Vizwiz Image Captioning based on AoANet with Scene Graph

Suwon Kim, HongYong Choi, JoongWon Hwang, JangYoung Song, SangRok Lee, TaeKang Woo
Modulabs AI
swkim0512@gmail.com, yong6600@gmail.com, angelic805@gmail.com,
sib621@naver.com, lsrock1@naver.com, wtk1101@gmail.com

Abstract

Recently, in the field of Vision, a recognition task study has been conducted to link Computer Vision with Natural Language Process. One of them is the Image Captioning study. In this paper, we propose an algorithm for Image Captioning and analyze the dataset. We introduce an open image captioning database composed of images taken by the blind. It uses the Attention method, which is a typical captioning algorithm. Specifically, it deals with the Attention on Attention for Image Captioning [2] algorithm in detail. AoANet effectively reflects images and word tokens using the transformer's self-attention structure. Simultaneously, use Scene Graph to further infer the relationship between objects. We call this AWS model. When this algorithm is applied to VizWiz, 60.5 CIDEr scores can be obtained. We publicly-share the dataset with captioning challenge instructions at https://vizwiz.org.

1. Introduction

1.1. Image Captioning

The importance of Visual Scene understanding in Computer Vision has steadily emerged. Among them, Image Captioning is the task of predicting appropriate captions for a given image. In other words, it is a task to describe an image. Natural Language Process and Computer Vision are used simultaneously. What is different from VQA, which is a similar task, is that VQA questions are given, and Image Captioning creates a description through images only. By creating a description of an image, the study can help the search area by allowing picture or image-based content to be sorted and requested in a new way

1.2. Our goal

We hope this study will mean something to the blind. It is hoped that this study will help people with visual impairments overcome their everyday visual problems or enhance their social accessibility. For example, the study may provide information about what is happening in front of the blind, or what information the blind person wants. It also hopes to educate many people about many technical needs for the visually impaired and at the same time contribute to developing technologies that can make the social accessibility of the visually impaired better

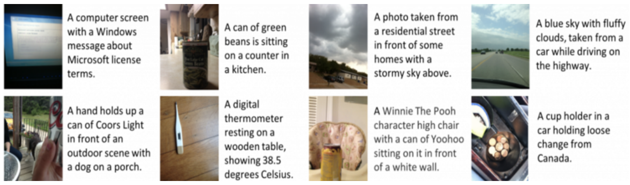
A computer screen with a Windows message about Microsoft license terms.
A can of green beans is sitting on a counter in a kitchen.
A photo taken from a residential street in front of some homes with a stormy sky above.
A blue sky with fluffy clouds, taken from a car while driving on the highway.
A hand holds up a can of Coors Light in front of an outdoor scene with a dog on a porch.
A digital thermometer resting on a wooden table, showing 38.5 degrees Celsius.
A Winnie The Pooh character high chair with a can of Yoohoo sitting on it in front of a white wall.
A cup holder in a car holding loose change from Canada.

Figure 1. Example of a Captured Image in Vizwiz Captions.

1.3. What's model

There have been countless studies for Image Captioning. Attention on Attention for Image Captioning [2] uses Attention Mechanism to improve Image Captioning performance through Attention on Attention Module. We grafted the Scene Graph here to better understand the association between image objects. The Caption was refined through the Scene Graph to add relativity in the form of a graph, so that it could focus more on certain parts and be described in more detail.

2. Related Work

2.1. Vizwiz Captions

This dataset, called Vizwiz Captions, is an image dataset taken by a blind person for about 10 years. This dataset consists of 39,181 images originating from people who are blind that are each paired with 5 captions. Most of the images are out of focus, objects are out of frame, too bright or too dark, etc. This is actually a problem that occurs because the visually impaired are photographed, and these problems are sufficiently real to occur. Therefore, we tried to train the model by using all the data as it is without preprocessing the image.

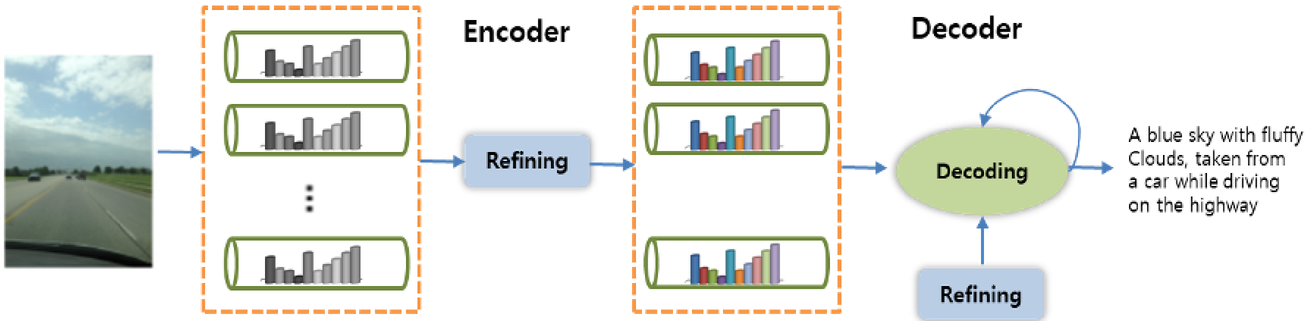

Figure 2. Overview of the encoder/decoder framework of AoANet with Scene Graph.

