AI Capstone Project Report (AIP490)

Group Code: GSP23AI04

Members: Le Trung Hieu (SE150560), Dao Cong Tuyen (SE150561)

Instructor: Nguyen Quoc Trung

Topic: SP23AI09 - Vietnamese Visual Question Answering

- Acknowledgment: We would like to express our deepest gratefulness to all those who provided us the aid to complete this project. A special gratitude we give to our instructor, Mr. Trung Nguyen, whose contribution in stimulating suggestions and encouragement helped us to coordinate my project, especially in writing the report. Furthermore, I would also like to acknowledge with much appreciation the crucial role of Mr. Khanh Quoc Tran, who gave me permission to access the dataset and the necessary materials to complete the project. We have to appreciate the guidance given by other supervisors as well as the panels, especially during our project review, their comments and advice improved our presentation skills.
- Author Contributions:

- Le Trung Hieu:
    - Research about the topics.
    - Implementing the vision modules.
    - Propose solutions.
    - Writing documents.
- Dao Cong Tuyen:
    - Propose solutions.
    - Implementing the language modules.
    - Running experiments.
    - Building demo.

- Abstract: A number of recent works have proposed attention models for Visual Question Answering (VQA) that generate spatial maps highlighting image regions and also part of question relevant to answering the question. In this project, we argue that although in English, there exist foundation models using the above structure that can adapt very well and achieve excellent results for visual question answering tasks, in Vietnamese, there gains very little attention and remains many limitations in terms of data and resources. Through this project, we aim to create a baseline for the Visual Question Answering task in Vietnamese , encourage further development in this

research area in the future, and also point out the necessary conditions to develop a successful Visual Question Answering system. We present a novel model for VQA that jointly reasons about image and question attention. In addition, our model reasons about the question (and simultaneously the image via the merge-attention mechanism) through a pyramid encoding fashion via the multi-head attention (MHA) layers. Our model improves the state-of-the-art on the ViVQA dataset from 34.96% to 62.01% accuracy, from 45.13 to 68.14 inWUPS 0.9 score , from 77.86 to 87.19 in WUPS 0.0 score.

I. Introduction.

1. Introduction to Visual Question Answering

The classification of images, object detection, and activity recognition are just a few of the computer vision tasks that have greatly benefited from recent advances in deep learning and computer vision research. Deep neural networks (DNNs) may do a certain action on par with humans when given adequate data. Similar results can be anticipated for other specialized computer vision issues as annotated datasets are growing in size quickly as a result of crowdsourcing. These issues, however, don't call for a comprehensive comprehension of images because of their specific nature, while humans are able to recognize the things in an image, comprehend their spatial relationships, infer their characteristics, and reason about each object's function in light of its surroundings.

It has long been believed that it is impossible to achieve the ambitious but unachievable aim of creating a computer vision system that can respond to arbitrary natural language questions regarding images. But since 2014, there has been a significant advancement in creating systems with these capabilities. The term **"Visual Question Answering" (VQA)** task refers to a multimodal task in which a system must infer the response to a text-based inquiry about an image. In computer vision, questions can be arbitrary and cover a wide range of sub-problems, e.g.

• Object recognition - What is in the image?
• Object detection - Are there any cats in the image?
• Attribute classification - What color is the cat?
• Scene classification - Is it sunny?
• Counting - How many cats are in the image?

Beyond this, a great deal more difficult queries can be posed, such as those involving spatial relationships between objects (e.g., what is between the cat and the sofa?) and inquiries involving common sense (e.g., why is the girl sobbing?). A strong VQA system must be able to reason about images and be capable of handling a variety of traditional computer vision tasks.

Figure 1: Samples for VQA task.

VQA has a wide range of potential uses. The most immediate use is as a tool for blind and visually impaired people, who can use it to get information about images both online and offline. A captioning system may, for instance, describe an image as a blind user travels through their social media feed, and the user could then use VQA to query the image to learn more about the scenario. In a broader sense, VQA might be utilized as a natural technique to question visual content to enhance human-computer interaction. Without employing image meta-data or tags, an image retrieval system can also be used. For instance, we can simply ask "Is it raining?" of all the images in the dataset to locate all the pictures taken in a rainy environment. Beyond applications, VQA is a significant area for basic research. It is possible to think of a VQA system as a part of an image understanding Turing Test because it needs to be capable of solving numerous computer vision challenges. A computer vision system is thoroughly tested using the Visual Turing Test to see whether it is capable of semantically analyzing images at the same level as a human. A system must be able to perform a wide range of visual activities in order to pass this test. VQA can be viewed as a type of visual Turing test that necessitates question comprehension but not necessarily more complex natural language processing. A significant portion of computer vision would likely be resolved if an algorithm could answer queries about images as good as or better than humans. However, this is only true if the *benchmarks and assessment methods are adequate to support such sweeping statements.*

VQA's main objective is to extract from the images question-relevant semantic data, which might range from the identification of minute details to the inference of abstract scene properties for the entire image, depending on the question. Although many computer vision issues require extracting data from the images, they are more constrained in scope and generality than VQA. Some effective methods use DNNs trained to categorize images into specific semantic categories to solve tasks like object recognition, activity recognition, and scene classification, the most popular of

these is object recognition, where computers are currently as accurate as humans. While those tasks are important computer vision problems that generalize object detection and recognition, they are not sufficient for holistic scene understanding.Label ambiguity is one of the main issues they deal with. The title is determined by the work. Furthermore, these methods by themselves do not comprehend an object's function in a larger context. In this illustration, designating a pixel as "bag" or "person" does not tell us whether the object is being carried by the person or if the person is sitting, running, or skateboarding. In contrast, VQA requires a system to respond to arbitrary queries regarding photos, which may necessitate deducing the connections between objects and the environment. *The appropriate label is specified by the question.*

