ViTAGRU: A ViT-GRU-Based System for Real-Time Image Captioning

Xin En Lim1, a), Tee Connie1, 2, b) and Michael Kah Ong Goh1, 2, c)

1Faculty of Information Science & Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia

2Centre for Image and Vision Computing, CoE for Artificial Intelligence, Multimedia University, Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia

b) Corresponding author: tee.connie@mmu.edu.my   
a) 1211103245@student.mmu.edu.my

c) michael.goh@mmu.edu.my

**Abstract.** Image captioning is the task of captioning the content of the image in natural language. The current state of image captioning has a couple of issues like inability to interpret complicated and ambiguous scenes, large amounts of training data, and inability to deal with unseen situations. This paper proposes a new system that combines real-time object recognition and natural language processing (NLP) that can be adapted to generate semantic scene descriptions. It uses a standard camera to capture real-time images and perform image feature extraction tasks to identify objects and their attributes such as size, colour, and location and finally generates a text-based description of the image. The proposed system provides ViTAGRU with a combination of Vision Transformer (ViT) and Gated Recurrent Units (GRU) architecture with Additive Attention layer to accomplish image feature extraction and captioning. This combination leverages ViT's superior ability to analyze images with global context and long-range visual relationships while GRU offers computational efficiency and effective sequential text generation with reduced vanishing gradient problems compared to traditional RNNs. The integrated attention mechanism further enables the model to focus on the most important image regions during the text generation process, enhancing overall performance. The model was tested on the benchmarks COCO 2017 dataset and got a promising score of 0.73 and 0.56 in BLEU-1 and BLEU-2 respectively.

## INTRODUCTION

Image captioning is a task consisting of producing sentences that explain the content of an image in natural language [1]. Computer vision and natural language processing allows machines to analyze, comprehend images and convert it to human readable form. The task has attracted a lot of attention as it is applied in various areas, including the management of digital assets, generation of automated reports, tagging of social media posts, description of products in online stores and communication with various AI systems. A captioning system can be used to explain the contents of a picture placed on an online store, or an overview of the surveillance video in security systems. Visual data translation into structured text enables the reduction in the manual effort and accelerates the information processing on the user systems which rely on the visual context. Image captioning has become a fundamental research problem in AI as the demand rises to have smarter systems with the ability to see and describe the surroundings.

Despite remarkable progress, current image captioning systems still have limitations, most of the models can easily describe a simple or frequently occurring image but fail on more complicated scenes with many objects, actions, or subtle contextual details. [2]. Generating the captions that describe not only the objects but also their characteristics such as the shape, the color, and the relative position is still a hard task to carry out. It also faces the common issue of relying on vast, labeled datasets and development is often expensive. Some of these data sets might not represent very rare or localized situations. Consequently, the model is not effective when faced with new scenes or situations. In addition, the standard captioning methods can result in repetitive and monotonous descriptions that do not include the diversity and significant information when applied to real-life settings. Because of these challenges, it can be difficult to use such systems in situations where precise understanding of images matters a lot.

Deep learning has boosted image captioning by allowing models to directly learn both visually and with text [2]. Building on these advances, this work proposes a real-time image captioning system called ViTAGRU that leverages hybrid architecture combining Vision Transformers (ViT) and Gated Recurrent Units (GRU). ViT extracts rich and globally aware visual features, while GRU sequentially generates captions. Besides, attention mechanisms allow these models to pay more attention to essential parts of the image and produce each word of the caption. Using these tools and technologies together allows developers to build a more advanced and intelligent model to address the limitations of traditional approaches and improve overall performance in practical applications.

## LITERATURE REVIEW

In recent years, much research has focused on environment sensing following the development of deep learning while the issue of scene captioning has not been majorly considered yet. Scene captioning is the transformation of any captured scene into a descriptive text. The captions are expected to be well-formed with correct grammar, the right order for prepositions, and to contain accurate and rich information about the image. Combination of environment sensing and scene captioning technologies are essential to address real-world challenges faced by the visually impaired community. Some related studies are listed here.

Najm et al. [3] suggested a method to identify the objects and provide them with audio output to help visually impaired individuals. The YOLOv3 model was trained to get weights that help it identify and locate objects with a high level of accuracy. The proposed model obtained an impressive accuracy of 90.17%. Paswan and Choudhary [4] proposed a new method to estimate approximate distances of the objects and the user in an indoor environment with auditory feedback. The YOLOv7 model and YOLOv7-tiny were used for object detection and a distance estimation formula using bounding box was used to coordinate the detected object. The YOLOv7 model has achieved 0.65, 0.46, and 0.49 for the precision, recall and mAP score respectively. Subramanian et al. [5] proposed a method by combining YOLOv3 model for object detection and OCR algorithms for text extraction. The description of the scene is created by determining the location of every object using the centers of the bounding boxes. As a result, this model achieved high accuracy of 93%. Bougheloum et al. [6] proposed an innovative approach by combining the YOLOv5 deep learning model and an audio response system using CSPDarknet53-like backbones. Consequently, the suggested model reached a mean average precision (mAP) of 0.7 with an overall accuracy of 83.9%, which is impressive. Jambhulkar et al. [7] proposed a novel solution to design a system that detects objects in real-time and provides audio guidance by utilizing YOLOv3 algorithms. As a result, the model obtained 90% of average accuracy throughout the experiments.

