Image Captioning for Building Damage Using Deep Learning: Explaining Context in Overlapping Structural Defects

Vedant Girish Dalvi[1], Jakob Martin[1] and Timur Weilbach-Eyüboglu[1]

[1]Hochschule München University of Applied Sciences, Munich, Germany

E-mail(s): vedant_girish.dalvi@hm.edu, jakob.martin0@hm.edu, timur.weilbach-eyueboglu@hm.edu

Abstract: Different types of damage are found in building structures, which vary depending on the materials used, such as steel, concrete, and glass. Deep learning-based vision models have been widely employed in recent years to classify, detect, and segment these damages. However, damages often overlap in real-world scenarios, creating complex image segmentations. The intricate nature of construction damages is effectively described using captions generated by deep learning-based Vision-Language Models (VLMs), which provide insight into the impact of these damages on structural stability. This paper proposes an image-captioning web application for automating the documentation of construction damages, thereby increasing the efficiency of safety inspections, workflows, and maintenance. The application utilizes a fine-tuned Bootstrapping Language-Image Pre-training (BLIP) architecture to generate captions that detail the overlapping damages in construction images, facilitating the analysis of damages. The open-source damage classification (dacl) dataset for semantic bridge damage inspections was filtered and modified according to a predefined criterion to include textual captions, enabling fine-tuning of the BLIP and BLIP2 models. The parameters considered for captioning the images were the location and nature of the damage. The developed application reduces the time and human effort required for safety inspections of concrete structures.

Keywords: Image-captioning, Deep Learning, Structural damages

DOI: 10.18154/RWTH-CONV-254876. Published in the conference proceedings of the 36. Forum Bauinformatik 2025, Aachen, Germany,

1 Introduction and Problem Statement

Artificial Intelligence (AI) has been adopted in the construction field for various applications, including site monitoring and performance evaluation via robotics and drones, tender evaluation, conflict resolution, sustainability assessments, waste management, safety, and hazard management [1]. Damage inspection is a key component of Structural Health Monitoring (SHM), used across various domains such as civil, mechanical, aerospace, and energy infrastructure, to evaluate structural integrity without physical intervention. It covers the detection of cracks, corrosion, wear, spalling, and other anomalies, often via non-destructive testing (NDT) methods such as ultrasonic, acoustic, magnetic flux leakage, and visual and optical techniques. Damage inspection has traditionally been humandriven and, therefore, requires time and is prone to human error. The deployment of AI applications accelerates and improves safety inspections by automating the workflows, thereby reducing human dependency and reducing inspection errors and delays.

AI, specifically Computer Vision (CV), has been widely used to create applications for safety monitoring, progress tracking, productivity measurement, and quality control [2]. Construction damages in reinforced concrete are often complex, with different overlapping damages. This makes it challenging for CV-based models to detect and segment these damages, resulting in complex and cluttered detections or segmentations. In the case of semantic segmentations of reinforced concrete images, the model often prioritizes one overlapping damage class over the other, resulting in an unclear understanding of the structural stability from the image. This paper proposes an alternative methodology by using textual captions generated by an image captioning model to explain the context of construction damage in an image, thereby eliminating the need for human intervention and enhancing the efficiency of safety inspections.

2 State of the Art: Captioning Damages to explain semantic context

Deep learning-based computer vision models have been widely used to classify, detect, and semantically segment construction damages from images. An example of this is the dacl challenge for segmentation of semantic bridge damage [3]. The task of this challenge was to develop an image segmentation model using deep learning to segment images of damage on reinforced concrete bridges.

Textual descriptions explain the nature of overlapping damages in comprehensive captions, thereby enhancing the semantic understanding of the overlapping damages in reinforced concrete. Deep learning-based image captioning models have been widely applied in various fields, including medical image analysis, semantic tagging, image retrieval, and content creation. In the construction domain, an existing study investigates whether deep learning techniques can generate accurate, descriptive captions for construction site images to support scene understanding and documentation [4]. The paper titled “Manifesting construction activity scenes via image captioning” [5] introduces a novel, automated framework that generates descriptive captions of real-world construction activity scenes using image captioning techniques. The applications of Large Language Models (LLMs) in the construction domain are limited [6]. There is no existing research on the implementation of image captioning for construction damage images to explain the interrelationships between these damages.