In addition to VQA, a lot of recent research has focused on the intersection of vision and language (Multimodal). Image captioning, where an algorithm's objective is to provide a natural language description of a given image, is one of the most researched techniques. In order to offer a thorough description of a picture, image captioning is a fairly broad endeavor that may entail explaining complicated properties and object interactions. The visual captioning task, however, has a number of issues, with the evaluation of captions posing a particular difficulty. In short, a captioning system is free to choose the amount of granularity of its picture analysis, as opposed to VQA, where the level of granularity is determined by the type of the question addressed.

2. Evaluation Metrics for VQA.

VQA has been posed as either an open-ended task in which an algorithm constructs a string to answer a query or as a multiple-choice question in which it selects from a set of options. Simple accuracy is frequently used to evaluate multiple-choice questions, with an algorithm obtaining an answer accurately if it chooses the proper choice. Simple accuracy can also be utilized for open-ended VQA. In this situation, the predicted answer string of an algorithm must exactly match the ground truth answer. However, accuracy can be overly strict because some errors are significantly worse than others. For example, if the question is 'What animals are in the photo?' and a system returns 'dog' rather than the right label 'dogs,' it is penalized just as severely as if it returned 'dogs'. Questions may also have many correct answers; for example, 'What is in the tree?' may have 'bald eagle' listed as the correct ground truth response, so a system that outputs 'eagle' or 'bird' as the answer would be penalized just as much as if it outputs 'yes' as the answer. Because of these concerns, numerous alternatives to perfect accuracy for evaluating open-ended VQA methods have been proposed.

Wu-Palmer Similarity (WUPS) was considered as a substitute to accuracy.It attempts to quantify how much a predicted answer differs from the ground truth based on semantic meaning differences. WUPS will assign a value between 0 and 1 based on the similarity of a ground truth answer and a predicted answer to a query. It accomplishes this by locating the least common subsumer between two semantic

senses and assigning scores based on how far down in the semantic tree the common subsumer must be found. WUPS penalizes semantically related but non-identical terms relatively less. However, WUPS tends to assign relatively high scores to even remote concepts. To address this, thresholding WUPS scores was proposed, where a score below a threshold is scaled down by a factor.

	Pros	Cons
Simple Accuracy	• Very simple to evaluate and interpret • Works well for small number of unique answers	• Both minor and major errors are penalized equally • Can lead to explosion in number of unique answers, • especially with presence of phrasal or sentence answers
Modified WUPS	• More forgiving to simple variations and errors • Does not require exact match • Easy to evaluate with simple script	• Generates high scores for answers that are lexically related but have diametrically opposite meaning • Cannot be used for phrasal or sentence answers
Consensus Metric	• Common variances of same answer could be captured • Easy to evaluate after collecting consensus data	• Can allow for some questions having two correct answers • Expensive to collect ground truth • Difficulty due to lack of consensus
Manual Evaluation	• Variances to same answer is easily captured • Can work equally well for single word as well as phrase or sentence answers	• Can introduce subjective opinion of individual annotators • Very expensive to setup and slow to evaluate, especially for larger datasets

Figure 2: Comparison of different evaluation metrics proposed for VQA.

3. Algorithms for VQA.

A large number of VQA algorithms have been proposed recently. However, all of those existing methods consist of 3 components:

1. An image encoder.
2. A question encoder.
3. An algorithm that combines these features to produce the answer (Fusion module).

For image features, most algorithms use DNNs that are pre-trained on ImageNet, with common examples being VGGNet, ResNet, Vision Transformer (ViT) and its variants. A wider variety of question featurizations have been also explored, including bag-of-words (BOW), long short term memory (LSTM) encoders, gated recurrent units (GRU), and BERT(Bidirectional Encoder Representations from Transformers) . To generate an answer, the most common approach is to treat VQA as a classification problem. In this framework, the image and question features are the input to the classification system and each unique answer is treated as a distinct category.

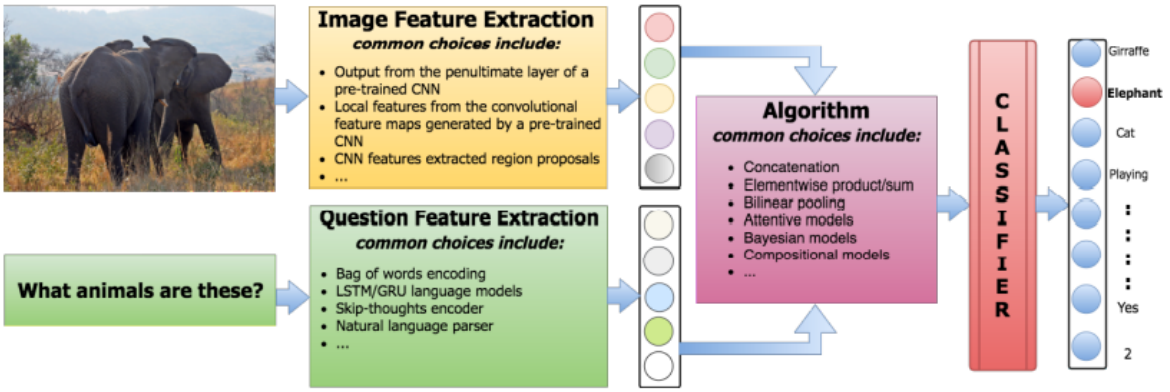


Figure 3: Simplified illustration of the classification-based framework for VQA.

