engineering proceedings

MDPI

Proceeding Paper

Preliminary Study on Image Captioning for Construction Hazards †

Wen-Ta Hsiao [1], Wen-Der Yu [1,*], Tao-Ming Cheng [1] and Alexey Bulgakov [2]

[1] Department of Civil and Construction Engineering, Chaoyang University of Technology, Taichung 413310, Taiwan; wdshiau@cyut.edu.tw (W.-T.H.); tmcheng@cyut.edu.tw (T.-M.C.)
[2] Department of Automation, Robotic and Mechatronic in Construction, South-Russian State Polytechnic University, Novocherkassk 346428, Russia; agi.bulgakov@gmail.com
* Correspondence: wenderyu@cyut.edu.tw
† Presented at the 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data, Taipei, Taiwan, 19–21 April 2024.

Abstract: Construction accidents are a major contributor to occupational fatalities. To tackle this issue, improved monitoring for hazard elimination is crucial. By introducing a deep-learning image captioning system, we identified hazards via closed-circuit television in construction sites. By leveraging Inception-v3 for feature extraction and the gated recurrent unit for caption generation, real-time hazard monitoring was enabled. Bilingual evaluation understudy scores were determined for continuous and effective hazard detection, helping construction managers enhance safety.

Keywords: construction safety; image captioning; deep learning; closed-circuit television (CCTV) monitoring

check for updates

Citation: Hsiao, W.-T.; Yu, W.-D.; Cheng, T.-M.; Bulgakov, A. Preliminary Study on Image Captioning for Construction Hazards. *Eng. Proc.* **2024**, *74*, 20. https://doi.org/10.3390/engproc2024074020

Academic Editors: Shu-Han Liao, Teen-Hang Meen and Cheng-Fu Yang

Published: 28 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Construction accidents are a major cause of occupational fatalities worldwide [1–3]. The most effective way to prevent accidents is to eliminate hazards [4]. However, current methods depend on manual inspections by limited safety personnel and are ineffective for real-time monitoring. An alternative to this method is the use of closed-circuit television (CCTV) monitoring [5]. Hazard identification from CCTV images is manually conducted by supervisors, which is impractical for continuous operation [6]. Thus, we explored an AI-based image-captioning approach using a deep learning model with an attention mechanism. It utilizes a pre-trained Inception-v3 network [7] for feature extraction and a gated recurrent unit (GRU) [8] for caption generation. Preliminary tests using bilingual evaluation (BLEU) scores [9] showed promising results regarding the model's effectiveness for captioning hazard images.

2. Literature Review

2.1. Hazard Identification in Construction

A "hazard" is defined as a dangerous condition potentially causing injury, property damage, or environmental impact [10]. According to Heinrich's Domino Theory, accidents can be prevented by removing underlying factors [11]. Traditional identification of hazard images adopts process hazard analysis (PHA) to identify possible hazardous scenarios [12]. However, this method is ineffective in dynamic construction sites where safety measures can be temporarily compromised.

Safety personnel use regular patrols and CCTV to monitor real-time hazards. Additionally, computer vision technologies have been employed to identify hazards [13]. Despite technological advancements, the use of CCTV for real-time hazard detection has been underexplored. Using CCTV effectively, existing safety equipment can be utilized to ensure continuous surveillance to reduce accident risks on construction sites [5].

2.2. Image Caption and Applications

Image captions are generated as descriptive sentences for images using various techniques [14]: (1) template-based methods, which use syntactic parse trees from datasets; (2) image search-based methods, comparing target images against dataset images; (3) deep learning-based methods, utilizing neural networks for multimodal transformations to convert images into text [15]. Recently, deep learning methods have been widely chosen for use over MS COCO image caption generation due to its inherent challenges [16]. The architecture for these models is the encoder–decoder framework, integrating a convolutional neural network (CNN) for image extraction and a recurrent neural network (RNN) for sentence generation [15]. Examples of deep learning-based models include the neural image caption (NIC) model [16], mask RCNN + LSTM with attention mechanisms [17], self-critical sequence training [18], and transformer [19].

Image captioning has shown success in various fields [15–19] but is relatively new in construction engineering. Liu et al. [20] and Wang et al. [17] applied it to describe construction scenes and extract semantic information from images and videos, respectively. Although promising, their results are simple descriptions with limited semantic detail, and thus, are inadequate for complex hazardous construction scenarios. This limitation highlights the importance of the current study.