Scene captioning consists of a semantic analysis of the video and the creation of a meaningful caption that best conveys the content in a natural language. The purpose of continuous research in this field is to allow captions to be more accurate and versatile.

Oion et al. [8] introduced a method named ItOD-AD to detect object and create the natural language audio description to the blind people. CNN was used to identify objects quickly and VGG-16 was used to classify objects well. After extracting detailed contextual information from images, text-to-speech technology generated audio captions. The model achieved an excellent training accuracy of 95.80%. Shanthi et al. [9] proposed a model by utilizing object detection, image captioning, and NLP technology. A CNN model which was used as the encoder worked to determine features of the image and the GRU which works as the decoder, putting into the obtained features into images and generating grammatically correct and semantically meaningful captions. Images were further analysed using InceptionV3 with audio output. For a result, the proposed InceptionV3 model performed on Flickr8K dataset achieved excellent Bilingual Evaluation Understudy (BLEU) scores ranging from 70% to 100%. Varma and Peter [10] introduced transformer-based video captioning architecture to help blind people. CNN-extracted frame features were position-encoded and processed by the transformer encoder via self-attention and feed-forward layers, while the decoder applies self- and cross-attention for caption generation. Using the MSVD dataset, the proposed model achieved BLEU score of 53.2 thereby outperforming the existing methods. Chharia and Upadhyay [11] proposed a scene description generator using deep recurrent architecture. Video frames were processed with a pre-trained VGG16 to extract 4096-dimensional feature vectors, which were then input into an LSTM to generate captions, later converted to audio. The model obtained an average BLEU score of 27.0 in the evaluation result.

# PROPOSED SOLUTION

## Overall Framework

The real-time environment sensing and auditory captioning system named ViTAGRU, designed for assisting blind individuals follows a streamlined process that can be seen in the flowchart below. First, the system starts with the step to detect each scene with a standard HD camera to capture real-time video stream from the user's surroundings. The input is essential for continuing to the next steps. Next, the system uses a pre-trained ViT model as feature extraction techniques to extract important features such as main objects, subjects, actions, and contextual elements. In the image captioning stage, the extracted features are fed into a deep learning model called GRU that acts as a decoder which writes a description of the detected objects and their features. Finally, the system moves to caption output and user interaction where the generated descriptive text is displayed to the user. Figure 1 shows the flowchart of the real-time image captioning system.

A diagram of a bird

AI-generated content may be incorrect.

**Figure 1.** Overall framework of image captioning

## Image Caption Preprocessing

In this study, text preprocessing including converting captions to lowercase, removing digits and special characters, standardizing spaces, and adding “startseq” and “endseq” tags is applied to each caption. Next, tokenization breaks sentences into smaller units. For example, the caption “A boy is playing” would be split into tokens like [“a”, “boy”, “is”, “playing”] and then converted into unique integer IDs. To ensure consistency, padding is applied to make sure all sequences have the same length by adding zeros to shorter captions. Finally, word embedding maps tokens into dense, lower dimensional vectors that capture the semantic relationships between words, placing similar words (like “cat” and “dog”) closer together in the embedding space, while unrelated words (like “cat” and “car”) are farther apart.

### **The Proposed ViTAGRU Architecture**

In this section, we introduce our proposed image captioning model called ViTAGRU which is a real-time assistive model for visually impaired people. The model employs a Vision Transformers for encoding and visualizing images, a GRU-based decoder for decoding a sequence, and a sophisticated Bahdanau (Additive) attention mechanism, which maps images and generates words at a higher level. ViTAGRU is developed based on a dual-stream approach to distinctively process the visual and text modalities.

