Enhancing Construction Project Management through a Cross-Modal Retrieval System

Stanford CS224N Custom Project

Jayadev Rajan
jayrajan@stanford.edu
Mentor: Shijia Yang

Abstract

The dynamic and complex nature of construction projects demands innovative solutions for effective project management. Traditional methods for managing visual information, relying on manual intervention, often fall short in terms of efficiency and accuracy. This project introduces a pioneering Cross-Modal Retrieval System designed to revolutionize the way construction images are classified and their textual descriptions are generated and matched. By utilizing cutting-edge multimodal embedding techniques, including CLIP, ViLT and Vision Encoder Decoder, along with a ResNet50-based baseline for image analysis, the system enables intricate queries like "images of excavator and dumpster" to yield precise results. Furthermore, it can generate accurate textual descriptions from construction site images, significantly enhancing safety compliance and project oversight. Findings from this project reveal a substantial improvement over existing methods, showcasing the potential of this system to transform construction project management practices.

1 Introduction

Construction project management is multi faceted, involving numerous activities that require meticulous coordination and monitoring. A critical aspect of this process is the management of visual information from construction sites, which is essential for ensuring safety compliance, quality assurance, and efficient resource utilization. The traditional methods, which often rely on manual inspection and documentation, are inefficient and prone to errors.

The primary challenge addressed in this project is the need for an automated system that can accurately match construction images with a few labels. Additionally, it's extended to generate captions from images. Such a system would streamline the process of identifying and retrieving relevant visual data, thereby enhancing the efficiency and effectiveness of construction project management.

Existing approaches to managing construction site images are largely manual and involve basic image tagging systems. These methods are not only time-consuming but also lack the integration needed to effectively match images with detailed textual descriptions. Furthermore, the scalability and adaptability of these solutions are often limited, making them unsuitable for the dynamic nature of construction projects.

This project introduces a Cross-Modal Retrieval System that utilizes advanced multimodal embedding techniques, such as CLIP(Contrastive Language-Image Pretraining) Radford et al. (2021), Vision Encoder Decoder Hug (2024c) and ViLT(Vision and Language Tranformer) Kim et al. (2021). The proposed system bridges the gap between visual and textual data, enabling efficient object detection and retrieval. Key features include handling construction image classifications and generating precise textual descriptions from construction site images, demonstrating its efficacy in improving safety compliance and project management. The system's ability to accurately identify key objects and conditions from images highlights its potential to revolutionize construction project management practices.

2 Related Work

Multimodal embeddings have gained significant attention in recent years due to their ability to learn joint representations of images and text. The CLIP Radford et al. (2021) model by Radford et al. (2021) is one of the seminal works in this area. CLIP demonstrates that aligning images and text in a shared embedding space allows for impressive zero-shot learning capabilities across various tasks.

Building on the principles of CLIP, the VisionTextDualEncoder Hug (2024a) model further explores the alignment of visual and textual data. The work by Kim et al. (2021) on ViLT Kim et al. (2021)(Vision-and-Language Transformer) integrates vision and language processing into a single model, enhancing the contextual understanding of image-text pairs. Vision Encoder Decoder models, such as those discussed in Transformer-based Optical Character Recognition with Pre-trained Models Li et al. (2022), extend this approach by employing encoder-decoder architectures to retrieve words from images, which is crucial for tasks requiring precise information extraction.

Several baseline approaches have been proposed for cross-modal retrieval in construction project management. Traditional methods, such as basic image tagging and keyword-based search, are often limited by their reliance on manual input and lack of contextual understanding. More recent approaches leverage computer vision techniques, but these typically fall short in integrating detailed textual descriptions with visual data. For example, Ren et al. (2015) introduced the Faster R-CNN He et al. (2015) model for object detection, which has been widely used as a baseline in various image retrieval tasks. However, this model does not inherently support generation of textual descriptions.

