Image Captioning for Visual Understanding

Caption Crafters
Team -14
H Sai Prakash, K Ugender, Srinjoy Mukherjee, Prashanth S, Ankit Kumar

1 Introduction

Image captioning, a groundbreaking fusion of computer vision and natural language processing, empowers machines to articulate detailed descriptions of visual content. This transformative technology bridges the gap between images and language, allowing computers to generate human-like captions for diverse visuals. It plays a pivotal role in making visual information accessible to all, aiding in content indexing, enhancing human-machine interactions, and finding applications in fields such as healthcare, autonomous vehicles, and e-commerce. By seamlessly translating visual narratives into textual form, image captioning opens avenues for enriched communication and a deeper understanding of the visual world through the lens of artificial intelligence. Image captioning plays a crucial role in making visual content accessible to individuals with visual impairments. By providing textual descriptions of images, it ensures that everyone, regardless of their abilities, can access and understand the information. Captioned images become more searchable and retrievable. This is particularly valuable in large image databases, social media platforms, or content archives where efficient searching based on image content is essential.

2 Methodology

In our project, we have implemented three distinct models to tackle the challenges of visual data analysis. The first model combines Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks, leveraging the spatial hierarchies captured by CNNs and the temporal dependencies modeled by LSTMs. This hybrid approach is particularly effective for tasks where both spatial and sequential information are crucial. Then, we tried to better this model by adding attention mechanism into CNN and LSTM, and observe its effectiveness.Finally, To remove the need of object detectors and train the image captioning system in an end-to-end manner, we leverage the pre-trained visual transformer (ViT) as encoder and language transformer (GPT2) as decoder, making up the basic single-modal module.

2.1 CNN + LSTMs

The proposed model utilizes a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) for the task of image captioning. The CNN serves as a powerful feature extractor, capturing hierarchical and spatial representations from the input images. By leveraging a pre-trained CNN, such as VGG16, the model can benefit from the learned visual features without the need for extensive training on a large dataset. On the other hand, LSTMs play a crucial role in understanding and generating sequential data, making them well-suited for constructing coherent and contextually relevant captions. The LSTM network takes as input the features extracted by the CNN and generates captions word by word, considering the temporal dependencies within the sequence. Combining CNNs and LSTMs capitalizes on the strengths of both architectures—CNNs for image feature extraction and LSTMs for sequential language modeling. This synergistic approach enables the model to effectively bridge the gap between visual and textual information, yielding a robust image captioning system capable of producing meaningful descriptions for a given image.

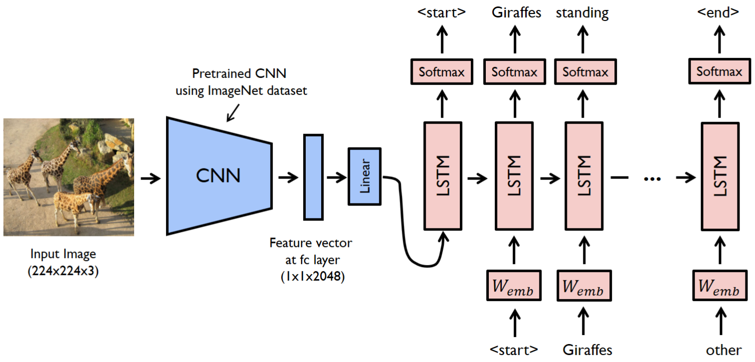

Figure 1: Architecture of CNN + LSTMs for Image Captioning

Name of Dataset	Train	Valid	Test
Flicker8k	6000	1000	1000

Table 1: Dataset Information

2.2 Image Captioning using Attention(CNN, LSTM, Attention Module)

The proposed method combines the above approach with Attention Concept. This work is motivated by the paper *Show, Attend and Tell*, accepted in 2015. The model automatically learns to attend to important portions of the image to augment the captioning process. The main novelty of the this approach is to understand the Attention mechanism . The motivation of this method stems from how humans look at images and understand captions. We use a pretrained Resnet50 network as CNN backbone to extract features from the image. The decoder is a LSTM based model which takes care of the sequential nature of the caption (since it is textual). We add a Attention module where at a particular instant , we take the hidden state of previous timestep and the extracted features from image and calculate attention weights.We implement the soft-attention mechanism introduced in the paper. We back this up by showing our result and the corresponding attention maps, and show how each word is looking at(or 'attending' to) certain parts of image.