In summary, existing AI applications in the construction domain for safety inspections primarily focus on Computer Vision to infer information from images. In contrast, the proposed approach, which uses a VLM fine-tuned on a captioning dataset, can potentially better define the context of complex and overlapping damages than semantic segmentation.

3 Methodology

Deep learning-based image captioning models aim to generate descriptive text for images automatically. These models traditionally used a Convolutional Neural Network (CNN) to extract visual features from an image, followed by a Recurrent Neural Network (RNN) to generate a sequence of words that form a caption. Recent models utilize transformer architecture, which employs a self-attention mechanism for both the encoder and decoder, focusing on relevant image regions during the captioning process. This approach has enabled significant progress in developing more accurate and contextually relevant descriptions. Bidirectional Encoder Representations from Transformers (BERT) [7], Vision transformer (ViT) [8], Bootstrapping Language-Image Pretraining (BLIP) [9], and Bootstrapping Language-Image Pretraining 2 (BLIP 2) [10] are some examples of transformer architecture-based image captioning models.

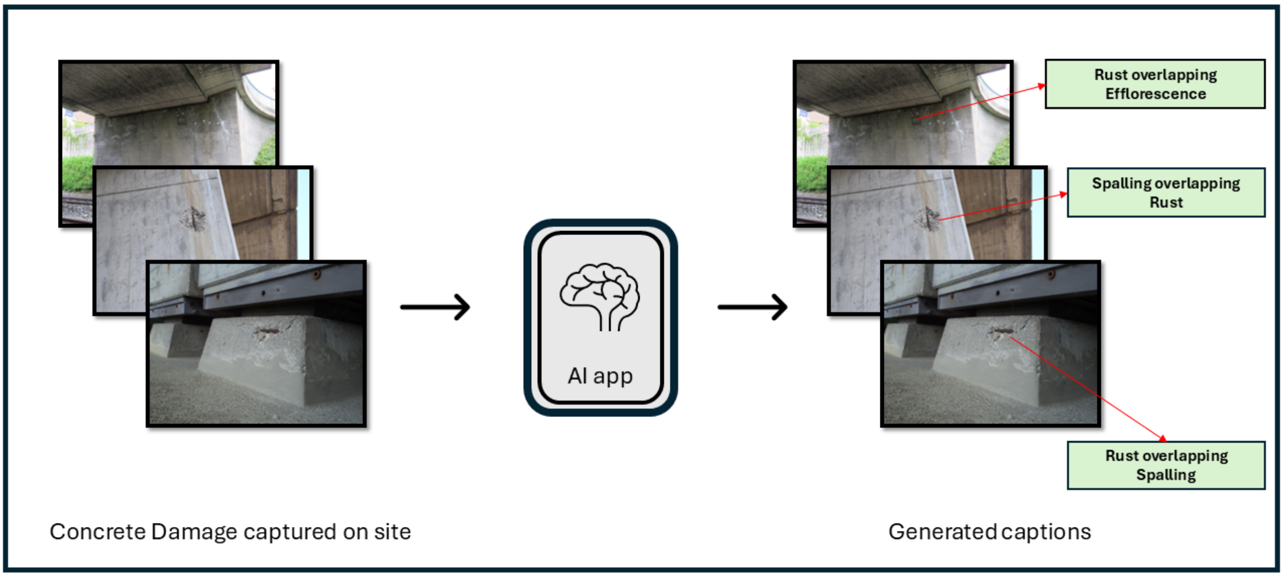

AI app
Rust overlapping Efflorescence
Spalling overlapping Rust
Rust overlapping Spalling
Concrete Damage captured on site
Generated captions

Figure 1: Proposed Methodology: The proposed approach is for a safety inspection application. The site engineer takes a picture of the reinforced concrete damage, which is then uploaded to an image captioning application. The application generates captions explaining the nature of the overlapping damages in the picture.

The BLIP architecture is particularly suitable for this project on construction damage captioning due to its ability to generate context-aware, detailed descriptions from visual input. Unlike traditional image captioning models that rely solely on paired image-text datasets, BLIP leverages both web-scale noisy data and high-quality image-text pairs through its bootstrapped training strategy, resulting in a model that generalizes well even with limited fine-tuning data. This makes it ideal for a domain like construction, where large annotated caption datasets are scarce. BLIP's architecture, as shown in Figure 2, combines a vision transformer encoder with a text decoder, enabling it to attend to fine-grained damage features, such as cracks, corrosion, or material overlaps, while generating coherent and meaningful captions. Its strong performance on multimodal tasks, combined with pre-trained checkpoints and easy adaptation, makes BLIP a robust foundation for generating captions that explain overlapping structural damages in real-world construction images.