Alternative approaches to cross-modal retrieval include the use of generative models and attention mechanisms. The work by Xu et al. (2015) on the Show, Attend and Tell model Xu et al. (2016) employs attention mechanisms to generate image captions, demonstrating the potential for detailed textual descriptions. Vision Encoder Decoder models further enhance this capability by using transformer-based architectures to encode visual inputs and decode them into textual outputs. Other notable approaches include the use of transformers for vision-language tasks, such as the work by Li et al. (2019) on VisualBERT Li et al. (2019), which integrates BERT (Bidirectional Encoder Representations from Transformers) with visual features for enhanced understanding of image-text pairs. These methods provide valuable insights but often lack the precision needed for specific tasks in construction project management.

The application of image-text retrieval technologies extends beyond typical media and advertising domains, impacting fields such as medical imaging and autonomous driving. For instance, Multimodal Multitask Deep Learning for X-Ray Image Retrieval Yu et al. (2021) explored the use of these technologies in diagnosing diseases from medical imaging and reports, showcasing the potential of multimodal embeddings in healthcare.

3 Approach

3.1 Baseline: ResNet50

The baseline model for this project is based on ResNet50 He et al. (2015), a widely used convolutional neural network (CNN) for image classification tasks. ResNet50's architecture, which includes 50 layers with skip connections, allows it to effectively handle deep learning tasks by mitigating the vanishing gradient problem. To establish a baseline performance with ResNet50, a random subset of 200 images was selected from the dataset. The ResNet50 model was applied to this subset to evaluate its performance in detecting objects within the images. The spatial coordinates of the inferece were ignored and the objects detected were treated as labels. The results from this baseline model provide a point of comparison for the more advanced models developed in this project.

3.2 CLIP from OpenAI

The CLIP (Contrastive Language-Image Pre-training) Radford et al. (2021) model from OpenAI is designed to understand and align images and texts within a shared embedding space. This model employs a contrastive learning strategy to effectively bridge the gap between visual and textual data.

The model comprises two main components: a visual encoder and a text encoder. The visual encoder is based on a Vision Transformer(ViT), such as `ViT-B/32`, which processes images to extract highlevel visual features. The text encoder utilizes a transformer architecture, similar to those used in models like GPT or BERT, to process textual input. These encoders jointly map their respective inputs into a shared embedding space, enabling the learning of cross-modal relationships.

The training procedure for CLIP involves using a contrastive loss function, which maximizes the cosine similarity between the correct pairs of images and texts while minimizing it for the incorrect pairs. This contrastive approach is critical in fine-tuning the model's ability to produce closely aligned representations of text and images, enhancing its accuracy in matching textual descriptions with corresponding images. The model adjusts the weights during training such that it can correlate image and text pairs that are related and distinguish that those are unrelated.

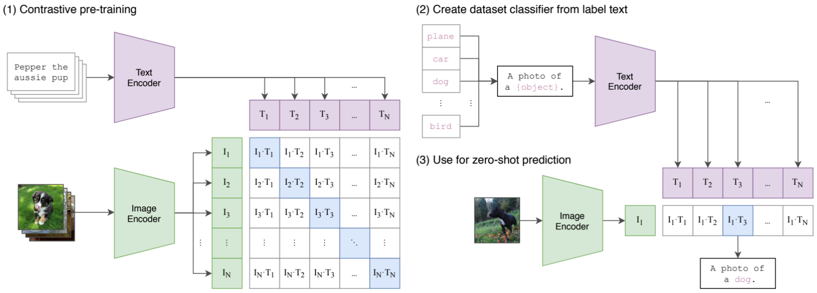
(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ... IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Figure 1: Architecture of CLIP Radford et al. (2021)

3.3 ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

ViLT (Vision-and-Language Transformer) Kim et al. (2021) simplifies the integration of visual and textual data by eliminating the need for convolutions and region supervision. Instead, it processes image patches and text tokens directly with a unified transformer model.

The architecture of ViLT uses a transformer model where both visual and textual inputs are tokenized and processed through a shared set of transformer layers. The visual inputs are divided into fixed-size patches, similar to the approach used in Vision Transformers (ViT), and these patches are embedded into the same space as the text tokens. This unified processing allows the model to learn complex relationships between visual and textual data.