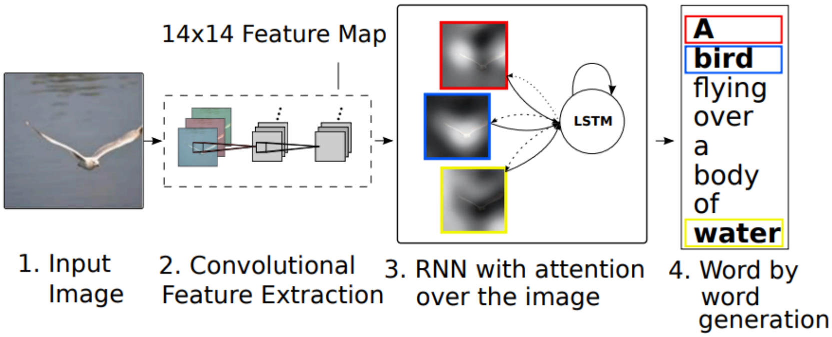
14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Figure 2: Architecture of CNN-LSTMs-Attention for Image Captioning

2.3 ViT + GPT2

ViT operates without the need for additional object detectors; instead, it directly processes a sequence of image patches as input and produces visual representation hidden states for each patch as output. In our framework, we initialize the visual encoder using a state-of-the-art pre-trained ViT model, which has demonstrated superior performance across various visual benchmarks. This approach allows our framework to leverage the outstanding visual representation capabilities inherent in the ViT model, contributing to enhanced performance in our project's visual analysis tasks.

we initialize the single-modal language decoder using the state-of-the-art language generation pre-trained model, GPT-2. Unlike many previous works that overlook the significance of single-modal generation capability, we recognize its importance. Neglecting this aspect often requires models to exert additional effort in learning how to effectively model language and generate coherent sentences, adding complexity to their training process. Leveraging the remarkable language generation ability of GPT-2, our design aims to alleviate the burden on the model during cross-modal training. This approach allows the model to dedicate more attention to aligning cross-modal information, enhancing its overall performance in handling multimodal tasks.

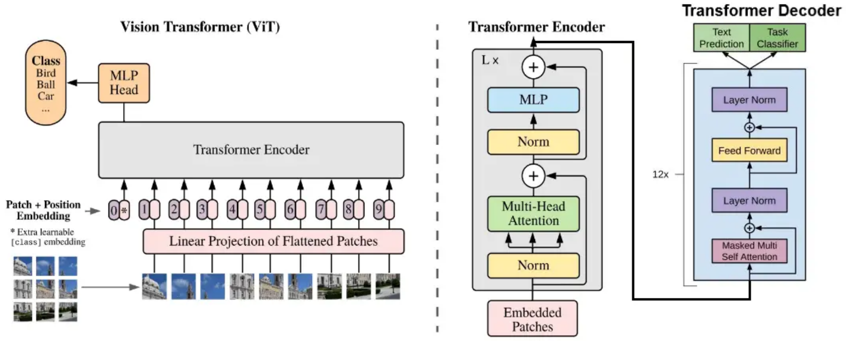

Figure 3: The Vision Encoder-Decoder architecture we'll use for image captioning

We used the pretrained model from Hugging Face https://huggingface.co/nlpconnect/vit-gpt2-image