For visual encoding, we use a pre-trained Vision Transformer (ViT) model from Dosovitskiy et al. [12]. Unlike traditional convolutional neural networks (CNNs) that extract local features through convolution and pooling operations, ViT processes an image as a sequence of fixed-size patches using self-attention mechanisms. The model outputs a 768-dimensional feature vector from the final hidden state of the special classification token [CLS], which serves as a global representation of the image: *I*∈*R*768. To enhance feature robustness, this vector is passed through a dropout layer, batch normalization, a dense layer that projects it to the embedding dimension d, and a layer normalization step as shown in Equation (1).

|  |  |
| --- | --- |
|  | (1) |

Each caption is tokenized and mapped to an embedding vector via a learnable embedding layer. The resulting embedded sequence is denoted as in Equation (2).

|  |  |
| --- | --- |
|  | (2) |

These embeddings pass through a dropout layer and are then fed into a single-layer GRU encoder, which returns the full sequence of hidden states and the final hidden state. This final hidden state, henc is used to initialize the decoder GRUs (see Equation (3).

|  |  |
| --- | --- |
|  | (3) |

The decoder consists of two stacked GRU layers, connected with residual links to improve gradient flow. The first GRU layer takes the embedded caption tokens as input and is initialized with the encoder’s final hidden state (refer to Equation (4)).

|  |  |
| --- | --- |
|  | (4) |

The second GRU layer processes the output of the first (shown in Equation (5)).

|  |  |
| --- | --- |
|  | (5) |

A residual connection is applied between the two layers to improve gradient flow and training stability (see Equation (6)).

|  |  |
| --- | --- |
|  | (6) |

Each GRU cell operates using the standard update rules (shown in Equations (7) – (10)).

|  |  |
| --- | --- |
|  | (7) |
|  | (8) |
|  | (9) |
|  | (10) |

where zt, rt are the update and reset gates, which control how much of the past state is preserved and determine how much past information to forget respectively. is the candidate activation which represents the new content to add to the state while is the final hidden state at time step *t*.

A diagram of a software process

AI-generated content may be incorrect.

**Figure 2.** ViTAGRU system architecture

Figure 2 illustrates the architecture of the proposed ViTAGRU model. The process begins with image inputs, which are split into patches and embedded using the ViT encoder to extract meaningful image features. Such features are subsequently normalized and regularized and sent to the decoder. On the text side, the input captions are embedded and fed to a single-layer GRU and finally processed by a two-layer GRU decoder with residual links. An improved Bahdanau attention mechanism fuses the image features with decoder outputs to generate a context vector, which is then concatenated with the decoder state and passed through fully connected layers to produce the final word predictions. To generate captions of images more precisely, the architecture places emphasis on effective regularization, residual learning, and enhanced attention.

## EXPERIMENT RESULTS AND DISCUSSION

The experiment was conducted using a benchmark dataset COCO 2017 [13]. The dataset contains more than 118k images. Each image contains five human-annotated captions. BLEU (Bilingual Evaluation Understudy) metric is used to assess the quality of the captions, as it is a popular way to measure machine-generated translations or text. BLEU evaluates the degree of n-gram overlap between a candidate (generated) sentence and one or more reference (human-written) sentences. The better the alignment with sentences written by a person, the higher the BLEU score which reflects greater fluency and similar meaning.

Our neural image caption generator demonstrates a notable capability to generate rich and descriptive captions for images in the test set. Training the model on rich COCO 2017 datasets has enabled us to obtain a system that is capable of perceiving and describing visual content to some degree. The generated captions show the capability of the model to leverage contextual as well as semantic features of images. The model achieved 0.73 of BLEU-1 score and 0.56 of BLEU-2 score, showcasing its performance across 5, 000 images in testing set.

As illustrated in Figure 3, the generated captions perfectly match the descriptions written by people. It recognizes the major objects, actions and environment appearing in each image. For example, image: a plane flying through the sky, human caption: a large passenger airplane flying through the air, generated caption: a large commercial airplane flying in the sky. These examples prove the model's ability to correctly interpret both straightforward and unambiguous images.



**Figure 3.** Sample caption outputs

When comparing the performance of models with several images, we see notable differences in caption generation. It correctly identifies objects and actions but struggles with more complex scenes, missing key details or generating slightly inaccurate actions in challenging scenarios. However, it only happens with a minor number of images.

To compare with existing study, we have conducted several experiments with different architectures and training strategies. We evaluated our proposed ViTAGRU model against established baselines by varying components such as the encoder backbone, the decoder structure, and the attention mechanism. Table 1 shows the results of existing methods alongside our experiment models, from which we selected the best-performing architecture as our final proposed model.

|  |  |  |
| --- | --- | --- |
| **TABLE 1.** *Comparison of Experimental Results*. | | |
| **Method** | **BLEU-1** | **BLEU-2** |
| VGG16 + LSTM | 0.52 | 0.30 |
| VGG16 + GRU | 0.54 | 0.32 |
| ViT + GRU + Multi-head Attention Layer | 0.69 | 0.46 |
| ViT + GRU + Adaptive Attention Layer | 0.68 | 0.46 |
| ViT + GRU + Hierarchical Cross Modal Attention Layer | 0.69 | 0.46 |
| ViT + GRU + (Additive, Dot-Product, Scaled Dot-Product Attention Layer) Sequentially | 0.64 | 0.41 |
| ViT + GRU + (Additive, Dot-Product, Scaled Dot-Product Attention Layer) Parallelly | 0.67 | 0.44 |
| ViT + GRU + Modified Additive Attention Layer **(proposed)** | **0.73** | **0.56** |