Baseline approaches aid in *determining the difficulty of a dataset and establishing the bare minimum of performance that more sophisticated algorithms should achieve*. The simplest baselines for VQA are random guessing and guessing the most repeated answers. After combining the picture and question features into a baseline classification system, a linear or non-linear classifier, such as a multi-layer perceptron (MLP), is applied to them. Another approach is using Bayesian and Question-Aware Models, because drawing inferences and modeling links between the question and the image are required for VQA. Once the questions and images have been customized, modeling co-occurrence statistics of the question and image attributes can be used to infer the correct responses. A significantly different Bayesian model took advantage of the fact that the type of answer can be predicted purely by the query, for example, the model should assign 'What color is the flower?' as a color question, then effectively reducing the open-ended problem into a multiple-choice one. However, this was accomplished through the use of a form of quadratic discriminant analysis, which modeled the probability of picture features given the question features and answer type.

Previous features using global features alone may obscure task-relevant input space spaces. Attentive models attempt to overcome this problem. These models learn to 'attend' to the most important areas of the input space. Other vision and NLP tasks, such as object identification, captioning, and machine translation, have shown considerable success with attention models. All of these models are based on the assumption that particular visual regions in an image and specific phrases in a question are more informative than others for answering a given topic. For example, for a system that answers the question 'What color is the umbrella?' The image region containing the umbrella is more informative than other image regions. Similarly, 'color' and 'umbrella' are the textual inputs that require more direct attention than the others. Global picture characteristics, such as a CNN's final hidden layer, and global text elements such as bag-of-words, skip-thoughts, and so on, may not be detailed enough to solve region-specific concerns. Before employing spatially attentive methods, an algorithm must represent visual properties across all spatial

regions rather than only at the global level. Then, depending on the question, local aspects from relevant regions can be given more importance. There are numerous approaches to achieving local feature encoding. One method is to impose a uniform grid over all picture locations, as illustrated in Figure 4, with the local image features present at each grid site. This is frequently accomplished by operating on the last CNN layer before the final spatial pooling that flattens the features. The question then determines the significance of each grid location. Another approach to implementing spatial attention is to generate region suggestions (bounding boxes) for an image, encode each of these boxes using a CNN, and then use the question to decide the relevance of each box's attributes. When it came to the Transformer era, things couldn't have been easier. Dividing images into patches and learning the embedding of patches based on its relationship to the rest of the image has made learned features more and more informative. Along with the multi-head attention mechanism attached to the encoding process, the most important information is learned based on the context of both the photo and the question, used for late decision-making. The use of joint attention for picture and question features is also investigated. The primary idea is to let picture and question attention assist one another, guiding attention to significant words and visual regions at the same time. To do this, visual and question input are represented jointly by a memory vector, which is utilized to forecast attention for both question and picture features at the same time. The attentive process generates updated picture and question representations, which are then used to update the memory vector iteratively. This recursive memory update approach can be repeated K times to refine attention in stages before fed to a Multi-Layer Perceptron (MLP) to generate the answer (in the open-ended VQA task).

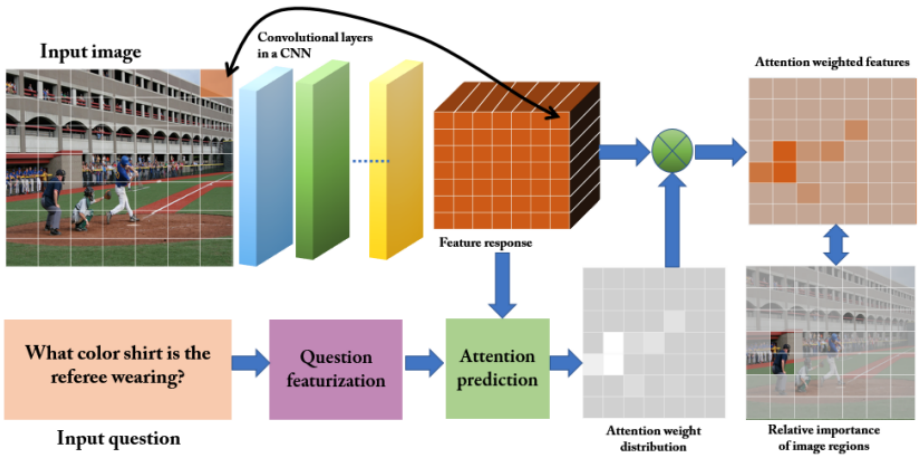


Figure 4: A way to incorporate attention into a VQA system.

Our concerns are also attention-based models. With the ability to extract the characteristics most relevant to the given problem, the model not only reduces parametric comprehension but also reduces the complexity of the task when

compared to using global features by limiting the variability of the input space, is a huge combined distribution between Image Domain and Language Domain. At the same time, the characteristics learned from the encoder modules are also very rich when expressing context, correlating with the rest of the image or question, facilitating the filtering process to take place easily.

4. Vietnamese Visual Question Answering.

Visual Question Answering (VQA) is a challenging task that involves processing both images and natural language questions to produce accurate answers. In recent years, there have been significant advancements in the development of deep learning models for encoding both visual and textual information. However, when working with languages other than English, the task becomes even more challenging.