3 Results

3.1 Dataset

We used Flicker8k dataset. The Flickr8k dataset is a widely used benchmark in the field of computer vision and natural language processing. It comprises 8,000 high-resolution images gathered from the photo-sharing platform Flickr, each paired with five human-generated descriptions. The dataset is designed for image captioning tasks, where the goal is to generate relevant and coherent textual descriptions for the images. These captions cover a diverse range of scenes, objects, and activities, making the dataset suitable for training and evaluating models that aim to understand the intricate relationship between visual content and natural language

Sample Image

Captions

- a man and a baby are in a yellow kayak on water .
- a man and a little boy in blue life jackets are rowing a yellow canoe .
- a man and child kayak through gentle waters .
- a man and young boy ride in a yellow kayak .
- man and child in yellow kayak

Figure 4: Sample Image from Dataset

3.2 Metrics

We Use blue score to evaluate the model responses. The BLEU (Bilingual Evaluation Understudy) score is a metric commonly used for evaluating the quality of machine-generated text, such as captions or translations. Originally designed for assessing machine translation outputs, BLEU has found application in various natural language generation tasks. It measures the similarity between a generated text (such as a machine-generated caption) and one or more reference texts (ground truth captions) based on the overlap of n-grams.

The calculation involves precision for different n-gram orders (unigrams, bigrams, trigrams, etc.) and a brevity penalty. Precision accounts for how many n-grams in the generated text also appear in the reference texts. The brevity penalty adjusts for cases where the generated text is shorter than the reference, penalizing overly concise outputs.

The BLEU score ranges from 0 to 1, with 1 indicating a perfect match between the generated and reference texts. Higher BLEU scores generally suggest better alignment and quality in the generated text. It's important to note that BLEU is just one of several metrics used for evaluating text generation, and while it proAdes a quantitative measure, it may not full capture the nuances of language fluency and coherence.

$$\text{n-gram precision} = \frac{\text{no.of n-gram word matches}}{\text{no.of n-gram words generated}}$$

CNN + LSTMs

BLEU	Value
BLEU@1	0.30
BLEU@2	0.18
BLEU@3	0.06
BLEU@4	0.02

Table 2: BLEU score for CNN + LSTMs

CNN+LSTMs+Attention

BLEU	Value
BLEU@1	0.52
BLEU@2	0.25
BLEU@3	0.09
BLEU@4	0.05

Table 3: BLEU score for CNN+LSTMs+Attention

ViT + GPT2

BLEU	Value
BLEU@1	0.39
BLEU@2	0.19
BLEU@3	0.10
BLEU@4	0.07

Table 4: BLEU score for ViT + GPT2

3.3 Model CNN+LSTM

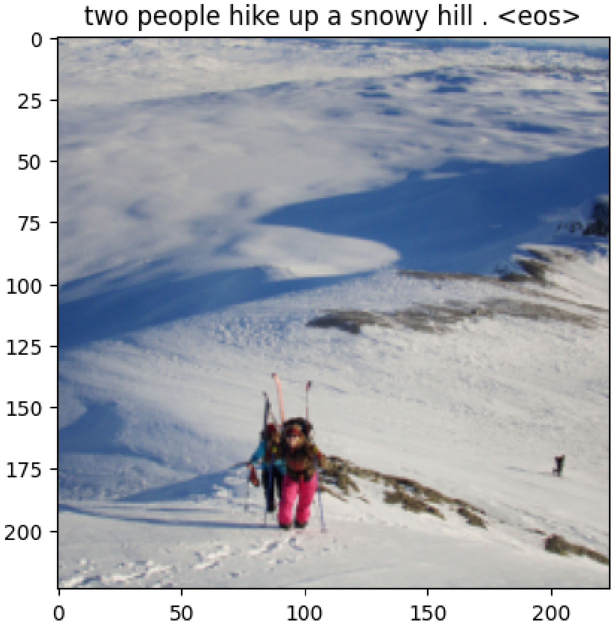

Figure 5: Test Image

Generated caption : Two people hike up a snowy hill.

