CHAPTER 1 INTRODUCTION

1.1. Background

Digital visual content, especially images, has become dominant in today's information age. Every day, billions of images are uploaded and shared through various online platforms, ranging from social media to scientific repositories and personal collections [1]. This massive volume presents significant challenges regarding content management, accessibility, and understanding. Manual descriptions for each image become impractical due to time constraints, costs, and potential subjectivity and inconsistencies between annotators [2]. Further, the absence of accurate textual descriptions hinders the accessibility of visual information for individuals with visual impairments, limiting them from fully participating in the digital world [3]. Therefore, developing automated systems capable of producing accurate and relevant textual descriptions (captions) for images, otherwise known as image captioning, is an urgent need with far-reaching implications.

The impact of the lack of compelling image descriptions is felt globally and across sectors. In the context of the web and social media, images without adequate alt-text reduce the user experience for the visually impaired [4] and decrease the visibility of the content in search engines, which impacts reach and Search Engine Optimization (SEO) [5]. This phenomenon suggests that the ability to automatically understand and describe visual content is not only a technical issue but also has significant practical implications in various aspects of digital life, driving the need for advanced solutions in image captioning. Rapid developments in Artificial Intelligence (AI), particularly Deep Learning (DL), have offered a promising solution to the challenges of image captioning. This task is inherently multimodal, requiring close integration between Computer Vision (CV) to understand the visual content of images and Natural Language Processing (NLP) to produce coherent and

grammatical textual descriptions [6]. Early models of image captioning generally adopted an encoder-decoder architecture, in which Convolutional Neural Networks (CNNs) such as VGG or ResNet were used as encoders to extract visual features from images, and Recurrent Neural Network (RNN) or their variants, such as Long Short-Term Memory (LSTM), were used as decoders to generate descriptive word sequences [7], [8].

Over time, various improvements have been proposed to improve the quality of the resulting captions. The use of LSTM and Gated Recurrent Unit (GRU) has proven to be more effective than standard RNN in handling long-term dependencies in text, resulting in more flowing and logical descriptions [9]. A significant breakthrough then came with the introduction of the attention mechanism. This mechanism allows the decoder model to dynamically focus on the most relevant parts of the image while generating each word in the description, resulting in more detailed and contextually accurate captions [6]. Recent advances in neural network architecture, particularly the Transformer model introduced by Vaswani et al. [10], have brought a new paradigm in sequential tasks, including image captioning. In contrast to RNNs that process input sequentially, Transformers utilize a self-attention mechanism to process the entire input sequence in parallel, allowing for better global dependency capture and higher computational efficiency during training. Several studies have shown that Transformer-based architectures can produce superior image descriptions than RNN/LSTM-based models with attention, especially in terms of coherence, detail, and understanding of the relationships between objects in the image for general image datasets [6]. This success prompted further exploration of the use of the Transformer for various aspects of image captioning as a state-of-the-art architecture.

Although Transformer-based image captioning models have shown excellent performance in producing accurate descriptions, their large size and computational complexity often lead to high inference times. This can be a

significant constraint for on devices with limited resources. To address this, our study proposes an effective image captioning model by integrating the Vision Transformer (ViT) as the image encoder and the Distilled Generative Pre-trained Transformer 2 (DistilGPT2) as the decoder. We selected ViT for its strong ability to capture global contextual information from images and DistilGPT2 for its advanced capabilities in generating fluent, coherent text while offering advantages over traditional RNN-based decoders in handling long-range dependencies.

The research will be conducted in two main phases. The first phase will focus on implementing and training the ViT-DistilGPT2 model. This involves using a ViT Image Processor for pre-processing visual inputs, which are then encoded by ViT to produce rich feature representations. These features are subsequently fed into the DistilGPT2 decoder to generate textual descriptions. The model will be trained on the Flickr8k dataset. In the second phase, once an optimal model is developed, we will apply post-training dynamic quantization to reduce the precision of the model's weights and activations.

The primary goal of this methodology is to significantly reduce inference time and model size while maintaining an acceptable level of description quality. The project will conclude with a comprehensive evaluation of the model both before and after quantization. We will use standard caption quality metrics, such as Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE), alongside direct measurements of inference time and model size to quantify the improvements.

1.2. Problem Formulation

The formulation of this research problem focuses on the development of image captioning models using ViT as an encoder and DistilGPT2 as a decoder, as well as investigating the effectiveness of post-

training dynamic quantization techniques in reducing model size and inference time while maintaining the quality of the resulting image description.

1.3. Purpose and Benefits

Based on the formulation of this research problem, the purpose of this research is to develop a transformer-based image captioning model, using ViT as image encoder and DistilGPT2 as a decoder, and investigating the effectiveness of post-training dynamic techniques in reducing model size and inference time while maintaining the quality of the resulting image description.

The benefit of this research is the development of technology. This research can be a reference for the development of accurate and efficient image captioning, which can be applied in various sectors. The results of this research can be used to help develop better image captioning methods.

1.4. Problem Limitations

Based on the formulation of the problem and the purpose of the research, to realize research that is in accordance with the problem, there are the limitations of the problem that are studied as follows.

- This research only focuses on the development of an image captioning model.
- The use of algorithms is limited only to transformer-based models such as Vision Transformer (ViT) and Distilled Generative Pre-Trained 2 (DistilGPT2).
- The dataset used is the Flikr8k Dataset.
- The quantization method used is limited to post-training dynamic quantization (PTDQ).

1.5. Research Methods

The research method used in this Final Project focuses on the design, implementation, and evaluation of the image captioning system. The main stages of this research include: first, a comprehensive literature study on image captioning, Vision Transformer (ViT) architecture, Distilled Generative Pre-Trained Transformer 2 (DistilGPT2), dynamic quantization techniques, and relevant evaluation and metrics. Second, the design of the ViT-DistilGPT2 model architecture for image captioning and dynamic quantization application schemes was carried out. Third, training the model with Seq2SeqTrainer on the Flickr8k dataset and implementing post-training dynamic quantization (PTDQ). Fourth, testing and evaluation of model performance before and after quantization were carried out using standard metrics such as BLEU, ROUGE, model size, and inference time to analyze the effectiveness and efficiency of the proposed method.