- One of the main issues when working with languages such as Vietnamese is the lack of large-scale datasets. This is particularly problematic when it comes to VQA, as the task requires a deep understanding of both visual and textual information. Without enough data to cover the full variability of text and images, it becomes difficult to develop models that can generalize well to new inputs. Moreover, the quality of the dataset is also a crucial factor that affects the performance of the models. Low-quality data with errors, inconsistencies, and biases can lead to poor generalization and inaccurate predictions.
- Another challenge when working with Vietnamese is the lack of pre-trained models that can effectively encode both visual and textual information. While there have been significant advancements in the development of deep learning models for image and text encoding, these models are typically designed and trained for English-language datasets. As a result, when working with languages such as Vietnamese, it becomes difficult to achieve optimal performance using these models. To address these challenges, researchers have proposed several solutions. One approach is to develop new models specifically tailored for the Vietnamese language. This involves collecting more data and building models that are optimized for the specific characteristics of the language. While this approach has shown promise in some cases, it requires a significant amount of resources and may not be feasible for all research groups. Another approach is to adapt existing models to work with Vietnamese-language datasets. This involves fine-tuning pre-trained models on Vietnamese data to improve their performance on the VQA task. However, the success of this approach depends heavily on the quality of the pre-trained models and the availability of large-scale Vietnamese-language datasets. A third approach is to use transfer learning to leverage existing models that have been pre-trained on English-language datasets. This involves using the pre-trained models as feature extractors and training new models on top of these features to perform the VQA task. While

this approach has shown promising results in some cases, it is still limited by the fact that the pre-trained models may not be optimized for Vietnamese-language data.

- Despite these challenges, there have been some recent advancements in the development of models for VQA in Vietnamese. For example, some researchers have proposed using multi-modal transformers, which are models that can jointly process both visual and textual information. These models have shown promising results in the VQA task and can be adapted to work with Vietnamese-language data. And this is the way we are adapting for the current task. The lack of both data and academic effort in problems related to large and multimodal language models in Vietnamese is a pressing issue that needs attention. To address this problem, we have chosen to focus on adapting and improving the results on benchmark datasets available in the Vietnamese VQA problem. Our goal is to encourage and motivate more research efforts in Vietnamese-related tasks. With improved results, we hope to showcase the potential and the current drawbacks of Vietnamese language models and raise awareness of the need for further development and research in this area. Through our efforts, we aim to inspire and support a community of researchers who are dedicated to advancing the field of Vietnamese language models.

II. Related works.

1. Vision Models.

a. CNN and Visual Attention Network.

The most fundamental difficulty in computer vision is determining how to effectively compute powerful feature representations. Convolutional neural networks (CNNs) use local contextual information and translation invariance features to considerably improve neural network effectiveness. Since AlexNet, CNNs have swiftly become the standard foundation in computer vision. To increase usability, researchers worked hard to make CNNs deeper and lighter.

The attention mechanism can be thought of as an adaptive selection process based on the input feature that is introduced into computer vision. It has aided in a variety of visual tasks, including image classification, object identification, and semantic segmentation. Channel attention, spatial attention, temporal attention, and branch attention, as well as their combinations such as channel & spatial attention, are the four core categories of attention in computer vision. In visual tasks, each type of attention has a different effect.

Self-attention is a type of attention mechanism that originated in NLP. It is becoming increasingly essential in computer vision due to its effectiveness in capturing long-term reliance and adaptation. Various deep self-attention networks (a.k.a., vision transformers) outperformed mainstream CNNs on various visual tasks, demonstrating the enormous potential of attention-based models. Self-attention, on

the other hand, was originally created for NLP. When it comes to computer vision tasks, it has three flaws.

(1) It considers images as 1D sequences, ignoring the 2D structure of images.

(2) For high-resolution photos, quadratic complexity is too expensive.

(3) It merely achieves spatial flexibility while ignoring channel dimension adaptation. Different channels frequently represent different objects in vision tasks.

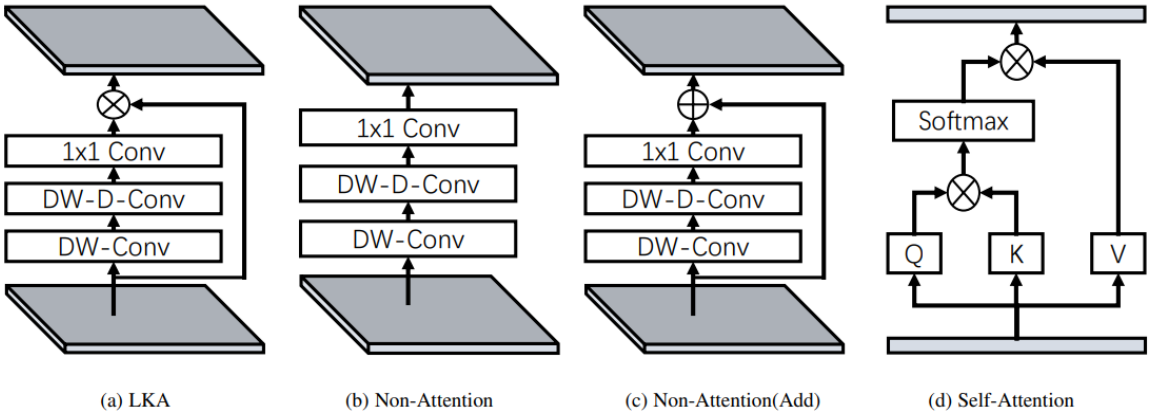


Figure 5: The structure of different modules: (a) the proposed Large Kernel Attention (LKA) for Visual Attention Network (VAN); (b) non-attention module; (c) replace multiplication in LKA with addition ; (d) self-attention. It is worth noting that (d) is designed for 1D sequences.

b. Vision Transformer.

