A Multimodal Transformer-based Architecture for Vietnamese Image Captioning focusing traffic scenes

Duc Nhan Bui   
Department of Information Technology  
HCMC University of Technology and EducationHo Chi Minh, Vietnam  
nhanbui15122003@gmail.com

Trieu Phi Dai   
Department of Information Technology  
HCMC University of Technology and EducationHo Chi Minh, Vietnam  
dtptrieuphidtp@gmail.com

Thien Nhan Huynh   
Department of Information Technology  
HCMC University of Technology and EducationHo Chi Minh, Vietnam  
159.thiennhan@gmail.com

Van Dung Hoang   
Department of Information Technology  
HCMC University of Technology and EducationHo Chi Minh, Vietnam  
dunghv@hcmute.edu.vn

*Abstract*—This study proposed a multimodal image captioning architecture and implementation pipeline for Vietnamese traffic dataset. The main objective is to generate high-quality captions that accurately describe objects and actions in complex traffic scenarios. The proposed approach combines Vision Transformer, BARTpho, and GPT-2 into one architecture that leverages both visual and language understanding. A method of building a Vietnamese image captioning dataset is proposed, with traffic-related images and web metadata being scraped based on keyword-based scraping and afterwards caption labelled using a large language model. In order to increase object diversity and generalization of the model, a set of image augmentation techniques are employed before training. The overall outcomes show that the model produces context-aware, coherent captions with increased relevance to the Vietnamese traffic environment. The proposed framework demonstrates the potential to be utilized in smart transportation systems and city surveillance through the automatic localized scene understanding. The proposed framework also has potential to be utilized in assistive technologies that can improve the traffic navigation capability of visually impaired individuals.

Keywords— Vietnamese Image Captioning, Multimodal Learning, Vision-Language Transformer, Deep Neural Network, Vietnamese Dataset

# Introduction

Image captioning has been a very popular topic in the computer vision field recently. Image captioning aims at describing the content of an image using descriptive text. It is an interesting and challenging problem because of its diverse applications in real life and also because of the difficulty in making computer vision systems understand the content of an image. Moreover, the creation of image captioning systems has tremendous potential to help visually impaired individuals "see" the world through descriptive text.

Computer Vision (CV) and Natural Language Processing (NLP) used to be two separate research domains. However, with the booming growth of multimedia data from mobile devices, IoT systems, and social media, researchers now tend to merge techniques from the two fields for improved understanding of multimodal content and more efficient extraction of useful information from visual data. Actually, automatic image annotation—a key task in image retrieval and computer vision—is meant to attach meaningful words or phrases to represent images. To this end, artificial intelligence approaches, typically with the help of pre-trained models, are used to discover mappings from low-level visual features to high-level semantic representations for generating meaningful captions for a given image.

Image captioning has been a flagship task at the intersection of computer vision and natural language processing, and top-down and bottom-up have been two prominent paradigms. Whereas the top-down approach takes an image, extracts the global visual features, and translates them into natural language, the bottom-up approach begins with identifying and describing salient regions in the image, which are integrated into consistent textual descriptions. Both approaches typically rely on advanced language models to supply fluency, coherence, and contextual relevance of the generated captions.

In this paper, we propose a novel image captioning model using the top-down method, enhanced with an Attention Mechanism within the encoder-decoder paradigm. With this, the model is capable of selectively focusing on the most informative visual cues when creating explanatory captions. Even though current models have achieved state-of-the-art results for English, there have been limited studies on Vietnamese image captioning in traffic scenes. To bridge this gap, we test the proposed model on a Vietnamese image captioning dataset of traffic scenes. Our work is aimed not only at advancing image captioning research but also at supporting real-world applications. We are targeting the development of assistive technology for the blind, with the particular aim of enhancing their situational awareness and traffic safety by describing images in Vietnamese automatically.

The remainder of this paper is structured as follows: the remainder of Chapter II discusses related works from traditional to state-of-art ones, specifying model's output goals and coverage; Chapter III explains the suggested approach to build Vietnamese image captioning dataset and multimodal framework underpinning the research; Chapter IV presents the final outcome and V for conclusion.

