Deep Visual-Semantic Alignments for Generating Image Descriptions

1. Introduction

We develop a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe. Our model associates the two modalities through a common, multimodal embedding space and a structured objective. We validate the effectiveness of this approach on image-sentence retrieval experiments in which we surpass the state-of-the-art.

We introduce a multimodal Recurrent Neural Network architecture that takes an input image and generates its description in text. Our experiments show that the generated sentences significantly outperform retrieval- based baselines, and produce sensible qualitative pre- dictions. We then train the model on the inferred correspondences and evaluate its performance on a new dataset of region-level annotations.

2. Related Work

Dense image annotations.

Generating descriptions.

Grounding natural language in images.

Neural networks in visual and language domains.

3. Our Model

**Overview.**

The ultimate goal of our model is to generate descriptions of image regions. During training, the input to our model is a set of images and their corresponding sentence descriptions (Figure 2). We first present a model that aligns sentence snippets to the visual regions that they describe through a multimodal embedding. We then treat these correspondences as training data for a second, multi- modal Recurrent Neural Network model that learns to generate the snippets.

3.1. Learning to align visual and language data

3.1.1 Representing images

3.1.2 Representing sentences

3.1.3 Alignment objective

Since the supervision is at the level of entire images and sentences, our strategy is to formulate an image-sentence score as a function of the individual region- word scores.

3.1.4 Decoding text segment alignments to images

3.2. Multimodal Recurrent Neural Network for generating descriptions

RNN training. The RNN is trained to combine a word (xt), the previous context (ht−1) to predict the next word (yt)

RNN at test time. To predict a sentence, we compute the image representation bv, set h0 = 0, x1 to the START vec- tor and compute the distribution over the first word y1. We sample a word from the distribution (or pick the argmax), set its embedding vector as x2, and repeat this process until the END token is generated.

3.3. Optimization

4. Experiments

Datasets.

Data Preprocessing.

4.1. Image-Sentence Alignment Evaluation

4.2. Generated Descriptions: Fulframe evaluation

4.3. Generated Descriptions: Region evaluation

4.4. Limitations

5. Conclusions