Deep Learning Image Captioning in Construction Management: A Feasibility Study

Bo Xiao[1], Yiheng Wang[2], and Shih-Chung Kang[3*]

1 Research Assistant Professor, Department of Building and Real Estate, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: eric.xiao@polyu.edu.hk

2 Ph.D. Student, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada, T6G 2R3. E-mail: yiheng6@ualberta.ca

3 Professor, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada, T6G 2R3. E-mail: sckang@ualberta.ca (corresponding author)

Abstract:

Deep learning image captioning methods are able to generate one or several natural sentences to describe the contents of construction images. By deconstructing these sentences, the construction object and activity information can be retrieved integrally for automated scene analysis. However, the feasibility of deep learning image captioning in construction remains unclear. To fill this gap, this research investigates the feasibility of deep learning image captioning methods in construction management. First, a linguistic schema for annotating construction machine images is established, and a captioning dataset is developed. Then, six deep learning image captioning methods from the computer vision community are selected and tested on the construction captioning dataset. In the sentence-level evaluation, the Tsfm-SCST method has obtained the best performance among six methods with the BLEU-1 score of 0.606, BLUE-2 of 0.506, BLEU-3 of 0.427, BLEU-4 of 0.349, METEOR of 0.287, ROUGE of 0.585, CIDEr of 1.715, and SPICE score of 0.422. In the element-level evaluation, the Tsfm-SCST method has achieved an average precision of 91.1%, recall of 83.3%, and an F1 score of 86.6% for recognition of construction machine objects by deconstructing the generated sentences. This research indicates that deep learning image captioning is feasible as a method of generating accurate and precise text descriptions from construction images, with potential applications in construction scene analysis and image documentation.

Keywords: Deep Learning; Image Captioning; Construction Machines; Feasibility Study; Vision-based Monitoring.

INTRODUCTION

Construction videos contain important visual information (e.g., working objects and their activities) that is of benefit for project management purposes (Xiao and Zhu 2018). By analyzing construction videos using vision-based methods, many applications can be developed to automatically monitor crew productivity, identify safety risks, and optimize working spaces (Yang et al. 2015). For example, Chen et al. (2020) developed a vision-based system to calculate excavator productivity in earthmoving. Kolar et al. (2018) proposed a vision-based method for identifying safety guardrail at construction sites in order to prevent workers from accessing hazardous site areas. Compared with other advanced technologies (e.g., laser scanners, radio frequency identification, global positioning systems), vision-based monitoring of construction sites offers the advantages of lower cost, simple deployment, and easy maintenance (Xiao and Kang 2021a).

Scene analysis refers to interpreting the construction image by identifying construction elements (e.g., objects, activities, and relationships between objects), which provides a holistic view of visual information pertaining in images to construction management (Liu et al. 2020). The applications of scene analysis include automated detection of safety rule violation (Wang et al. 2019), roadway asset evaluation (Balali and Golparvar-Fard 2015), and construction hazards identification (Fang et al. 2020). However, existing scene analysis methods are template-based, which refer to retrieving construction elements separately by object detection (Xiao et al. 2021b) or activity recognition (Golparvar-Fard et al. 2013) and combining all information into a pre-defined template, and have two limitations: 1) the different scene information in construction images is retrieved separately, making it a time-consuming process;

and 2) the retrieved scene information is usually combined based on pre-defined orders to generate a set of words or a sentence as the final output, but these results are prone to error or incompleteness.

Image captioning refers to the automatic generation of one or several sentences to describe the contents of an image; it is a disciplinary technology rooted in computer vision and natural language processing (Hossain et al. 2019), which can be potentially used for construction scene analysis. Recently, deep learning methods have enjoyed considerable success, as they are capable of extracting high-level features automatically from images for various applications in computer vision, natural language processing, reinforcement learning, and so on (Lecun et al. 2015). By incorporating deep learning, image captioning methods can generate precise and concise text descriptions from images by training on annotated image datasets. A typical deep learning image captioning method is the encoder–decoder architecture (Mao et al. 2014), where a convolutional neural network (CNN) encoder extracts embedding features from the images and a recurrent neural network (RNN) decoder generates text based on the embedded features. By adopting deep learning image captioning, the scene information in construction images or videos can be retrieved integrally in the form of a natural sentence.

Although deep learning image captioning has achieved considerable success within the computer vision community, few studies have adopted these methods in construction scenarios, and the feasibility of deep learning image captioning has yet to be substantiated in the following respects: 1) a method of integrating deep learning image captioning methods for generating construction text descriptions is not available—most studies of deep learning image captioning are conducted in daily life scenarios; 2) a linguistic schema for annotating construction images has yet to be established, while an annotated dataset of construction images is needed in order to train deep learning image captioning algorithms for construction management purposes; and 3) studies within the construction research domain have typically employed only the basic

CNN-RNN method, while advanced deep learning methods (e.g., attention) have not yet been tested in construction applications.

The overall objective of the research described herein is to investigate the feasibility of deep learning image captioning in construction scenarios. To achieve this goal, we propose a method for integrating deep learning image captioning in construction scenarios that consists of three main steps—dataset development, model establishment, and experimental evaluation. The results of this research demonstrate the practical applicability of deep learning image captioning and underscore its potential in the context of construction management. Currently, studies of deep learning image captioning are still in the early stage in the field of construction management. The results of this research are expected to help the researchers in our field to decide if and how to use deep learning image captioning methods in their future construction studies.

LITERATURE REVIEW

Automatic analysis of construction images or videos using vision-based methods is an important emerging avenue of construction research with potential project management benefits in terms of increasing crew productivity, reducing safety risks, and monitoring project progress. In this section, existing studies of applying vision-based methods in construction sites are firstly reviewed in terms of methods and applications. Then, relevant works on deep learning image captioning are reviewed in a comprehensive way. Finally, the research gaps and the objectives of the present study are identified.

Vision-based Monitoring in Construction

Existing vision-based methods adopted in construction include object detection, object tracking, activity recognition, and scene analysis. Many construction applications (Roberts and Golparvar-Fard 2019; Son et al. 2019; Wang and Zhu 2021) have been developed based upon

these four types of vision-based methods. Object detection methods retrieve the localization and categorical information of construction objects (e.g., machines, workers, materials) from construction images or videos (Kulchandani and Dangarwala 2015), which is the fundamental step for many vision-based studies in construction. For instance, Yin et al. (2020) have adopted the deep learning detection method YOLO-v3 (Redmon and Farhadi 2018) to detect defects in sewer pipes. Kim et al. (2018) have proposed a region-based fully convolutional network for detecting construction machines that achieved a mean average precision of 96.33% in their experiments. Object tracking aims to retrieve and interpret the movements of construction objects in continuous frames, which produces the trajectory information by assigning an identification number (ID) to each object. For example, Zhu et al. (2017) have integrated object detection for the purpose of tracking construction workers and machines from jobsite videos. Konstantinou et al. (2019) have developed a vision-based method for tracking construction workers by integrating an appearance model with a filtering model. Xiao et al. (2021) have adopted a deep learning illumination enhancement method for tracking multiple construction machines in a nighttime environment.

Activity recognition has been applied in construction monitoring to identify the activities of construction workers and machines in continuous videos, a task that has traditionally relied upon the results of object recognition and object tracking with indicating the activity and localization information of a specific construction object. For example, Yang et al. (2016) have proposed a cutting-edge video interpretation method based on dense trajectories for recognizing the activities of workers. Luo et al. (2018b) have adopted the two-stream CNN method for recognizing workers' activities in site surveillance videos, achieving 80.5% accuracy. Furthermore, Chen et al. (2020) have adopted the 3D ResNet method (Hara et al. 2018) for recognition of excavator activity in earthmoving. Luo et al. (2018a) have proposed a two-step method that uses CNN for recognition of construction machine activities

from still images from a construction site. What these and other studies available in the literature demonstrate is that activity recognition can be used effectively for advanced vision-based applications including safety control, productivity analysis, and so on.

The goal of scene analysis is to interpret the entire construction image by identifying all construction elements including objects (e.g., machines, workers, and materials), activities of objects, and relationships between objects. Existing scene analysis methods, it should be noted, usually consist of two steps: 1) recognition of construction objects and activities from images by adopting object detection and activity recognition methods, respectively; and 2) identification of relationships between objects based on construction rules. For example, Kim et al. (2016) have proposed a data-driven scene analysis method based on a scene dense alignment module for extracting information from construction images. Rahimian et al. (2020) have proposed a game-like hybrid application to simulate real construction activities in a Building Information Modeling (BIM) environment by analyzing construction scenes from videos. Wang et al. (2019) have developed a crowdsourcing approach to parse the construction scene from images in order to monitor safety compliance in construction scenarios. So far, an integrated method of automatically generating complete scene information for construction scene analysis, though, has yet to be developed.

Deep Learning Image Captioning

Image captioning, which generates one or several sentences from an image to describe the scene information appearing in the image, is an interdisciplinary research area combining computer vision and natural language processing (Huang et al. 2019). Unlike other vision-based methods (i.e. object detection, object tracking, and activity recognition), the goal of image captioning is to interpret an image by natural sentences from a holistic view instead of providing localization information of construction objects appearing in the image. The image captioning framework generally consists of an “encoder” and a “decoder”, where the “encoder”

150 extracts features from images and the “decoder” generates text from the retrieved features
151 (Hossain et al. 2019). Deep learning methods (e.g., CNN and RNN) have performed well in
152 image captioning applications. Mao et al. (2014) have proposed a deep learning image
153 captioning method that uses CNN as the “encoder” and RNN as the “decoder”; this approach
154 has been built upon in the “show and tell” method by only inputting the visual features at the
155 first time-step of RNN (Vinyals et al. 2017).