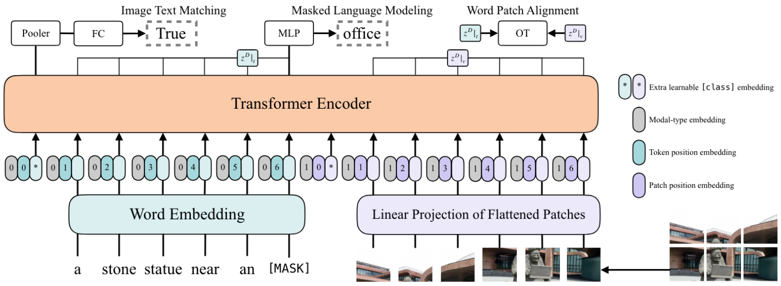

Figure 2: Architecture of ViLT Kim et al. (2021)

3.4 Vision Encoder Decoder Model

The Vision Encoder Decoder Hug (2024c) model combines the strengths of encoder-decoder architectures with advanced vision and language processing techniques. This model is particularly effective for generating detailed textual descriptions from images.

The Vision Encoder Decoder model comprises two main parts: the encoder and the decoder. The encoder is typically a CNN or a transformer-based vision model that processes the image and extracts relevant features. The decoder is usually a transformer-based language model that takes the encoded visual features and generates corresponding textual descriptions. This architecture allows for a detailed and coherent transformation of visual information into text.

Training the Vision Encoder Decoder model involves optimizing the model to generate accurate and relevant descriptions for given images. The training process uses a combination of cross-entropy loss for language generation and other auxiliary losses to improve the alignment between visual features and generated text. The model is trained on large datasets of image-text pairs to learn the nuances of describing complex visual scenes accurately.

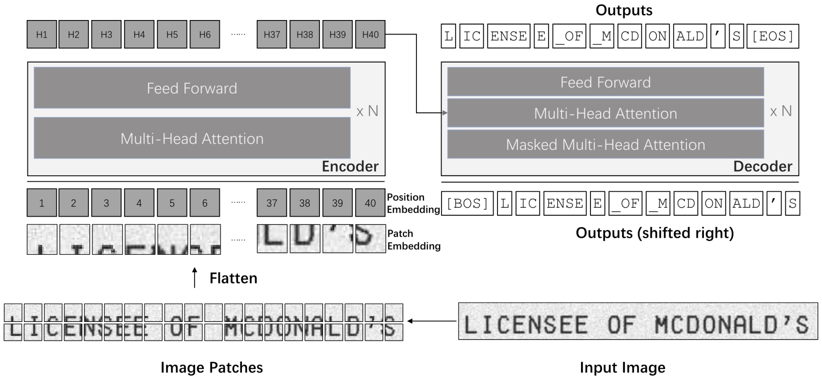

Figure 3: Architecture of Vision Encoder Decoder Li et al. (2022)

4 Experiments

4.1 Data

The dataset utilized in this project is a custom collection of images from various construction sites. Each image has been meticulously annotated with the objects present in the scene, a process completed by a third-party annotation service. This annotation effort has produced a comprehensive dataset comprising over 5,000 images for training and 2,000 images for evaluation.

For the purpose of this project, the identified objects within each image are treated as labels. These annotations include a diverse range of objects commonly found in construction sites, such as `FORMWORK_PANEL` and `MOBILE_CRANE`. There are 43 such classes in the dataset. To ensure a focused and relevant annotation for training, only the top three to five largest objects in each image are considered as labels.