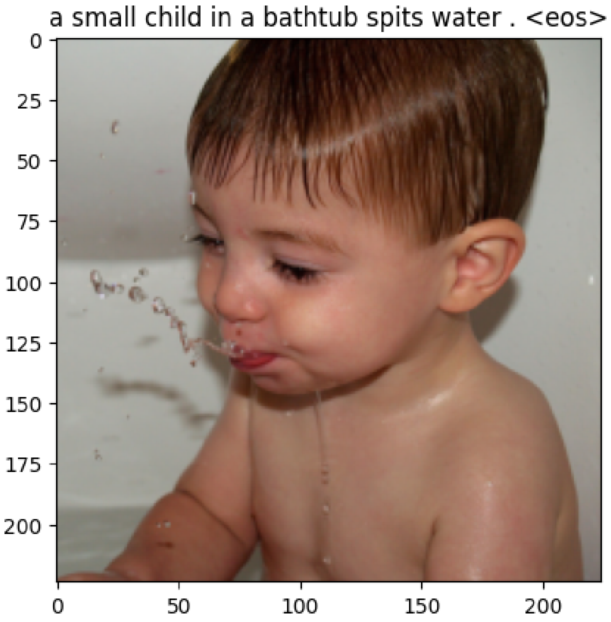

Figure 6: Test Image

Generated caption : A small child in a bathtub spits water .

