ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0327
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Automatic Generation of Bridge Defect Descriptions Using Image Captioning Techniques

Chengzhang Chai[1, +], Yan Gao[1, +], Haijiang Li[1, *], Guanyu Xiong[1]

[1] BIM for Smart Engineering Centre, School of Engineering, Cardiff University, Cardiff, UK,
E-mail address: chaic1@cardiff.ac.uk; gaoy74@cardiff.ac.uk; lih@cardiff.ac.uk; xiongg@cardiff.ac.uk
* Corresponding author.
+ These authors contributed equally to this paper.

Abstract: Bridge inspection is crucial for infrastructure maintenance. Current inspections based on computer vision primarily focus on identifying simple defects such as cracks or corrosion. These detection results can serve merely as preliminary references for bridge inspection reports. To generate detailed reports, on-site engineers must still present the structural conditions through lengthy textual descriptions. This process is time-consuming, costly, and prone to human error. To bridge this gap, we propose a deep learning-based framework to generate detailed and accurate textual descriptions, laying the foundation for automating bridge inspection reports. This framework is built around an encoder-decoder architecture, utilizing Convolutional Neural Networks (CNN) for encoding image features and Gated Recurrent Units (GRU) as the decoder, combined with a dynamically adaptive attention mechanism. The experimental results demonstrate this approach's effectiveness, proving that the introduction of the attention mechanism contributes to improved generation results. Moreover, it is worth noting that, through comparative experiments on image restoration, we found that the model requires further improvement in terms of explainability. In summary, this study demonstrates the potential and practical application of image captioning techniques for bridge defect detection, and future research can further explore the integration of domain knowledge with artificial intelligence (AI).

Key words: Image captioning, Bridge engineering, Defect detection, Deep learning, Visual inspection

1. INTRODUCTION

Bridge inspection is crucial for infrastructure maintenance. In recent years, bridge inspection using computer vision has been extensively researched; however, these studies are often limited to the simple detection of cracks or corrosion [1-2]. The results of such detection can only serve as preliminary references for bridge inspection reports, and the generation of detailed reports still requires on-site engineers to describe the structural conditions through lengthy textual descriptions. These manual processes are typically time-consuming, costly, and prone to human errors. Furthermore, with the increasing number of aging bridges, the demand for analyzing a vast amount of bridge imagery also imposes significant work pressure on on-site engineers. Therefore, in light of this situation, we pose an important question: How can we improve the efficiency and accuracy of bridge inspections while simultaneously reducing the burden on professionals?

The automatic generation of image captions is a complex task because of the need to explain and describe various visual elements in different contexts accurately. As a result, this problem has attracted much attention and led to significant research work in computer vision and natural language processing [3-4]. While existing techniques have made significant progress in natural image analysis, there is still much space for research to generate descriptions of bridge defects in the infrastructure automatically. Particularly, when dealing with bridge defects of great variety and complexity, there is a need for a method that can accurately capture and explain these complex visual features.

This study addresses these challenges by proposing an innovative framework to bridge the research gap. The framework combines a deep learning-based encoder-decoder architecture with an attention mechanism. Our goal is to automate the generation of descriptions in bridge inspection. By doing so, we improve the accuracy and efficiency of bridge inspections and significantly reduce the site engineers' workload. Furthermore, this lays the groundwork for automatically generating detailed bridge inspection reports. Subsequently, considering the well-known black-box property of deep learning, we also design comparative experiments based on image restoration to explore the accuracy and reliability of the generated descriptions. In conclusion, our approach will provide new ideas for the comprehensive application of artificial intelligence (AI) in the broader infrastructure field.

