Mind's Eye: A Recurrent Visual Representation for Image Caption Generation

Xinlei Chen[1], C. Lawrence Zitnick[2]
[1]Carnegie Mellon University. [2]Microsoft Research Redmond.

A good image description is often said to "paint a picture in your mind's eye." The creation of a mental image may play a significant role in sentence comprehension in humans [3]. In fact, it is often this mental image that is remembered long after the exact sentence is forgotten [5, 7]. As an illustrative example, Figure 1 shows how a mental image may vary and increase in richness as a description is read. Could computer vision algorithms that comprehend and generate image captions take advantage of similar evolving visual representations?

Recently, several papers have explored learning joint feature spaces for images and their descriptions [2, 4, 9]. These approaches project image features and sentence features into a common space, which may be used for image search or for ranking image captions. Various approaches were used to learn the projection, including Kernel Canonical Correlation Analysis (KCCA) [2], recursive neural networks [9], or deep neural networks [4]. While these approaches project both semantics and visual features to a common embedding, they are not able to perform the inverse projection. That is, they cannot generate novel sentences or visual depictions from the embedding.

In this paper, we propose a bi-directional representation capable of generating both novel descriptions from images and visual representations from descriptions. Critical to both of these tasks is a novel representation that dynamically captures the visual aspects of the scene that have already been described. That is, as a word is generated or read the visual representation is updated to reflect the new information contained in the word. We accomplish this using Recurrent Neural Networks (RNNs). One long-standing problem of RNNs is their weakness in remembering concepts after a few iterations of recurrence. For instance RNN language models often find difficultly in learning long distance relations without specialized gating units [1]. During sentence generation, our novel dynamically updated visual representation acts as a long-term memory of the concepts that have already been mentioned. This allows the network to automatically pick salient concepts to convey that have yet to be spoken. As we demonstrate, the same representation may be used to create a visual representation of a written description.

We demonstrate our method on numerous datasets. These include the PASCAL sentence dataset [8], Flickr 8K [8], Flickr 30K [8], and the Microsoft COCO dataset [6] containing over 400,000 sentences. When generating novel image descriptions, we demonstrate results as measured by BLEU, METEOR and CIDEr . Qualitative results are shown for the generation of novel image captions. We also evaluate the bi-directional ability of our algorithm on both the image and sentence retrieval tasks. Since this does not require the ability to generate novel sentences, numerous previous papers have evaluated on this task. We show results that are better or comparable to previous state-of-the-art results using similar visual features.

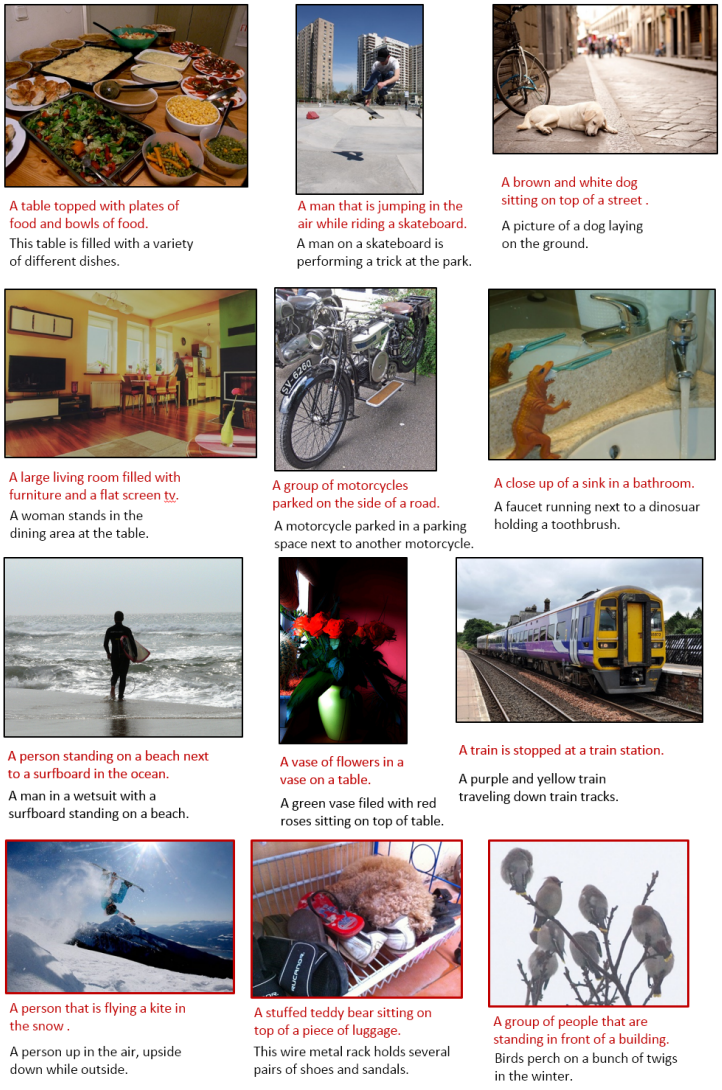

Figure 1: Several examples of generated captions (red) and human generated captions (black). Last row shows failure cases.

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

[2] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.(JAIR)*, 47:853–899, 2013.

[3] Marcel Adam Just, Sharlene D Newman, Timothy A Keller, Alice McEleney, and Patricia A Carpenter. Imagery in sentence comprehension: an fmri study. *Neuroimage*, 21(1):112–124, 2004.

[4] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014.

[5] Lewis R Lieberman and James T Culpepper. Words versus objects: Comparison of free verbal recall. *Psychological Reports*, 17(3):983–988, 1965.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[7] Allan Paivio, Timothy B Rogers, and Padric C Smythe. Why are pictures easier to recall than words? *Psychonomic Science*, 11(4):137–138, 1968.

[8] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using Amazon's mechanical turk. In *NAACL HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[9] Richard Socher, Q Le, C Manning, and A Ng. Grounded compositional semantics for finding and describing images with sentences. In *NIPS Deep Learning Workshop*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.