3.4 Model CNN+LSTM+Attention

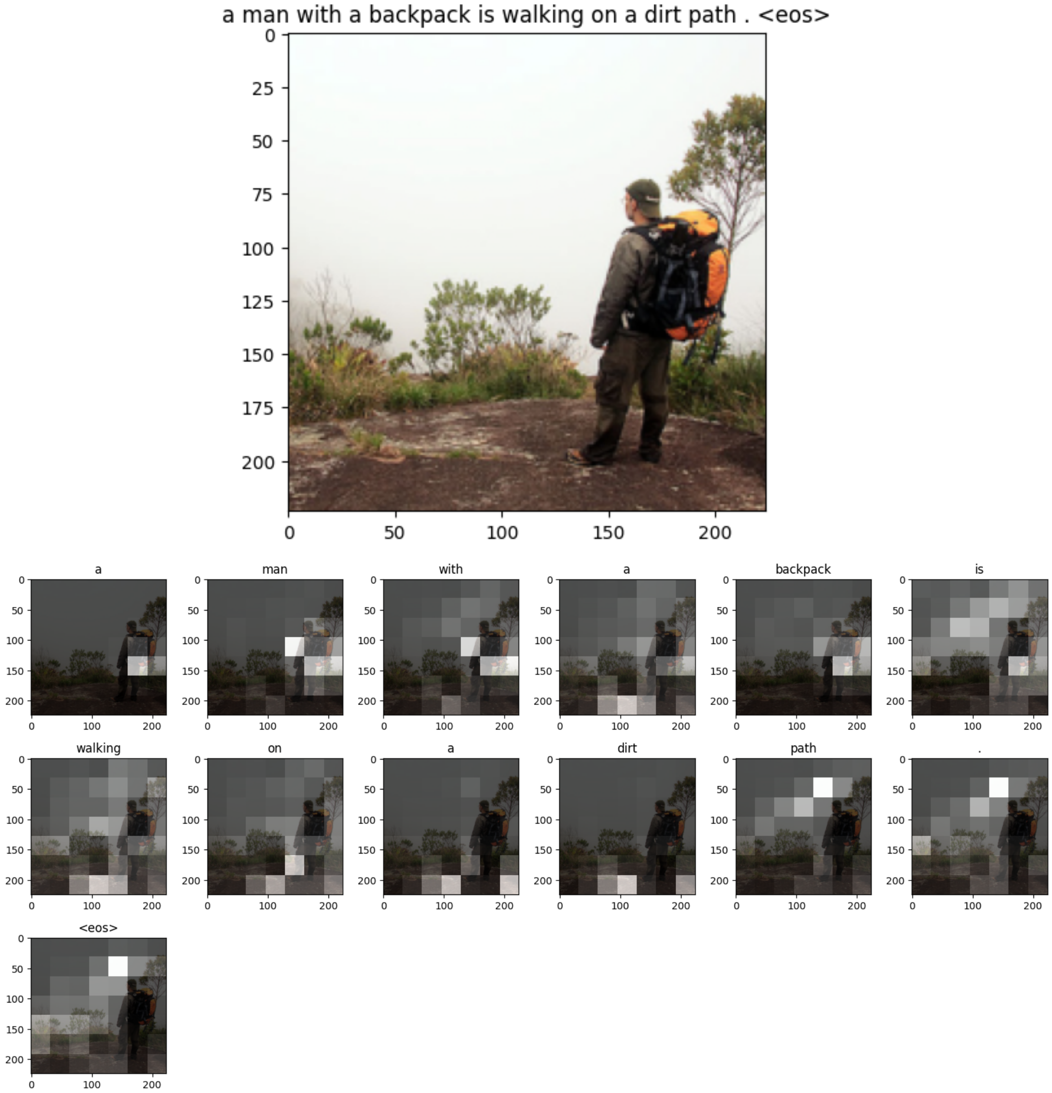

Figure 7: Test Image along with its Attention Maps

Generated caption : A man with a backpack is walking on a dirt path.