Because it was designed for NLP, the typical Transformer model received a one-dimensional series of word embeddings as input. When used in computer vision for image classification, the input data to the Transformer model is provided in the form of two-dimensional images.The input image then is split up into smaller two-dimensional patches for the aim of arranging the input image data in a manner that mirrors how the input is structured in the NLP domain (in the sense of having a series of distinct words). Each image patch is then flattened into a vector. A sequence of embedded image patches is generated by mapping the flattened patches to dimensions, with a trainable linear projection. The patch embeddings are finally augmented with one-dimensional positional embeddings , hence introducing positional information into the input, which is also learned during training. In order to perform classification, we feed at the input of the Transformer encoder, which consists of a stack of multi-head attention layers to learn deeper representations. The Multi-head Attention Network (MSP) is a network that generates attention maps using embedded visual tokens. These attention maps assist the network in focusing on the image's most critical regions, such as object(s). The concept of attention maps is similar to that of saliency maps and alpha-matting in traditional computer

vision literature. Layer Norm keeps the training process on track and allows the model to adapt to differences between training photos.

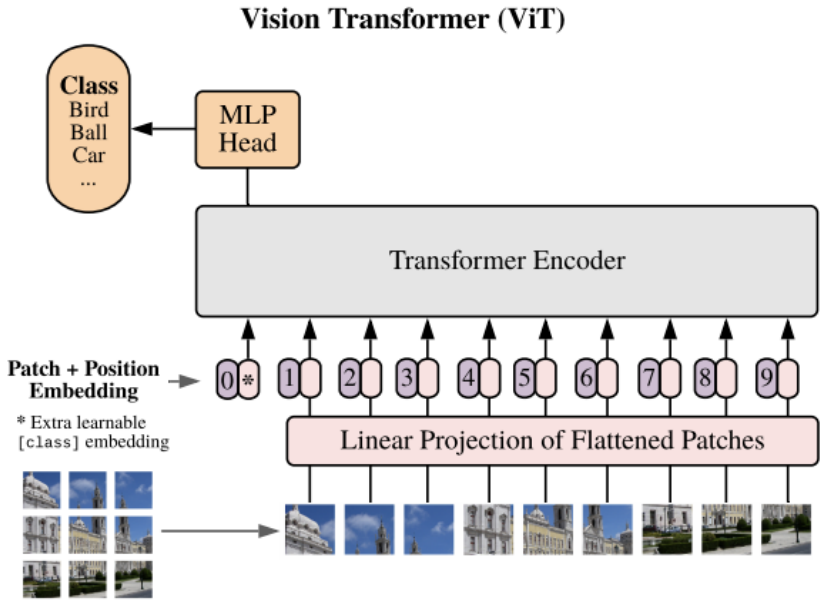


Figure 6: ViT architecture.

There are 6 novel features of ViT Architecture compared to CNN:

- When compared to CNNs, ViT shows more similarity between representations obtained in shallow and deep layers.
- In contrast to CNNs, ViT receives the global representation from the shallow levels, but the local representation from the shallow layers is also crucial.
- Skip connections in ViT are considerably more influential than in CNNs and have a significant impact on representation performance and similarity.
- CNNs maintain less spatial information than ViT.
- With vast volumes of data, ViT can learn high-quality intermediate representations.
- MLP-Mixer's representation is more similar to ViT than CNNs.

c. Swin Transformer.

To remedy the shortcomings of the original ViT, the Swin Transformer included two fundamental concepts: hierarchical feature maps and shifted window attention. Swin Transformer derives its name from "Shifted Window Transformer." Swin Transformer creates 'hierarchical feature maps', which is a considerable departure from ViT. To begin, 'feature maps' are just the intermediate tensors produced by each succeeding layer. In this case, 'hierarchical' refers to the fact that feature maps are merged from layer to layer (more on that in the following section), thus lowering the spatial dimension (i.e. downsampling) of the feature maps from one layer to another.