3. Method

We employed a conventional architecture that combines an Inception v3 deep CNN encoder to extract image features and a GRU recurrent network decoder to generate hazard descriptions. The input images were sourced from CCTV cameras widely used on construction sites.

3.1. Feature Extraction

The proposed model's feature extraction process adopts a pre-trained Inception v3 network, implemented via a MATLAB 2023 program, as detailed in Figure 1. Initially, the model gathers information on the network's input size, constructs a layer graph of the architecture, and removes the classification layers. In the second phase, it introduces a new image input layer matching the original network's dimensions and sets the output size to [8 8 2048], defining the feature maps' dimensions and depth. Finally, the model creates a deep learning network based on Inception v3, specifically adapted for feature extraction by omitting the final classification layers, enhancing its suitability for high-level feature extraction across various computer vision tasks.

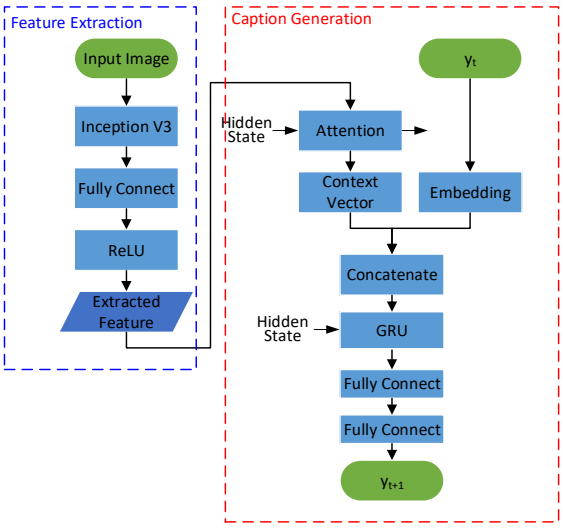


Figure 1. Procedure of proposed method.

3.2. Caption Generation

The process of generating image captions is implemented through a MATLAB program with the following functionalities (Figure 1). The program sets the model's parameters, such as word embedding dimension (256) and the number of hidden units (512), and then initializes encoder and decoder components, including fully connected layers, word embeddings, GRU weights, and attention mechanism weights. Next, it defines two critical functions: one for processing image data through fully connected layers and rectified linear unit (ReLU) activation, and another for generating predictions and attention weights for captioning. The program then sets training parameters and determines the number of epochs, the mini-batch size, and options for visualizing progress, followed by executing a custom training loop that includes data shuffling, preprocessing, loss calculation, gradient computing, and parameter updating via the Adam optimizer. The loop continues through the designated number of epochs, with data shuffling and mini-batch processing occurring during each epoch.

3.3. Performance Evaluation

Four key metrics are commonly used to evaluate natural language processing (NLP) tasks [17]: (1) BLEU, which measures the precision of n-grams in the generated caption and the reference caption; (2) ROUGE, assessing the overlap of n-grams between generated and reference captions; (3) CIDEr, calculating similarity with the reference caption by considering consensus across multiple references; (4) SPICE, evaluating captions by parsing them into semantic propositions and comparing these with reference captions. Each metric has specific strengths and weaknesses.

In this study, we used BLEU to evaluate the accuracy of hazard information extraction and the linguistic quality of the captions in images. We chose BLEU because it effectively evaluates the presence of key terms crucial for describing hazardous scenarios. These descriptions involve key terms including objects (e.g., workers, vehicles, materials), the accident medium (e.g., scaffolding, openness), and location (e.g., top floor, roof). BLEU is outlined in Equation (1), as follows:

$$\text{BLUE} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

where n denotes the n-gram, w_n denotes the weight of the n-gram, BP denotes the short sentence penalty factor (brevity penalty), and p_n denotes the coverage of the n-gram.

4. Preliminary Testing Results

To assess the hazard image caption method, the research team gathered CCTV videos from Chaoyang University of Technology (CYUT) and created a dataset of images and captions using their construction hazard description system (CHDS) [5]. The Matlab® Deep Learning Toolbox's "Image Captioning Using Attention" program was employed to train and test the method on these datasets.

4.1. Image Caption Dataset Generation

Figure 2 illustrates the construction of hazard image caption datasets. The process starts with the caption from an image from a CCTV video, which is then inputted into the construction hazard description system (CHDS) for captioning. Within CHDS, a domain expert marks hazard scenarios with a yellow square frame, identifying and classifying associated objects such as workers and formwork to determine the hazard type. The system generates a caption describing the hazard. In total, we have captioned 1138 hazard images, encompassing 1544 site objects. For each hazard image, we generated four synonymous captions using ChatGPT, resulting in a total of 5690 captions for the 1138 hazard images. These datasets were employed to train the image caption program.