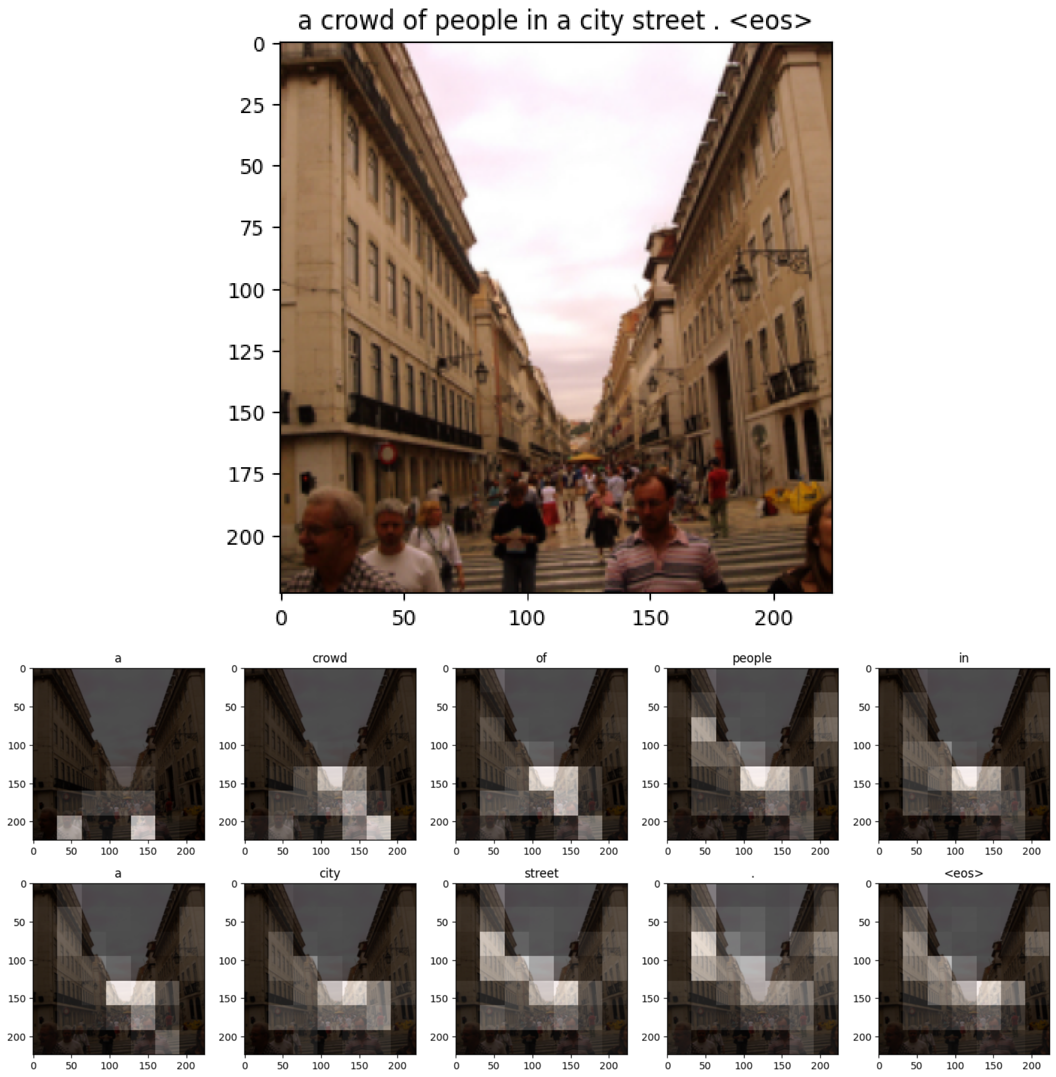

Figure 8: Test Image along with its Attention Maps

Generated caption : A crowd of people in a city street.

3.5 Model (ViT + GPT2) Generated Captions

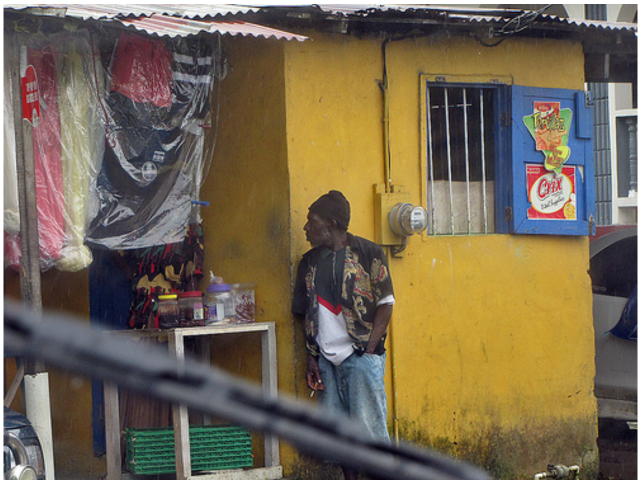

Figure 9: Test Image

Ground Truth captions

- The man is standing next to a yellow building with a blue window .
- A man standing in front of a yellow building
- A man is standing outside of a yellow building .
- A man is standing outside a yellow building .
- A dark skinned man standing outside a small yellow building which is setup to sell items .

Generated caption : A man standing in front of a yellow building

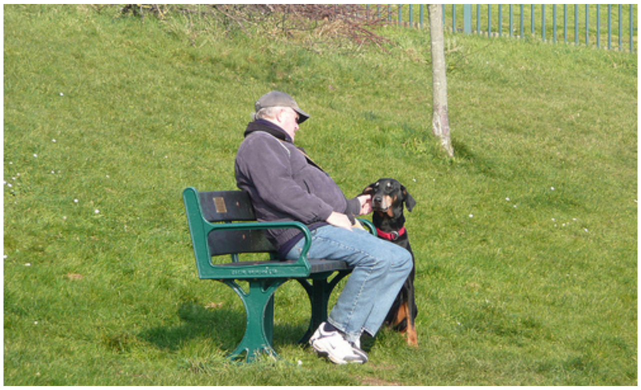

Figure 10: Test Image

Ground Truth captions

- A man is sitting on a green bench located on a grassy hill while playing with a black dog .
- A man on a bench feeds a dog .
- A man on a bench pets a dog .
- Man wearing a grey jacket sitting on a green bench petting a black dog wearing a red collar .
- Old man sitting on the park bench with a black dog .