# Related Works

Image captioning has been aggressively researched in recent years, where various models have taken advantage of vision-language pretraining and attention mechanisms. "Show, Attend and Tell" by Xu et al. [1] is one of the early works, where the method of soft visual attention was used to dynamically focus on the regions of an image while predicting the caption. This piece of work brought forth the spotlight for attention-based image captioning architecture. Building on that, Anderson et al. [2] created a Bottom-Up and Top-Down Attention mechanism which demultiplexes object-level feature extraction and caption generation so that the model can attend to important regions in a more meaningful way. Similarly, Huang et al. [3] reinforced the attention mechanism by the Attention on Attention (AoA) module, which enhances the representation ability by considering the attention weights and attended features.

Transformer models, originally introduced in the "Attention Is All You Need" paper by Vaswani et al. [4], have been a strong force in image captioning research. Several works have adopted this paradigm. Of interest is Cornia et al. [5]'s Meshed-Memory Transformer, which employs meshed connections and memory vectors to allow greater information exchange across layers. Pan et al. [6] introduced X-Linear Attention Networks, which leverage higher-order interactions between visual features with bilinear pooling, setting another state-of-the-art. Herdade et al. [7] proposed a new object-to-word alignment method, directly converting object detections into semantic tokens, that bridged the gap between textual and visual modalities. At the same time, the rise of large-scale vision-language pretraining has opened up new possibilities. Zhou et al. [8] presented a joint pretraining framework for image captioning and visual question answering (VQA), while Zhang et al. [9] presented VinVL, which re-examined visual representations from more sophisticated features and better object detection. More recently, the Florence-2 model of Xiao et al. [10] demonstrated the power of a single representation for numerous vision tasks, including image captioning, with data and computing scaling general-purpose pretraining architecture.

Although state-of-the-art image captioning methods have achieved stunning performances, they are primarily designed for English and large-scale English-based datasets such as MSCOCO. Conversely, Vietnamese image captioning is significantly underresearched amidst the rapidly increasing demand in its applications, especially in the important but less explored field of traffic scene understanding. In this work, we propose a multimodal Transformer-based model for Vietnamese image captioning. Our model combines Vision Transformer (ViT) [11] for vision feature extraction, BARTpho tokenizer for Vietnamese language robustness, and GPT-2 decoder for context-aware caption generation fluency. Apart from the model, we also present a novel contribution towards a methodology to construct a Vietnamese traffic scene dataset to supplement the lack of large-scale domain-specific annotated corpora in Vietnamese. While some datasets such as UiT-ViIC [12] and KTViC [13] with general-purpose image-caption pairs exist, none of them focus on traffic scenes. Our dataset fills this gap, enabling the model to learn linguistically accurate and contextually relevant captions specific to Vietnamese traffic scenes

# Materials and Methods

There are some available image datasets of traffic scenes that are typically built for general-purpose usage such as object detection or scene segmentation and tend to be primarily focused on well-structured data captured in ideal conditions. Within the specific case of traffic scenes, open-source captioning corpora remain absent and those available are typically limited in contextual information, typically supplying short object-based descriptions rather than semantically rich stories that include real-world interaction and environmental signals. Besides, there is also the absence of high-quality Vietnamese-language traffic image captioning datasets while significant progress has been made in creating image captioning datasets, most of the existing resources are targeted towards general domains such as indoor scenes, everyday objects, or social events with a strong English-language caption preference. This language disparity represents a major stumbling block in developing AI systems that can communicate with non-English-speaking users, particularly in critical applications such as accessible technology for the visually impaired, where accurate and context-aware scene perception is critical.

To address these limitations, we propose the development of a domain-specific, high-quality image captioning dataset—Traffic Pictures Captioning dataset [14] or TPC37k—captures the complexity and variability of real-world traffic environments. This dataset includes a wide range of traffic-related visual content, covering various object categories (vehicles, pedestrians, traffic signs, lights, sidewalks, roads), lighting conditions (daytime, nighttime, dusk, weather variations), perspectives (wearable cameras, traffic surveillance), and diverse traffic scenarios (intersections, one-way streets, residential areas, etc.).