BLIP-2 is an advanced multimodal model that improves on the original BLIP by efficiently bridging vision and language understanding. Unlike BLIP, which directly processes images and text together, BLIP-2 uses a lightweight image encoder combined with a frozen large language model (LLM), connected via a learnable query network. This design significantly reduces computational costs while maintaining strong performance on image captioning, visual question answering, and cross-modal

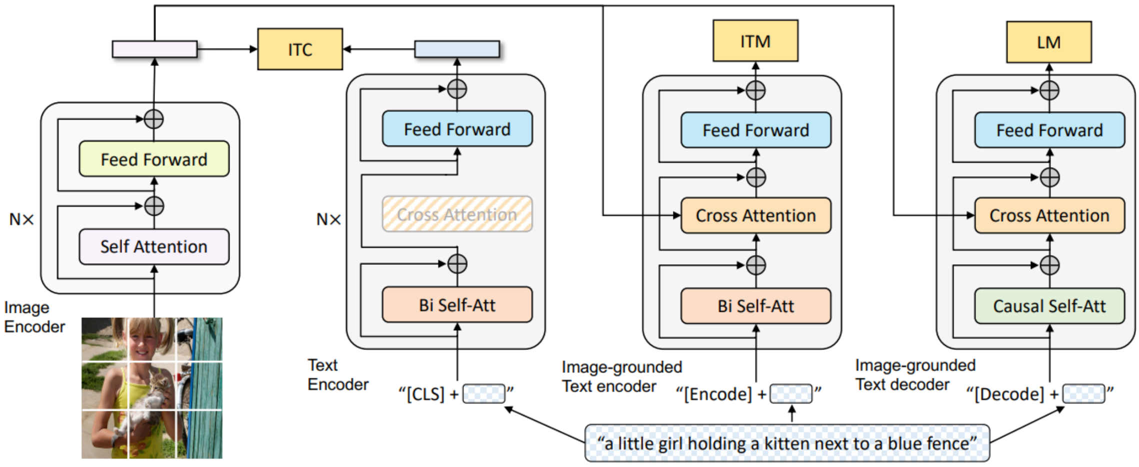


Figure 2: BLIP Architecture [9]

reasoning. Its advantages over BLIP include faster training, lower memory requirements, and the ability to leverage powerful pre-trained LLMs for richer, more coherent text generation from images.

Evaluation metrics for image captioning models assess how the generated captions align with human-written references. The fine-tuned model is evaluated using various metrics like BLEU, METEOR, and ROUGE.

BLEU (Bilingual Evaluation Understudy)

BLEU is a metric to evaluate the quality of machine-generated translations or captions against reference translations or captions. It measures the overlap between the generated and reference captions based on n-grams (continuous sequence of n words) present in both. BLEU score ranges from 0 to 1; a score of 1 means a perfect match, while a score of 0 means no overlap between the generated and reference captions. BLEU's primary focus on exact word matches often results in a lack of correlation with human judgment, especially in cases where synonyms or paraphrasing occurred [11].

METEOR (Metric for Evaluation of Translation with Explicit ORdering)

This metric considers exact, stem, synonym, and paraphrase matches, balancing precision and recall, and often better correlates with human judgment than BLEU. The METEOR score ranges from 0 to 1, with a higher score indicating better caption quality [12].

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

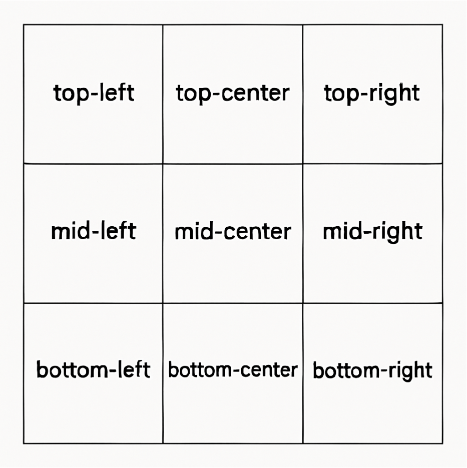
ROUGE focuses on recall-based n-gram and sequence matching, a technique commonly used in text summarization and applicable to captioning. ROUGE ranges from 0 to 1, with higher scores indicating greater similarity between the generated caption and the reference [13].