156 Recently, the attention mechanism has been widely applied in deep learning image
157 captioning, as it allows the neural network to focus on its subset of inputs in selecting specific
158 features. Xu et al. (2015) introduced the attention mechanism in image captioning, and it has
159 performed well on the Common Objects in Context (COCO) captioning dataset (Lin et al.
160 2014). The COCO captioning is a public image captioning dataset in the computer vision
161 community, which contains over one and a half million captions describing over 330,000
162 images collected from common-life scenarios. Gao et al. (2019), meanwhile, have proposed an
163 adaptive attention approach to image captioning that considers both visual information and
164 language context information for caption generation purposes. Closely related to this,
165 transformer attention (Vaswani et al. 2017), though originally proposed for language
166 translation purposes, was soon adopted in image captioning studies (Vig 2019; Zhang et al.
167 2019), where it achieved reliable performance. While the attention mechanism has been
168 successfully applied in the computer vision community, though, its feasibility in construction
169 scenarios has not yet been investigated.

170 In construction management, image captioning can be used for automated scene
171 analysis. By analyzing the sentences generated by image captioning methods, the major objects,
172 activities, and interactions of objects can be retrieved integrally, which reduces the processing
173 time and improves the scene analysis completeness. For example, Liu et al. (2020) have
174 adopted the CNN and long short-term memory (LSTM) neural networks for captioning

175 construction images to manifest construction worker activity scenes. In their research, a
176 linguistic schema used for annotating images of construction workers is proposed and three
177 experiments are conducted to illustrate the feasibility of image captioning in construction.
178 However, that study only focused on construction workers, while the captioning images of
179 construction machines is not included. Bang and Kim (2020) have applied image captioning to
180 drone images for vision-based monitoring of construction sites, achieving a mean average
181 recall (mAP) of 45.52%. However, that study only investigated the use of basic CNN-RNN for
182 image captioning. Although image captioning can provide versatile information for
183 construction scene analysis, it should be noted that image captioning cannot provide effective
184 localization information as to what object detection, object tracking, and activity recognition
185 can offer. Therefore, image captioning can be a promising alternative for construction scene
186 analysis but it cannot object detection or activity recognition for vision-based monitoring of
187 construction sites.

188 **Research Gaps and Objectives**

189 As reviewed above, deep learning image captioning has enormous potential as a method
190 for construction scene analysis, providing an integrated solution for describing construction
191 objects and their activities as captured in images. The following research gaps are identified
192 and addressed in this research:

193 • A robust method incorporating deep learning image captioning to generate construction
194 text descriptions is currently not available. Such a method can be expected to provide
195 a standard approach for developing construction image captioning dataset, selecting the
196 proper image captioning technique, and evaluating image captioning models for
197 construction management.
198 • A linguistic schema for annotating construction machine images with professional text
199 descriptions is currently lacking. Such a schema would be capable of bridging the gap

between deep learning image captioning and construction management. Although there exist linguistic schema (Liu et al. 2020) for construction worker images, these cannot be directly applied to construction machine images.

- The performance of deep learning image captioning methods in construction scenarios needs to be investigated. Most studies (Bang and Kim 2020; Liu et al. 2020) in construction only adopt the basic CNN-RNN method for image captioning, while the use of advanced techniques (e.g., attention and transformer) in construction has not been validated.

To fill these gaps, this research proposes a feasibility study of deep learning image captioning in construction. To achieve this goal, the following objectives are pursued:

- Propose a method for integrating deep learning captioning methods in construction scenarios. The proposed method demonstrates a standard approach to applying deep learning image captioning techniques to construction scenarios. It should be noted that the present research, though it focuses on construction machine images, can be expanded to images of other construction objects (e.g., workers) in future works.
- Develop a linguistic schema for annotating construction machine images. A linguistic schema instructs the annotator to label construction images with professional descriptions that are precise and accurate and that will be useful for construction management purposes. In this context, deep learning models can be trained to generate professional descriptions for construction scene analysis.
- Compare the performance of the state-of-the-art deep learning image captioning methods built upon different mechanisms in construction scenarios. The comparison results show the feasibility of deep learning image captioning in construction scenarios. Meanwhile, researchers in the construction community can select the appropriate

captioning method for a given application based on the comparative results of this research.

METHEDOLOGY

Our methodology of integrating deep learning captioning methods in construction scenarios is introduced in this section. First, the overview of the proposed method is introduced. Then, the details of the dataset development and model establishment steps are described.

Methodology Overview

Figure 1 illustrates the overview of the proposed methodology for incorporating deep learning image captioning methods in construction to generate professional text descriptions from construction images. The method is divided into three main steps: dataset development, model establishment, and experimental evaluation. In the dataset development step, construction images are first collected from different viewpoints, illumination conditions, and construction stages and projects. Then, the collected images are manually filtered to select qualified images in a manner that ensures the diversity of the dataset and mitigates potential overfitting issues. Meanwhile, a linguistic schema for construction machine images is proposed for instructing image annotation. Each machine image is annotated with several sentences to describe what is occurring as captured in the given image. Finally, an image captioning dataset is developed in this step.

In the model establishment step, six deep learning image captioning techniques are identified based on their reliable performance in the computer vision community. Five evaluation metrics are then selected for automatic evaluation of the deep learning image captioning techniques. The implementation of each technique is also demonstrated for model establishment purposes. In the experimental evaluation step, the captioning dataset is divided into a training set and a validation set for testing the six image captioning techniques. They are compared at both the sentence-level and the element-level in order validate the feasibility of

their use in construction management applications. The dataset development and model establishment steps are described in greater detail in the following subsections, while the experimental evaluation and results are described in the next section.

The technical innovations of this study are two-fold: 1) a standard method is introduced to illustrate how to integrate deep learning image captioning methods in construction, which has three steps including dataset development, model establishment, and experimental evaluation. By following this three-step method, other novel image captioning methods can be integrated and evaluated in construction scenarios conveniently; 2) a linguistic schema is proposed to annotate construction machine images with professional text descriptions, which fills the gap of applying deep learning image captioning in construction management. The captioning of other types of construction images (e.g., workers) can be accomplished by following the linguistic schema by updating the object and activity types.

Dataset Development

The development of an annotated image dataset is the fundamental step in applying deep learning image captioning in construction management. To develop the dataset, three major steps need to be accomplished: image collection, selection, and annotation. In this subsection, the details of image collection and selection and linguistic schema and image annotation are described, and a comprehensive summary of the captioning dataset is provided.

Image collection and selection

Collection of construction images with a range of visual characteristics in terms of size, color, shape, and illumination level improves the robustness and generalizability of deep learning image captioning methods in construction scenarios. In our previous work (Xiao and Kang 2021b), we developed an image dataset of construction machines referred to as Alberta Construction Image Dataset (ACID). Images of ten types of construction machines are included in ACID: excavator, compactor, dozer, grader, dump truck, concrete mixer truck, wheel loader,

backhoe loader, tower crane, and mobile crane. The images in ACID are adopted in the present research for further image annotation.

In this previous work, 124,500 construction images were compiled from online sources (i.e., Google Images, Naver, and YouTube), while 37,500 construction images captured by smartphones, fixed-position cameras, and drones from site visits in Edmonton, Canada, and Xi'an City, China, were also included. Then, the 162,000 images collected were manually processed following four steps: 1) removal of duplicate images; 2) removal of low-resolution images; 3) removal of oversized and undersized images; and 4) blurring of all human faces appearing in the images (for privacy reasons). This process yielded 10,000 construction images that qualified for inclusion in the ACID dataset, where Figure 2 shows some sample images from the ACID dataset. In the present research, 4,000 construction images are randomly selected from ACID for the purpose of developing the image captioning dataset.

Linguistic schema and image annotation

For the image captioning dataset, each image needs to be manually annotated with several sentences to describe the contents of the image. The linguistic schema informs the process of annotating the construction images, which in turn play an important role in the application of deep learning image captioning in construction. In typical computer vision applications, annotators are required to describe images in their own words because the target images are captured from daily life. In construction, the text annotations of images must be professional and precise for construction management purposes. As such, the annotators must use specific and correct terms to describe construction objects, activities, and working contents using the linguistic schema rather than simply using their own words as in some other computer vision applications.

Figure 3 shows the linguistic schema used in this study for annotating construction machine images. First, the following elements must be deconstructed from the construction

image according to the linguistic schema: 1) the primary machine object; 2) the machine object cooperating with the primary object; 3) the working contents (e.g., dirt, stone, and construction materials) of the primary object; 4) the activities of the primary machine; and 5) supplementary information, such as color, count, and weather conditions. Then, the correct terms must be matched with each element deconstructed in the previous step. Finally, a logical and correct sentence is formed using words to describe what is occurring in the construction image.

Since the construction images under study are drawn from the ACID dataset, the primary object and cooperating object terms must be selected from among the ten designated construction machine types: excavator, compactor, dozer, grader, dump truck, concrete mixer truck, wheel loader, backhoe loader, tower crane, and mobile crane. As a further measure to ensure the accuracy and precision of the annotations, we develop a list of activities (as shown in Table 1) for each type of construction machine to serve the activity selection in the linguistic schema, where "general activities" are those that are applicable to any type of machine. Annotators are encouraged to select the activity of the primary object from the options listed in Table 1, although they are also permitted to use other activity terms based on their construction knowledge/background if needed.