## Building Dataset

The TPC37k dataset [14], as depicted in Figure 1, was constructed through a systematic pipeline (refer to Figure 1). Traffic-related images were initially collected from web pages using SerpApi with traffic-specific keywords for topic diversity and relevance. Data preprocessing was subsequently performed to verify image URLs, remove duplicates (approximately 10%), and remove low-resolution images, resulting in a clean dataset of 9,300 images. Automatically generated initial captions came from the Gemini 2.0 Flash API and were subsequently edited manually (80% of captions) to ensure linguistic accuracy and contextual relevance. Finally, various augmentation techniques (e.g., pixel shift, spatial transformation, adding noise) expanded the dataset to 37,056 images, adding data richness and enabling strong model training.

A diagram of a software development process

AI-generated content may be incorrect.

1. The overall pipeline for building TPC37k dataset
2. Token and word count in TPC37k

|  |  |  |
| --- | --- | --- |
| **Type** | **Total** | **Unique** |
| Token | 1,160,004 | 1,731 |
| Word | 1,267,648 | 2,356 |

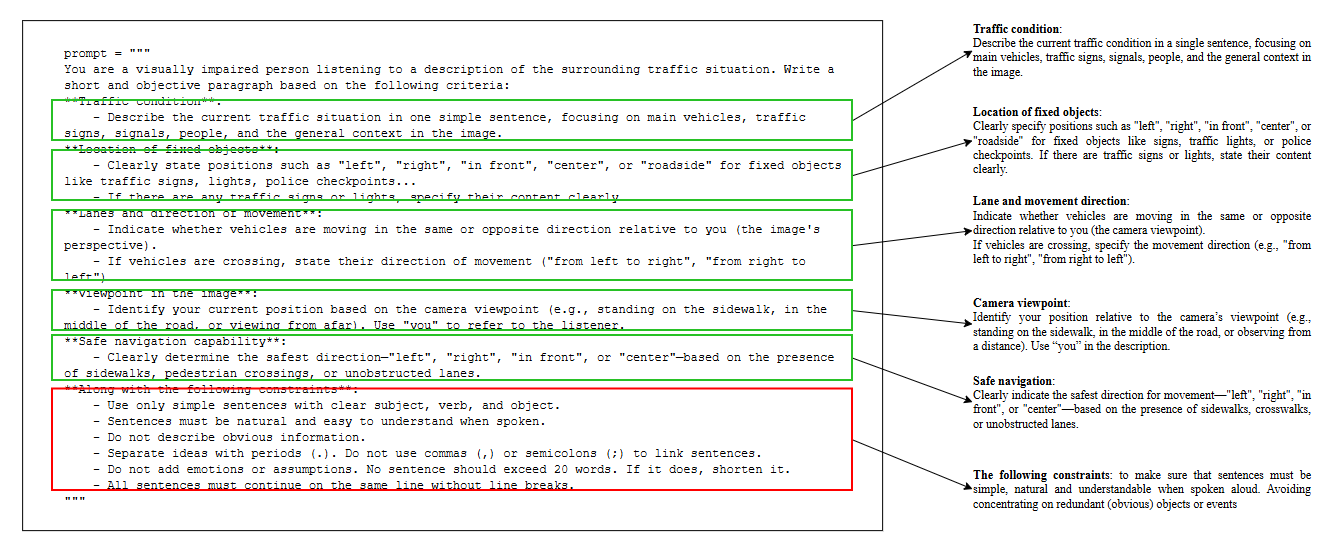
1. The splits in TPC37k

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Split** | **Images** | **Captions** | **Images/Caption** | **Max length** | **Avg. length** |
| **Train** | 29,644 | 7266 | 4 | 71 | 31,34 |
| **Val** | 3,706 | 922 | 4 | 64 | 31,12 |
| **Test** | 3,706 | 920 | 4 | 58 | 31,16 |

### Data Augmentation: To enhance the diversity and scalability of the dataset, various image augmentation techniques were applied using the Albumentations library, a high-performance framework widely adopted in computer vision tasks:

* Pixel-level transformations: Each input image had a 50% chance of being altered in brightness and contrast (within ±20%) or undergoing color jitter with variations in brightness, contrast, saturation (±20%), and hue (±0.1). This step aims to simulate changes in lighting and color conditions.
* Spatial transformations: To simulate different perspectives and slight geometric variations, either a shift-scale-rotate operation (shift and scale within ±10%, rotation limit of 15°) or affine transformation (scaling between 0.9 and 1.1, translation within ±10%, and rotation between –10° and 10°) was applied with a probability of 50%.
* Noise and blur: To model imperfections in real images, Gaussian noise (variance 10–30) or Gaussian blur (kernel size 3–5) was randomly added with a probability of 30%.
* Random cropping and resizing: A random resized crops were performed with a scale range of 0.8 to 1.0 and aspect ratio of 0.9 to 1.1 to simulate occlusions and partial views. All the resulting images were then resized to a fixed resolution of 512×512 pixels for uniformity across the dataset.

### Caption Label Generation: To generate caption labels for the images, we employed the Gemini 2.0 Flash model using its API. Captions were generated in Vietnamese, ensuring concision (maximum of 40 words per caption), natural language flow, and proper generalization of image content. Prompt engineering was carried out iteratively to fine-tune the generation behavior such that the captions were in keeping with human-like descriptions and preserved essential semantic meaning.



1. Prompt template for caption label generating

### Data Preprocessing: In order to ensure the quality and consistency of input data for image captioning modeling, we designed a systematic data processing pipeline (see Figure 3) that involves multiple interdependent stages, as depicted in Figure 1. The pipeline is designed to process both image and caption data in parallel and make them clean, standardized, and semantically consistent before being used for model training.

A diagram of a software company

AI-generated content may be incorrect.

1. Data preprocessing pipeline for caption images to training structure

It begins with image crawling for raw image-caption pairs acquisition, which are sanitized to remove noise and formatting anomalies. The image-augmentation operation of resizing, cropping, and blurring is used to increase model robustness by introducing visual variation from the same source material.

These are preprocessed using the ViT processor (google/vit-base-patch16-224), which rescales input to the size and splits them up into patches. Pixel intensities are standardized using ImageNet statistics, returning tensors of the shape that conform to the input format of the ViT encoder. Besides, captions are tokenized with the BARTpho word-level tokenizer (vinai/bartpho-word), which includes special tokens that are needed and word-to-vocabulary-ID mappings. Sequences are padded or truncated to 315 tokens at most and become tensors of shape *(batch\_size, max\_length)* for GPT-2.

Both image and caption tensors are aligned inside the combiner module, which allows proper matching. They are thereafter passed to the encoding stage where abstractions are extracted from features to serve as inputs to the captioning model. This modularity allows for clean and uniform multimodal preparation for training.

## Proposed Architecture for Image Captioning PVG

Image captioning involves two key components: extracting semantic information from visual content and generating grammatically correct natural language descriptions. In this work, we propose an image captioning framework based on the widely adopted Encoder-Decoder architecture. As can be seen in Figure 4, our approach leverages pretrained models, including Vision Transformer (ViT) [11] as the image encoder and GPT-2 as the language decoder—both of which are Transformer-based architectures. Because our focus is on the Vietnamese language, we utilize the BARTpho tokenizer [15], a word-level tokenizer specifically pre-trained on Vietnamese high-scale datasets. Even though originally designed for BART-based models, BARTpho tokenizer can be used with Transformer-based decoders such as GPT-2 and supports effective tokenization tailored to Vietnamese linguistic patterns. The resulting model consists of two main modules: a visual encoder and a caption decoder. We refer to our proposed architecture as BartPho-ViT-GPT2, and for brevity, we will denote it as *PVG* in the remainder of this paper.

A diagram of a diagram

AI-generated content may be incorrect.

1. Multimodal architecture following Encoder-Decoder of PVG

The ViT and GPT-2 pretrained models, with the assistance of the BARTpho tokenizer, constitutes an interesting multimodal approach to image captioning, particularly for Vietnamese text generation.