All of these metrics offer complementary perspectives, and combining them provides a more comprehensive picture of the captioning performance of the fine-tuned model.

4 Implementation

The project was developed in two phases. In the first phase, the dacl dataset was filtered and captioned using predefined criteria. In the second phase, BLIP and BLIP2 models have been finetuned on the overlapping damages-captions dataset, and then evaluated and deployed in a web app.

Of the 19 total available classes, eight damage classes (crack, wetspot, efflorescence, rust, rock pocket, spalling, cavity, and exposed rebar) were selected based on their impact on the structural stability of reinforced concrete bridges. The captions have been automatically generated by a custom filtering and captioning script for all training images using the polygonal annotations from the dacl dataset. The criterion of the script for filtering is to filter and retain images from the original dataset that contain the selected eight damage classes and include at least two damage classes that overlap with each other. The resulting 1,500 images were used to fine-tune the image captioning model. Additionally, 200 images were used for evaluating the fine-tuned model. The images were captioned with parameters indicating the damage classes present in the image, their locations, and whether any overlapping damage was also present. To determine the position of the damage in the image, the image is divided into nine equal sections, as shown in Figures 3 and 4. Each caption is structured with the image name, image path, and the caption itself as shown in Algorithm 1. The training caption specifies the presence of each damage instance overlapping another damage instance, and the captions of all the images are stored in a nested JSON file.



top-left	top-center	top-right
mid-left	mid-center	mid-right
bottom-left	bottom-center	bottom-right

Figure 3: Image is divided into nine sections, which are used in the captioning schema

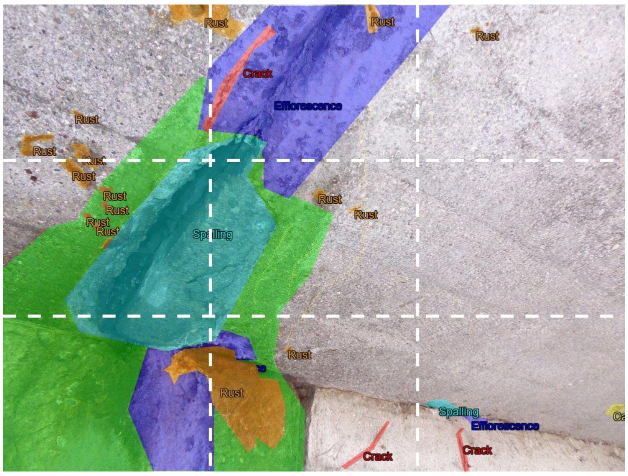


Figure 4: Example training image showcasing the various overlapping damages

Algorithm 1: Example Caption

```
{
  "image_id": "dacl10k_v2_train_0179",
  "image_path": "output/train_overlays/dacl10k_v2_train_0179.jpg",
  "caption": "wetspot overlapping efflorescence damage at mid-left; wetspot
     overlapping spalling damage at mid-left; wetspot overlapping
     efflorescence damage at mid-left; wetspot overlapping crack damage at mid
     -left; wetspot overlapping rust damage at mid-left; wetspot overlapping
```

```
        rust damage at mid-left; wetspot overlapping rust damage at mid-left;
        wetspot overlapping rust damage at mid-left; wetspot overlapping rust
        damage at mid-left; wetspot overlapping rust damage at mid-left; wetspot
        overlapping efflorescence damage at mid-left; efflorescence overlapping
        spalling damage at top-center; efflorescence overlapping crack damage at
        top-center; efflorescence overlapping rust damage at top-center; spalling
         overlapping efflorescence damage at mid-left; spalling overlapping rust
        damage at mid-left; spalling overlapping efflorescence damage at mid-left
        ; efflorescence overlapping rust damage at bottom-left."
  }
```

The BLIP model was fine-tuned on the fine-tuning dataset using a batch size of 2 and a learning rate of 5e-5, while the BLIP2 model was trained with the same batch size and learning rate of 10e-5. The developed model was then evaluated and deployed in a web-based application using Gradio, which is an open-source Python library for quickly creating an interactive frontend.

