Look Deeper See Richer: Depth-aware Image Paragraph Captioning

Ziwei Wang* · Yadan Luo* · Yang Li
Zi Huang · Hongzhi Yin

School of Information Technology and Electrical Engineering,
The University of Queensland, Australia

Problem

Describing an image by a full paragraph involves organising sentences **orderly**, **coherently** and **diversely**, inevitably leading higher complexity than by a single sentence.

Existing image paragraph captioning approaches limit themselves in 2D recognition, and therefore lack cognition on locative relationship between objects in realistic 3D space. Besides, object-detection based methods require extra tedious and expensive human labelling, and over-lapped bounding boxes may result in redundant representations of a same object.

Therefore, **order**, **coherence**, **diversity** and **relevance** of generated paragraph captions still need to be improved.

Framework

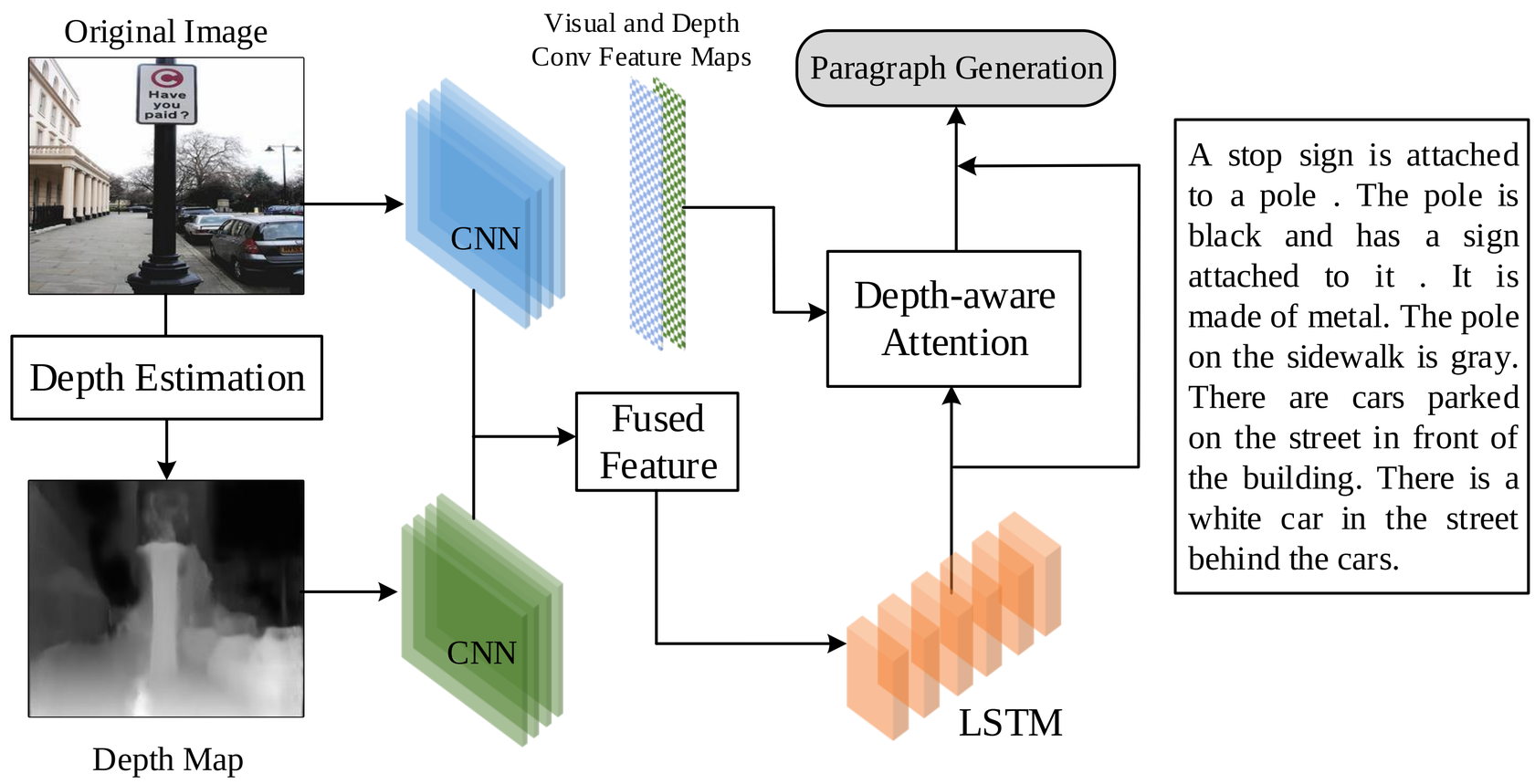

Figure 1. An overview of the framework. Given an image, its depth map is generated after the process of depth estimation. The visual representation and depth representation of the image are generated simultaneously based on its raw RGB image and the depth map through a dual-stream CNN. A fused feature is further generated as a depth-aware visual representation, which is fed into an LSTM model. A novel depth-aware attention mechanism is designed to select an attended area at each time-step by taking into account both the visual and depth information.