Generated caption : A man sitting on a bench with a dog .

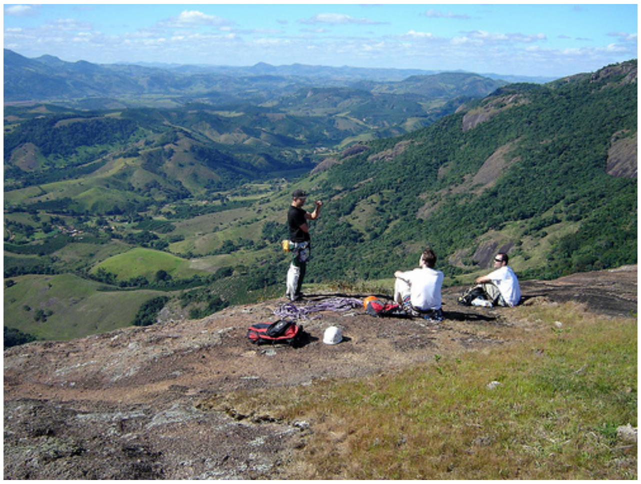

Figure 11: Test Image

Ground Truth captions

- People sit on the mountainside and check out the view .
- Three people are on a hilltop overlooking a green valley .
- Three people hang out on top of a big hill .
- Three people overlook a green valley .
- Three people rest on a ledge above the mountains .

Generated caption : Three people sitting on top of a rocky hillside

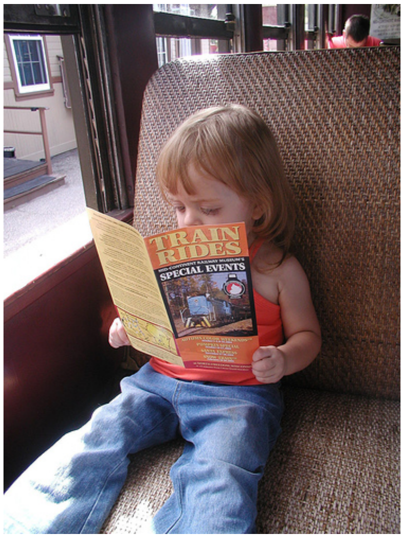

Figure 12: Test Image

Ground Truth captions

- A little girl looking at a brochure on train rides .
- A young blond girl with a magizine in her hands .
- A young girl on a train reads a book about train rides .
- A young girl sits on a seat and looks at a train pamphlet .
- Child sitting down looking at train ride brochure .

Generated caption : A child sitting on a bench reading a book

4 Conclusion

Our image captioning model stands out for its proficiency in crafting clear and concise descriptions for images featuring uncomplicated scenes. It excels in scenarios where the visual content is straightforward, sidestepping the need for intricate textual details or complex object interactions. This makes the model particularly effective in providing easily comprehensible captions for images with evident visual elements, simplifying the interpretation of depicted scenes. To optimize its applicability to our specific context, we leverage a pretrained model initially trained on diverse, generic data. While this model provides a foundational understanding of visual information, we further enhance its performance by fine-tuning it with our proprietary dataset. This fine-tuning process involves adjusting the model's parameters to align more closely with the nuances and details unique to our domain. The result is a tailored image captioning solution that excels not only in generating straightforward captions but also in delivering contextually relevant and nuanced descriptions within our specific use case.

To optimize our model's effectiveness for our specific application, we utilize a pretrained model initially trained on general data. Recognizing the unique features of our domain, we fine-tune the model using our own dataset. This involves adjusting the model's parameters and updating its understanding based on the intricacies of our specific domain, ultimately ensuring superior performance tailored to our particular use case.