Layering Method for Generating Image Paragraphs

Lina Zhu

June 16 ,2018

Abstract

In this article we introduced the latest developments in image captioning to make it possible to generate a new sentence language that describes images in nature, but compressing the image into a single sentence can only describe visual content in rough details. And a new subtitle method, dense subtitles, may use more subtitles to describe the area within the image of the finer detail image, which in turn is a coherent story of an image that cannot be produced. We have overcome this by describing the image by generating the entire paragraph and displaying a detailed, unified story. We develop their components that decompose images and paragraphs into model models, detect semantic regions in the image, and use hierarchical recursive neural networks to reason about language. Language analysis confirmed its complexity segment generation tasks and conducted thorough experiments to demonstrate the effectiveness of our approach on new data sets of image and paragraph pairs.

1. Introduction

Vision is the main sensory form of human perception. Language is the most powerful tool and world we communicate. Build systems that can perform at the same time understand visual stimuli and describe them in nature so that language is a core issue and AI in computer vision. With the advent of large datasets pairing images with natural language descriptions,it has recently become possible to generate new sentences describing images[1]. Although the success of these methods is encouraging, they all share a key constraint: the details. By only describing the image there is only one high-level sentence, which is the upper limit of the number and quality of the underlying information methods that can be produced. The most recent alternative to subtitle selection is the condensed captioning task[2], which overcomes this limitation by detecting many regions of interest in the image, describing each one with a phrase. By extending task object detection including natural language descriptions, condensed subtitles describe images in more detail than standard image captions. However, this is at cost: the description generated for intensive subtitles is not consistent, ie they do not form a cohesive whole describing the entire image.

In this paper, we address two traditional shortcomings of image subtitles and recently proposed compact image subtitles, and describe the images richly by introducing the task of generating paragraphs (Figure 1). Producing paragraphs for images is challenging and requires both delicate image understanding and long-term language reasoning. In order to overcome these challenges, we propose a model that decomposes images and paragraphs into a model of their components: we break down the image into semantics by detecting meaningful interest in the object and other regions, and we use language to analyze language recurrence. Neural networks break down paragraphs into their corresponding sentences. In addition, we have also demonstrated that this is the first time that the ability to transfer visual and linguistic abilities has come from subtitles in large regions, where we have demonstrated the ability to improve paragraph generation.

2. Related Work

Image titles establish connections between visions and text data is always a long-term target vision in the computer. One line of work treats the problem as a ranking task, using images to retrieve relevant titles from the database and vice versa [3]. Due to the nature of the constituent language, any database may not contain all possible image titles; therefore, another line of work focuses on generating subtitles directly. Earlier work used handwriting templates to generate language[4], and more recent methods trained on neural network language models based on image features[1] and generated text from their samples. Similar methods have also been applied to generate titles for videos[5]. A few methods of image subtitles do not apply only to the entire image, but also to the image area. Our approach is about the semantics of the regions in the image, they are all able to pass information from these regions and lead to more explanation.

We quantify these observations as well as other various

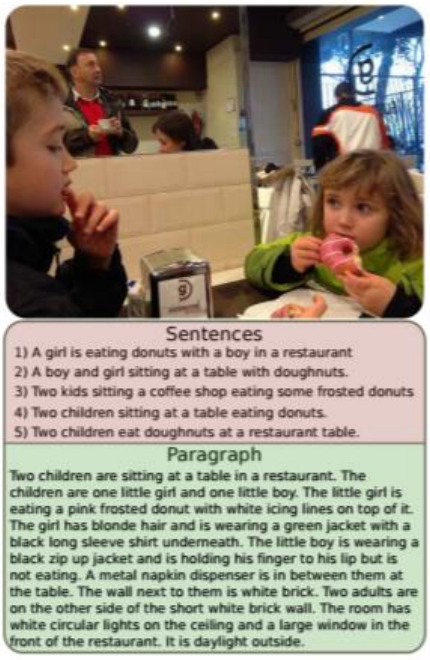

Figure 1. Paragraphs are longer, more informative, and more linguistically complex than sentence-level captions.

observational language statistics in Table 1. For example, we found that each paragraph is approximately six times the sentence heading of the average time, and that each sentence in a single sentence is equivalent to a sentence-level subtitle. In order to test the problem of sentence diversity, a paragraph was collected between each image and each individual sentence. Through this observational measure, the difference in diversity is that the sentence in the astonishing paragraph is more diversified than the sentence subtitle. Quantitative evidence shows that the sentence provides more information about the image in the paragraph. Incorporate part of speech into higher-level language categories. Table 1 gives some common parts. As a ratio, paragraphs have more verbs and pronouns,the frequency of adjectives is quite similar, and there are fewer nouns. In view of the nature of the paragraph, this makes the description of a longer term beyond the existence of a few significant objects and contains information and relationships about its attributes. We also noticed, but did not quantify these paragraphs to show more complex linguistic frequency phenomena, such as the commonality that occurred in Figure 1, we believe these type of long-range phenomenon is a descriptive paragraph of basic attributes and human language. It cannot be fully explored with sentence-level titles.

3. Conclusion

In this article, we introduce the task of describing images with long descriptive paragraphs, and propose a layered approach for generation using the combined image and language structure. We have shown this segment to generate different images with traditional subtitling and customizing our model to accommodate these differences. Experimentally, we have proved its advantages over the traditional image subtitling method and demonstrate how regional knowledge can be effectively transferred to paragraph titles.

Table 1. Statistics of paragraph descriptions, compared with sentence-level captions used in prior work.

	Sentences	Paragraphs
Description Length	11.30	67.50
Sentence Length	11.30	11.91
Diversity	19.01	70.49
Nouns	33.45%	25.81%
Adjectives	27.23%	27.64%
Verbs	10.72%	15.21%
Pronouns	1.23%	2.45%

References

[1] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015. 1

[2] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 1

[3] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 1

[4] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, 2011. 1

[5] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, 2016. 1