https://doi.org/10.7840/kics.2023.48.8.950

Image Captioning Network for Disaster Dataset Based on Transfer Learning and Disaster Convolution Block

Jeong-hun Hong◆, Dong-hun Lee*, Han-gyul Baek*, Byungjun Bae**, Sang-hyo Park°

ABSTRACT

Image captioning has been widely studied using various deep learning models on extensive and well-prepared datasets. Accurately captioning images of serious and sudden disasters is important; however, the study of disaster image captioning is yet to be thoroughly investigated compared to natural image captioning. Furthermore, existing image captioning models may need to perform better in generating captions for disaster images because there are fewer disaster images in popular datasets than non-disaster images. To address these problems, we refine and propose a cleaned disaster dataset and an image captioning model optimized for the dataset. Experimental results showed that our proposed model outperformed the existing model in terms of generating accurate captions for disaster images.

Key Words : image captioning, dataset, transfer learning, test retrieval, deep learning

Ⅰ. Introduction

Image caption generation is one of the main fields of computer vision, involving object recognition in an image and representing the relationship between objects through natural language. Image captioning, which automatically generates sentences that describe images, poses significant challenges, as it requires learning both image recognition and natural language expression. To address this problem, a neural image caption (NIC)[1] generator was proposed, which is an end-to-end system that uses a convolutional neural network (CNN) as an encoder and a recurrent neural network-based module as a decoder. In addition, self-critical sequence training (SCST)[2] using global CNN features was proposed, which is a fully connected layer model that encodes images using ResNet-101[3] while retaining the original dimensions. Extensive research [1,2] has been conducted on image captioning. Transformer-based large-scale models, such as contrastive language-image pretraining (CLIP)[4], which provide images and text as a pair of inputs and trains multi-modal embedding space, are emerging. Therefore, the use of advanced neural network models and large datasets[1,2,4] have improved image captioning significantly. Despite these advancements, previous image captioning models do not generate accurate captions on several topics, especially in disaster situations[5]. In addition, the lack of datasets representing disaster situations are one of the underlying causes of low performance.

We propose an image captioning model optimized for disaster-situation images to overcome these limitations. Based on disaster datasets generated from five datasets, we implemented an NIC-based model. Our model showed relatively high performance in captioning images in disaster datasets compared with previous models.

Disaster image captioning can improve disaster preparedness and recovery efforts by providing real-time information about affected areas and infrastructure. Developing accurate and efficient disaster image captioning techniques is critical for

※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00083, Development of customized disaster media service platform technology for the vulnerable in disaster information awareness).
◆ First Author : School of Computer Science and Engineering, Kyungpook National University, invhun@knu.ac.kr, 학생회원
° Corresponding Author : School of Computer Science and Engineering, Kyungpook National University, s.park@knu.ac.kr, 정회원
* School of Computer Science and Engineering, Kyungpook National University, hy05205@naver.com; qorgksruf123@gmail.com
** Media Research Division, Electronics and Telecommunications Research Institute, 1080i@etri.re.kr, 정회원
논문번호 : 202305-099-A-LU, Received May 11, 2023; Revised May 28, 2023; Accepted May 28, 2023

disaster response and management efforts that benefit stakeholders, such as emergency responders, public officials, and the general public, including people with visual disabilities.

Ⅱ. Proposed method

2.1 Dataset refining

We trained the model using a dataset with "disaster" as a keyword to create a specialized model for disaster data. However, no specialized public data exist related to disasters. Therefore we generated our own dataset using the following process (see Fig. 2). First, we set disaster-related keywords, which were synonyms for words that effectively express disaster situations such as wildfires, floods, and storms. Using these keywords, we extracted images and captions that contain sentences related to disasters to create data by referring to the annotation files of five datasets: MS-COCO[6], Flickr30K[7], VizWiz[8], ADE20K[9,10], and Open Images dataset v6[11,12]. Despite this filtering, many outlier data points included disaster keywords unrelated to a disaster situation. Therefore, the remaining outliers were manually removed.

Another inevitable problem was the variation in the number of captions from one data point to another because the data were imported from several datasets. To address this issue, we duplicated each image in the dataset based on the number of captions per image and numbered the duplicated images to ensure each image had a caption. The datasets was divided into training, validation, and test sets at an 8:1:1 ratio, ensuring that images duplicated from the same image but with different captions were included in the same set. As a result, a disaster dataset with 2,130 data points was generated.