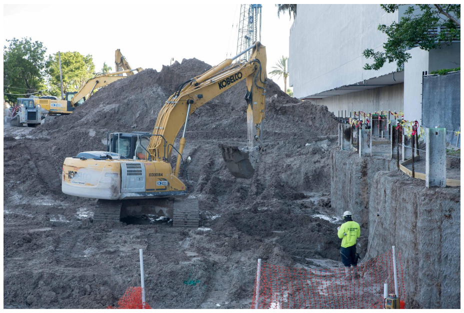

Figure 4: An image of excavator and skid steer

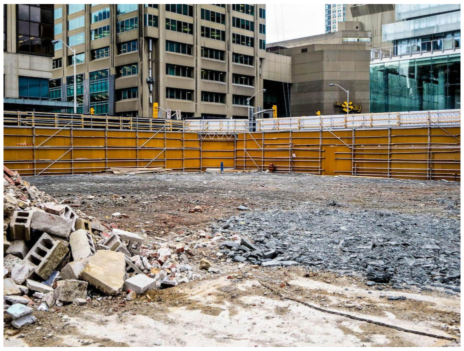

Figure 5: An image of lumber bundle and cone

4.2 Evaluation Method

The performance of the models is evaluated using standard metrics for image-text retrieval, providing a comprehensive understanding of their effectiveness in accurately matching construction images with textual descriptions. The evaluation is based on the following metrics:

- **Accuracy**: Measures the proportion of correctly identified items out of the total items. This metric evaluates how accurately the model identifies objects within images and matches them with the correct textual descriptions. It is crucial for assessing the reliability of the retrieval system in practical applications.
- **Mean Reciprocal Rank (MRR)**: This metric evaluates the average inverse rank of the first correct answer returned by the system. MRR is particularly useful for understanding at what position the correct result appears, on average, in the list of responses the system provides. It is an indicator of the effectiveness of the retrieval system in returning relevant results as its top choices.
- **Top-k Accuracy**: Measures the accuracy of the system in having the correct answer within the top 'k' positions of its ranked output list. This metric helps to gauge the model's precision in retrieving relevant images or descriptions within the top results, reflecting the user's experience by showing the likelihood that a correct and relevant result is immediately visible without needing to look further.

Another useful metric was average time for an inference. The models are either labelling or generating captions. The faster the inference, the helpful it is for the user to accept the suggestion in real time.

4.3 Experimental Details

- **Baseline: ResNet50**: An out-of-the-box ResNet50 model was used for this task. The model was applied directly to the dataset to evaluate its performance in detecting objects within images. This baseline serves as a comparative benchmark to evaluate the enhanced capabilities of the more sophisticated models used in later experiments.
- **CLIP (Contrastive Language-Image Pre-training)**: The CLIP Hug (2024b) model from OpenAI utilizes a contrastive learning approach to effectively align text and image representations. I used the pre-trained version of CLIP available at (`openai/clip-vit-base-patch32`). The model was fine-tuned on our dataset for 5 epochs with a carefully selected learning rate of `5e-5`. A very low weight decay of `.001` was used in order to stabilize the training which otherwise was stuck in local minima. The betas were set to `0.9,0.98`. Batch size was set to 50 and trained on a 1 GPU with a memory of 24GB. This fine-tuning was specifically targeted at improving the model's capability to generate and match accurate textual descriptions with corresponding images, leveraging its powerful cross-modal understanding.

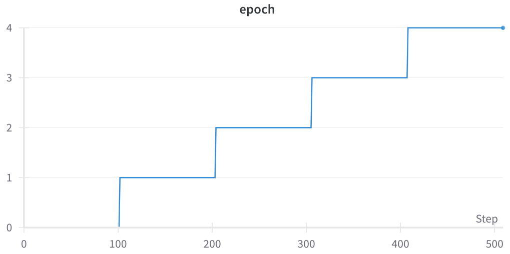

Figure 6: Epochs

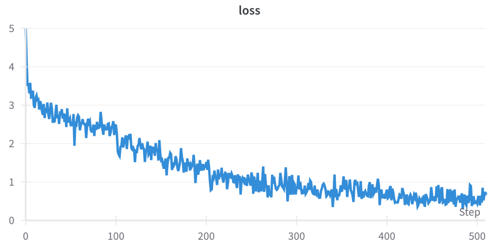

Figure 7: Loss Over time

- **ViLT**: The Vision-and-Language Transformer (ViLT) model, with its architecture of 12 layers, 768 hidden units, and 12 attention heads, processes image patches and text tokens through a shared transformer structure. This model excels in the integrated processing of visual and textual data, allowing for a highly efficient workflow that minimizes the need for extensive preprocessing. The ViLT model employed in this project is configured specifically for enhanced image and text retrieval tasks.