2. RELATED WORK

2.1. Defect detection

In recent years, computer vision-based defect detection methods have been extensively studied and have demonstrated their significant value in several fields, especially in industry, infrastructure maintenance, and production quality control. These methods utilise advanced image processing techniques and machine learning algorithms. In particular, a typical deep learning model, convolutional neural network (CNN), can learn complex features and patterns from many labelled images and has achieved remarkable success in automatically identifying and classifying various defects. Cha et al. [5] proposed a method to detect concrete cracks using the deep architecture of CNN to learn image features automatically without traditional image processing techniques, demonstrating up to 98% accuracy in various complex real-world scenarios. Zhang et al. [6] proposed an improved single-stage you only look once (YOLOv3) detector by introducing novel migration learning and enhancement algorithms, which achieves efficient and accurate detection of multiple damage types on concrete bridge decks, surpassing traditional deep learning methods in terms of intersection over union (IoU) evaluation metrics. Mundt et al. [7] solved the problem of efficiently identifying multiple defects in bridge concrete by employing MetaQNN and efficient neural architecture search techniques. They innovatively introduced a convolutional neural network architecture with fewer parameters and higher accuracy. Forkan et al. [8] proposed CorrDetector, an integrated AI framework based on CNN models for infrastructure structure identification and corrosion detection. These studies show that due to CNN's significant local feature learning capability, using them as feature extractors or variant models can yield good detection results in defect identification.

However, most of these detections focus on crack segmentation or corrosion detection and fail to provide detailed descriptions for bridge images. This implies that when generating detailed bridge inspection reports, the defect detection results still necessitate significant secondary processing by engineers. For engineers who are already facing a workforce shortage, this results in high work pressure.

2.2. Image captioning techniques

Image captioning techniques combine computer vision and natural language processing to recognize image content and automatically generate descriptive text. Traditional image captioning models mainly include retrieval and template-based methods [9-10]. With the development of deep learning, especially the application of CNN and recurrent neural network (RNN), image captioning technology has been dramatically enhanced [11]. Vinyals et al. [12] first introduced the encoder-decoder framework for generating captions. Xu et al. [13] introduced the attention mechanism, focusing on salient objects in the image to further improve the quality of image captions. Liu et al. [14] proposed using textual attention mechanisms to increase the completeness of the conveyed information and combined label generation with textual attention and image captioning.

Despite significant progress in image captioning technology, it still faces some limitations. Firstly, most existing models are mainly trained on general datasets like MS-COCO and Flickr30k, which significantly limits their effectiveness and generalizability in specific domains when the general training data and the unique requirements of the application domain differ. Secondly, although image captioning technology has been widely applied in medicine and remote sensing [15-16], only a few studies are being conducted in infrastructure. There is considerable research space for bridge defect detection and the automatic generation of inspection reports. Future work needs to explore training datasets more suitable for specific domain requirements, develop models that can accurately understand and describe the complex characteristics of these domains, and create more intelligent and accurate image captioning systems to meet the diverse needs of practical applications.

3. METHODOLOGY

3.1. Overall framework

The proposed overall framework is shown in Fig. 1, which can automatically generate appropriate textual descriptions from bridge images. The core of the framework is the encoder-decoder architecture, seamlessly integrated with the attention mechanism to provide fine-grained analysis of images. The model is trained and fine-tuned to accurately identify and describe image defects, facilitating a more streamlined and efficient defect evaluation process.