Figure 2. Image caption data generation with CHDS.

4.2. System Training

The "Image Captioning Using Attention" program from the Matlab® Deep Learning Toolbox® was used to test the proposed method. In the testing experiment, model parameters were initialized with 512 hidden units and a word embedding dimension of 256. The fully connected network's weights were initialized using the Glorot initializer. The output size was matched with the decoder's embedding dimension (256), and the input size was aligned with the "mixed10" layer's 2048 output channels from the pre-trained Inception-v3 network. The mini-batch size was set at 128, with a maximum of 30 epochs, and other parameters are detailed in Table 1.

Table 1. Parameters of captions.

Initial Learn Rate	Mini Batch	L2 Regularization	Optimizer
0.0001	30	0.01	rmsprop

The hardware for the training experiment includes (1) CPU: Intel Xeon E5-2620v4 @ 2.10 GHz, (2) RAM: 40 GB at 2400 MHz, (3) Operating System (OS): Microsoft Windows 10, and (4) GPU: NVIDIA Quadro P2000 (5 GB). The training process lasted over 21.5 h, encompassing 30 epochs, and culminated in a final BLEU score of 0.3851. This BLEU score suggests a moderately feasible outcome.

4.3. System Testing

The trained image caption model was evaluated using site hazard images. The caption generation process of the program is illustrated in Figure 3. The initial step involves the identification of site objects, including workers and formwork. Subsequently, the program detects the associated hazard scenario within the image, marked by the yellow-dot square area. Finally, it proceeds to generate a caption, word by word, to describe the hazard scenario.

Figure 4 shows a test result of a site image, where a worker standing beside a mobile crane, without protective barriers, is at risk of being struck.

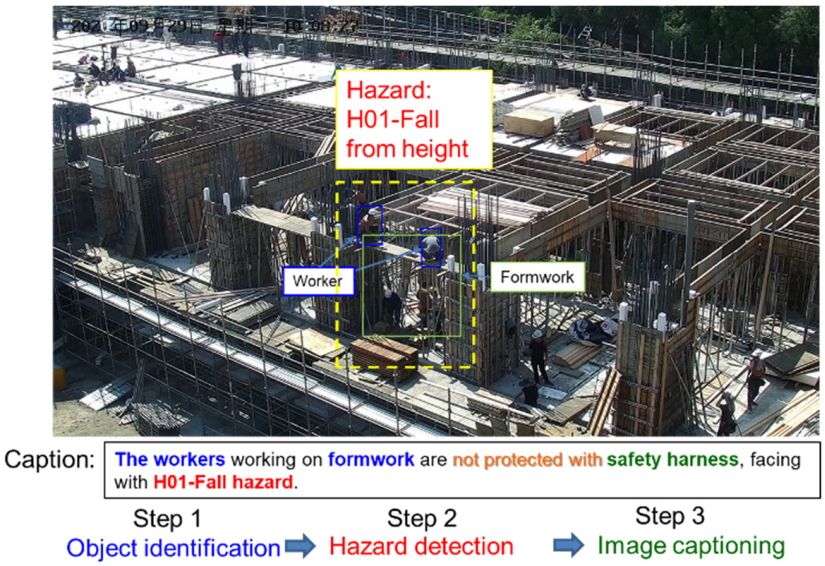


Figure 3. Image caption generation process.



Figure 4. Sample testing hazard image.

5. Conclusions

We explored the generation of image captioning for the construction industry. The approach involved a traditional encoder–decoder architecture, where an Inception-v3 model was utilized for image feature extraction, while a GRU recurrent network was used for caption generation. The method was tested on a dataset comprising 1138 hazard images, 1544 site objects, and 5690 captions, resulting in a moderate BLEU score of 0.3851. While the results are promising, limitations exist. Firstly, the performance of the pre-trained Inception-v3 model for object recognition needs refinement. Despite an improvement from an initial accuracy of 74.4 to 92.53%, enhancement is still demanded. Additionally, it is essential to enhance the caption generation accuracy of the GRU model, which necessitates a more extensive image caption dataset surpassing the current 5690 captions. Furthermore, the feature extraction capability of the encoder requires enhancement to increase its performance.

