Exploring Associative Localization for Image-Text Retrieval

Noor Mohamed Ghouse
UW Madison
mohamedghous@wisc.edu

Shashank Verma
UW Madison
sverma28@wisc.edu

Deepan Das
UW Madison
ddas27@wisc.edu

Abstract

Image Retrieval captures how well the model has learnt to semantically bridge the text and image modalities. Some previous works have demonstrated that a pooled representation over an entire input image used for training can be un-pooled for localized analysis. We believe that such retrieval tasks will pave the way for unsupervised text-visual associative localizations. A recent work has been able to derive such multimodal associative localizations across the audio-visual modalities, and we intend to extend it to the text-visual modalities.

1. Introduction

In Computer Vision, Image-Text Retrieval is the problem of searching images in a database using a text descriptor or vice versa. This is different from simply using meta-data like tags and other keywords to search a database in the sense that it involves analyzing the semantic content of the concerned image. Some previous works [1] have used a two-branch matching network to jointly represent different modalities like text and image in the same embedding space, which has a computable distance metric. This metric can be used to evaluate the similarity across different image-text pairs, which has further applications like cross-modal retrieval, object localization, visual grounding, etc. These methods belong to the category of Supervised Learning methods and depend on the quality of the embeddings produced for the given data. The underlying advantage in using such methods is the fact that we do not need any distance measure in the input space. Moreover, these methods learn a globally coherent non-linear function that maps the data to the output manifold and can learn mappings that are invariant to certain transformations of the input data [2]. Elaborate CNNs have enabled us to extract such complex non-linear representations of the provided image data. Similarly, on the other hand, we have efficient representations for text data as well. When new samples not seen during the training process are introduced to such a two-branch cross-modal model, it is able to provide a close approximation of the similarity across the text and image data provided.

2. Proposed Method

We generate two different non-linear representations for each image-text pair using a two-branch network. However, unlike previous methods that map the entire image and text to fixed points in an embedding space, we learn representations that are distributed spatially and sequentially, along the span of the text. We believe that this would enable the model to directly co-localize within both modalities [3]. We take inspiration from metric learning literature that try to optimize a ranking based criterion so that images and captions that form a positive pair produce a higher similarity score than the negative image-caption pair. We intend to train our network by using appropriate sampling strategies and the ranking objective. Similarity scores can be evaluated over this co-localized space, instead of the Euclidian distance metric, which has been used previously over a fixed embedding space in previous works. We intend to avoid the loss of spatially localized stimulus caused due to the flattening of the last conv layer in a given CNN model, by retaining the last convolutional activation maps as the ultimate representation of our images. The evaluation of a similarity score now involves the development of a 3rd order tensor called the Matchmap, which captures the similarity between the cross-modal representations. Since it is unrealistic to match all words in a caption simultaneously with an image, we consider the case where we can match each frame of the caption with the most similar image patch and then average over the caption frames. We intend to use the Flickr30K and Flickr8K dataset which has associated captions and tags with each image. Moreover, there is a possibility of extending this work to an unsupervised learning of object localization.

References

[1] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning Two-Branch Neural Networks for Image-Text Matching Tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. c, p. 1, 2018.

[2] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality Reduction by Learning an Invariant Mapping."

[3] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input."