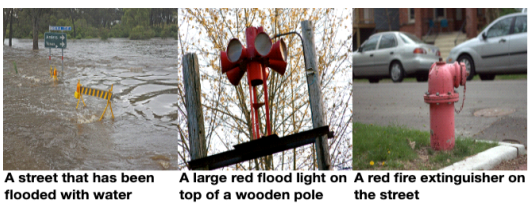

Fig. 1. Example of disaster data.

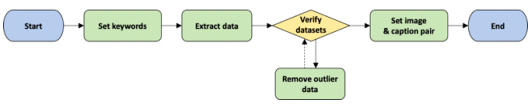

Fig. 2. Dataset refining flowchart.

2.2 Architecture

2.2.1 Base model

In this study, we used an NIC with visual attention[13] as the base model. The NIC model has been widely used in image captioning and has performed well. It consists of a CNN-based encoder that extracts image features and a long short-term memory (LSTM) network decoder[1,14], which generates captions one word at a time. In addition, an attention[15] mechanism is added to the LSTM to focus on salient features. The attention mechanism enables the model to selectively focus on relevant image regions during caption generation, thereby improving the quality of the generated captions. This model was chosen as the base model because of its simplicity and strong performance, making it a suitable starting point for the proposed method.

2.2.2 Model detail

The training process was divided into two steps. First, we trained the base model using the MS-COCO dataset. Subsequently, the trained base model was used as a pretrained model for transfer learning on the disaster dataset. To improve the performance, we tested several CNN image feature extraction modules during training and finally selected ResNet-101 as the encoder for the base model because it showed better performance (see Table 1). As shown in the first image of Fig. 1, unlike common images, disaster images are characterized by similar visual features throughout the image. Consequently, we introduced a new disaster convolution block (D-conv) to the encoder of the base model for transfer learning. D-conv consisted of three 3 × 3 convolution layers with increasing numbers of filters, followed by batch normalization and ReLU. Finally, we used a 1 × 1 convolution layer at the end to adjust the number of channels. Fig. 3 illustrates the overall model structure, including D-conv.

Table 1. Performance comparison on disaster datasets between baseline and proposed methods.

Encoder model	BLEU-1(↑)	BLEU-2(↑)	BLEU-3(↑)	BLEU-4(↑)	KS (↑)
Pretrained Baseline +ResNet[3]	0.0457	0.0061	0	0	0.125
Disaster-trained Baseline[13]	0.3000	0.1682	0.0949	0.0541	0.0769
Disaster-trained Baseline +EfficientNetV2 [16]	0.2478	0.1360	0.0737	0.0407	0.0868
Disaster-trained Baseline +ResNet[3]	0.2847	0.1685	0.1032	0.0596	0.1971
Disaster-trained Baseline +ResNet[3] +D-conv	0.3269	0.1892	0.1171	0.0732	0.3028

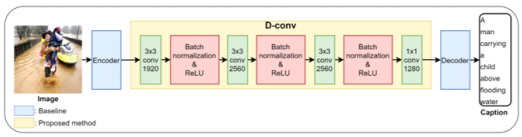

Fig. 3. Overview of the proposed captioning network for disaster.

Ⅲ. Experiments

3.1 Environments

In our experiments, we used the MS-COCO dataset for pretraining and the disaster dataset for transfer learning. MS-COCO consisted of 82,783 training images with five reference sentences per image. The disaster dataset contained 1,714 training images and one caption per image. The experiment was conducted on Ubuntu 18.04 with one RTX3060 GPU, taking 20 h for pretraining and 40 min for transfer training.

The results were reported using the BLEU metric and keyword scores. The BLEU score[17] is a precision calculation value of word n-grams between a generated caption and reference sentence and is widely used as an index in general image captioning. However, BLEU may be insufficient for evaluating the relevance of generated captions to the disaster; therefore, we devised a new metric: the keyword score. The keyword score represents the percentage of generated captions containing disaster-related keywords from the image. We computed the score S_i for the i-th sample using the following equation:

$$S_i = \begin{cases} 1, & if \ \ k_i \in c_i \\ 0, & if \ \ k_i \notin c_i \end{cases}, \tag{1}$$

where k_i denotes the disaster keyword of i-th ground truth, and c_i denotes the generated caption of the i-th image. Consequently, the keyword score per dataset, KS, is defined as follows:

$$KS = \frac{\sum_{n=1}^{N} S_n}{N}, \tag{2}$$

where N denotes the dataset size.