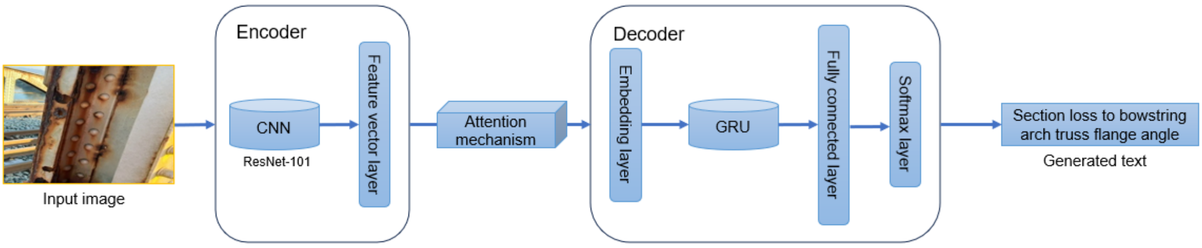


Figure 1. The proposed overall framework

3.2. Encoder-Decoder architecture

The encoder uses the ResNet-101 model as a pre-trained network in this architecture. It is a variant of CNN known for its depth and ability to extract detailed feature vectors from large amounts of visual data. This pre-training enhances the model's ability to accurately locate salient features indicative of potential defects in bridge images. After feature extraction, the output of the encoder acts as a rich coded input for the decoding stage. The decoder expands this input into a sequence of words using gated recurrent units (GRU), while the attention mechanism directs the focus to specific regions of the image. This process ensures that each word generated in the description text is related to the corresponding visual feature in context. After the fully connected layer, a softmax layer is introduced to finalize the textual output. It helps to convert the numerical output of the GRU into a lexicographic probability distribution, thus efficiently selecting the most probable words at each step and finally forming a coherent descriptive text.

3.3 Attention mechanism

The soft-attention mechanism is a key component of this framework, linking the high-level visual features extracted by the ResNet-101 encoder to the word generation process of the GRU-based decoder. As each word is generated, the attention mechanism computes a set of weights corresponding to the different pixels in the feature map, ensuring that the sum of these weights is 1, resulting in a normalized attention distribution. This probabilistic approach keeps the sum of attention consistent across the image for each decoding time step. Moreover, ultimately ensures that each word generated in the description text is associated with the corresponding visual feature in the context.

3.4 Loss function

The loss function is one of the crucial factors in determining the effectiveness of model learning. We use a dual loss function strategy to update the model weights. Cross entropy loss is used to minimize the difference between the actual word labels and the probability distribution predicted by the model. More accurate descriptions are generated by encouraging the model to increase the probability of predicting the correct word. Moreover, using bi-stochastic attentional regularisation helps solve the potential problem of focusing on local regions and ignoring global information in encoder-decoder models. By penalizing the allocation of attention to images that are not fully covered, the model is encouraged to spread its attention and generate accurate and comprehensive descriptions.

4. EXPERIMENT VALIDATION

4.1. Dataset preparation

The training of the model requires the preparation of the appropriate dataset. Considering that there are currently no publicly available image-text description datasets directly usable in bridge inspection. Therefore, we constructed the dataset based on regular bridge inspection reports provided by Centregreat Rail (CGR), an industry company in the ongoing DigiBridge project. Specifically, this dataset will include images of the daily inspections and corresponding descriptions.

We performed preprocessing work on the text part to better use the dataset for training. The cleaning method of removing stopwords is mainly used. The constructed dataset has a sample size of 500 pairs of images and texts. Then, 80%, 10%, and 10% are selected as the training, validation, and test sets.

4.2. Model evaluation metrics

In this study, a set of evaluation metrics that are widely used in the field of natural language generation were used. They are BLEU [17], METEOR [18], ROUGE-L [19], and CIDEr [20]. These metrics focus on different aspects of text generation. Specifically, BLEU and METEOR are mainly used to evaluate machine translation tasks, ROUGE-L is widely used for text summarisation tasks, and CIDEr is mainly used for image captioning tasks. For these metrics, higher values indicate higher quality of generated descriptions. The scores for BLEU-1 to BLEU-4, METEOR, and ROUGE-L range from 0 to 1, while the scores for CIDEr range from 0 to 10.

4.3 Implementation details

All experiments were executed in Python 3.8 and Pytorch 2.0.1 environments with an Intel(R) Core (TM) i7-10700KF CPU@ 3.80 GHz processor and 48 GB of RAM. It also has an NVIDIA GeForce RTX 3060 Ti to ensure high computational efficiency and robustness.

The neural network was configured with a maximum of 50 epochs for training. The size of the recognized image is 384*384, the batch size of the input data is set to 8, and a loss rate of dropout=0.5 is used to cope with overfitting. The word vectors are embedded in 512 dimensions, and both the attention mechanism and decoder are assigned 512 dimensions to capture the complexity of the visual-to-text translation process.