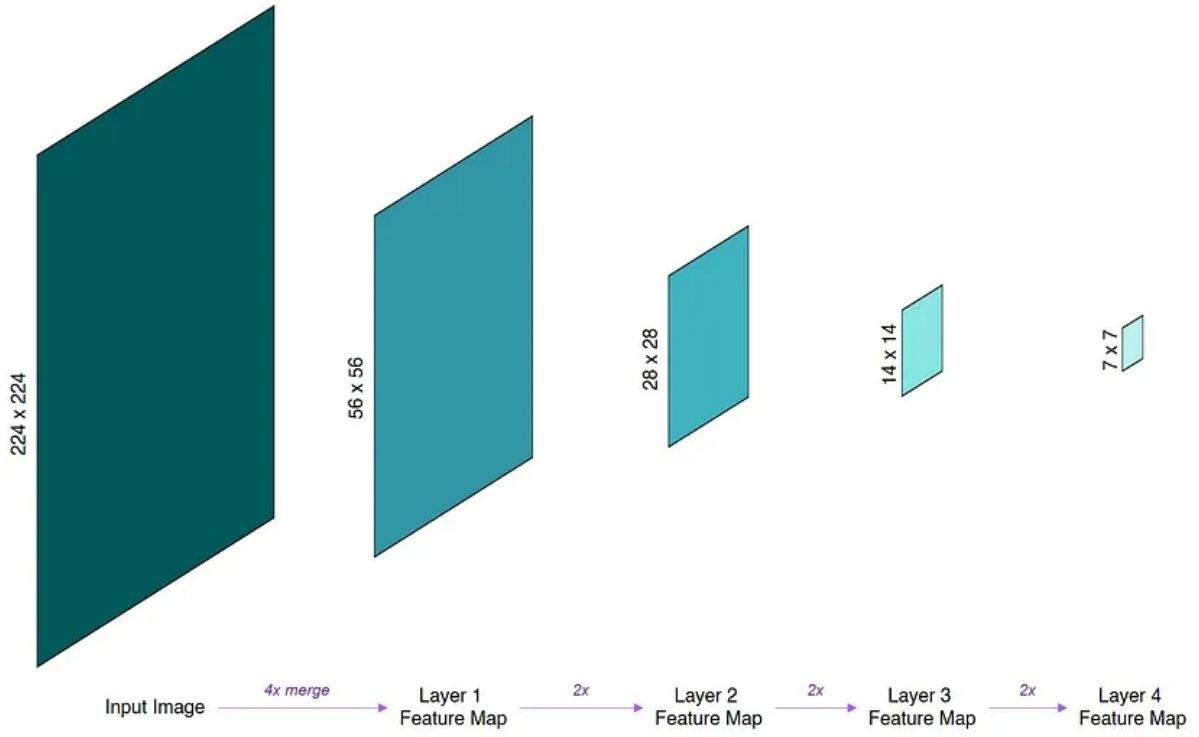


Figure 7: Hierarchical feature maps in Swin Transformer. Feature maps are progressively merged and downsampled after each layer, creating feature maps with a hierarchical structure.

More crucially, the Swin Transformer may be used in domains where fine-grained prediction is necessary, such as semantic segmentation, thanks to these hierarchical feature maps. In contrast, the ViT employs a single, low-resolution feature map throughout its architecture.

Patch merging is the technique employed in Swin Transformer for convolution-free downsampling. Swin Transformer's transformer block substitutes ViT's conventional multi-head self-attention (MSA) module with a Window MSA (W-MSA) and a Shifted Window MSA (SW-MSA) module. The standard MSA used in ViT does global self-attention, and the relationship between each patch and all other patches is computed. As a result, the complexity increases quadratically with the number of patches, rendering it unsuitable for high-resolution photos. Swin Transformer addresses this issue using a window-based MSA technique. A window is nothing more than a collection of patches, and attention is only computed within each window. Because the window size remains constant across the network, the complexity of window-based MSA is linear with respect to the number of patches (i.e. the size of the image), which is a significant improvement over the quadratic complexity of standard MSA. However, one clear limitation of window-based MSA is that limiting self-attention to each window reduces the network's modeling power. Swin Transformer addresses this by adding a Shifted Window MSA (SW-MSA) module after the W-MSA module. This shifting window technique introduces critical

cross-connections across windows and has been shown to boost network performance.

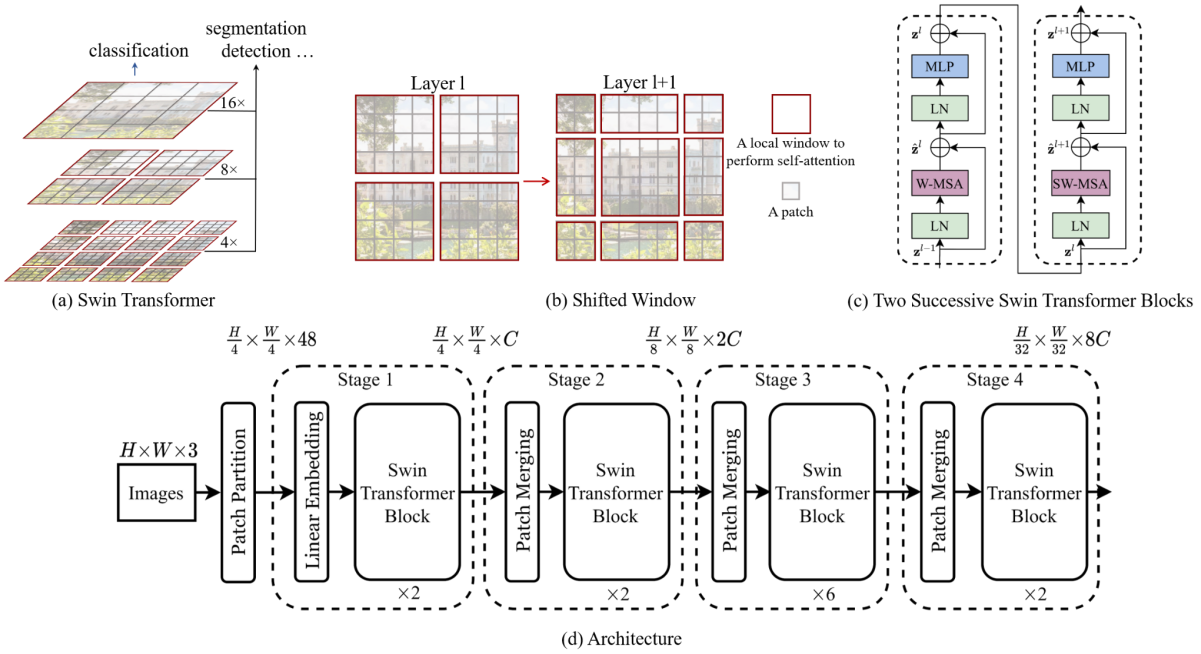


Figure 8: Swin Transformer Architecture.

d. Convolutional vision Transformer (CvT).