2.2. Attention on Attention

The attention Mechanism [1, 4] is widely used in encoder and decoder Framework such as Machine translation and QA task. However, decoder knows little about or how well attention results is related to query. The results expected by the attention result and those expected by the decoder are different, which may lead to incorrect results. To solve this problem, AoA (Attention on Attention) [2] is proposed, which extends the existing Attention Mechanism by adding another Attention. Apply AoA to both encoder and decoder to better model the relationship between objects in the image.

2.3. Self-critical Sequence Training for Image Captioning

Self-critical Sequence Trainin [3] uses reinforcement learning to optimize Image Captioning. The main idea of the SCST is to set the result of the algorithm used in the test as the basis for the enhanced learning algorithm. Using own test- time algorithms, only positive weights are given to samples of models with better performance than the current test time system, and inferior samples are suppressed.

$$\hat{w}_t = \arg\max_{w_t} p(w_t|h_t) \tag{1}$$

Generalize this to minimize the impact of the baseline on training time through the test time reasoning algorithm, and to train the system to optimize for fast and greedy decoding at test time. Using own test-time algorithms, only positive weights are given to samples of models with better performance than the current test time system, and inferior samples are suppressed. This allows them to train more efficiently at mini-batches.

2.4. Scene Graph

Recently, beyond simply obtaining the existence and location information of objects through Object Detection, reasoning the nature and relationship of objects has emerged.

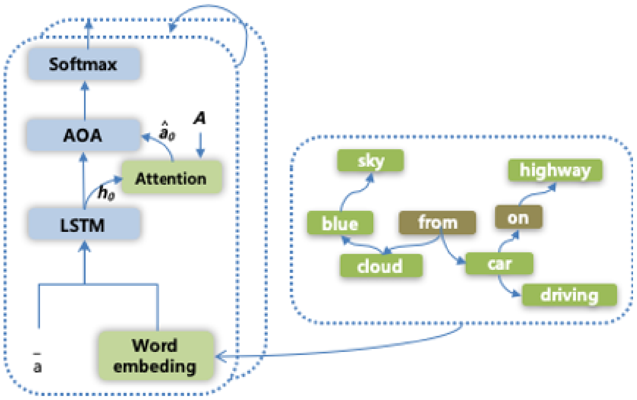

Figure 3. Decoder of AoANet with Scene Graph.

It expressed this in a manner that is easy to express and intuitively understand through Graph. This structure is called the Scene Graph [6].

3. Method

We applied the association between results and query by expanding the existing Attention Mechanism using AoA Module. In addition, create a graph of the relationship between objects and objects based on Caption. Create an image captioning with the shape of the generated Scene Graph.

3.1. Feature extraction

We first extract feature factor $A = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ based on CNN or R-CNN. $\alpha_i \in R^D k$ is the number A and D is the dimension of each.

3.2. Encoder

Refine the features through the refine network of the AoA module based on the extracted features. Refined feature vectors look for interactions between objects in the image and measure how well they are related.

Model	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	METEOR	CIDEr	SPICE
BUTD with COCO	75.58	59.69	46.05	35.48	55.91	26.62	110.5	
AWS with Vizwiz	65.89	47.43	33.28	23.15	46.64	19.95	60.50	15.16

Table 1. Test results using coco dataset and vizwiz dataset.

3.3. Decoder

We first extractDecoder generates Caption y with encoded vector A. Context Vector c_t is modeled to calculate conditional probabilities for vocabulary.

$$P(y_t|y_{1:t-1}, I) = softmax(W_p c_t) \tag{2}$$

Where $W_p \in \mathbb{R}^{D\times|\Sigma|}$is the size $|\sum|$ and weight parameter of vocabulary. In addition, we used Caption, which derived relationships based on Graph in Caption, as Word Embedding input. Through Scene Graph in Caption, we want to derive a relationship between better objects to achieve better Captioning results.

3.4. Training

We first extractFirst, train AoANet by optimizing Cross Entropy.

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^*|y_{1:t-1}^*)) \tag{3}$$

The CIDEr is then optimized through Self-critical Sequence Training.

$$L_{RL}(\theta) = -E_{y_1:T^{\sim p(\theta)}}\left[r(y_{1:T})\right] \tag{4}$$

CIDEr [5] is used reward .

$$\nabla_\theta L_{RL}(\theta) \approx -(r(y_{1:T}^s) - r(y_{\hat{1:T}}))\nabla_\theta \log p_\theta(y_{1:T}^s) \tag{5}$$

This allows us to calculate the gradient. web page for a discussion of the use of color in your document.

4. Experiment

We experimented with COCO Dataset and Vizwiz Captions. This dataset, called Vizwiz Captions, differs from the existing datasets. It consists of 39,181 images, and this set of data is captured by the blind, (1) often of poor quality, and (2) paired with five captions. We tried to solve the Image Captioning problem using this data The proposed methods were evaluated and compared with other methods using SPICE, METEOR, ROUGE-L and CIDEr, including BLEU.

5. Conclusions

We first extract We offer Image Captioning Algorithm that can help blind people understand their visual environment. This was achieved by applying the Scene Graph to AoANet.

Acknowledge

This work was supported by crowdworks.

References

[1] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

[2] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643, 2019.

[3] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[4] Ronald A Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

[5] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[6] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.