  These configurations enable the model to directly handle and analyze the input data, optimizing the retrieval process and enhancing the overall performance in matching images with their corresponding textual descriptions.
- **Vision Encoder Decoder**: The Vision Encoder Decoder model features an advanced encoder-decoder architecture tailored for image-to-text applications. Uti-

lizing a Vision Transformer (ViT) for the encoder, specifically pretrained with `google/vit-base-patch16-224-in21k`, this model excels in processing visual inputs from construction site images, capturing detailed and relevant features. The decoder, powered by a BERT-like transformer model `google-bert/bert-base-uncased`, translates these features into coherent textual descriptions. The training of this model was conducted over four epochs with an AdamW optimizer. Batch size was set to 4, and a learning rate of `5e-5` was chosen, focusing on optimizing cross-entropy loss for textual output accuracy and applying additional regularization techniques. The model was also trained on 1 GPU with 24GB memory. This rigorous training regimen ensures that the generated descriptions are both accurate and contextually relevant, enhancing the system's utility in real-world construction management scenarios.

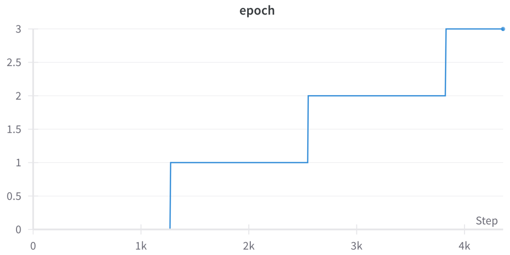

Figure 8: Epochs

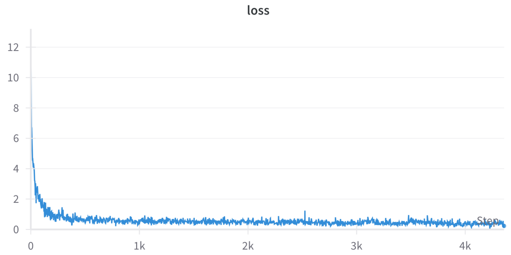

Figure 9: Loss Over time

4.4 Results

All the models were evaluated on the same metrics. Code for baseline and training code for the models can be found at https://github.com/jay-rajan/cross-modal-retrieval. Each of the models were put through a set of 200 images selected from validation set to assess different metrics. Below is the summary of all the metrics collected from the models.

Table 1: Performance Metrics of Construction Image-Text Retrieval Models

Model	Accuracy	Mean Reciprocal Rank (MRR)	Top-k Accuracy
Baseline(ResNet50)	0.11	0.08	Top-3: 0.11
ViLT(Pretrained)	0.60	0.475	Top-3: 0.60
CLIP(Finetuned)	0.72	0.597	Top-3: 0.72
Vision Encoder Decoder(Finetuned)	0.83	0.7395	Top-5: 0.83

In addition, the average inference time is also noted from these models.

Table 2: Average Inference Times of Construction Image-Text Retrieval Models

Model	Average Inference Time (seconds)
ResNet50	0.05943
ViLT	0.17505
CLIP	0.09567
Vision Encoder Decoder	1.16482

As can be seen in the Table 1, the finetuned Vision Encoder Decoder model surpasses all other models in performance. However, it exhibits slower inference times as shown in Table 2, which may render it unsuitable for applications that require real-time caption generation on mobile or other platforms. Conversely, the finetuned CLIP model offers lower latency and delivers satisfactory performance, making it a viable alternative when accuracy is less critical. The pre-trained ViLT model achieves adequate accuracy with a reasonable rate of inference. Yet, as the number of classes increases, its inference time also increases, potentially limiting its scalability in certain scenarios. Overall all these models perform significantly better than the baseline.

