A Multi-modal Vision and Language Model for Descriptive Image Captioning

Pranav Donepudi, Nikhil Banerjee, Shelton Zhou
Department of Computer Science, Rice University
{pd53,nb60,sz105}@rice.edu

Link to Project Notebook: Multi-Image Captioning

Abstract

Image captioning has made significant strides in the field of Computer Vision and Natural Language Processing, with Deep Learning models demonstrating a remarkable ability to generate descriptive captions for standalone images. While most of the currently existing approaches focus on generating a single caption for a standalone image, generating individual captions for a sequence of related images based on a prompt remains a challenge. In our project, we intend to use state-of-the-art multi-modal Large Language Models(LLMs) for generating descriptive image captions while observing and analyzing the advancements in this field and comparing them with existing implementations that were based on Long-short-term memory (LSTM) and Gated Recurrent Units (GRU) to observe how today's models fare when compared to the older methodologies.

1. Introduction

The introduction of transformers created a catalyst that made possible significant advancements in the realm of deep learning. Deep learning witnessed a remarkable turn of events from the state of being outperformed by Support Vector Machines (SVMs) using kernels to solving problems in the multi-modal space.

Understanding and describing a sequence of images, while maintaining the context across is a complex task for computers. It requires extracting visual information from each image and weaving them into a cohesive narrative. Recent breakthroughs in deep learning have introduced powerful multi-modal large language models(LLMs) that excel at various tasks that are at the intersection of Computer Vision and Natural Language Processing(NLP) like image captioning.

The project utilizes InstructBLIP, a cutting-edge multi-modal LLM to generate descriptive captions based on a given prompt. It combines a Vision Transformers, ViT to analyze images and a Language Transformer, Flan-T5-xl to generate captions. This encoder-decoder structure allows the model to efficiently capture image features and produce coherent and contextually relevant captions.

We evaluate our model on the VIST subset of the MMInstruction/M3IT [1] dataset. This dataset features collections of image sequences with corresponding textual descriptions. We would assess the model's performance using the METEOR [2] metric which measures caption quality by considering both accuracy and completeness. BLEU [3] which compares the machine-generated translation to one or more human-created reference translations and calculates the degree of overlap. ROUGE [4] which compares an automatically generated summary against a set of reference summaries. By harnessing the strengths of both computer vision and natural language processing techniques, we seek to push the boundaries of multi-modal learning and enable more natural and engaging interactions between humans and machines.

