Interactive Image Search System Based on Multimodal Analogy

Kosuke Ota, Keiichiro Shirai, Hidetoshi Miyao, and Minoru Maruyama(✉)

Computer Science and Engineering, Shinshu University, Nagano, Japan
maruyama@cs.shinshu-u.ac.jp

Abstract. We propose an image search system based on multimodal analogy, which is enabled by using a visual-semantic embedding model. It allows us to perform analogical reasoning over images by specifying properties to be added to/subtracted with words such as *[a image of a blue car] - 'blue' + 'red'*. The system mainly consists of the following two parts: (i) an encoder that learns image-text embeddings and (ii) a similarity measure between embeddings in a multimodal vector space. As for the encoder, we adopt a CNN-LSTM encoder proposed in [1], which was reported that it can learn multimodal linguistic regularities. We also introduce a new similarity measure based on the difference between additive and subtractive query. It gives us reasonably better results than the previous approach at qualitative analogical reasoning tasks.

Keywords: Multimodal learning · Neural networks

1 Introduction

Most of the modern image search systems only determine images that have properties specified by a given query. They still work most of the time. However, it is often the case that search results are not quite right for our need. Suppose, for example, we have an image that is perfect except for the color of one object. In this case, it is desirable that we can perform arithmetic operations such as *[a image of a blue car] - 'blue' + 'red'* so that we can get the ideal image. Such a system could extend the possibilities of image searching. This could lead to a more interactive system, for instance, which can search for images through verbal interactions with voice assistants.

To address this problem, we propose an image search system based on multimodal analogy. A visual semantic embedding model, which forms the core of the system, enables us to perform analogical reasoning over images by specifying properties to be added to/subtracted from the current results with words.

Additionally, in order to search for an image specified by the aforementioned arithmetic operation, it is necessary to introduce an appropriate similarity measure. We therefore propose a measure based on the difference between additive and subtractive queries. We show the effectiveness of the measure by experiment.

C. Stephanidis (Ed.): HCII Posters 2017, Part II, CCIS 714, pp. 591–596, 2017.
DOI: 10.1007/978-3-319-58753-0_83

2 Multimodal Learning

2.1 Visual-Semantic Embedding Models

To operate arithmetic between images and text, they need to be represented in a shared vector space. Much research has been done on learning joint embeddings of images and text. A well-known approach is to learn a function that maps both image and word embeddings into a common vector space [1–3]. The learned image-text embeddings are often called *visual-semantic embedding* since semantic relationships between images and text can be obtained through its training process.

2.2 Multimodal Linguistic Regularities

Kiros et al. [1] reported that multimodal linguistic regularities were found in an image-text embedding space, while the main focus of their work is on image captioning. They qualitatively investigated properties of the multimodal vector space and the results indicate that linguistic regularities [4] carry over to the joint space.

They also proposed the visual-semantic embedding learned on the image-text encoder. The image-text encoder (Fig. 1) consists of convolutional neural network (CNN) [5] and long short-term memory (LSTM) [6]. The CNN and LSTM take images and sentences as input, respectively. In the training phase, the network is optimized to minimize a pairwise ranking loss:

$$\begin{aligned}\min_{\boldsymbol{\theta}} \quad & \sum_{\boldsymbol{X}}\sum_{k}\max\{\,0,\ \alpha - S(\boldsymbol{X},\boldsymbol{V}) + S(\boldsymbol{X},\boldsymbol{V}_k)\,\} + \\ & \sum_{\boldsymbol{V}}\sum_{k}\max\{\,0,\ \alpha - S(\boldsymbol{V},\boldsymbol{X}) + S(\boldsymbol{V},\boldsymbol{X}_k)\,\},\end{aligned} \tag{1}$$

where $\max\{\cdot,\cdot\}$ returns the larger value, θ denotes the model parameters, α is a margin, and cosine similarity is used as a scoring function $S(\cdot,\cdot)$. $\boldsymbol{V}_k$ and $\boldsymbol{X}_k$ are, respectively, contrastive embeddings for image embeddings $\boldsymbol{X}$ and sentence embeddings $\boldsymbol{V}$. Intuitively, the loss function trains the network to assign high scores to correct pairs of images and text, while it gives incorrect pairs low scores.