Author Contributions: Research conceptualization, W.-D.Y.; conceptualization refinement, T.-M.C.; methodology, W.-D.Y. and W.-T.H.; CHDS program preparation and performance, W.-T.H.; validation, W.-D.Y. and W.-T.H.; formal analysis, W.-T.H.; investigation, W.-T.H.; resources, T.-M.C.; data curation, W.-T.H.; writing—original draft preparation, W.-D.Y. and W.-T.H.; writing—review

and editing, T.-M.C. and A.B.; visualization, W.-D.Y.; supervision, T.-M.C.; project administration, W.-D.Y.; funding acquisition, W.-D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This project (MOST 111-2221-E-324-011-MY3) was funded by the National Science and Technology Council of Taiwan. The authors gratefully acknowledge this support.

Institutional Review Board Statement: Not applicable, for this study does not involve humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available upon request to the corresponding author.

Acknowledgments: The case study construction site was provided by Chaoyang University of Technology. The authors would like to express their sincere gratitude to the provider.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Kang, K.; Ryu, H. Predicting types of occupational accidents at construction sites in Korea using random forest model. *Saf. Sci.* **2019**, *120*, 226–563. [CrossRef]
2. Chiang, Y.-H.; Wong, F.K.-W.; Liang, S. Fatal Construction Accidents in Hong Kong. *J. Constr. Eng. Manag.* **2018**, *144*. [CrossRef]
3. TOSHA. 2020 Labor Inspection Annual Report. Occupational Safety and Health, Taiwan. 2020. Available online: https://www.osha.gov.tw/1106/1164/1165/1168/34345/ (accessed on 24 October 2022).
4. Yu, W.D.; Wang, K.C.; Wu, H.T. Empirical Comparison of Learning Effectiveness of Immersive Virtual-Reality-Based Safety Training for Novice and Experienced Construction Workers. *J. Constr. Eng. Manag.* **2022**, *148*, 04022078. [CrossRef]
5. Yu, W.D.; Hsiao, T.; Cheng, T.M.; Chiang, H.S.; Chang, C.Y. Describing Construction Hazard Images Identified from Site Safety Surveillance Video. In Proceedings of the 3rd International Civil Engineering and Architecture Conference (CEAC 2023), Kyoto, Japan, 17–21 March 2023.
6. Ding, L.; Fang, W.; Luo, H.; Love, P.E.D.; Ouyang, X. A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Autom. Constr.* **2018**, *86*, 118–124. [CrossRef]
7. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2023**, arXiv:1512.00567v3.
8. Cho, K.; Van Merri, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
9. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
10. Heinrich, H.W. *IFRCRCS Public Awareness and Public Education for Disaster Risk Reduction*; International Federation of Red Cross and Red Crescent Societies: Geneva, Switzerland, 16 June 2021; pp. 7–9. Available online: https://www.ifrc.org/sites/default/files/2021-06/04-HAZARD-DEFINITIONS-HR.pdf (accessed on 11 November 2023).
11. Heinrich, H.W. *Industrial Accident Prevention*; McGraw-Hill: New York, NY, USA, 1931.
12. Cameron, I.; Mannan, S.; Németh, E.; Park, S.; Pasman, H.; Rogers, W.; Seligmann, B. Process hazard analysis, hazard identification and scenario definition: Are the conventional tools sufficient, or should and can we do much better? *Process Saf. Environ. Prot.* **2017**, *110*, 53–70. [CrossRef]
13. Zhang, L.; Wang, J.; Wang, Y.; Sun, H.; Zhao, X. Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge. *Autom. Constr.* **2022**, *142*, 104535. [CrossRef]
14. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image caption. *ACM Comput. Surv.* **2019**, *51*, 1–36. [CrossRef]
15. Li, S.C. Image Caption with Object Detection and Self-Attention Mechanism Integration and Model Analysis. Master's Thesis, Graduate Program in Artificial Intelligence Technology and Applications, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, July 2021. (In Chinese).
16. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. *arXiv* **2015**, arXiv:1411.4555.
17. Wang, Y.; Xiao, B.; Bouferguene, A.; Al-Hussein, M.; Li, H. Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image caption. *Adv. Eng. Inform.* **2022**, *53*, 101699. [CrossRef]
18. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical Sequence Training for Image Caption. *arXiv* **2016**, arXiv:1612.00563.

19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
20. Liu, H.; Wang, G.; Huang, T.; He, P.; Skitmore, M.; Luo, X. Manifesting construction activity scenes via image caption. *Autom. Constr.* **2020**, *119*, 103334. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.