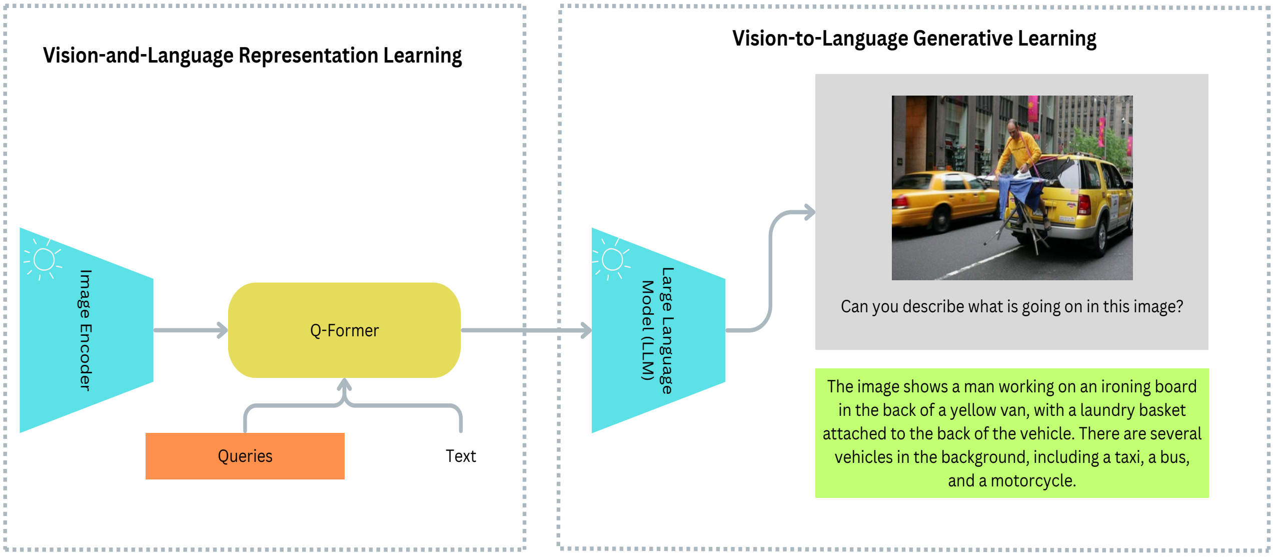
2. Related Work

Through our project, we intend to draw a comparison on one of the early captioning implementations done that date back to 2016 which proposed a sequence-to-sequence Recurrent Neural Network (RNN) architecture for contextual captioning [4] An image encoder RNN processes the image sequence in reverse order, and its final hidden state initializes the story decoder RNN, which generates the story word by word using softmax loss during training and both the encoder and decoder use the Gated Recurral Units (GRUs).

2.1. Early influential papers

Vinyals et al.[9] use a neural network architecture that consists of a CNN encoder, and an LSTM decoder for image

[1] Multi-Modal Multilingual Instruction Tuning Dataset
[2] Metric for Evaluation for Translation with Explicit Ordering
[3] Bilingual Evaluation Understudy
[4] Recall-Oriented Understudy for Gisting Evaluation


Vision-and-Language Representation Learning
Image Encoder
Q-Former
Queries
Text
Vision-to-Language Generative Learning
Large Language Model (LLM)

Can you describe what is going on in this image?

The image shows a man working on an ironing board in the back of a yellow van, with a laundry basket attached to the back of the vehicle. There are several vehicles in the background, including a taxi, a bus, and a motorcycle.

Figure 1. BLIP-2 Architecture with frozen transformers and Q-former.

captioning training on image-caption pairs. This encoder-decoder architecture has been widely adopted and built upon. Karpathy & Li [6] introduced an alignment model that learns to map image regions to corresponding words in the caption, enabling more fine-grained associations between visual and textual elements. It also discusses training multi-modal RNNs on aligned image regions and snippets to generate descriptions for unseen image regions.

2.2. Attention-based models

Xu et al.[10] consists of an attention mechanism in the decoder, which allows the model to dynamically focus on relevant image regions & generate every word of the caption. This approach has been a guiding force, with many following works building upon this idea of visual attention for image captioning. Anderson et al.[1] proposed a model that combined bottom-up attention, which extracted features from salient image regions, with top-down attention that focuses on regions that the language model thinks are important.

2.3. Improvements to the language model

Jozefowicz et al.[5] showed the significance of a language model in image captioning tasks, showing that notable gains can be attained by training on larger amounts of text data. This highlighted the potential benefits of large-scale language data to produce captions of better linguistic quality. Lu et al.[7] established an adaptive attention mechanism that decides when to focus on the image and when to rely on the language model thus minimizing needless attention to visual features when generating non-visual words.

2.4. Models incorporating additional information

Gulcehre et al.[3] presented a new approach to deal with out-of-vocabulary words in image captioning, allowing the model to point to novel objects in the images that were not visible during training. This approach tackles a major weakness of many earlier models which struggle to generate captions with rare or unseen words. Fisch et al.[2] introduced a framework for generating image captions based on a user-specified context such as answering a query or explaining a part of the image. This work demonstrates the possibility of more adaptable and purpose-driven image captioning.

3. Model

In this project, we leverage transformer models. Transformers provide thousands of pre-trained models to perform tasks on different modalities such as text, vision, and audio.

3.1. Overview of InstructBLIP

In our project, we utilize the InstructBLIP model, an advanced multi-modal framework that integrates image and language processing capabilities to generate image captions. InstructBLIP is a visually-instructed version of the BLIP (Bootstrapped Language-Image Pre-training) model, designed to respond effectively to textual prompts while referencing specific visual content.