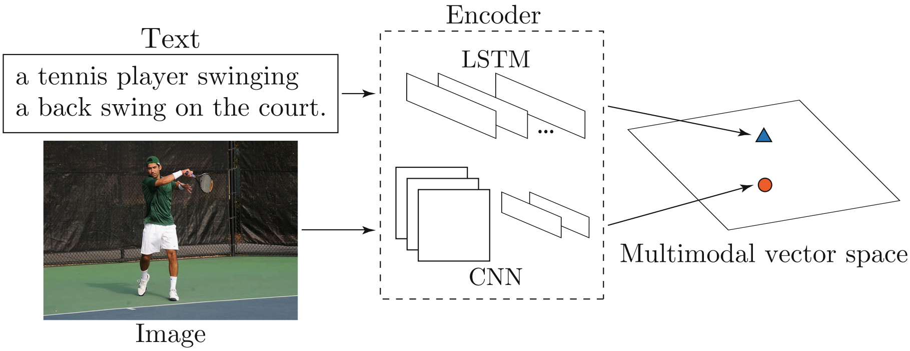


Fig. 1. Image-text encoder of the visual-semantic embedding model

3 Similarity Measure Based on the Difference Vector

Given a base query image and words that specify additive/subtractive properties, Kiros et al. [1] uses the following similarity measure:

$$S(\boldsymbol{X},\ q_{\mathrm{img}} - q_{\mathrm{sub}} + q_{\mathrm{add}}), \tag{2}$$

where q_{img}, q_{add}, q_{sub} are vector representations of queries in the multimodal vector space and $S(\cdot, \cdot)$ is cosine similarity. With the similarity measure (2), we try to find vector $\boldsymbol{X}$ that is closest to $q_{\mathrm{img}} - q_{\mathrm{sub}} + q_{\mathrm{add}}$, with respect to cosine similarity.

A desirable target $q_{\mathrm{img}} - q_{\mathrm{sub}} + q_{\mathrm{add}}$ can be represented as "q_{img} + difference." The difference vector is specified by direction and magnitude. The similarity measure in (2) constrains both the direction and the magnitude of the difference from the base image. However, the arithmetic "base − sub + add" is just a qualitative one. Usually, it could be hard for us to specify the magnitude of the difference vector by only giving additive/subtractive words.

In our method, instead of (2), we use the following measure:

$$S(\boldsymbol{X} - q_{\mathrm{img}},\ q_{\mathrm{add}} - q_{\mathrm{sub}}). \tag{3}$$

With this similarity measure, we try to find $\boldsymbol{X}$ such that the difference from the base image is similar to "add − sub" (see Fig. 2).

