Review 2 - 1

A Hierarchical Approach for Generating Descriptive Image Paragraphs

Summary

This paper provides a model to generate entire paragraphs for image caption. The paragraphs are more unified and descriptive than sentence-level captions. This occurs because sentence-level captions always describe the image as a whole, which might omit the details about scene elements. However, the model proposed by this paper detects regions of interest first and then descript regions with different sentences respectively, whose language has better diversity and is more human-like.

In Section 1, the author argues what the limitations of sentence-level captions are and how paragraph-level caption can address these problems. In Section 2, the paper presents the different methods of images captions, state-of-the-art sub-region based image captions and different Hierarchical Recurrent Networks which are used to generate language in both sentence and word level. Section 3 analyzes the difficulties and advantages of generating paragraphs. In Section 4, the details about the model are described. Firstly, by using transfer learning, this paper extracts features from images and detects objects by a pre-trained Region Proposal Network. Secondly, the model uses max-pooling to project several region feature vectors to the same feature space. This pooled vector is used to train a hierarchical recurrent network to generate paragraphs. The hierarchical recurrent network contains two level RNNs. One is Sentence RNN which is used to predict the number of sentences and provide a topic for each sentence. The other is Word RNN which learns how to generate one relative sentence based on the topic. Section 5 describes benchmarks, experiments and quantitative results conducted by the authors. It concludes the language generated by the model described in this paper not only performs very well in image-caption metrics, but also is more human-like.

Main Contribution

1. Propose a model that decomposes both images and paragraphs into their constituent parts, detecting semantic regions in images and using a hierarchical recurrent neural network to reason about language.
2. Prove paragraph-level captions are more descriptive and human-like.
3. Anticipate further opportunities for knowledge transfer in tasks at the intersection of vision and language
4. Demonstrated the benefits of our model in interpretability, showing how to generate descriptive paragraphs using only a subset of image regions.

Pros and Cons

Pros: 1. Generative paragraph-level captions which are more detailed and human-like 2. Leverage pre-trained region detection network and word-embedding weights. 3. Prove the interpretability of the model.

Cons: I personally think there should a parameter can used to scale “description-range”. For example, if this parameter is small, the model only generates description from top few regions, vice versa. After implement that, we can analyze the performance of the model with different “description-range” parameter to see if larger “description-range” gives a better performance.

Review 2 - 2

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Summary

This paper provides an attention-based model for image caption, which means this model can not only describe the image with natural language, but also can provide information about which part of the images is responsible to each word in the caption sentences. The paper also proves the model has better performance and interpretability.

In Section 1, aside from introducing methods in image captions, the paper also explains one of the most important meanings of attention-based model is that attention can extract vital information when the images are full of clutter. Section 2 briefly explains the reason why this paper doesn't use object detection methods is learning latent alignments from scratch allows the model to go beyond basic concept of objects and learn to attend to abstract concept. In Section 3, this paper gives details about their model. For the "Encoder" part, the model uses pre-trained CNN to extract several features from one raw image. More importantly, they extract features from previous low-level CNN layers instead of the final fully-connected layers. This allows the decoder to selectively focus on certain parts of an image by selecting a subset of all the feature vectors. As for the "Decoder" part, they use a designed LSTM-RNN to generate language based on the image features and the context. The challenge is how to determinate which location has large responsibility for a specific word in captions. To address this problem, the paper proposes two versions of mechanisms to calculate the context vector. Section 4 goes deep to how to learn attention weights of different locations' feature vectors. For the **stochastic "Hard" attention**, in which the attention weights can be interpreted as the probability that this location is the right place to focus for producing the next word, the papers presents a method by learning the vibrational lower bound of likelihood of observing word sequences to estimate the gradient. Further, the model can learn the weights better by adding an entropy term on the multinouilli distribution. For the **deterministic "Soft" attention**, the paper takes the expectation of the context vector directly by add all the product of feature vectors and weights. In the paper, the authors also optimize the learning process with skillful constrain on the weights. In Section 5, the papers validate that their model gets to state of the art in image caption area. Moreover, they visualize how the model gazes on salient objects, which is very meaningful for further research.

Main Contribution

1. Introduce two attention-based image caption generators under a common framework. Both of them are proved to be theoretically efficient.
2. Show how we can gain insight and interpret the results of this framework by visualizing "where" and "what" the attention focused on.
3. Quantitatively validate the usefulness of attention in caption generation with state of the art performance

Pros and Cons

Pros: 1. Describe important info by focusing on "attention". 2 Doesn't use object detection, but learn to attend to abstract concept. 3. By using many generalization and training speed optimizations, the training should be faster. 4. Cons: 1. Didn't ensemble the results from different feature extractors.