Thirty volunteer annotators with engineering background from the University of Alberta having been recruited to participate in the annotation task, the annotators are first given a half-hour presentation to introduce the research, including an overview of deep learning image captioning, annotation tasks, and the linguistic schema. The volunteers are then assigned a series of construction images and prompted to write one sentence describing the contents of each image, with further instruction and feedback provided by the research in reference to the annotations as needed. The annotation task having been completed by the volunteers, the authors of this research manually check the annotation results and resolve any errors identified.

Summary of captioning dataset

As mentioned above, 4,000 images from the ACID dataset are annotated and a total of 8,226 captions produced, meaning that each construction image was annotated by two annotators on average. Figure 4 shows the element distribution in the captioning dataset, including the machine terms and activity terms. It can be observed that the excavator and dump truck are the two object terms appearing most frequently in the captioning dataset, while loading and dumping are the two most frequently used activity terms. Table 2 outlines the top 20 n-grams ($n = 2, 3, 4$) in the captioning dataset where the n-gram is defined as the contiguous sequence of n words from natural sentences. As shown in Table 2, the most frequent terms are found to be "loader is" (2069), "wheel loader is" (1151), and "a wheel loader is" (768) for 2-gram, 3-gram, and 4-gram, respectively. The captioning dataset is divided into a training set (80%) and a validation set (20%) for the experimental evaluation step.

Model Establishment

Most of the existing image captioning techniques in use within the construction domain have been built using the CNN-RNN approach, while the use of other advanced techniques in construction has yet to be investigated. In what follows we describe in greater detail the selection of the six deep learning image captioning techniques from the computer vision community, introduce the five metrics employed for automatic evaluation of these techniques, and give an account of the implementation of the resulting image captioning models.

Method selection

Baseline method (Base): The baseline method consisting of CNN and RNN networks is selected for evaluation. In the baseline method, the ResNet101 network (He et al. 2016) is employed as the encoder and the LSTM network is adopted as the decoder. Figure 5 shows the architecture of the baseline method. It should be noted that most studies in construction (Bang

and Kim 2020; Liu et al. 2020) have adopted baseline methods for captioning construction images.

Attention method (Att): In this study, the attention method is selected as the decoder for testing the construction images (ResNet101 having been selected as the encoder). The attention decoder (Xu et al. 2015) allows neural networks to look at different parts of the image at different steps in the sequence, and this approach has enjoyed considerable success within the computer vision community. The architecture of the attention method is shown in Figure 6. Typically, an attention decoder functions as a small neural network added to an LSTM neural network that takes the hidden state as input (meaning it will look at the sequence generated thus far), and outputs a set of weights for the image feature that indicates what areas the LSTM should focus on (based on a heavy weight). The weights are applied to the image features in order to obtain the feature content; the content is then sent back to the LSTM to help generate the output. In contrast to the baseline method, in the attention method LSTM functions as an attention network, while the encoder networks remain the same.

Transformer method (Tsfm): The transformer decoder (Vaswani et al. 2017) is a multi-head attention mechanism that has achieved better performance than the attention decoder in computer vision applications. We implement the transformer method by integrating the ResNet101 encoder with a transformer decoder. As shown in Figure 7, the transformer decoder consists of multi-head attention layers, normalization layers, and feed-forward layers. The multi-head attention layers are a set of parallel attention networks that calculate the attention weights, while the feed-forward layers (e.g., LSTM) are responsible for conducting the bulk of the decoding work.

Self-critical sequence training (SCST): It should be noted that applying specific training strategies will improve the performance of deep learning image captioning methods. In this regard, the self-critical sequence training (SCST) strategy (Rennie et al. 2017) is

integrated in the present study. The SCST adopts reinforcement learning for the purpose of training deep learning image captioning methods, and it uses a non-differentiable task metric for optimization. In SCST, two sequences have been estimated in the inference testing procedure where one is sampled from the softmax distribution and another is greedily sampled. The rewards of two sequences are combined for "self-critical" as the final loss, which makes SCST can be more effectively trained in deep learning image captioning. In the present study, the SCST strategy is applied to all three of the methods described above (i.e., Base, Att, and Tsfm).

To sum up, six deep learning image captioning methods (Base, Base-SCST, Att, Att-SCST, Tsfm, Tsfm-SCST) having been selected for testing on the developed captioning dataset. In the captioning process, the construction image is firstly processed by the same encoder neural networks ResNet101 (the architecture is described in Table 3) to extract a feature vector with the size of 2048 by 1, and then the extracted feature vector is processed by different decoder neural networks in different image captioning methods to generate natural sentences. Moreover, Table 4 illustrates the technical details of six deep learning image captioning methods in terms of decoder neural networks, and the training strategy.

As shown in Table 4, the number of training epochs is 30 for Base, Att, and Tsfm. For Base-SCST and Att-SCST, the models are first trained for 30 epochs by optimizing the cross-entropy loss and then training for 20 epochs using the SCST strategy. For Tsfm-SCST, the model is trained for 15 epochs for the traditional strategy (optimizing the cross-entropy loss) and for 5 epochs for the SCST strategy. To be noted, the training epoch is a hyperparameter determined by finetuning in the implementation stage. In this study, the epoch number of Tsfm-SCST is smaller than Base-SCST and Att-SCST because the authors observed that the Tsfm-SCST converged faster in the training process. This is resulting from: 1) Tsfm methods have a more complicated neural network architecture than Base and Att models. Larger networks are

naturally easier to converge; and 2) the SCST increased the converging speed of training larger neural networks.

Evaluation metric selection

At present, no single general metric for evaluation of image captioning methods in computer vision is available. For the purpose of this study, five automatic evaluation metrics are adopted in order to assess the performance of deep learning image captioning methods in the sentence-level by comparing the ground truth sentences and the generated sentences. These evaluation metrics are Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Metric for Evaluation of Translation with Explicit ORdering (METEOR), Consensus-based Image Description Evaluation (CIDEr), and Semantic Propositional Image Caption Evaluation (SPICE). For these metrics, a higher value indicates better captioning performance. It should be noted that the scale of values for CIDEr is 0 to 10, while the scale for the other four metrics is 0 to 1.

BLEU (Papineni et al. 2001) measures the overlap between the predicted single word or *n*-gram (sequence of *n* adjacent words) and a set of reference sentences. BLEU only measures the word match and sentence length match and does not take the semantic meaning of the words into account. Variations of BLEU include BLEU-1, BLEU-2, BLEU-3, and BLEU-4, where the number appearing after the hyphen signifies the number of words used up to *n*-grams (the variations listed here are the ones adopted in the present study).

ROUGE (Chin-Yew 2004) uses *n*-grams to measure the recall score of the generated sentences relative to the reference sentences. The most widely used version of ROUGE, ROUGE-L, is adopted in the present study. ROUGE-L computes the recall and precision of the longest common subsequences between the candidate and reference sentences.

METEOR (Lavie and Agarwal 2007) introduces semantic matching for automatic evaluation. It includes lexical match, stemmed word match, synonym match, and paraphrase match. The METEOR score is calculated by mapping the unigrams of the candidate and reference sentences and measuring their alignment with one another.

CIDEr (Vedantam et al. 2014) first converts the words in both the candidate and reference sentences into their root forms and then measures the co-existence frequency of the *n*-grams in both sentences. The most commonly used version is CIDEr-D, as it is capable of preventing outlier scores resulting from poor human judgment. The present study adopts the CIDEr-D version.

SPICE (Anderson et al. 2016) calculates the score by measuring the similarity between the scene graph tuples of the candidate and reference sentences. The scene graph includes objects, their attributes, and relationships extracted from the sentence.

Model implementation

The developed deep learning image captioning models are trained on the developed image captioning training set. All images are resized to 256×256 and normalized based on a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225]. Moreover, the sentence annotations are tokenized in order to divide them into lists of single words without punctuation. As part of the tokenization process, the <start> and <end> label are added to the beginning and ending of each token list to indicate the start and end of each annotation.

All six deep learning image captioning methods are implemented in the Python language. The encoder (i.e., ResNet101) and decoders (i.e., LSTM, attention, and transformer) adopted in this study are implemented using the Pytorch library. The ResNet101 is pretrained on the ImageNet dataset, while the Opencv library is employed for image input/output. In terms of hardware, the evaluation is conducted on a computer that features two NVIDIA GTX 1080

444 Ti GPUs (11 GB each), an Intel Core i9-7920X@ 2.90 Hz CPU with 12 cores, and two 32 GB
445 memory cards. The testing environment uses the Ubuntu 16.04 system.

446 **EXPERIMENTAL EVALUATION AND RESULTS**

447 The trained deep learning image captioning models are validated on the validation set
448 for the purpose of experimental evaluation. Evaluations are carried out at both the sentence-
449 level and the element-level. The sentence-level evaluation follows the standard process
450 employed within the computer vision community for image captioning evaluation, which
451 calculates the similarity between the predicted sentences and the manually labeled sentences.
452 However, the sentence-level evaluation is not intuitive for construction management since
453 these metrics only focus on the holistic similarity of sentences. To address this limitation, the
454 element-level evaluation has been conducted in experiments, which aims to evaluate the
455 performance of image captioning methods of retrieving the scene element information (e.g.,
456 objects and activities) from construction images.