Consistent with the two-tier architecture, separate learning rates are used for the encoder and decoder. The encoder is fine-tuned with a learning rate of 1e-4, while the learning rate of the decoder is set to a slightly higher 4e-4 to promote faster convergence of the model parameters. This differential learning rate strategy helps balance the learning rate between the two network components.

Furthermore, improvements in the BLEU-4 score employed an early stopping mechanism with a threshold of 5 epochs. This approach ensures that the training halts if the BLEU-4 score does not improve for five consecutive epochs. BLEU-4, a crucial indicator of phrase-level textual accuracy, reflects the model's proficiency in generating coherent and relevant descriptions. This early-stopping strategy prevents overfitting and optimizes computational resources by terminating training when minimal improvements occur.

4.4. Experimental result of captioning generation

a. Quantitative results

Experiments were conducted in two separate scenarios with and without the attention mechanism to assess the model's ability to generate descriptions automatically. The evaluation metrics generated using the performance metrics presented in Section 4.2 are shown in Table 1. The table includes seven assessment metrics: BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr.

Table 1. The evaluation metrics of descriptions generation

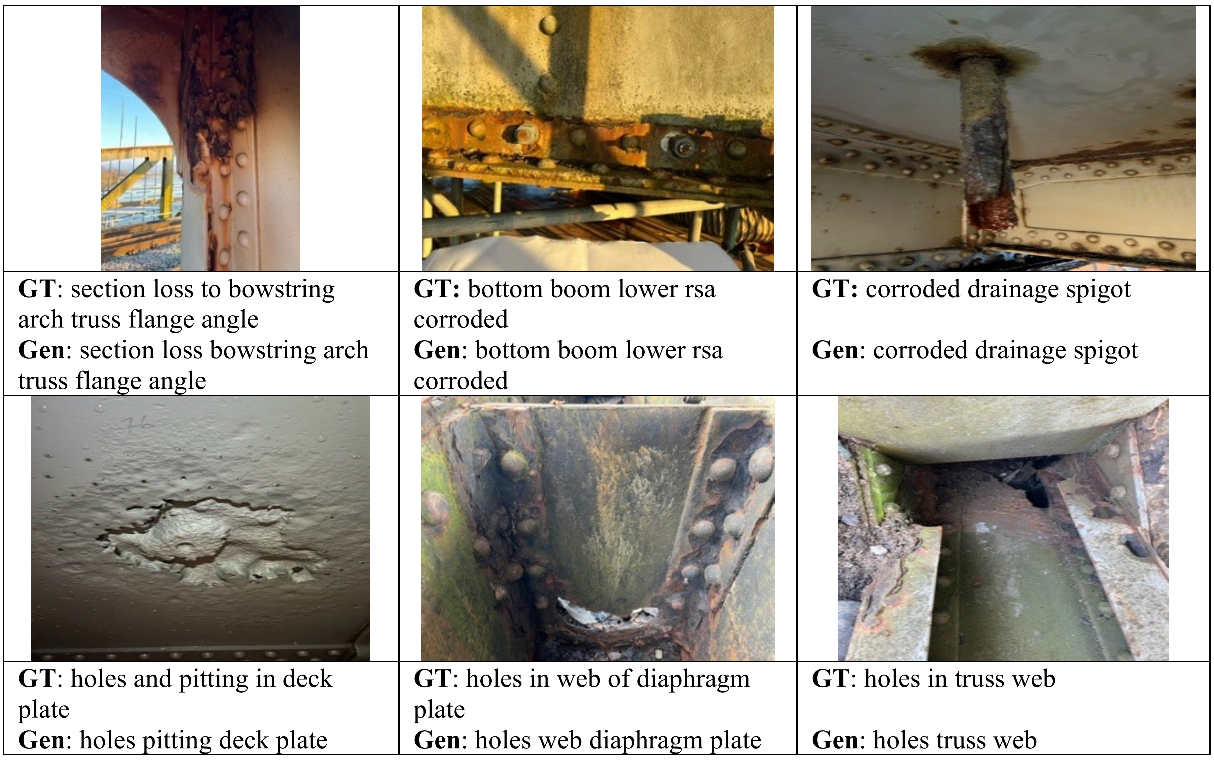
No.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Attention	0.703	0.638	0.602	0.575	0.399	0.722	5.583
Without	0.655	0.589	0.554	0.535	0.371	0.675	5.202

