

Stacked Cross Attention for Image-Text Matching

Kuang-Huei Lee¹, Xi Chen¹, Gang Hua¹, Houdong Hu¹, and Xiaodong He²

¹ Microsoft AI and Research

fkuailee, chnxi, ganghua, houhuG@microsoft.com

² JD AI Research

xiadong.he@jd.com

Abstract. In this paper, we study the problem of image-text matching. Inferring the latent semantic alignment between objects or other salient stuff (e.g. snow, sky, lawn) and the corresponding words in sentences allows to capture fine-grained interplay between vision and language, and makes image-text matching more interpretable. Prior work either simply aggregates the similarity of all possible pairs of regions and words without attending differentially to more and less important words or regions, or uses a multi-step attentional process to capture limited number of semantic alignments which is less interpretable. In this paper, we present Stacked Cross Attention to discover the full latent alignments using both image regions and words in a sentence as context and infer image-text similarity. Our approach achieves the state-of-the-art results on the MS-COCO and Flickr30K datasets. On Flickr30K, our approach outperforms the current best methods by 22.1% relatively in text retrieval from image query, and 18.2% relatively in image retrieval with text query (based on Recall@1). On MS-COCO, our approach improves sentence retrieval by 17.8% relatively and image retrieval by 16.6% relatively (based on Recall@1 using the 5K test set). Code has been made available at: <https://github.com/kuanghuei/SCAN>.

Keywords: Attention, Multi-modal, Visual-semantic embedding

1 Introduction

In this paper we study the problem of image-text matching, central to image-sentence cross-modal retrieval (i.e. image search for given sentences with visual descriptions and the retrieval of sentences from image queries).

When people describe what they see, it can be observed that the descriptions make frequent reference to objects and other salient stuff in the images, as well as their attributes and actions (as shown in Figure 1). In a sense, sentence descriptions are **weak annotations**, where words in a sentence correspond to some particular, but unknown regions in the image. Inferring the latent correspondence between image regions and words is a key to more interpretable image-text matching by capturing the fine-grained interplay between vision and language.

Work performed while working at Microsoft Research.