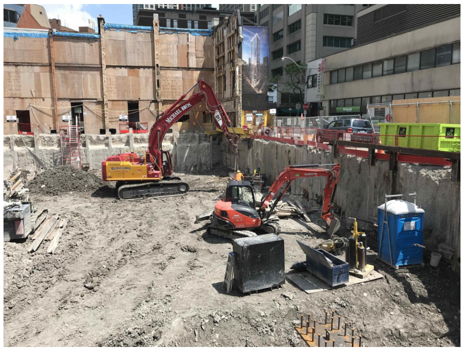

Figure 10: CLIP Model Predictions: MINI_EXCAVATOR, EXCAVATOR, DRILLRIG, TIEBACK_RIG, DOZER; Visual Encoder Decoder generation: an image of excavator tieback rig power generator.

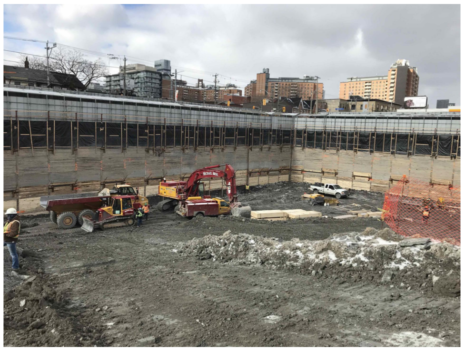

Figure 11: CLIP Model Predictions: TIEBACK_RIG, SKID_STEER, EXCAVATOR, DOZER, TELESCOPIC_HANDLER; Visual Encoder Decoder generation: an image of telescopic handler with lumber bundle.

5 Analysis

The baseline model, utilizing the ResNet50 architecture, generally exhibits suboptimal performance when applied to this construction-specific dataset. This deficiency is primarily attributed to the model's training on general imagery rather than on construction-specific visual content, which significantly hinders its efficacy in relevant contexts.

The ViLT model demonstrates notable improvements in handling diverse datasets, thanks to its unified transformer architecture that processes both image patches and text tokens simultaneously. Despite its training on the comprehensive COCO dataset, the model's performance remains only moderately effective in construction contexts due to the absence of nuanced construction terminology in its training regime.

The CLIP model, particularly when fine-tuned, shows considerable robustness across a broad spectrum of images and texts. It is exceptionally proficient in aligning complex textual descriptions with pertinent images, including those where visual cues are subtle. However, the model occasionally encounters difficulties in edge cases where the relationship between text and image is tenuous or the textual references are inherently ambiguous.

The Vision Encoder Decoder model excels at producing accurate and detailed textual descriptions from images, proving highly capable of interpreting intricate scenes and converting them into coherent text. Nevertheless, its performance tends to diminish when confronted with highly abstract or non-standard images that lack traditional visual cues, presenting significant challenges in accurate depiction.

The outputs of these models were visually checked for a wide variety of images. Overall the Vision Encoder Decoder performs well for the use case I have in mind. It has an acceptable inference rate and is scalable in training as long as we have the training data.

6 Conclusion

This project aimed to improve the effectiveness and accuracy of image-text retrieval systems within construction project management by leveraging advanced models such as ResNet50, ViLT, CLIP, and the Vision Encoder Decoder. The findings revealed that while the baseline ResNet50 model was capable for basic image classification, it fell short in accurately identifying key objects in the construction sector. Conversely, the pretrained ViLT model, enhanced by its transformer architecture and training on the COCO dataset, showed considerable improvements but still faced challenges with construction-specific jargon. The finetuned CLIP model was particularly robust, achieving a 72% accuracy with superior alignment capabilities. The Vision Encoder Decoder model outperformed all, achieving an 83% accuracy and demonstrating its effectiveness in producing precise textual descriptions from images. However, the project encountered significant challenges, including the need for large-scale datasets for training, which demanded substantial computational resources and time. Another major hurdle was obtaining meaningful captions that align with professional standards in construction, necessitating in-depth industry knowledge and agreements with users for using their captions for training. This limitation restricted our ability to produce captions that truly reflect professional insights. Future efforts could focus on gathering more diverse data and training the model to generate more descriptive captions that better represent construction imagery. Additionally, expanding the range of images and extending the labels or captions could encompass a broader spectrum of construction activities, including horizontal and civil engineering projects.