457 **Sentence-level Evaluation Results**

458 Table 5 summarizes the sentence-level evaluation results for deep learning image
459 captioning applications in construction. As shown in the table, among the six methods, the
460 Tsfm-SCST is found to achieve the best performance in captioning construction images,
461 attaining the BLEU-1 score of 0.606, BLUE-2 of 0.506, BLEU-3 of 0.427, BLEU-4 of 0.349,
462 METEOR of 0.287, ROUGE of 0.585, CIDEr of 1.715, and SPICE score of 0.422,
463 underscoring the feasibility of the transformer decoder and SCST strategy in deep learning
464 image captioning. In computer vision, it should be noted, the leading scores on the COCO
465 captioning are 0.795 on BLEU-1, 0.635 on BLEU-2, 0.485 on BLEU-3, 0.363 on BLEU-4,
466 0.573 on ROUGE, 0.277 on METEOR, 1.196 on CIDEr and 0.213 on SPICE, and these metrics
467 are close to the performance observed in the present study. This indicates that the deep learning
468 image captioning methods under consideration attain comparable results in construction

applications to those observed in the computer vision community. Figure 8 shows example captioning results generated by six methods in the validation set; it can be seen from these examples that the deep learning image captioning methods under study are capable of correctly describing the contents of construction images in most cases.

The Base method is found in this experiment to rank second in performance, achieving a BLEU-1 score of 0.587, a BLEU-2 of 0.477, a BLEU-3 of 0.398, a BLEU-4 of 0.320, a METEOR of 0.274, a ROUGE of 0.560, a CIDEr of 1.499, and a SPICE score of 0.394, outperforming the attention-based method (Att) and the transformer-based method (Tsfm) in construction scenarios. In computer vision, in contrast, the Att and Tsfm achieve better performance than the Base method. This result demonstrates the degree of difficulty of image captioning in construction is lower than that in traditional computer vision applications. The implications behind this finding are two-fold: 1) currently, the differences of captioning performance among existing methods are insignificant in construction scenarios; and 2) when considering both captioning performance and the processing speed, the Base method are recommended in construction because it achieved comparable performance than other advanced methods (i.e. Att and Tsfm) with consuming less computational resources.

Element-level Evaluation Results

Image captioning provides a holistic solution for understanding the scene information (i.e., objects, activities, and relationships between objects) in an image. By analyzing the generated sentences, this scene information can be retrieved for various uses related to construction object and activity recognition. To validate the feasibility of deep learning image captioning methods for construction scene analysis, the Tsfm-SCST method is evaluated at the element-level in terms of its ability to recognize machine objects and activities in images.

As with the sentence-level evaluation, in the element-level evaluation the Tsfm-SCST method is trained on the training set and validated on the validation set. In the evaluation

process, first the machine objects and activities are extracted from the generated sentences. Then, the retrieved elements are compared with the elements appearing in the ground truth sentences. In the present study, it should be noted, the precision, recall, and F1 score are used for element-level evaluation. The calculation of these metrics is illustrated in Equations 1 to 3.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1\ score = 2 \times \frac{Precison \times Recall}{Precision + Recall} \quad (3)$$

where TP (true positive) is the number of machine objects appearing in both the generated sentences and the ground truth sentences, FP (false positive) is the number machine objects appearing in the generated sentences but not appearing in the ground truth sentences, and FN (false negative) is the number of machine objects appearing in the ground truth sentences but not in the generated sentences.

Table 6 shows the element-level evaluation results for the Tsfm-SCST method in terms of its ability to recognize machine objects. As can be seen in the table, Tsfm-SCST is found to achieve, on average, a precision of 91.1%, recall of 83.3%, and an F1 score of 86.6% for the validation set, meaning that it has close but slightly lower performance with state-of-the-art object detection methods in construction scenarios. The Tsfm-SCST method achieves its highest precision (94.6%) in recognizing dozer objects and its highest recall (92.9%) in recognizing grader objects. Table 7 summarizes the element-level evaluation results produced by Tsfm-SCST in terms of recognizing machine activities. It should be noted that this research is limited in scope to the three activities appearing most frequently in the captioning dataset: loading, dumping, and excavating. As shown in the table, the Tsfm-SCST achieves an average precision of 42.5%, recall of 45.4%, and F1 score of 41.2% in terms of construction machine activity recognition.

There are two reasons why the deep learning image captioning performs relatively poor in construction activity recognition: 1) the limited number of training data. n the proposed study, a captioning dataset that contains 4,000 construction images and 8,226 captions has been developed for training and validation, which is relatively small compared with the datasets in computer vision. When the number of training data increases, the performance of recognizing construction activities will be improved; and 2) the limited types of construction activities are included. Currently, we only considered few types (less than five) of activities for each construction machine and many of activities are overlapped, which increases the difficulty of using deep learning image captioning for construction activity recognition.

The above results suggest that further development is needed in future work before deep learning image captioning can be considered as a replacement method for activity recognition in construction scenarios. To be noted, the image captioning is developing fast in computer vision community and the authors believe this technique will satisfy the requirements of construction management in the close future. When more powerful image captioning models are developed, our proposed method will help researchers in our field to quickly adopt new image captioning models for construction management applications.

DISCUSSIONS

The experimental results indicate the successful achievements of the objectives underlying this research. The specific research findings and limitations are discussed as follows:

- For construction scene analysis, image captioning methods have two advantages over existing construction methods (performing object detection, activity recognition separately and combining all information together), which are: 1) a faster inference speed. The inference speed of image captioning methods is around 30 frames per second (fps). while the inference speed of existing scene analysis methods is less than 1 fps (Kim et al. 2016); and 2) a simpler training process. The image captioning

543 methods only require one annotated dataset and can retrieve multiple scene information
544 from the captioning results, while existing scene analysis methods require multiple
545 datasets for training object detection and activity recognition. To be noted, it does not
546 mean image captioning is more effective than object detection and activity recognition
547 for vision-based monitoring in construction. In fact, they are three different tasks and
548 cannot be compared directly. Meanwhile, object detection and activity recognition
549 cannot be replaced by image captioning since they can provide clear localization and
550 activity information of specific construction objects that cannot be provided by image
551 captioning.

552 • In construction monitoring, the practical implications of image captioning include
553 automated image documentation, evaluation of infrastructure damages, and efficient
554 image querying. For image documentation, image captioning methods can generate
555 natural sentences from images/videos to describe what happened in construction sites.
556 Currently, project managers still need to manually prepare daily reports. And the image
557 captioning has the potential to automatically document the project progress and safety
558 issues in the text format, which will reduce manual documentation efforts. Moreover,
559 image captioning can be used to evaluate infrastructure damages since the generated
560 sentences contain explanatory information. The image captioning can generate an
561 explanatory sentence to identify the damage type and level from infrastructure images.
562 For example, Chun et al. (2021) has adopted image captioning for evaluating the
563 damages of bridges. Meanwhile, image captioning can be employed for efficient image
564 querying in construction management. A large number of images has been accumulated
565 in construction projects, the generated sentences from image captioning can serve the
566 role of “text-index”, which is able to provide versatile information of individual
567 construction images.

- The SCST strategy is shown to improve the performance of image captioning methods in construction. In the evaluations, the three methods in the present study adopting SCST obtain an average BLEU-1 of 0.576, BLEU-2 of 0.483, BLEU-3 of 0.410, BLEU-4 of 0.336, METEOR of 0.271, ROUGE of 0.578, CIDEr of 1.655, and a SPICE of 0.399. For the methods not adopting SCST, the average performance is 0.581 for BLEU-1, 0.471 for BLEU-2, 0.392 for BLEU-3, 0.313 for BLEU-4, 0.271 for METEOR, 0.555 for ROUGE, 1.445 for CIDEr, and 0.388 for SPICE. With the exception of BLEU-1 and METEOR, the methods adopting SCST are found to outperform the methods without SCST. The results indicate that applying specific strategies in training can improve the performance of image captioning methods.
- Figure 9 shows some example captioning errors produced by the Tsfm-SCST method. The top-3 most frequent captioning errors encountered in this study are: 1) mis-recognition of primary machines (e.g., in failure #3, the primary machine wheel loader is misidentified as a grader); 2) mis-recognition of activities (e.g., in failure #4, the wheel loader is driving on the road, whereas it is misidentified as being engaged in dumping soil; and 3) mis-recognition of working contents in failure #7, the Tsfm-SCST fails to recognize the working contents and yields the unreasonable sentence "a mobile crane is lifting the construction site"). There are three main reasons of the captioning errors: 1) the limited number of training images restricts the learning ability of image captioning methods, which results in the mis-categorization of objects and activities; 2) the limited types of construction activities included in the captioning datasets increases the captioning difficulty of construction activities; and 3) the existing image captioning methods are built upon the encoder-decoder mechanism, which focuses on the holistic image features instead of fully considering the sequence relationships of the construction elements in the captioning process.

- The present research has three limitations that need to be addressed in future research: 1) the quantity of the captioning image dataset needs to be increased. The present research develops a captioning image dataset comprising 4,000 images and 8,226 captioning sentences. Compared with datasets in computer vision, the quantity of the proposed dataset is relatively small; 2) the present research only investigates "encoder–decoder" image captioning methods, whereas there are some deep learning image captioning methods available built upon other mechanisms (e.g., captioning by detection (Johnson et al. 2015; Lu et al. 2018)) whose application to construction warrants investigation in future work; and 3) when using image captioning for machine object recognition, the captioning method can only retrieve the primary and cooperating machine objects, while the small machine objects in the background are not captured. Object detection methods, on the other hand, are capable of recognizing all machine objects appearing in the given image or video.