5 Results

Both the BLIP and BLIP2 models fine-tune quickly on the relatively small dataset. Training and validation losses decrease exponentially and then stabilize after a few epochs. This indicates a good fit of the model to the training data. It is also observed that when fine-tuned for longer epochs (>20), the model begins to overfit to the dataset, as observed by the slightly increasing validation loss after epoch 30.

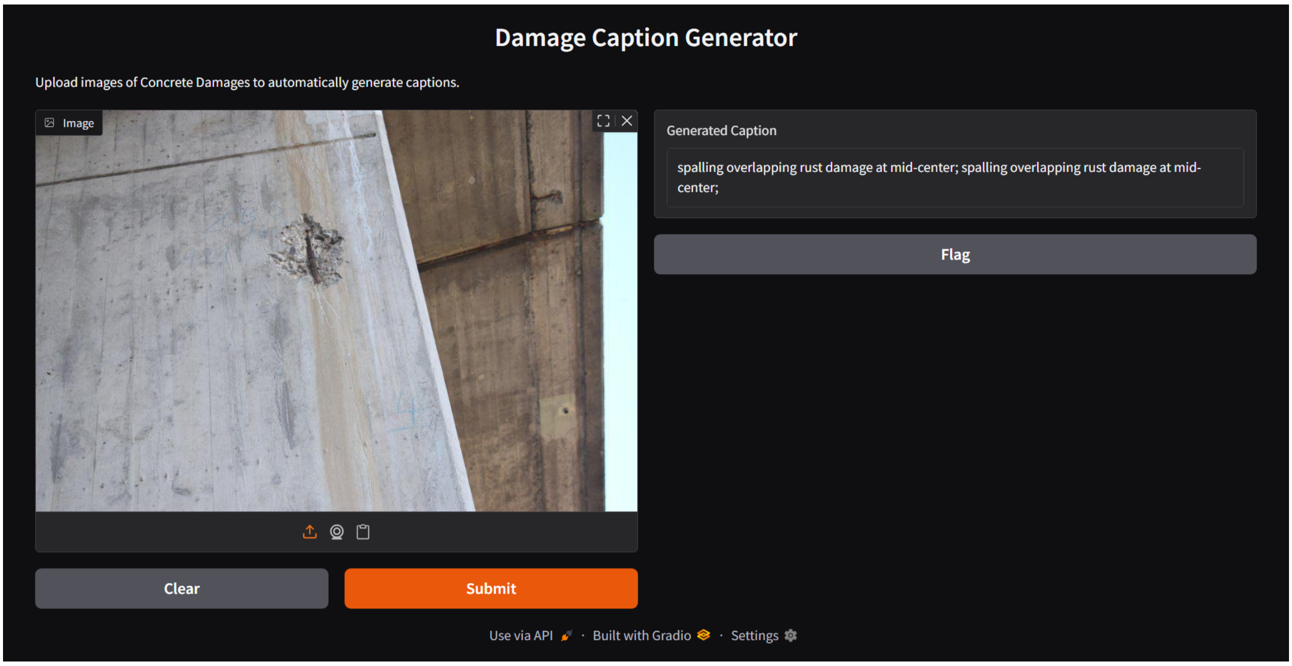
Table 1: Image Captioning Model Evaluation Metrics indicate better scores for BLIP2 model as compared to the BLIP model.

Sr. No.	Model	BLEU	METEOR	ROUGE
1	BLIP	0.19	0.17	0.21
2	BLIP2	0.23	0.45	0.27

The fine-tuned models are evaluated on the validation dataset of 200 images. The generated captions are stored in a nested JSON file, and the average BLEU, METEOR, and ROUGE scores are computed for the models. The scores are tabulated below in Table 1. The results indicate that the BLIP2 model performs better than the BLIP model. It is observed that the captions generated by BLIP2 are in the correct schema and coherent as compared to the BLIP captions. Therefore, the BLIP2 model has learned to recognize damage in concrete images from captions and can describe them on new, unseen images to some extent. The BLIP2 model sometimes generates inaccurate or repetitive captions. Hence, there is still some room for improvement in the caption generation capability of the model, which is discussed in section 7. The BLIP2 model is then deployed in the web application using Gradio as seen in Figure 5.