## CONCLUSION

In this study, we proposed ViTAGRU which is an image captioning model, and it combines a Vision Transformer for processing images, a single-layer GRU encoder, a two-layer GRU decoder with added residual connections for language modeling, and an improved additive attention mechanism to improve the link between the image and caption. Through extensive experiments on the COCO 2017 dataset, ViTAGRU demonstrated superior performance compared to traditional CNN-based encoder-decoder models, highlighting the effectiveness of combining transformer-based visual features with a strengthened sequential decoder. The proposed model achieved high accuracy in detecting and describing scenes with BLEU-1 and BLEU-2 scores of 0.73 and 0.56 respectively. This highlights the model’s capability to learn important relationships between images and words which result in captions that are both meaningful and correctly formed. More work can be done to improve the model by including bigger and wider user data, adjusting the model structure and experimenting with transformers for image captioning. Additionally, real-world testing with wearable cameras and feedback from users can help refine the system's usability and reliability.

## REFERENCES

1. B. Arystanbekov, A. Kuzdeuov, S. Nurgaliyev, and H. A. Varol, “Image captioning for the visually impaired and blind: A recipe for low-resource languages,” in Proc. 2023 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC) (IEEE, Sydney, Australia, 2023), pp. 1–4.
2. N. Haque, I. Labiba, and S. Akter, “FaceGemma: Enhancing image captioning with facial attributes for portrait images” (2024).
3. H. Najm, K. Elferjani, and A. Alariyibi, “Assisting blind people using object detection with vocal feedback,” in Proc. 2022 IEEE 2nd Int. Maghreb Meeting Conf. Sci. Tech. Automatic Control Comput. Eng. (MI-STA) (IEEE, Sabratha, Libya, 2022), pp. 48–52.
4. V. K. Paswan and A. Choudhary, “Camera based indoor object detection and distance estimation framework for assistive mobility,” in Proc. 2022 IEEE Int. Conf. Service Oper. Logist. Informatics (SOLI) (IEEE, Delhi, India, 2022), pp. 1–6.
5. R. R. Subramanian, L. Ravikiran, K. V. P. Teja, K. V. Reddy, and K. N. Reddy, “Voice guided object detection: Enabling independence for the visually impaired,” in Proc. 2024 3rd Int. Conf. Intell. Tech. Control, Optim. Signal Process. (INCOS) (IEEE, Krishnankoil, Virudhunagar District, Tamil Nadu, India, 2024), pp. 1–6.
6. L. Bougheloum, M. B. Salah, and M. Bettayeb, “Real-time obstacle detection for visually impaired people using deep learning,” in Proc. 2023 6th Int. Conf. Signal Process. Inf. Secur. (ICSPIS) (IEEE, Dubai, UAE, 2023), pp. 51–56.
7. A. R. Jambhulkar, A. R. Gajera, C. M. Bhavsar, and S. Vatkar, “Real-time object detection and audio feedback for the visually impaired,” in Proc. 2023 3rd Asian Conf. Innov. Technol. (ASIANCON) (IEEE, Ravet, India, 2023), pp. 1–5.
8. M. S. R. Oion, A. T. Prokriti, S. S. Hoque, S. Akter, S. Ara, Md. R. A. Shazid, and M. T. Mahmud, “A machine learning based image to object detection and audio description for blind people,” in Proc. 2023 26th Int. Conf. Comput. Inf. Technol. (ICCIT) (IEEE, Cox’s Bazar, Bangladesh, 2023), pp. 1–6.
9. N. Shanthi, J. Gowthami, N. Aravindhraj, J. S., K. R., and B. M., “Deep learning based audio description of visual content by enhancing accessibility for the visually impaired,” in Proc. 2023 Int. Conf. Sustain. Commun. Netw. Appl. (ICSCNA) (IEEE, Theni, India, 2023), pp. 1234–1240.
10. S. Varma and J. D. Peter, “Deep learning-based video captioning technique using transformer,” in Proc. 2022 8th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS) (IEEE, Coimbatore, India, 2022), pp. 847–850.
11. A. Chharia and R. Upadhyay, “Deep recurrent architecture based scene description generator for visually impaired,” in Proc. 2020 12th Int. Congr. Ultra Modern Telecommun. Control Syst. Workshops (ICUMT) (IEEE, Brno, Czech Republic, 2020), pp. 136–141.
12. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv:2010.11929 (2021).
13. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Computer Vision – ECCV 2014, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Lecture Notes in Computer Science Vol. 8693 (Springer, Cham, 2014), pp. 740–755.