Firstly, some common trends are reflected in both cases. For example, when an attention mechanism is present, the BLEU scores decrease from 0.703 (BLEU-1) at the word level to 0.638 and 0.602 (BLEU-2 & BLEU-3) at the phrase level and finally to 0.575 (BLEU-4) for the overall structure. That is, the BLEU scores all decrease as the n-gram length increases. This decrease reflects the increasing difficulty of the model in capturing the subtle structure of longer phrases and sentences as the sequences lengthen. In addition, the higher scores for METEOR & ROUGE-L, 0.399 & 0.722, reflect the consistency of the model-generated descriptions with the reference descriptions regarding semantic matching and similarity in sentence structure. In contrast, the high CIDEr score, 5.583, indicates the model's ability to generate highly relevant descriptions of a given task.

In contrast, models without an attention mechanism scored lower on all metrics than those with an attention mechanism. This suggests that the attention mechanism aids the model in capturing finer details at the word and phrase level and maintaining the coherence and relevance of the generated descriptions over longer text spans. Thus, the quantitative results highlight the model's effectiveness in automatically generating bridge defect descriptions. The importance of attentional mechanisms in image captioning models also allowed for a better detailed and contextualised understanding of complex visual scenes.

b. Qualitative result

To visually demonstrate the method's effectiveness, we provide some examples to evaluate the qualitative performance of our image caption generation model, as shown in Fig. 2. Each example consists of an image, a ground truth description (GT), and a generated description (Gen). GT is authentically derived from field engineers' records and can be used as a reference standard for evaluating the Gen. Observation from Fig. 2. This reveals that the model generated almost exactly correct descriptions in these examples. It is worth noting that the disappearing stopwords in the Gen are mainly due to the cleaning method of stopword removal used in constructing the dataset. Therefore, we can assume that the model has been able to generate corresponding descriptions of actual images to a large extent and will be adaptable and useful in several common bridge detection scenarios.



GT: section loss to bowstring arch truss flange angle
Gen: section loss bowstring arch truss flange angle

GT: bottom boom lower rsa corroded
Gen: bottom boom lower rsa corroded

GT: corroded drainage spigot
Gen: corroded drainage spigot

GT: holes and pitting in deck plate
Gen: holes pitting deck plate

GT: holes in web of diaphragm plate
Gen: holes web diaphragm plate

GT: holes in truss web
Gen: holes truss web

Figure 2. Examples of generating descriptions (GT: ground truth, Gen: generated)

4.5. Model explainability analysis

Deep learning models often need help in explainability due to their complexity and multi-layer structure. In particular, models with multiple hidden layers and many parameters often have a "black box" decision-making process that is not intuitively understood. In this case, a heatmap of the attention mechanism can visualise which parts of the input data the model pays more attention to when making decisions. It will help to understand the model's decision-making rationale by highlighting important features. As shown in Fig. 3, (a), (c), and (e) are the original images, respectively. (b), (d) and (f) are then the corresponding heatmaps of the attention mechanism.