CONCLUSIONS AND FUTURE WORKS

In this research we conducted a feasibility study to investigate the performance of deep learning image captioning in construction scenarios. We proposed a linguistic schema to deconstruct construction machine images into primary objects, cooperating objects, activities, working contents, and supplementary information. Using the linguistic schema, professional descriptions can be annotated for the purpose of training deep learning image captioning methods. A captioning dataset was developed containing 4,000 images and 8,226 sentences. In turn, six deep learning image captioning methods were tested on the captioning dataset in the sentence-level evaluation. In the element-level evaluation, the Tsfm-SCST method achieved F1 scores of 86.6% and 41.2% for recognizing scene objects and activities, respectively, in construction images.

The contributions of this research are three-fold. First, a three-step method has been proposed for integrating deep learning image captioning methods in construction. Using the proposed method, the feasibility of image captioning methods in construction can be investigated. Second, a linguistic schema for annotating construction machine images has been developed. Using this schema, an annotated image dataset has been developed for training deep learning methods for the purpose of captioning images featuring construction machines. Third, six deep learning image captioning methods have been compared using automatic metrics in computer vision community. Meanwhile, an analysis of the efficacy of various image captioning methods in identifying construction objects and activities has been conducted using the Tsfm-SCST method, demonstrating the potential of applying image captioning in scene element analysis.

Future work will focus on expanding the quantity of the captioning image dataset. Currently, only 4,000 images in the ACID dataset are annotated, with an average of two ground truth sentences for each image. In the future, all 10,000 images in the ACID dataset will be annotated, with five ground truth sentences per image. Meanwhile, combining of image captioning methods with advanced training strategies (e.g., SCST and data augmentation) in construction scenarios is another important future work. By adopting these strategies, the performance of image captioning can be significantly improved when the number of training data is limited. We will also work on integrating image captioning methods into advanced construction applications, such as a construction image querying system and auto-generation of daily reports.

DATA AVAILABILITY STATEMENT

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request (e.g. captioning datasets and models).

ACKNOWLEDGEMENTS

The authors would like to thank Mr. Zicong Huang and Ms. Dilyara Tulegenova for assisting with the development of the image captioning dataset.

REFERENCES

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). "SPICE: Semantic propositional image caption evaluation." *2016 European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, 382–398.

Balali, V., and Golparvar-Fard, M. (2015). "Segmentation and recognition of roadway assets from car-mounted camera video streams using a scalable non-parametric image parsing method." *Automation in Construction*, 49, 27–39.

Bang, S., and Kim, H. (2020). "Context-based information generation for managing UAV-acquired data using image captioning." *Automation in Construction*, 112, 103116.

Chen, C., Zhu, Z., and Hammad, A. (2020). "Automated excavators activity recognition and productivity analysis from construction site surveillance videos." *Automation in Construction*, 110, 103045.

Chin-Yew, L. (2004). "ROUGE: A package for automatic evauluation of summaries." *Text Summarization Branches Out*, ACL, 74–81.

Chun, P., Yamane, T., and Maemura, Y. (2021). "A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage." *Computer-Aided Civil and Infrastructure Engineering*.

Fang, W., Ma, L., Love, P. E. D., Luo, H., Ding, L., and Zhou, A. (2020). "Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology." *Automation in Construction*, 119, 103310.

Gao, L., Li, X., Song, J., and Shen, H. T. (2019). "Hierarchical LSTMs with adaptive attention for visual captioning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1112–1131.

667 Golparvar-Fard, M., Heydarian, A., and Niebles, J. C. (2013). “Vision-based action recognition
668 of earthmoving equipment using spatio-temporal features and support vector machine
669 classifiers.” *Advanced Engineering Informatics*, 27(4), 652–663.
670 Hara, K., Kataoka, H., and Satoh, Y. (2018). “Can spatiotemporal 3D CNNs retrace the history
671 of 2D CNNs and imageNet?” *2018 IEEE/CVF Conference on Computer Vision and*
672 *Pattern Recognition*, IEEE, 6546–6555.
673 He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep residual learning for image recognition.”
674 *2016 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas,
675 USA, 770–778.
676 Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). “A comprehensive survey
677 of deep learning for image captioning.” *ACM Computing Surveys*, 51(6), 1–36.
678 Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). “Attention on attention for image
679 captioning.” *2019 IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul,
680 Korea, 4633–4642.
681 Johnson, J., Karpathy, A., and Fei-Fei, L. (2015). “DenseCap: Fully convolutional localization
682 networks for dense captioning.”
683 Kim, H., Kim, H., Hong, Y. W., and Byun, H. (2018). “Detecting construction equipment using
684 a region-based fully convolutional network and transfer learning.” *Journal of Computing*
685 *in Civil Engineering*, 32(2), 04017082.
686 Kim, H., Kim, K., and Kim, H. (2016). “Data-driven scene parsing method for recognizing
687 construction site objects in the whole image.” *Automation in Construction*, 71, 271–282.
688 Kolar, Z., Chen, H., and Luo, X. (2018). “Transfer learning and deep convolutional neural
689 networks for safety guardrail detection in 2D images.” *Automation in Construction*, 89,
690 58–70.
691 Konstantinou, E., Lasenby, J., and Brilakis, I. (2019). “Adaptive computer vision-based 2D

692 tracking of workers in complex environments." *Automation in Construction*, 103, 168–
693 184.
694 Kulchandani, J. S., and Dangarwala, K. J. (2015). "Moving object detection: Review of recent
695 research trends." *2015 International Conference on Pervasive Computing*, IEEE, Louis,
696 USA, 1–5.
697 Lavie, A., and Agarwal, A. (2007). "METEOR: An automatic metric for MT evaluation with
698 high levels of correlation with human judgments." *Second Workshop on Statistical*
699 *Machine Translation*, ACL, Prague, Czech Republic, 228–231.
700 Lecun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning." *Nature*, 521(7553), 436–444.
701 Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick,
702 C. L. (2014). "Microsoft COCO: Common objects in context." *2014 European*
703 *Conference on Computer Vision*, Springer, Zurich, Switzerland, 740–755.
704 Liu, H., Wang, G., Huang, T., He, P., Skitmore, M., and Luo, X. (2020). "Manifesting
705 construction activity scenes via image captioning." *Automation in Construction*, 119,
706 103334.
707 Lu, J., Yang, J., Batra, D., and Parikh, D. (2018). "Neural Baby Talk." *ArXiv*, ID: 1803.09845.
708 Luo, X., Li, H., Cao, D., Dai, F., Seo, J., and Lee, S. (2018a). "Recognizing diverse
709 construction activities in site images via relevance networks of construction-related
710 objects detected by convolutional neural networks." *Journal of Computing in Civil*
711 *Engineering*, 32(3), 04018012.
712 Luo, X., Li, H., Cao, D., Yu, Y., Yang, X., and Huang, T. (2018b). "Towards efficient and
713 objective work sampling: Recognizing workers' activities in site surveillance videos with
714 two-stream convolutional networks." *Automation in Construction*, 94, 360–370.
715 Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. L. (2014). "Explain images with
716 multimodal recurrent neural networks." *arXiv*, ID: 1410.1090.

717 Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). “BLEU: A method for automatic
718 evaluation of machine translation.” *40th Annual Meeting on Association for*
719 *Computational Linguistics*, ACL, Morristown, USA, 311.
720 Pour Rahimian, F., Seyedzadeh, S., Oliver, S., Rodriguez, S., and Dawood, N. (2020). “On-
721 demand monitoring of construction projects through a game-like hybrid application of
722 BIM and machine learning.” *Automation in Construction*, 110, 103012.
723 Redmon, J., and Farhadi, A. (2018). “YOLOv3: An incremental improvement.” *ArXiv*, ID:
724 1804.02767.
725 Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). “Self-critical sequence
726 training for image captioning.” *2017 IEEE Conference on Computer Vision and Pattern*
727 *Recognition*, IEEE, Honolulu, USA, 1179–1195.
728 Roberts, D., and Golparvar-Fard, M. (2019). “End-to-end vision-based detection, tracking and
729 activity analysis of earthmoving equipment filmed at ground level.” *Automation in*
730 *Construction*, 105, 102811.
731 Son, H., Seong, H., Choi, H., and Kim, C. (2019). “Real-time vision-based warning system for
732 prevention of collisions between workers and heavy equipment.” *Journal of Computing*
733 *in Civil Engineering*, 33(5), 1–14.
734 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and
735 Polosukhin, I. (2017). “Attention is all you need.” *arXiv*, ID: 1706.03762.
736 Vedantam, R., Zitnick, C. L., and Parikh, D. (2014). “CIDEr: Consensus-based image
737 description evaluation.” *arXiv*, ID: 1411.5726.
738 Vig, J. (2019). “A multiscale visualization of attention in the transformer model.” *arXiv*, ID:
739 1906.05714.
740 Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2017). “Show and tell: Lessons learned
741 from the 2015 MSCOCO image captioning challenge.” *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 39(4), 652–663.