3.2 Result

To evaluate the performance of the proposed method in disaster image captioning, we compared it with several models, including the baseline, using a disaster dataset. Table 1 shows that our method achieved at least a 0.013 improvement in terms of BLEU, particularly a significant 0.04 improvement in terms of BLEU-1. Additionally, our method improved the keyword score by 0.1, indicating that it generated captions similar to the reference sentence and more relevant to the disaster. To emphasize the necessity of a disaster dataset, we compared the MS-COCO-pretrained model with the model trained using the disaster dataset. The model trained using additional disaster data consistently outperformed the MS-COCO-pretrained model across all evaluation metrics. Furthermore, according to Fig. 4, our method

Fig. 4. Disaster image captioning examples in the following order of captions: 1) ground truth, 2) proposed method, 3) disaster-trained baseline with ResNet, and 4) pretrained baseline with ResNet, from top to bottom.

outperformed the baseline model, which was observed by comparing the image captioning results for two specific disaster dataset images and generating captions that accurately describe disaster situations. By contrast, the pretrained model may fail to do so.

Ⅳ. Conclusion

In this study, we proposed a new disaster dataset and an optimized model for capturing disaster images. To capture the disaster image accurately, we created a D-conv block that combined 3 × 3 and 1 × 1 convolution layers, resulting in better extraction of the characteristics of such types. The results showed that the proposed method improves caption generation performance for disaster images. Therefore, the proposed method can help capture disaster images more accurately and possibly generate captions for the images.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3156-3164, Apr. 2015. (https://doi.org/10.48550/arXiv.1411.4555)

[2] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in *Proc. IEEE conference on computer vision and pattern recognition* (*CVPR)*, pp. 7008-7024, Nov. 2017. (https://doi.org/10.48550/arXiv.1612.00563)

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE conference on computer vision and pattern recognition* (*CVPR)*, pp. 770-778, Dec. 2015. (https://doi.org/10.48550/arXiv.1512.03385)

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger, “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn., (PMLR)*, vol. 139, pp. 8748-8763, Jul. 2021. (https://doi.org/10.48550/arXiv.2103.00020)

[5] Pung-Hwi Ye, and Chang-Hwan Son, “Image Captioning via Semantic Visual Feature Matching in Heavy Rain Condition,” *The Journal of Korean Institute of Information Technology (JKIIT)*, pp. 19-29, May. 2021. (https://dx.doi.org/10.14801/jkit.2021.19.5.19)

[6] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” *Computer Vision – ECCV 2014: 13th Eur. Conf.*, pp. 740-755, Zurich, Switzerland, Sep. 2014. (https://doi.org.10.1007/978-3-319-10602-1_48)

[7] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Assocation for Computational Linguistics (TACL)*, vol. 2, pp. 67-78, Feb. 2014. (https://doi.org/10.1162/tacl_a_00166)

[8] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, “Captioning images taken by people who are bind,” *Computer Vision-ECCV 2020: 16th Eur. Conf.,* pp. 417-434, Glasgow, UK, Aug. 2020. (https://doi.org/10.48550/arXiv.2002.08565)

[9] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, “Scene Parsing through ADE20K Dataset,” in *Proc. IEEE conference of computer vision and pattern recognition (CVPR)*, pp. 633- 641, Jul. 2017. (https://doi.org/10.1109/CVPR.2017.544)

[10] B. Zhou, H Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through ADE20K dataset,” *International Journal of Computer Vision (IJCV)*, pp. 302-321, Oct. 2018. (https://doi.org/10.48550/arXiv.1608.05442)

[11] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali,

S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision (IJCV)*, Feb. 2020. (https://doi.org/10.48550/arXiv.1811.00982)

[12] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," *Computer Vision-ECCV 2020: 16th Eur. Conf.,* pp. 647-664, Glasgow, UK, Aug. 2020. (https://doi.org/10.48550/arXiv.1912.03098)

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 38th Int. Conf. Mach. Learn., (PMLR)*, vol. 37, pp. 2048-2057, Feb. 2015. (https://doi.org/10.48550/arXiv.1502.03044)

[14] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2625-2634 Nov. 2014. (https://doi.org/10.48550/arXiv.1411.4389)

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Int. Conf. on Learning Representations, ICLR*, May 2015. (https://doi.org/10.48550/arXiv.1409.0473)

[16] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn., (PMLR)*, vol. 139, pp. 10096-10106, Apr. 2021. (https://doi.org/10.48550/arXiv.2104.00298)

[17] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th annual metting of the Assocaition of Computational Linguisitcs*, pp. 311-318, Jul. 2002. (https://doi.org/10.1145/3485766)