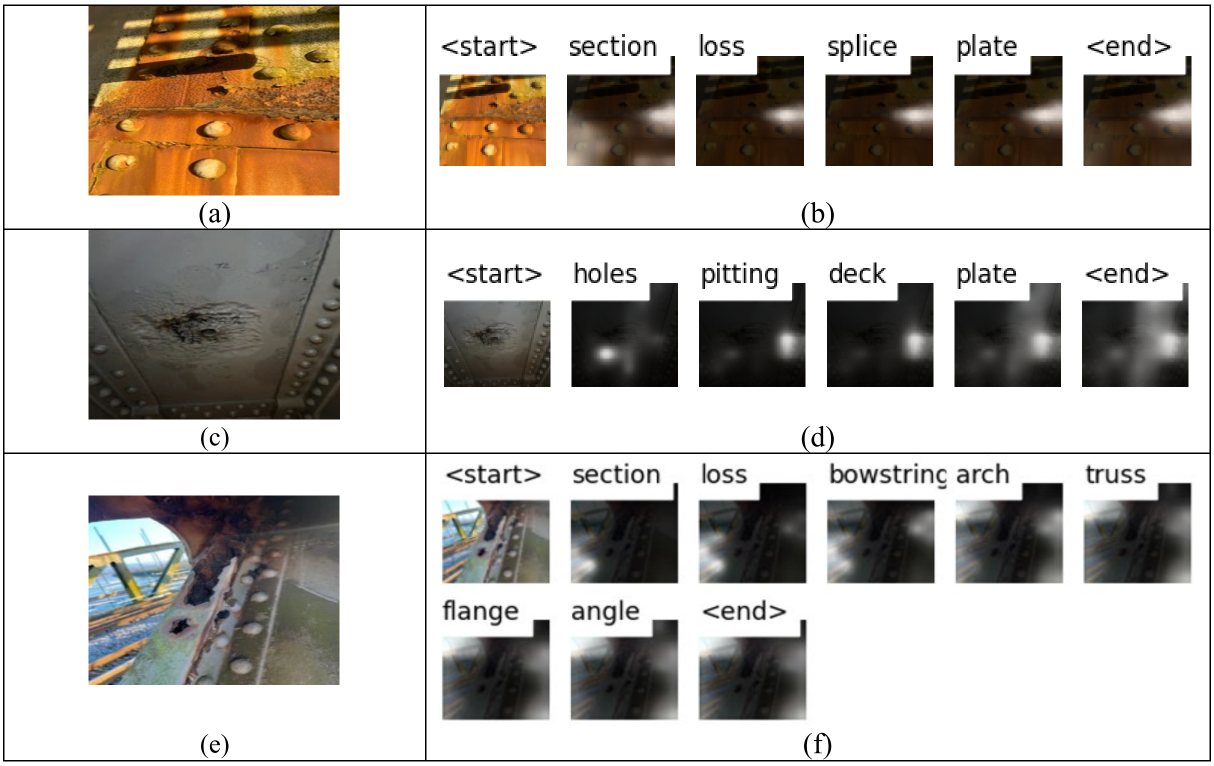
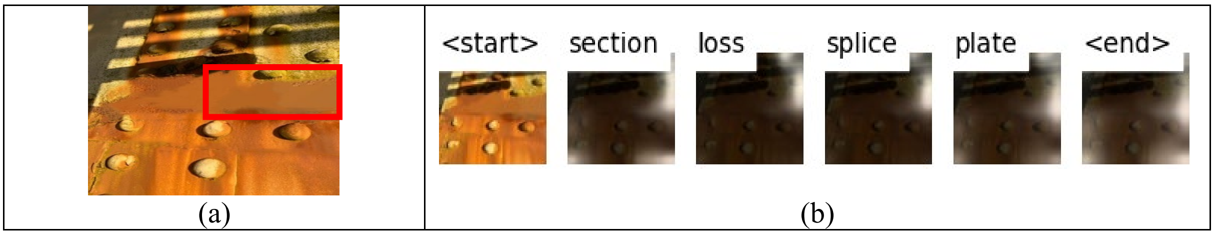


Figure 3. Attention mechanism heatmaps for original image generated descriptions

The observation reveals that the "section loss" generated in (b) correctly attends to the location of the defect in (a). However, the "holes" and "pitting" generated in (d) did not focus on the accurate location in (c). The "section loss" generated in (f) also does not focus on the accurate position in (e). Therefore, a research question arises: Are the regional features used by the model accurate during learning?