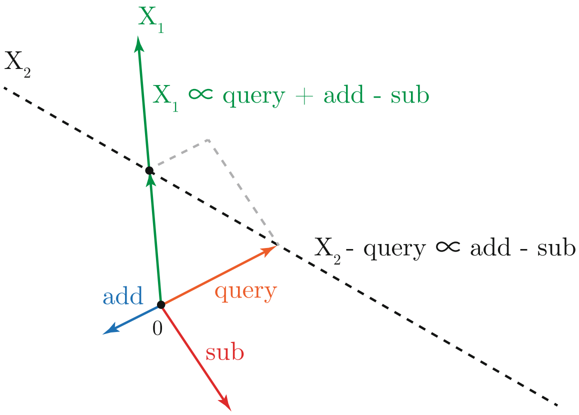
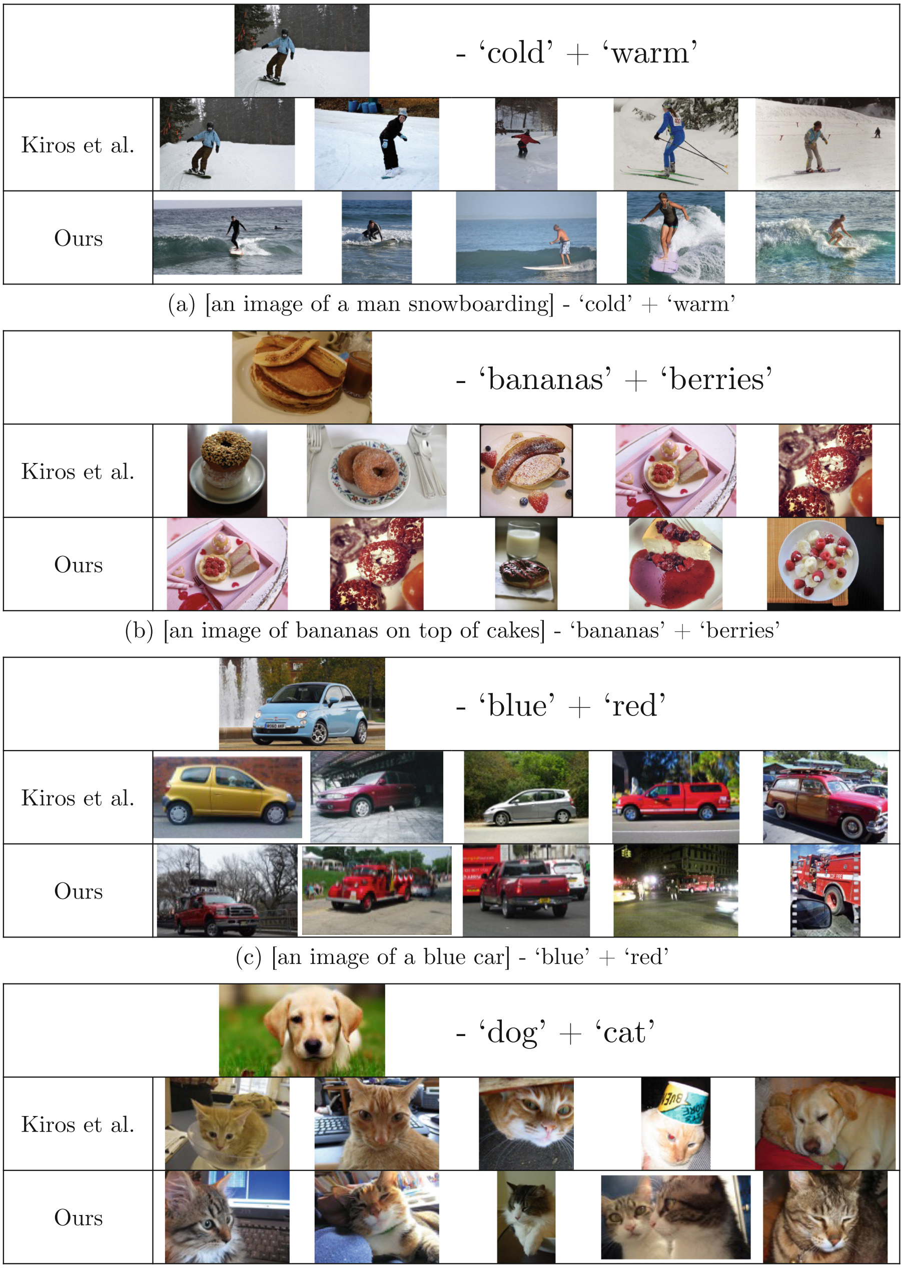


Fig. 2. Relations among our similarity measure and the previous one

4 Experiments

We conduct experiments in the same manner as [1] to compare our results to the previous method [1]. We used the Microsoft COCO dataset [7] to train the encoder.

[1] We reproduced their results using code available on https://github.com/ryankiros/visual-semantic-embedding.