Unlike conventional CNN-RNN architecture, the Vision Transformer (ViT) encoder captures global spatial relationships across the entire image via self-attention, instead of relying on local receptive fields. This enables a deeper understanding of the visual context, which is crucial for image captioning. Moreover, ViT’s modular design, which avoids handcrafted components like anchor boxes, facilitates seamless integration with Transformer-based decoders. The visual input I is encoded into a latent representation where each represents the feature vector of an image patch. The resulting visual embedding with being the number of patches and the hidden dimension, is used as input to the language model. The model then predicts each caption token based on prior tokens and the visual context, as shown in equation .

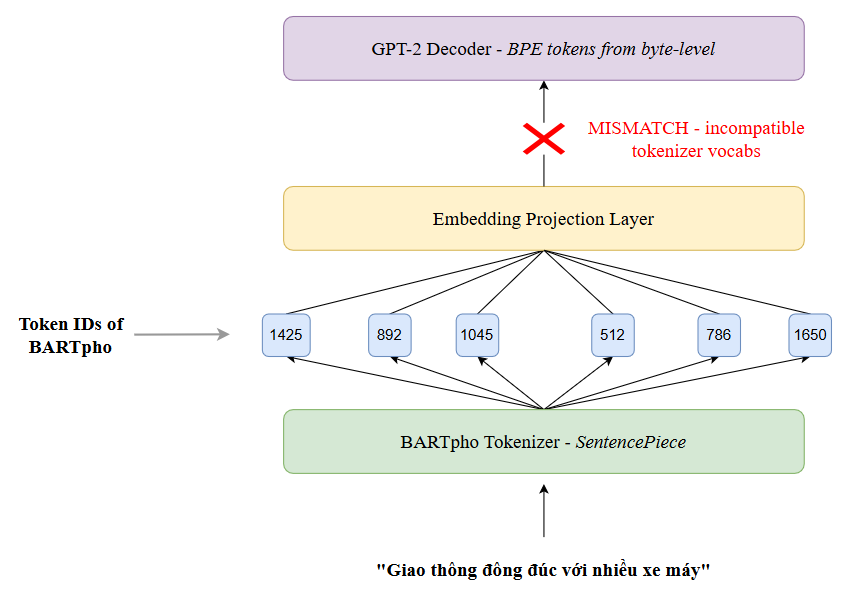
Additionally, ViT integrates easily with Transformer-based decoders such as GPT-2 due to shared architectural principles. The output from the [CLS] token is linearly projected and passed through a Tanh activation to generate the input embedding for the decoder

Where . This transformation outputs the final image embedding used as input for the decoder.

We use the BARTpho-word tokenizer to process Vietnamese text. BARTpho operates at the word level and is pre-trained on Vietnamese data, helping to reduce word segmentation errors—a common issue in subword-level tokenization methods like BPE. This leads to more linguistically appropriate word representations for Vietnamese, which supports the decoder training process.

After tokenizing the captions, the resulting input\_ids and attention\_mask are fed into a self-attention layer to model intra-caption dependencies. The output of this self-attention module serves as the query in the cross-attention mechanism. Simultaneously, the image features *(pixel\_values)* extracted by the encoder (ViT) are used as both the . This cross-attention layer enables the decoder to align and integrate visual information with the caption context, effectively learning the relationships between image regions and textual tokens. These can be formulated as Equations :

This architecture is built by integrating pre-trained components, leveraging the individual strengths of each model for the image captioning task. Its modular design enables efficient transfer of knowledge from large-scale pretraining, reduces training costs, and ensures smooth integration due to shared Transformer-based architecture. However, a notable limitation is that BARTpho-word operates at the word level, whereas GPT-2 was pre-trained on byte-level inputs. This mismatch may affect the compatibility between the tokenizer and the decoder (refer to Figure 5).