To explore this question further, we set up a comparison experiment based on image restoration. After manually selecting critical defective regions on the original image, an image processing algorithm using texture restoration techniques was used to obtain the restored image. The specific restored image areas are shown in red boxes in Fig. 4 (a), (c), (e). Subsequently, the restored image is provided to the model again to generate a new predictive description. In this way, we can perform a comparison test. If the model generates a different description on the restored image, the model relies more on the critical region when generating the description. Conversely, it may indicate that the model relied more on other regions in the image or learned features unrelated to the defect.



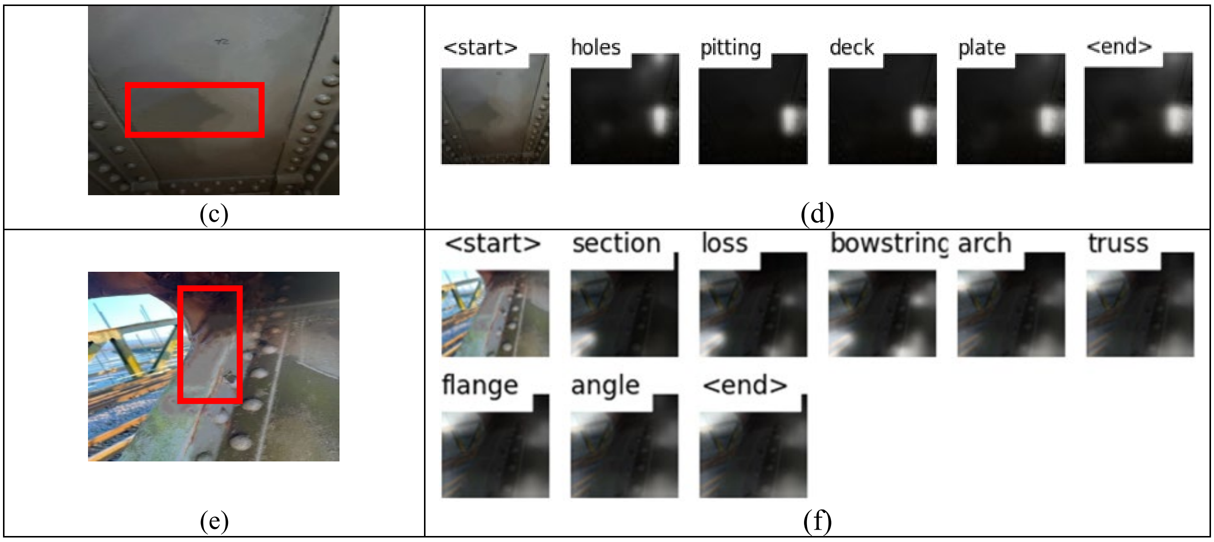


Figure 4. Attention mechanism heatmaps for restored image generated descriptions

As can be seen in (b), (d), and (f) of Fig. 4, the restored image generates the same description as the original image. This indicates that the model's understanding of the image content may be too superficial and needs a deeper recognition of the defect's salient features. The model relies more on learning contextual cues of the environment in the image without effectively focusing on identifying critical regions. In addition, the misaligned regions on the attention heatmap reveal the model's limitations in terms of architecture or attention mechanisms. In summary, the model needs further refinement in terms of explainability. Improvements can be continued by expanding the sample size and replacing the attention mechanism.

5. CONCLUSION

This study successfully implemented a deep learning-based image description method for automatically generating bridge defect descriptions. The model employs an advanced encoder-decoder architecture and attention mechanism to demonstrate its ability to explain and describe various bridge detection images. The main contribution of this research lies in the innovative application of image captioning techniques in infrastructure maintenance, particularly in the context of bridge inspections. This approach can be directly used to automatically generate defect descriptions in bridge inspection reports, enhancing the efficiency and accuracy of bridge inspections while significantly reducing manual workload. The evaluation results of the experiments demonstrate the ability of the model to generate detailed and relevant defect descriptions, which is mainly confirmed by the BLEU, METEOR, ROUGE-L, and CIDEr scores. The introduction of the attention mechanism also helps in generating more accurate descriptions. In addition, it is worth noting that through the comparative experiments on image restoration, we also found that the model has limitations regarding explainability. This is mainly manifested in the inability to fully recognize the features of salient regions, which requires further exploration.