Wang, X., and Zhu, Z. (2021). "Vision-based hand signal recognition in construction: A feasibility study." *Automation in Construction*, Elsevier B.V., 125(February), 103625.

Wang, Y., Liao, P.-C., Zhang, C., Ren, Y., Sun, X., and Tang, P. (2019). "Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision-based workplace safety." *Advanced Engineering Informatics*, 42, 101001.

Xiao, B., and Kang, S. (2021a). "Vision-Based Method Integrating Deep Learning Detection for Tracking Multiple Construction Machines." *Journal of Computing in Civil Engineering*, 35(2), 04020071.

Xiao, B., and Kang, S. (2021b). "Development of an Image Data Set of Construction Machines for Deep Learning Object Detection." *Journal of Computing in Civil Engineering*, 35(2), 05020005.

Xiao, B., Lin, Q., and Chen, Y. (2021a). "A vision-based method for automatic tracking of construction machines at nighttime based on deep learning illumination enhancement." *Automation in Construction*, 127, 103721.

Xiao, B., Zhang, Y., Chen, Y., and Yin, X. (2021b). "A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation." *Advanced Engineering Informatics*, 50, 101372.

Xiao, B., and Zhu, Z. (2018). "Two-dimensional visual tracking in construction scenarios: A comparative study." *Journal of Computing in Civil Engineering*, 32(3), 04018006.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). "Show, attend and tell: Neural image caption generation with visual attention." *arXiv*, ID: 1502.03004.

Yang, J., Park, M.-W., Vela, P. A., and Golparvar-Fard, M. (2015). "Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and

767 the future." *Advanced Engineering Informatics*, 29(2), 211–224.
768 Yang, J., Shi, Z., and Wu, Z. (2016). "Vision-based action recognition of construction workers
769 using dense trajectories." *Advanced Engineering Informatics*, 30(3), 327–336.
770 Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., and Kurach, L. (2020). "A
771 deep learning-based framework for an automated defect detection system for sewer pipes."
772 *Automation in Construction*, Elsevier, 109(August 2019), 102967.
773 Zhang, B., Titov, I., and Sennrich, R. (2019). "Improving deep transformer with depth-scaled
774 initialization and merged attention." *arXiv*, ID: 1908.11365.
775 Zhu, Z., Ren, X., and Chen, Z. (2017). "Integrated detection and tracking of workforce and
776 equipment from construction jobsite videos." *Automation in Construction*, 81, 161–171.
777

List of Tables

Table 1. List of suggested activities for construction machines

Construction Machine	**Customized Activity**
Excavator	swinging/dumping/excavating/loading/etc.
Compactor	compacting/etc.
Dozer	grading/stripping/loosening/pushing/etc.
Grader	grading/stripping/loosening/pushing/etc.
Dump Truck	dumping/hauling/transferring/etc.
Concrete Mixer Truck	dumping/transferring/loading/etc.
Wheel Loader	dumping/excavating/loading/transferring/etc.
Backhoe Loader	dumping/excavating/loading/transferring/etc.
Tower Crane	lifting/transferring/swinging/etc.
Mobile Crane	lifting/transferring/swinging/etc.
General Activity	travelling/waiting/idling/driving/parking/etc.

Table 2. Statistics of top 20 *n*-grams of the captioning dataset

Index	2-Gram	Count	3-Gram	Count	4-Gram	Count
1	loader is	2069	wheel loader is	1,151	a wheel loader is	768
2	dump truck	1575	a dump truck	979	a backhoe loader is	615
3	wheel loader	1392	a wheel loader	894	a mobile crane is	411
4	excavator is	1319	backhoe loader is	885	concrete mixer truck is	347
5	on the	1025	an excavator is	880	into a dump truck	315
6	an excavator	1014	a backhoe loader	665	a compactor is compacting	313
7	backhoe loader	988	a grader is	618	a dump truck is	270
8	a dump	983	mobile crane is	597	a concrete mixer truck	269
9	truck is	916	a compactor is	508	an excavator is excavating	258
10	a wheel	898	a mobile crane	468	wheel loader is loading	243
11	grader is	896	dump truck is	445	a grader is grading	213
12	crane is	761	compactor is compacting	441	compactor is compacting the	196
13	compactor is	722	concrete mixer truck	401	the wheel loader is	184
14	mobile crane	715	a dozer is	387	in a construction site	168
15	a backhoe	676	excavator is excavating	369	an excavator is dumping	166
16	a grader	645	mixer truck is	350	waiting to be loaded	266
17	the soil	625	loader is loading	319	is travelling on the	163
18	dozer is	590	the dump truck	318	the soil on the	155
19	is loading	585	into a dump	318	dump truck is waiting	152
20	the road	558	is travelling on	316	loading a dump truck	148

Table 3. The architecture of ResNet101

Layer name	**Output size**	**Architecture (kernel size, depth)**
conv1	$112 \times 112 \times 64$	7×7,64; stride 2
conv2_x	$56 \times 56 \times 64$	3×3 max pool; stride 2
		$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	$28 \times 28 \times 256$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4_x	$14 \times 14 \times 512$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$
conv5_x	$7 \times 7 \times 1024$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	$1 \times 1 \times 2048$	Average Pool

Table 4. Technical information of six deep learning image captioning methods

Method	**Encoder**	**Decoder**				**Training Strategy**			
		Decoder type	**# of LSTM layer**	**# of hidden nodes in LSTM**	**# of hidden nodes in Attention**	**Learning rate**	**# of training epoch**		**Batch size**
							Cross-entropy loss epoch	**SCST epoch**	
Base	ResNet101	LSTM	1	512	NA	0.0005	30	NA	10
Base-SCST	ResNet101	LSTM	1	512	NA	0.00001	30	20	10
Att	ResNet101	LSTM + Attention module	1	512	512	0.0005	30	NA	10
Att-SCST	ResNet101	LSTM + Attention module	1	512	512	0.00001	30	20	10
Tsfm	ResNet101	Transformer	NA	NA	NA	0.0005	30	NA	10
Tsfm-SCST	ResNet101	Transformer	NA	NA	NA	0.00001	15	5	10

Note: NA represents Not Appreciable

Table 5. Results of deep learning image captioning method evaluation

Method	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**	**SPICE**
Base	0.587	0.477	0.398	0.320	0.274	0.560	1.499	0.394
Base-SCST	0.546	0.460	0.390	0.318	0.253	0.567	1.549	0.358
Att	0.570	0.463	0.385	0.308	0.267	0.549	1.408	0.382
Att-SCST	0.576	0.484	0.412	0.340	0.274	0.582	1.702	0.418
Tsfm	0.586	0.474	0.392	0.311	0.273	0.556	1.427	0.388
Tsfm-SCST	**0.606**	**0.506**	**0.427**	**0.349**	**0.287**	**0.585**	**1.715**	**0.422**

Note: The best performance is denoted in bold.

Table 6. Element-level evaluation results for Tsfm-SCST for machine object recognition

Machine objects	**Precision**	**Recall**	**F1 score**
Excavator	87.6%	89.9%	88.7%
Compactor	92.9%	90.8%	91.9%
Dozer	94.6%	76.8%	84.8%
Grader	92.9%	92.9%	92.9%
Dump truck	79.6%	89.9%	84.4%
Concrete mixer truck	91.9%	61.8%	73.9%
Wheel loader	93.1%	77.7%	84.7%
Backhoe loader	90.9%	85.7%	88.2%
Tower crane	100.0%	82.1%	90.2%
Mobile crane	87.0%	85.7%	86.3%
Average	91.1%	83.3%	86.6%

Table 7. Element-level evaluation results of Tsfm-SCST for machine activity recognition

Machine activities	**TP**	**FN**	**FP**	**Precision**	**Recall**	**F1 score**
Load	38	123	57	40.0%	23.6%	29.7%
Dump	60	21	95	38.7%	74.0%	50.8%
Excavate	37	59	39	48.7%	38.5%	43.0%
Average	N/A	N/A	N/A	42.5%	45.4%	41.2%