Convolutional Vision Transformer (CvT) is a novel architecture that improves the performance and efficiency of Vision Transformer (ViT) by incorporating convolutional neural networks (CNNs) into the ViT architecture. This hybrid model leverages the benefits of both CNNs and Transformers, resulting in a highly effective and efficient network that can tackle a wide range of vision tasks.

The first key change in CvT is the use of a Transformer hierarchy with a new convolutional token embedding. This new token embedding incorporates local spatial information into the network, enabling the model to better capture shift, scale, and distortion invariance. By doing so, CvT is able to achieve better performance on tasks that require spatial reasoning, such as object detection and segmentation. The second key change in CvT is the use of a convolutional Transformer block with a convolutional projection. This enables the network to better capture local interactions between features, which is critical for tasks that require precise spatial localization, such as object detection. The convolutional projection also helps to reduce the overall computational cost of the network, enabling faster training and inference times. By incorporating both CNNs and Transformers into the architecture, CvT is able to achieve the benefits of both models. CNNs are highly effective at capturing local spatial information and feature interactions, while Transformers are able to capture global context and improve generalization. By combining the two, CvT is able to achieve superior performance on a wide range of vision tasks.

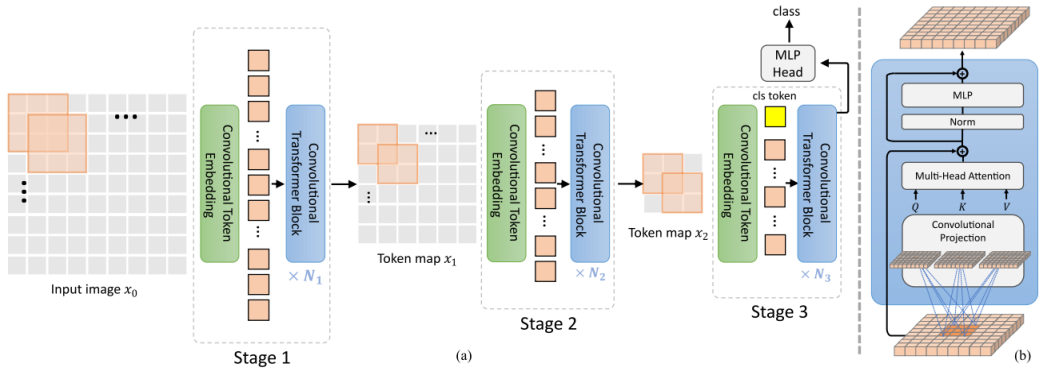


Figure 9: (a) Overall CvT architecture. (b) Details of the Convolutional Transformer Block, which contains the convolution projection as the first layer.

e. CLIP - ViT.

Although deep learning has revolutionized computer vision, current approaches have several major flaws: typical vision datasets are labor intensive and expensive to create while teaching only a limited set of visual concepts; standard vision models are good at one task and one task only, and require significant effort to adapt to a new task; and models that perform well on benchmarks perform disappointingly poorly on stress tests, casting doubt on the entire deep learning paradigm. CLIP (Contrastive Language–Image Pre-training) is a foundation network model aimed at addressing these issues: it is trained on a wide range of images with a wide range of natural language supervision readily available on the internet. The network is designed to be directed in natural language to complete a wide range of classification benchmarks without directly optimizing for the benchmark's performance, also called “zero-shot”.

CLIP models are substantially more versatile and general than existing ImageNet models in computer vision since they learn a wide range of visual concepts directly from plain language. All we need to do to apply CLIP to a new task is "tell" CLIP's text-encoder the names of the task's visual ideas, and it will output a linear classifier of CLIP's visual representation. Clip embedding can also be used for multimodal downstream tasks in English (such as VQA) because that embedding understands the context of the connection between 2 data distribution areas. CLIP enables users to create their own classifiers and eliminates the requirement for task-specific training data. The way these classes are created can have a significant impact on both model performance and model biases.

In our experiments, we reuse the trained visual encoder model of CLIP (ViT architecture), called CLIP-ViT. Although it has been trained and proven effective for understanding the link between embedding between 2 data distributions in images and English, we also hope that it can be adapted to Vietnamese through fine-tuning on the new dataset, because the language is generally structured data.