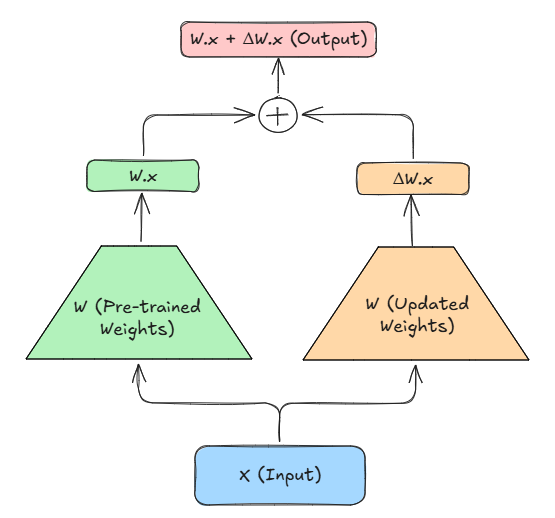


1. Illustrate of tokenizer-decoder mismatch when combining BARTpho with GPT2

Despite the mismatch between Tokenizer and Decoder, our model demonstrates strong performance in generating accurate and fluent Vietnamese captions. This observation can be attributed to two main factors.

First, while the token embeddings may not perfectly match due to the tokenization scheme difference, much of pre-trained GPT-2 structure—particularly its multi-head self-attention layers—remains intact and still plays a vital role in how the model is capable of producing language. To enhance training effectiveness and reduce the number of parameters that can be trained, Low-Rank Adaptation (LoRA) [16] is applied to the layers of the self-attention and cross-attention modules in the decoder. Instead of adapting the full weight matrix , LoRA freezes W and applies a trainable low-rank update:

Where , with . is a learnable low-rank matrix that approximates the weight update. W is the pre-trained (frozen) weight matrix (refer to Figure 6). LoRA works by injecting learnable low-rank matrices into these projection layers, enabling efficient fine-tuning with minimal parameter updates while preserving the pre-trained backbone. As a result, the model retains most of its general linguistic knowledge acquired during pretraining, while effectively adapting to new token distributions. This approach significantly reduces memory consumption, accelerates convergence, and maintains overall performance.



1. Illustrate of fine-tuning process with LoRA

Second, the Vietnamese caption training set used to fine-tune is of comparatively high quality with coherent structural models. The majorities of captions have analogous syntactic templates as well as similar lexical choices to use when interpreting ordinary traffic contexts, such as object locations, behaviors, and environments. Such uniformity reduces the burden on the model to generalize over extremely inconsistent sentence forms, allowing the decoder to more readily learn the visual-to-language correspondence. Therefore, even with potential token embedding discrepancies introduced by tokenizer differences, the decoder can leverage the caption structure regularity to generate accurate and coherent Vietnamese captions.

# Experiment Results

## Quantitative Evaluation

We evaluate model performance on BLEU, ROUGE-L, and CIDEr. BLEU estimates n-gram precision but does not suit Vietnamese due to its flexible syntax and rich expression and has a tendency to penalize semantically correct captions. ROUGE-L values content recall and is suitable for evaluating whether essential information is retained, especially beneficial in structured, factual captions. CIDEr, which considers both term frequency and semantic saliency across multiple references, best reflects the quality of informative and context-sensitive descriptions. Since our model is interested in generating short, well-organized captions for assisting visually impaired users in perceiving traffic scenes, CIDEr and ROUGE-L are the most indicative evaluation metrics for the same.

1. Data statistics amongst three datasets using for other models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Images | Captions | Captions/images | Avg. length |
| TPC37k [14] | 37,056 | 9108 | 1:4 | 31.21 |
| UIT-ViIC [12] | 4,000 | 20,000 | 5:1 | 12.19 |
| KTVIC [13] | 4327 | 21,635 | 5:1 | 10.97 |

1. Performance Comparison of Model Variants Using Evaluation Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Dataset | BLEU-1 | BLEU-4 | ROUGE-L | CIDEr |
| PVG | TPC37k | 54.7 | 34.1 | **84.7** | **142.5** |
| ResNetXt-152 (X152) PhoBERT [17] | UIT-ViIC |  | 47.0 | 63.97 | 128.8 |
| VinVL ResNeXt-152 PhoBERT [17] | UIT-ViIC |  | **47.3** | 64.83 | 137.1 |
| NIC [12] | UIT-ViIC | 68.2 | 32.7 | 59.9 | 81.8 |
| ResNet101 LSTM [13] | KTVIC | 53.0 | 15.5 | 42.3 | 21.8 |
| GRIT [13] | KTVIC | **74.7** | 40.6 | 59.7 | 136.0 |