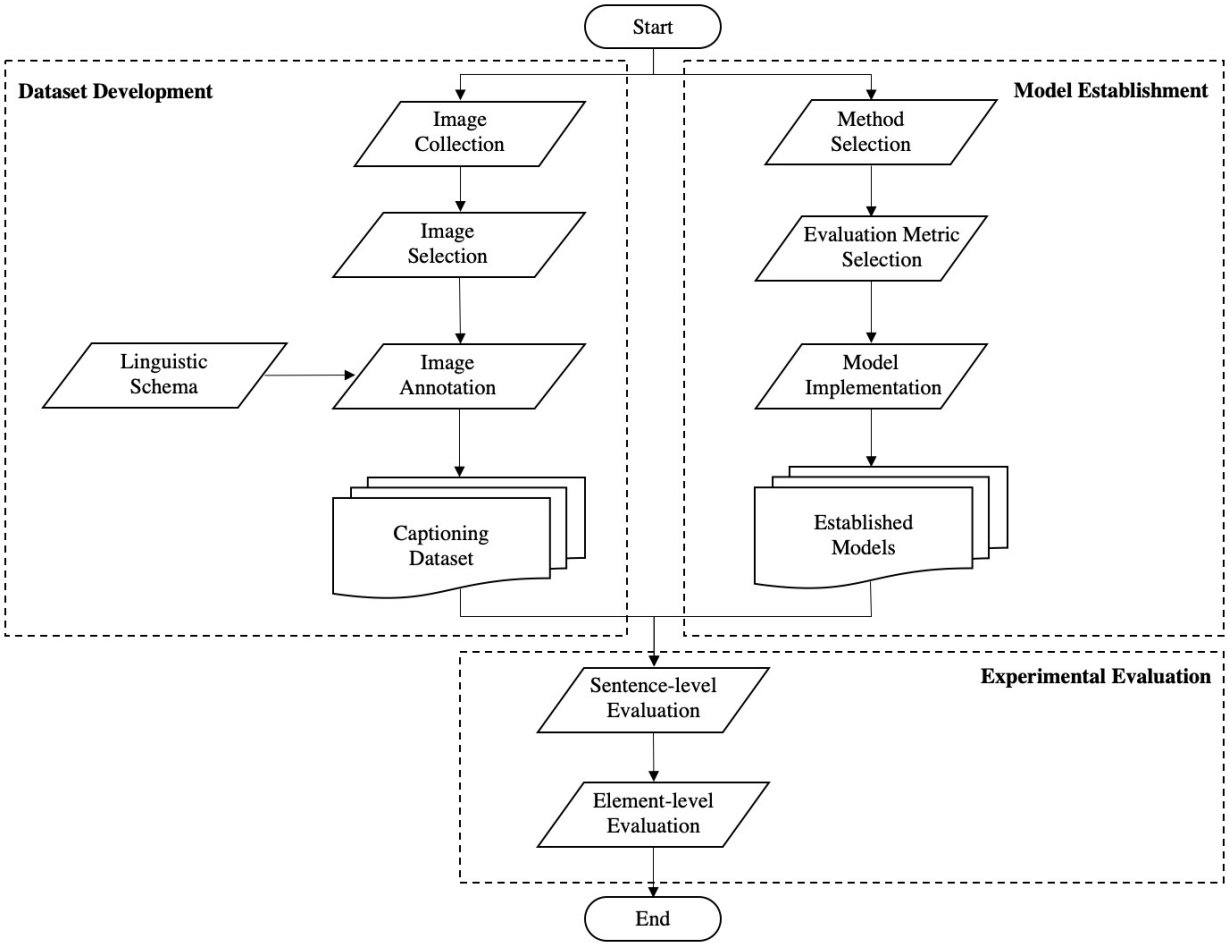

Figure 1. Overview of the proposed methodology

Figure 2. Example construction images in ACID (figure from (Xiao and Kang 2021b), used with permission)