3.2. Architecture

InstructBLIP employs a transformer-based design that masterfully integrates a Vision Transformer (ViT), a Language Model (LM), and a Querying Transformer (Q-Former) as shown in Figure 1. Here's how it works: The Vision Transformer dives into the image inputs, extracting a tapestry of feature representations that vividly capture different facets of the visual content. These features are then handed off to the Language Model, which is expertly crafted to craft text that not only resonates with the image but also fits perfectly with the given instructional prompt. The Q-Former acts as a bridge, seamlessly connecting the visual and textual modalities and enhancing their alignment. This sophisticated coordination has shown impressive results on benchmark datasets, underscoring its effectiveness.

- **Vision Transformer:** The ViT component of InstructBLIP is pre-trained on large-scale image datasets, enabling it to extract and prioritize visual features crucial for understanding and describing images.
- **Querying Transformer:** Q-Former is a transformer-based architecture with two sub-modules: (1) an image transformer that interacts with the visual features from the frozen image encoder and (2) a text transformer that can encode and decode texts. It uses a set of learnable querying vectors to extract relevant visual features that capture the most informative part of the text that goes with the image.
- **Language Model:** The LM component, based on the Flan-T5-xl architecture, excels in generating coherent and contextually appropriate captions. This model is fine-tuned with a focus on aligning language generation with visual inputs and instructional prompts.

3.3. Integration of Modalities

The integration of visual and textual modalities is facilitated through a co-attention mechanism that allows the Language Model to attend to specific parts of the image as described by the Vision Transformer's outputs. This interaction ensures that the generated captions are not only descriptive but also precise in detailing specific elements highlighted by the visual prompts.

3.4. Customization for M3IT Dataset

For our project, we tailored the InstructBLIP model to specifically address the challenges posed by the dataset, which consists of sequences of images that need cohesive and contextually linked descriptions.The MMInstruction/M3IT Dataset compiles diverse tasks of classical vision-language tasks, including captioning, visual question answering (VQA), visual conditioned generation, reasoning, and classification from which we have opted for vist subset which consists of 5000, 4315, and 4350 with image-caption data for training, validation, and test sets respectively.

3.5. Advantages and Limitations

The InstructBLIP model brings several advantages, including its robustness in handling diverse image contexts and its ability to generate detailed captions. However, its performance is contingent on the quality and diversity of the training data, and the model requires significant computational resources for training and inference, which could be a limitation for real-time applications.

4. Experimental Setup

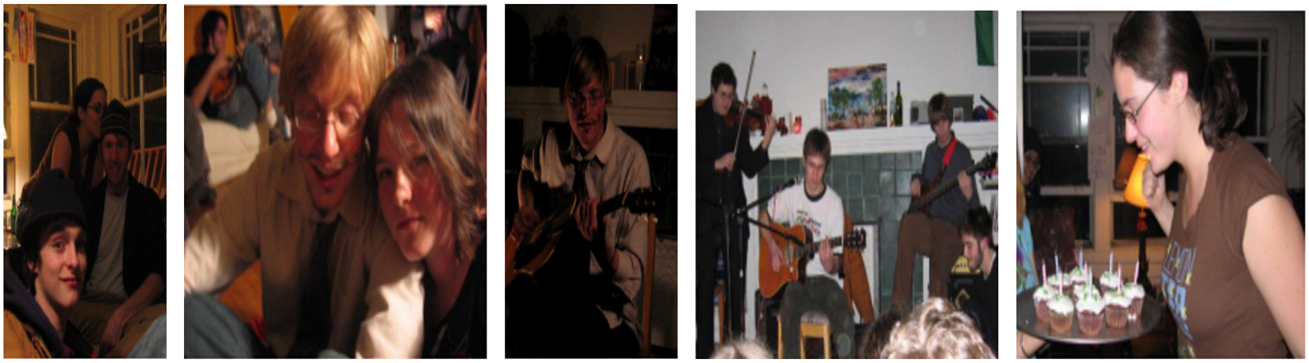
In subsequent sections, we will detail the experimental setup, including data pre-processing, model configurations, and the evaluation metrics used to assess model performance on the VIST subset of the MMInstruction/M3IT dataset. The dataset used is loaded using the datasets library, specifically from a source named MMInstruction/M3IT with a subset vist. This dataset contains image-text pairs, useful for tasks that involve understanding and generating content based on visual inputs. The class 'M3ITdataset' handles different subsets of dataset(train,validation,test) providing mechanisms to access individual data points (__getitem__) and the dataset's size(__len__).Before being used for training or inference, images are transformed using a series of operations. A Compose of Resize, ToTensor, and Normalize is typically applied to the images. This standardizes the images into a consistent format and scale, converting them into tensor objects with normalized values. This is crucial for ensuring that the model receives input data in a consistent format, optimizing learning or inference. Our script includes functionality to visualize images from the dataset, which can be helpful for verification of data loading and transformation processes or for qualitative analysis during experiments. The review by [8] demonstrates the importance of systematic data preprocessing and augmentation, suggesting these are indispensable for the success of any advanced data analysis project, particularly in areas requiring robust feature extraction and noise reduction.

