both Flickr30K (Plummer et al., 2015) and Microsoft COCO (Lin et al., 2014). A subset of 1000 images from Flickr30K set is randomly selected and used for caption generation. These are the ones that will be used as possible results for the final answer to the query. Word representations in vector space are trained on part of Google News dataset (about 100 billion words). The final model contains 300-dimensional vectors for 3 million words and phrases.

Figure 1 shows the top result of searching for the word *smoke*. Not any of the generated captions show the word *smoke* among their results. Actually none of the captions contain the word smoke so the result of the direct-match approach is just a random ordering of the images with no sense at all. However our semantic based matching approach is able to detect the high similarity between fire and smoke and rank most of the pictures with fire on it with a higher similarity value. This way we can infer from the captions things that have a high probability of being in the picture even though they are not directly written there.

Figure 3 shows the results for the query *look for a pet in a river*. This query is longer and contains more words to evaluate. As a result we can see the direct match could find some *river* matches but probably not that much for pet. However our algorithm was able to evaluate the semantic relation between the word *pet* and some animals.

4 Conclusions and Future Work

Our system generates captions from images and expands their semantic scope using word representations in vector spaces. We have shown that weighting the words semantics relation of a query to the captions can significantly improve the results of an image search application. The system can even provide meaningful results when queried with words that don't even appear on the captions. Still different types of distances and weighting can be tested and compared in order to try to improve the results of the final image ranking. Further analysis ether of the query or the captions can be done with different natural language processing tools to determine the importance of the words and weight accordingly. Finally different ways of semantical relation among words can be also explored to extend and compare the results among the different relating approaches.

References

Besser, Howard. 1990. Visual access to visual images: the uc berkeley image database project. *Library Trends*, 38(4):787–798.

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

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *In SIGMOD Conference*, pages 1247–1250.

Christoph, LOFI. 2016. Measuring semantic similarity and relatedness with distributional and knowledge-based approaches.

Fellbaum, Christiane. 1998. *WordNet*. Wiley Online Library.

He, Xiaodong, Rupesh Srivastava, Jianfeng Gao, and Li Deng. 2015. Joint learning of distributed representations for images and texts. *CoRR*, abs/1504.03083.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Landauer, Thomas K and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*. Springer, pages 740–755.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Plummer, Bryan, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

Zhila, Alisa, Wen-tau Yih, Christopher Meek, Geoffrey Zweig, and Tomas Mikolov. 2013. Combining heterogeneous models for measuring relational similarity. In *HLT-NAACL*, pages 1000–1009.