6 Conclusion

This paper proposed the development of a novel image captioning model for explaining overlapping concrete damages. A captioning dataset was constructed according to predefined criteria from the



Damage Caption Generator

Upload images of Concrete Damages to automatically generate captions.

Generated Caption

spalling overlapping rust damage at mid-center; spalling overlapping rust damage at mid-center;

Figure 5: Damage Captioning Generator: Upload an image with concrete damage to the app, and it generates the corresponding caption as illustrated.

dacl dataset, and a structured captioning schema was developed to annotate the data systematically. This dataset was utilized to fine-tune BLIP and BLIP2 models, which were then evaluated using various metrics, yielding promising initial results. The better-performing BLIP2 model was then deployed in a local web application to generate captions explaining the damage in the images. Currently, the model performs satisfactorily, although it lacks the precision to be used in a commercial application. With further fine-tuning on a better captioned dataset, the model can give improved results.

This application significantly reduces the time and effort required for real-world safety inspections. It automates the inspection workflows and minimizes the errors in inspection checks.

7 Future Work

The model's performance can be further improved by utilizing a more effective captioning schema and a more advanced model, such as GPT2 (Generative Pre-Trained Transformers), or by employing a CNN-RNN architecture-based model. The captioning schema can be extended to include additional parameters, such as a description of the severity of the damage. In this way, the captioning dataset for the BLIP model can be further refined to produce better captions. Therefore, there is still substantial scope for improving the captioning schema and developing optimized fine-tuning strategies for the model to enhance the accuracy of the generated captions.

In this project, the model has been deployed in a web-based application to automatically and efficiently describe the context between overlapping damages in reinforced concrete structures, thereby reducing the time and effort required for the task. The generated captions can be further stored in a database and linked to structural elements in a digital twin or IFC model of the concrete structure. This application could additionally be used for automatic image captioning to annotate pictures of reinforced concrete structures, creating a new caption dataset.

Data availability statement

The data supporting this study's findings are available at https://gitlab.lrz.de/000000003B9CC712/fbi-2025-image-captioning-concrete-damages, ensuring that they are Findable, Accessible, Interoperable, and Reusable (FAIR).

Acknowledgments

The authors gratefully acknowledge the support of the Nemetschek Innovation Foundation as part of the AInspect2Twin project, enabling our research and its practical application.

References

[1] S. O. Abioye et al., "Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges", *Journal of Building Engineering*, vol. 44, p. 103 299, 2021. DOI: https://doi.org/10.1016/j.jobe.2021.103299 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352710221011578

[2] S. Paneru and I. Jeelani, "Computer vision applications in construction: Current state, opportunities & challenges", *Automation in Construction*, vol. 132, p. 103 940, 2021. DOI: https://doi.org/10.1016/j.autcon.2021.103940 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580521003915

[3] J. Flotzinger, P. J. Rösch, and T. Braml, *Dacl10k: Benchmark for semantic bridge damage segmentation*, 2023. arXiv: 2309.00460 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2309.00460

[4] B. Xiao, Y. Wang, and S.-C. Kang, "Deep learning image captioning in construction management: A feasibility study", *Journal of Construction Engineering and Management*, vol. 148, no. 7, p. 04 022 049, 2022.

[5] H. Liu, G. Wang, T. Huang, P. He, M. Skitmore, and X. Luo, "Manifesting construction activity scenes via image captioning", *Automation in Construction*, vol. 119, p. 103 334, 2020. DOI: https://doi.org/10.1016/j.autcon.2020.103334 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580520309146

[6] A. Saka et al., "Gpt models in construction industry: Opportunities, limitations, and a use case validation", *Developments in the Built Environment*, vol. 17, p. 100 300, 2024. DOI: https://doi.org/10.1016/j.dibe.2023.100300 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666165923001825

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 `[cs.CL]`. [Online]. Available: https://arxiv.org/abs/1810.04805

[8] A. Dosovitskiy et al., *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2010.11929

[9] J. Li, D. Li, C. Xiong, and S. Hoi, *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, 2022. arXiv: 2201.12086 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2201.12086

[10] J. Li, D. Li, S. Savarese, and S. Hoi, *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*, 2023. arXiv: 2301.12597 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2301.12597

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation”, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL, 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040.pdf

[12] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments”, in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72. [Online]. Available: https://www.cs.cmu.edu/~alavie/METEOR/meteor-1.5.pdf

[13] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries”, in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013.pdf