5. Experiments and Results

We utilized the PyTorch framework to implement our project idea and adopted the InstructBLIP model available from Hugging Face as our starting point of reference [5]. To evaluate the performance of the InstructBLIP model on image captioning, a series of experiments were conducted using the M3IT dataset. The dataset was split into training, validation, and test sets. These experiments were designed to assess the model's ability to generate accurate and coherent descriptions of images based on the given prompt instruction and, when applicable, a previous context of the

[5] Hugging Face Documentation

Input Sequence of Images:



Prediction:

The image depicts a group of people sitting in a living room, with one person sitting on the couch and another person sitting on the chair. The man and woman in the image are posing for a photo while sitting in a living room. The man in the image is playing a guitar in a dark room. The group of people in the room are playing music together. The woman in the brown shirt is holding a tray of cupcakes for her birthday.

Figure 2. Descriptive Image Captioning for a Sequence of Images using InstructBLIP.

image.

Four different subsets of images from the M3IT dataset were selected for experiments. Each subset consisted of a sequence of five input images, the model was tasked with generating a summary for each of the images based on the provided input prompt and the output generated for the previous image in the sequence. The generated summaries were then compared to the ground truth references using different evaluation metrics.

Implementation	METEOR	BLEU	ROUGE
InstructBLIP-Flan T5-xl	0.2338	0.0198	0.2696
RNN and GRU. [4]	0.187	0.073	-

Table 1. Comparison of METEOR, BLEU metrics for our implementation and the base paper implementation.

The evaluation metrics we've used for our experiments are as follows and the final consolidated results are in Table 1.

1. **METEOR:** A metric that calculates the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also considers stemming and synonymy matching.
2. **BLEU:** A widely used metric in machine translation tasks that measures the n-gram overlap between the generated text and the reference text. In this work, the BLEU score was calculated using n-grams up to 4-grams.
3. **ROUGE:** A set of metrics that measure the overlap of n-grams between the generated text and the reference text. In this work, the ROUGE-L metric, which calculates the longest common subsequence between the generated and reference texts, was used.

BLEU score is precision-focused while ROUGE score is recall-focused, the METEOR metric on the other hand was designed to address issues encountered in BLEU, ROUGE metrics while simultaneously producing a good correlation with human judgment at the sentence or segment level.[6]

We see that our implementation generates a fairly improved METEOR Score with a 0.2238 score indicating a moderate amount of capturing of information from the reference captions. Our model generated coherent captions for a series of five images while maintaining a healthy context between each successive caption. We managed to randomly sample the validation set and verified the relevance of the captions to the images and got fairly descriptive captions as seen in Figure 2 that provided a greater deal of context as compared to what the traditional models gave us.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] A. Fisch, K. Jiang, Y. Li, A. Iyer, V. Radhakrishnan, Q. Do, M. Luo, J. Gao, and U. Alon. Capwap: Captioning with a purpose. In *Proceedings of the 2020 Conference on Em-*

[6] https://arize.com/glossary/meteor-score/

pirical Methods in Natural Language Processing (EMNLP), pages 8755–8768, 2020.

[3] C. Gulcehre, S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 140–149. Association for Computational Linguistics, 2016.

[4] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California, 2016. Association for Computational Linguistics.

[5] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[6] A. Karpathy and F.-F. Li. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[7] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

[8] K. Maharana, S. Mondal, and B. Nemade. A review: Data pre-processing and data augmentation techniques. *Journal of Data Science*, 42(3):101–120, 2024.

[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.