Based on the comparative results displayed in Table X, it can be observed that the designed PVG model, when trained on the domain-specific datasets of TPC, performs better consistently under different evaluation metrics. Based on the TPC37k dataset, PVG achieves the highest CIDEr (142.5) and ROUGE-L (84.7) values among all the compared models. These evaluation metrics are established to capture semantic relevance and content fidelity in caption generation more effectively, especially under structured or context-specific conditions like traffic imagery.

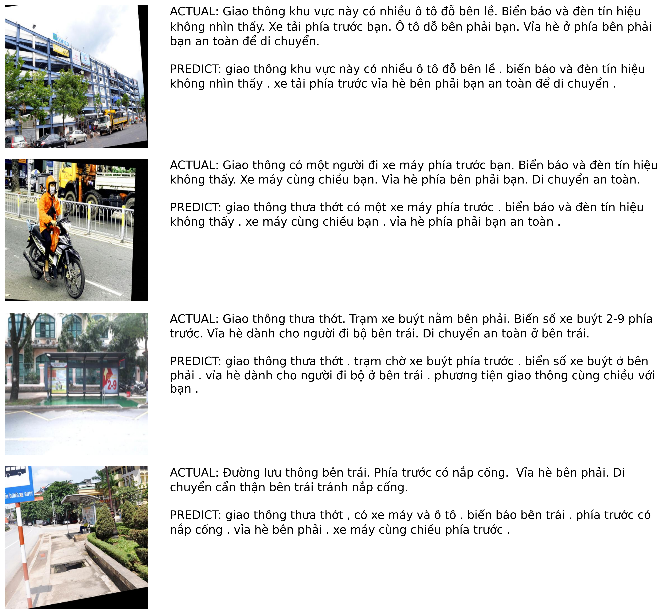
Interestingly, whereas both VinVL ResNeXt-152 and ResNeXt-152 (X152) trained on the UIT-ViIC dataset are decent in BLEU-4 value (47.3 and 47.0, respectively), this may or may not carry over to general caption quality. BLEU punishes n-gram overlaps, which may not necessarily indicate semantic similarity, particularly for free-form captioning tasks like those found in UIT-ViIC. On the other hand, the PVG model—thanks to its template-guided caption structure and domain-specific extension over TPC data—has stronger performance in metrics that emphasize content accuracy and contextual depth.

Interestingly, though the NIC model with the best BLEU-1 score of 68.2 is deficient in CIDEr (81.8) and ROUGE-L (59.9), it shows that shallow word matching alone cannot guarantee high caption quality. This further suggests that task-oriented modeling and alignment of the dataset, as done in PVG, can be significantly enhanced, particularly when the domain requires structured descriptions.

Briefly, the results prove the power of applying domain expertise and data augmentation methods in captioning images. PVG's performance on the TPC benchmarks demonstrates the power of adjusting model architecture and training data to the task-specific context requirements.

## Generated Caption Analysis

As shown in Figure 7, the captions are extremely fluent and semantically sound, rightly captioning principal visual elements in Vietnamese traffic scenes. The model, on average, requires 6 to 8 seconds to output a caption for each image. Qualitatively, the predictions are concise, contextually accurate, and grammatically correct, with strong attention to principal traffic elements such as cars, road signs, pedestrians, and lane markings. In dense situations, the model shows good prioritization, frequently detecting important objects (e.g., pedestrians, motorbikes, buses) while retaining clear sentence coherence. This shows in dense scenes or in scenes with rich traffic activity. Although these are positive, there are some weaknesses. In specific cases, the model occasionally fails to detect fine-grained visual cues, such as traffic sign glyphs or subtle visual warnings. Additionally, directional information, particularly the "left"-"right" dichotomy—is also subject to unreliability in certain situations, giving rise to the threat of ambiguity in spatial descriptions. These issues pose a need for enhanced spatial reasoning and observation of symbolic features in future revisions of models.