7 Ethics Statement

Images collected from various job sites may inadvertently capture workers, potentially leading to privacy concerns by revealing personally identifiable information (PII). Often, these images are obtained without the workers' knowledge, who may not anticipate being recorded by cameras. Although our primary interest lies in object detection and labeling within these images, not the individuals themselves, addressing privacy is crucial. A practical approach to mitigate this issue is to anonymize the images by masking the faces of any personnel depicted before they are used in training and inference processes. Furthermore, it is essential to inform users about our data handling practices through a clear agreement that explains the purpose of image collection. This agreement should also offer an option for individuals to opt-out of having their images used for training or inference.

Another ethical concern involves the generation of captions for images, typically provided by construction professionals who draw on their expertise to describe detailed scenes. Training a model to generate such captions could potentially aggregate and utilize the collective expertise of these professionals, thereby creating outputs that might surpass the knowledge level of personnel within the company. While this could be advantageous for some companies, it might not provide equal benefits to more established companies, whose expertise could be leveraged without direct compensation. To address this, we propose segmenting the training data according to the tiers of companies, such as by their construction volume, and developing distinct models tailored to different company tiers. Additionally, compensating professionals for their contributions to the training data could further mitigate ethical concerns, ensuring fair use of their expertise.

References

2024a. Hugging face vision text dual encoder doc. https://huggingface.co/docs/transformers/model_doc/vision-text-dual-encoder.

2024b. Huggingface clip training. https://huggingface.co/docs/transformers/v4.41.0/en/model_doc/clip.

2024c. Huggingface vision encoder decoder. https://huggingface.co/docs/transformers/model_doc/vision-encoder-decoder.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2022. Trocr: Transformer-based optical character recognition with pre-trained models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention.

Yang Yu, Peng Hu, Jie Lin, and Pavitra Krishnaswamy. 2021. Multimodal multitask deep learning for x-ray image retrieval. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 603–613, Cham. Springer International Publishing.

A Appendix

A.1 Vision Transformer Architecture

The Vision Transformer (ViT) architecture, introduced by Dosovitskiy et al., represents a significant shift in image processing technology. Unlike traditional convolutional neural networks (CNNs), ViT utilizes a transformer structure that processes images by dividing them into fixed-size patches, which are then embedded and treated similarly to tokens in natural language processing.

Each image patch is linearly transformed, and positional embeddings are added to preserve the positional context. The core of the architecture is the self-attention mechanism, which enables the model to dynamically focus on different parts of the image based on the context provided by other parts. This global approach allows for a more flexible interpretation compared to the local window of perception utilized by CNNs.

ViT has demonstrated superior performance on several benchmarks, particularly in settings where large-scale training data is available. It is increasingly being used in diverse applications ranging from basic image classification to complex tasks involving multimodal data integration.

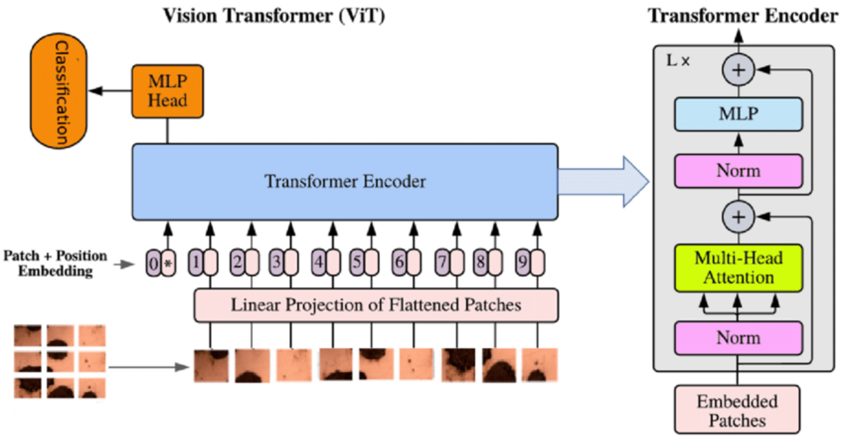

Figure 12: Vision Transformer Architecture Kim et al. (2021)