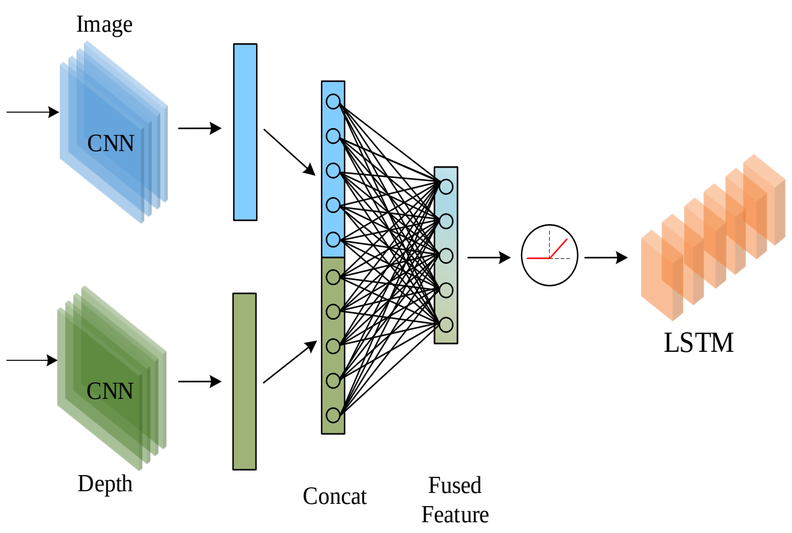

Figure 2. Feature Fusion Layer

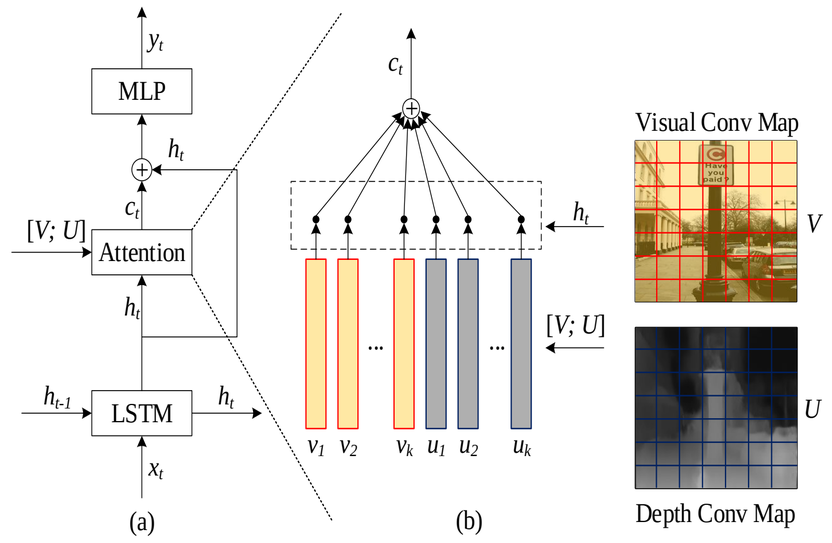

Figure 3. Depth-aware Attention Model

Contribution

This paper introduces a **Depth-aware Attention Model (*DAM*)** to generate paragraph captions for images.

The **depth map** assists the model to recognise the subtle and locative relationships between objects and generates well-organised paragraph in a logical and coherent way.

By incorporating the **attention mechanism**, the learned model swiftly shifts the sentence focus during paragraph generation, whilst avoiding verbose descriptions on a same object.

Conclusion

We propose a deep **depth-aware** framework to strengthen image paragraph captioning by enriching raw data with extra geometric information.

Acknowledgement

This work is partially supported by ARC FT130101530 and NSFC No. 61628206.

Experiments

Evaluation experiments on the benchmark image paragraph annotation dataset by comparing our method with the baseline and state-of-the-art models. A comprehensive user study is also conducted to discuss the performance of captioning achieved by different models from the user perfective in terms of five criteria.

Model	M	C	B-3	B-4
Sentence-Concat (Neuraltalk)	12.1	6.8	7.6	4.0
Sentence-Concat (NIC)	9.3	7.1	4.9	2.3
Image-Flat (NIC)	13.4	14.7	10.9	6.0
Region-Hierarchical	**14.8**	11.3	8.5	4.0
Depth-aware Basic	13.4	15.9	11.6	**6.7**
Depth-aware Attention	13.9	**17.3**	**11.7**	6.6
Human	19.2	28.6	15.6	9.7

Table 1. Evaluation using METEOR, BLEU and CIDEr on VG dataset.

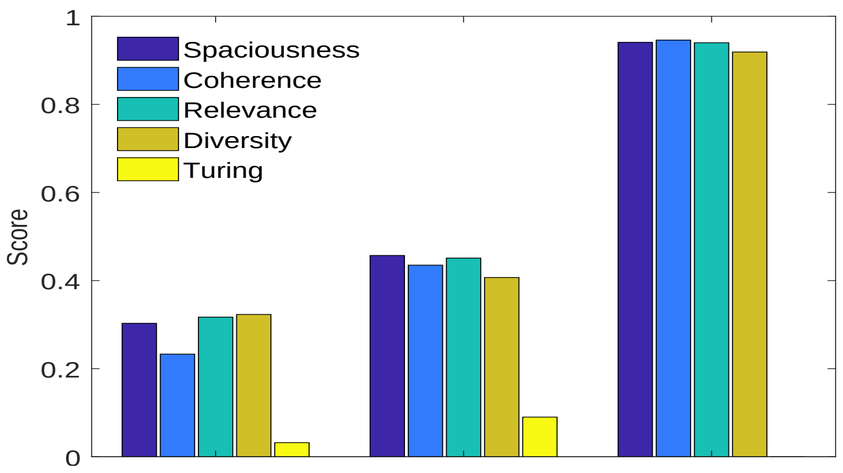

Figure 4. User Study

***Co-first Authors**