1. Examples of inference images from the TPC37k dataset with our proposed method (PVG)

# CONCLUSION

This work introduces PVG, a multimodal image captioning model that combines ViT for visual encoding, GPT-2 for language decoding, and BARTpho for enhanced Vietnamese language comprehension. To address the tokenizer–decoder mismatch between BARTpho (SentencePiece) and GPT-2 (BPE), we utilize LoRA to selectively fine-tune solely the final decoder layers, preserving pretrained knowledge with improved flexibility. Furthermore, captions on our specially formatted TPC37k dataset also follow a semantic template format that helps stabilize decoding and maintain relevance in the traffic scenario. Experiment results show PVG–Augment performs exceedingly well with 142.5 CIDEr and 84.7 ROUGE-L metrics, outperforming existing baselines. Results demonstrate the power of combining architecture-level adaptation, tokenizer alignment practices, and task-specific data structure.

# References

|  |  |
| --- | --- |
| [1] | K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Computer Vision and Pattern Recognition*, 2016. |
| [2] | 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 *Computer Vision and Pattern Recognition*, 2018. |
| [3] | L. Huang, W. Wang, J. Chen and X.-Y. Wei, "Attention on Attention for Image Captioning," *Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 4634-4643, 2019. |
| [4] | A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017. |
| [5] | M. Cornia, M. Stefanini, L. Baraldi and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," p. 10578–10587, 2020. |
| [6] | Y. Pan, T. Yao, Y. Li and T. Mei, "X-Linear Attention Networks for Image Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. |
| [7] | S. Herdade, A. Kappeler, K. Boakye and J. Soares, "Image Captioning: Transforming Objects into Words," *Advances in Neural Information Processing Systems,* vol. 32, 2019. |
| [8] | L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso and J. Gao, "Unified Vision-Language Pre-Training for Image Captioning and VQA," *Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition,* p. 13041–13049, 2019. |
| [9] | P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi and J. Gao, "VinVL: Revisiting Visual Representations in Vision-Language Models," *Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition,* p. 579–5588, 2021. |
| [10] | B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu and L. Yuan, "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," in *Conference on Computer Vision and Pattern Recognition*, 2023. |
| [11] | A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn and X. Zhai, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021. |
| [12] | Q. H. Lam, Q. D. Le, V. K. Nguyen, N. Luu and T. Nguyen, "UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning," in *Computational Collective Intelligence*, 2020. |
| [13] | C. A. Pham, Q. V. Nguyen, H. T. Vuong and T. Q. Ha, "KTVIC: A Vietnamese Image Captioning Dataset on the Life Domain," arXiv, 2024. |
| [14] | T. P. Dai, "Kaggle," 4 2025. [Online]. Available: https://www.kaggle.com/datasets/trieuphi/traffic-pictures-captioning. |
| [15] | N. L. Tran, D. M. Le and D. Q. Nguyen, "BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese," in *INTERSPEECH*, 2022. |
| [16] | E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2023. |
| [17] | K. NGUYEN, "EMPIRICAL STUDY OF FEATURE EXTRACTION APPROACHESFOR IMAGE CAPTIONING IN VIETNAMESE," *Journal of Computer Science and Cybernetics,* vol. 38, p. 327–346, 2022. |
| [18] | R. khan, M. S. Islam, K. Kanwal, M. Iqbal, M. I. Hossain and Z. Ye, "A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism," *nformation Technology and Control,* p. 16, 2022. |
| [19] | A. M. Rinaldi, C. Russo and C. Tommasino, "Automatic image captioning combining natural language processing and deep neural networks," *Results in Engineering,* p. 14, 2023. |
| [20] | G. O. d. Santos, E. L. Colombini and S. Avila, "CIDEr-R: Robust Consensus-based Image Description Evaluation," in *Computer Vision and Pattern Recognition*, 2021. |