(a) [an image of a man snowboarding] - ‘cold’ + ‘warm’

(b) [an image of bananas on top of cakes] - ‘bananas’ + ‘berries’

(c) [an image of a blue car] - ‘blue’ + ‘red’

(d) [an image of a dog] - ‘dog’ + ‘cat’

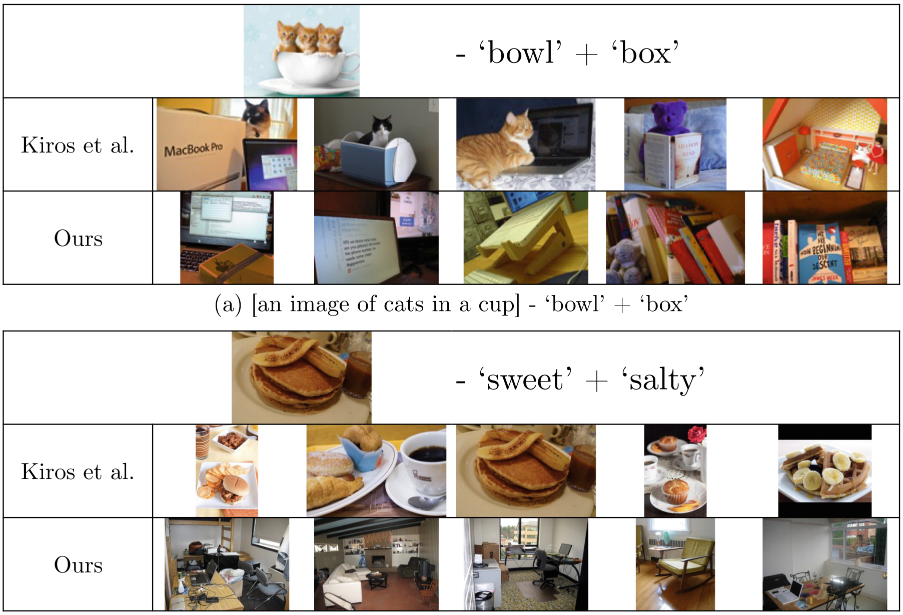
Fig. 3. Examples that our measure performs better than the previous method

The dataset contains about 83,000 images and each image is accompanied by 5 descriptive sentences.

Figures 3 and 4 both shows (2) vs. (3) comparison on multimodal analogical reasoning tasks. Figure 3 illustrates some examples that our measure performs better than the previous one in terms of ranking order. Note that our measure also performs well when query words are not visually obvious as shown in the first example in Fig. 3, while the previous approach struggles to find plausible images.

On the other hand, Fig. 4 shows cases of our poor results. The system gives us irrelevant images. We speculate this is most likely due to the insufficiency of training data and it prevents the network from learning semantic relationships. However, we still need to investigate the cause further.

Taking all these results into consideration, we consider our measure to be more suitable to search for images based on multimodal analogical reasoning.



(a) [an image of cats in a cup] - ‘bowl’ + ‘box’

(b) [an image of pancakes with bananas] - ‘sweet’ + ‘salty’

Fig. 4. Examples that our measure shows irrelevant images

5 Conclusion

We proposed an image search system based on multimodal analogy that allows us to perform analogical reasoning over images and text. Our difference-based similarity measure gives us reasonably better results than the previous method at qualitative analogical reasoning tasks. The system provides us flexibility that would be useful when searching for images through verbal interactions with voice assistants, which is gaining more and more attentions recently, as well as traditional web image search systems.

Acknowledgement. M. Maruyama was supported by JSPS Kakenhi Grant Number JP26330249.

References

1. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
2. Dong, J., Li, X., Snoek, C.G.M.: Word2visualvec: cross-media retrieval by visual feature prediction. arXiv preprint arXiv:1604.06838 (2016)
3. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: a deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems (NIPS) (2013)
4. Mikolov, T., Yih, W.T., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2013)
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations (2015)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997)
7. Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft coco captions: data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)