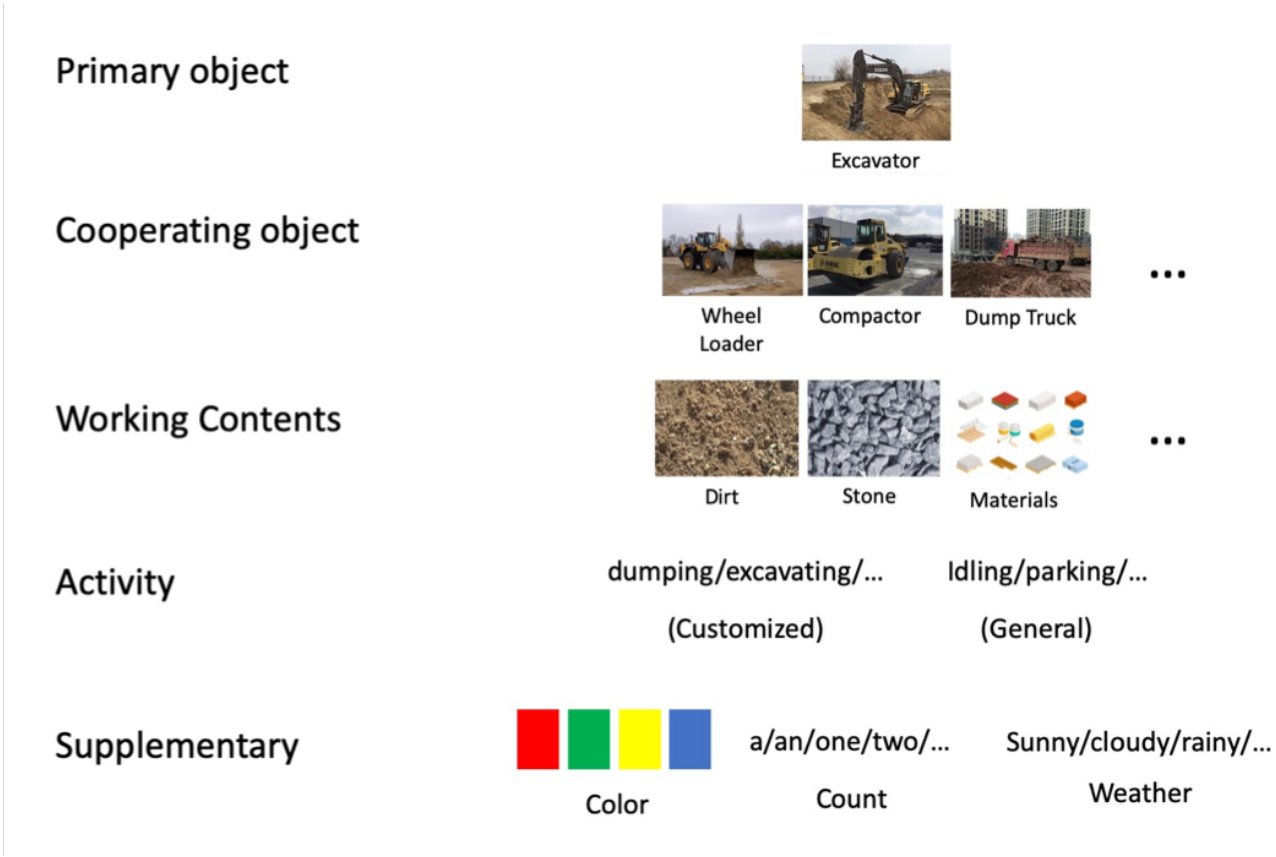

Figure 3. Illustration of linguistic schema

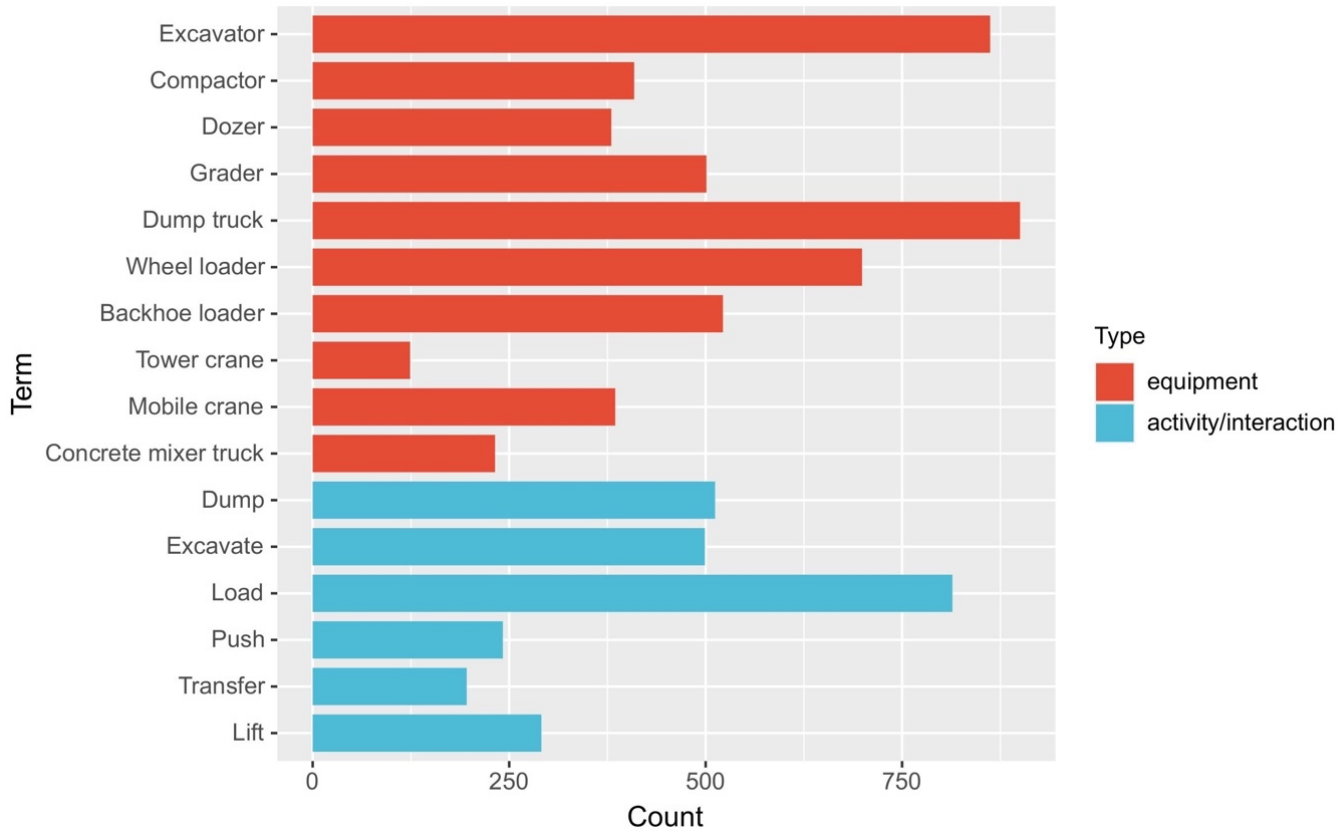

Figure 4. Element distribution of captioning dataset

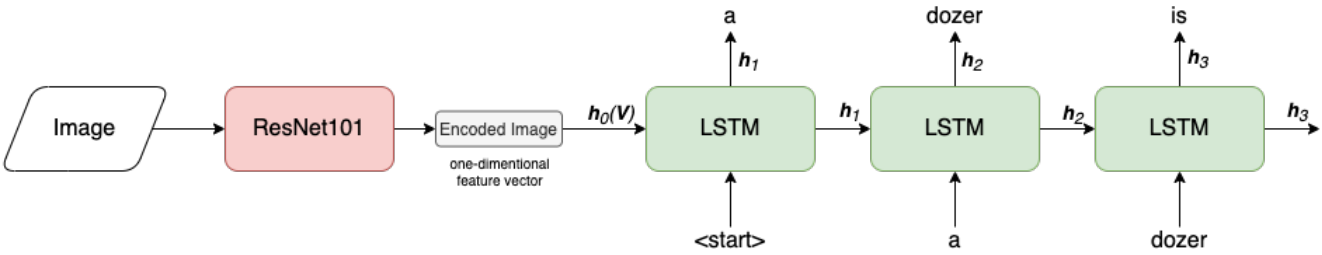

Figure 5. Architecture of the Base method

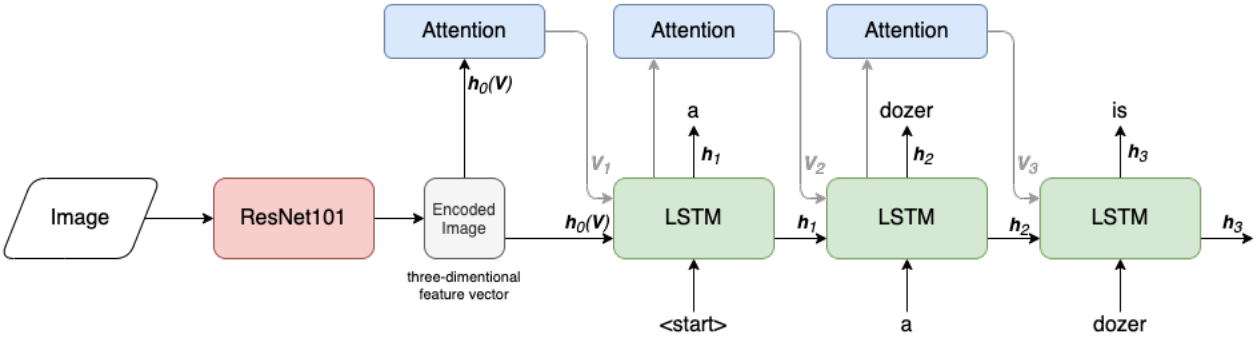

Figure 6. Architecture of the Att method

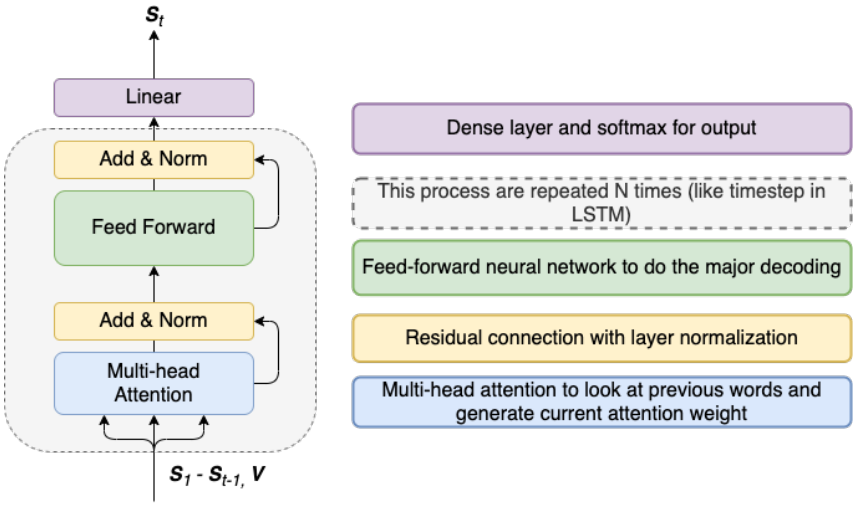

Figure 7. Illustration of transformer decoder

base: a backhoe loader is moving on the site
base-sc: a backhoe loader is travelling on a road
att: a backhoe loader is loading rocks
att-sc: a backhoe loader is moving on the road
tsfm: a backhoe loader is parked at a site
tsfm-sc: a backhoe loader is excavating the soil

base: a wheel loader is loading a dump truck with dirt
base-sc: a wheel loader is loading dirt
att: a wheel loader is loading dirt to a dump truck
att-sc: a wheel loader is loading a dump truck
tsfm: a wheel loader is loading soil into a dump truck
tsfm-sc: a wheel loader is dumping soil into a dump truck

base: a backhoe loader is moving on the site
base-sc: a backhoe loader is travelling on the site
att: a backhoe loader is moving on the road
att-sc: a backhoe loader is moving on the site road
tsfm: a backhoe loader is parked on the road
tsfm-sc: a backhoe loader is excavating soils

base: an excavator is loading a dump truck
base-sc: an excavator is excavating soil and a dump truck is waiting to be loaded
att: an excavator is dumping soil into a dump truck
att-sc: an excavator is excavating soil into a dump truck
tsfm: an excavator is dumping soil into a dump truck
tsfm-sc: two excavators are dumping soil into dump trucks

base: an excavator is loading a dump truck
base-sc: an excavator is dumping soil into a dump truck
att: an excavator is dumping into a dump truck
att-sc: an excavator is dumping soil into a dump truck
tsfm: an excavator is dumping dirt to a dump truck
tsfm-sc: an excavator is dumping dirt to a dump truck and several dump trucks are waiting

base: a wheel loader is loading a dump truck
base-sc: a wheel loader is loading a dump truck with dirt
att: a wheel loader is loading a dump truck
att-sc: a wheel loader is excavating
tsfm: a wheel loader is dumping soil to a dump truck
tsfm-sc: a wheel loader is excavating soil

base: a wheel loader is loading a dump truck with dirt
base-sc: a wheel loader is travelling on a construction site
att: a wheel loader is driving on the ground
att-sc: a wheel loader is travelling on a site
tsfm: a wheel loader is driving
tsfm-sc: a wheel loader is driving on the site

base: a grader is grading the road
base-sc: a grader is grading the soil
att: a grader is removing snow
att-sc: a grader is grading the snow
tsfm: a grader is grading on the snow
tsfm-sc: a grader is grading the snowy ground

Figure 8. Example captioning results in the validation set

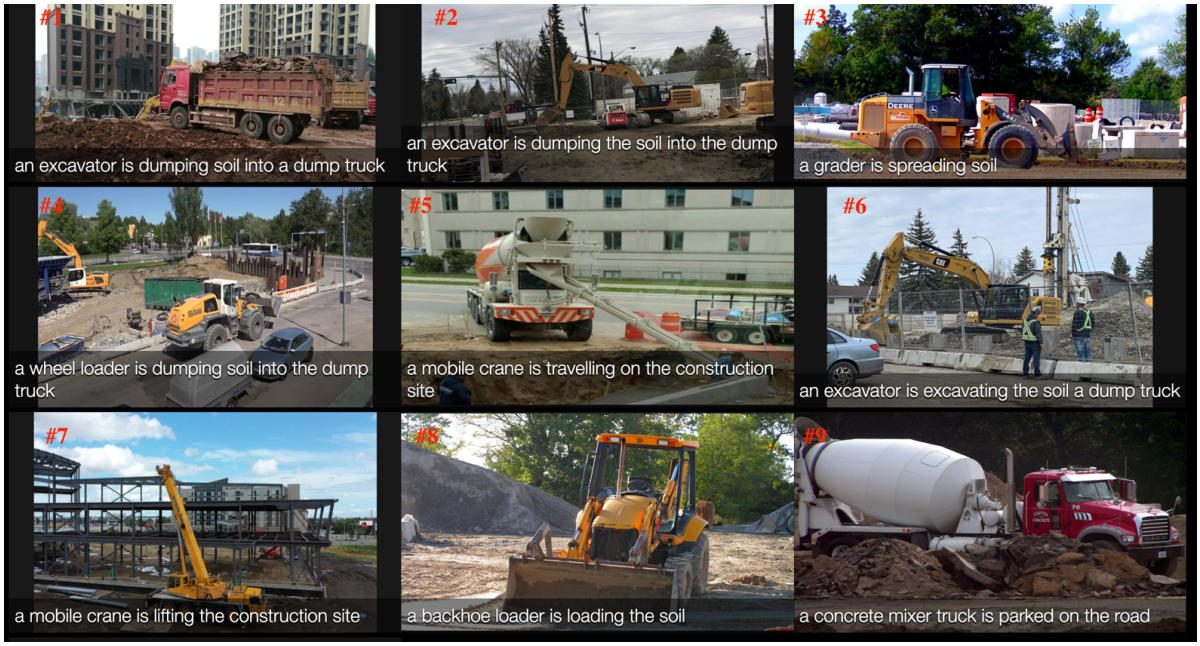

Figure 9. Example captioning errors committed by Tsfm-SCST