In conclusion, this study makes a practical contribution to bridge inspection in the infrastructure field and lays the foundation for automating bridge inspection report generation. Future research can explore more advanced models of attention mechanisms and introduce a more comprehensive domain-specific knowledge base to improve the utility of AI-driven bridge inspection further.

ACKNOWLEGEMENTS

This work was supported by the DigiBridge KTP (Grant Reference Number 10003208) and dataset was contributed by Jarrod Richards, Centregreat Rail Ltd.

REFERENCES

[1] Dabous, S.A. and Feroz, S., 2020. Condition monitoring of bridges with non-contact testing technologies. Automation in Construction, 116, p.103224.

[2] Abdallah, A.M., Atadero, R.A. and Ozbek, M.E., 2022. A state-of-the-art review of bridge inspection planning: Current situation and future needs. Journal of Bridge Engineering, 27(2), p.03121001.

[3] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[4] Hossain, M.Z., Sohel, F., Shiratuddin, M.F. and Laga, H., 2019. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6), pp.1-36.

[5] Cha, Y.J., Choi, W. and Büyüköztürk, O., 2017. Deep learning ‐ based crack damage detection using convolutional neural networks. Computer ‐ Aided Civil and Infrastructure Engineering, 32(5), pp.361-378.

[6] Zhang, C., Chang, C.C. and Jamshidi, M., 2020. Concrete bridge surface damage detection using a single ‐ stage detector. Computer ‐ Aided Civil and Infrastructure Engineering, 35(4), pp.389-409.

[7] Mundt, M., Majumder, S., Murali, S., Panetsos, P. and Ramesh, V., 2019. Meta-learning convolutional neural architectures for multi-target concrete defect classification with the concrete defect bridge image dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11196-11205).

[8] Forkan, A.R.M., Kang, Y.B., Jayaraman, P.P., Liao, K., Kaul, R., Morgan, G., Ranjan, R. and Sinha, S., 2022. CorrDetector: A framework for structural corrosion detection from drone images using ensemble deep learning. Expert Systems with Applications, 193, p.116461.

[9] Mason, R. and Charniak, E., 2014, June. Nonparametric method for data-driven image captioning. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 592-598).

[10] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C. and Berg, T.L., 2013. Babytalk: Understanding and generating simple image descriptions. IEEE transactions on pattern analysis and machine intelligence, 35(12), pp.2891-2903.

[11] Huang, L., Wang, W., Chen, J. and Wei, X.Y., 2019. Attention on attention for image captioning. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4634-4643).

[12] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[13] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015, June. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

[14] Liu, M., Li, L., Hu, H., Guan, W. and Tian, J., 2020. Image caption generation with dual attention mechanism. Information Processing & Management, 57(2), p.102178.

[15] Ayesha, H., Iqbal, S., Tariq, M., Abrar, M., Sanaullah, M., Abbas, I., Rehman, A., Niazi, M.F.K. and Hussain, S., 2021. Automatic medical image interpretation: State of the art and future directions. Pattern Recognition, 114, p.107856.

[16] Zhao, R., Shi, Z. and Zou, Z., 2021. High-resolution remote sensing image captioning based on structured attention. IEEE Transactions on Geoscience and Remote Sensing, 60, pp.1-14.

[17] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

[18] Denkowski, M. and Lavie, A., 2014, June. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation (pp. 376-380).

[19] Lin, C.Y., 2004, July. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[20] Vedantam, R., Lawrence Zitnick, C. and Parikh, D., 2015. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).