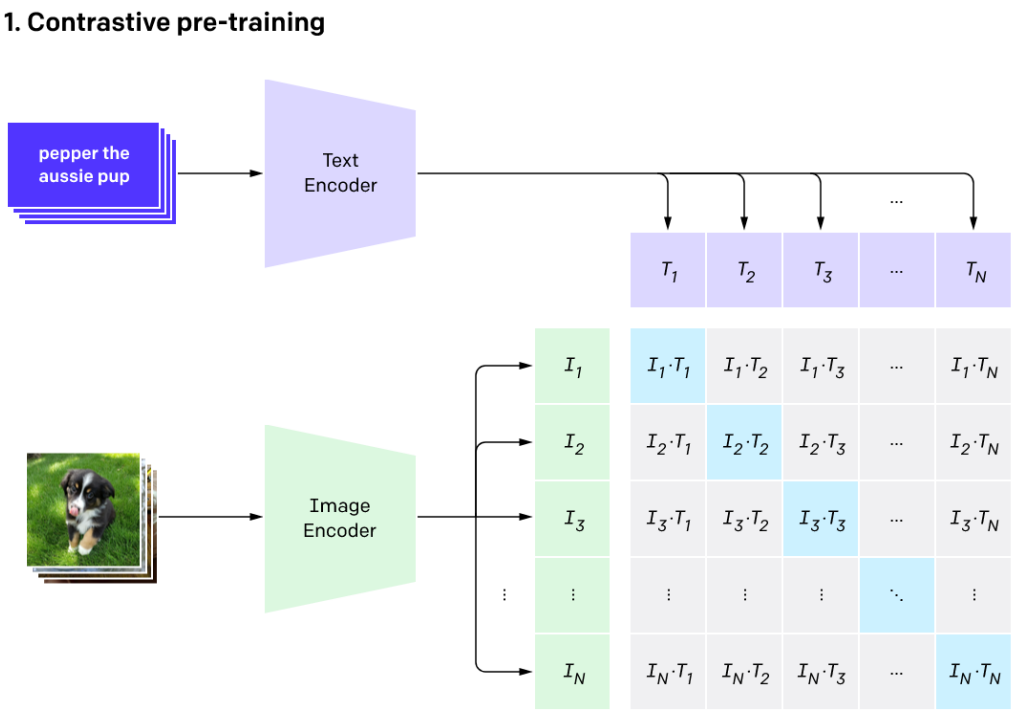


Figure 10: Training CLIP.

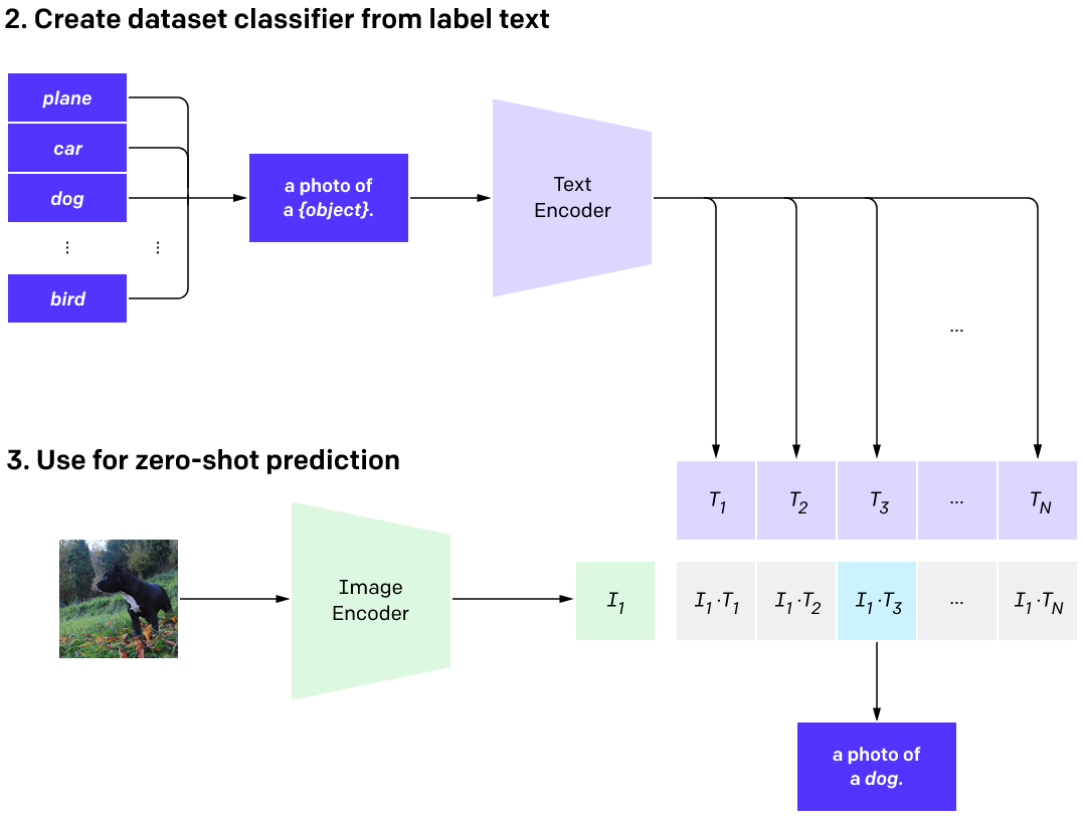


Figure 11: Zero-shot Classification with CLIP.

2. Language Model.

a. PhoW2V.

Before the advent of the word embedding method using word2vec, to represent words, people often chose ways like one-hot encoding. However, this method has too many weaknesses such as the length of the vector representing words as long as the number of words in the dictionary, which consumes a lot of storage space and the semantic relationship between words has not been shown. Since then word2vec was born using a neural network capable of representing words with a vector dimension much lower than the dictionary length. Based on the idea that two words appearing in the same context have the same meaning and from the semantics of the surrounding words the model can guess the central word. Word2vec has two models: Skip-gram and Continuous Bag of Words (CBOW) which all have in common is using a neural network to embed the semantics of words.

- Skip-gram: In the context of this study, it is postulated that a sliding window of predetermined dimensions traverses a sentence, where the word positioned in the center of the window is referred to as the "target," while the words situated to its left and right, within the confines of the sliding window, are designated as the "context" words. The skip-gram model, employed for training purposes, endeavors to estimate the probabilities associated with a word being designated as a context word for a given target word.

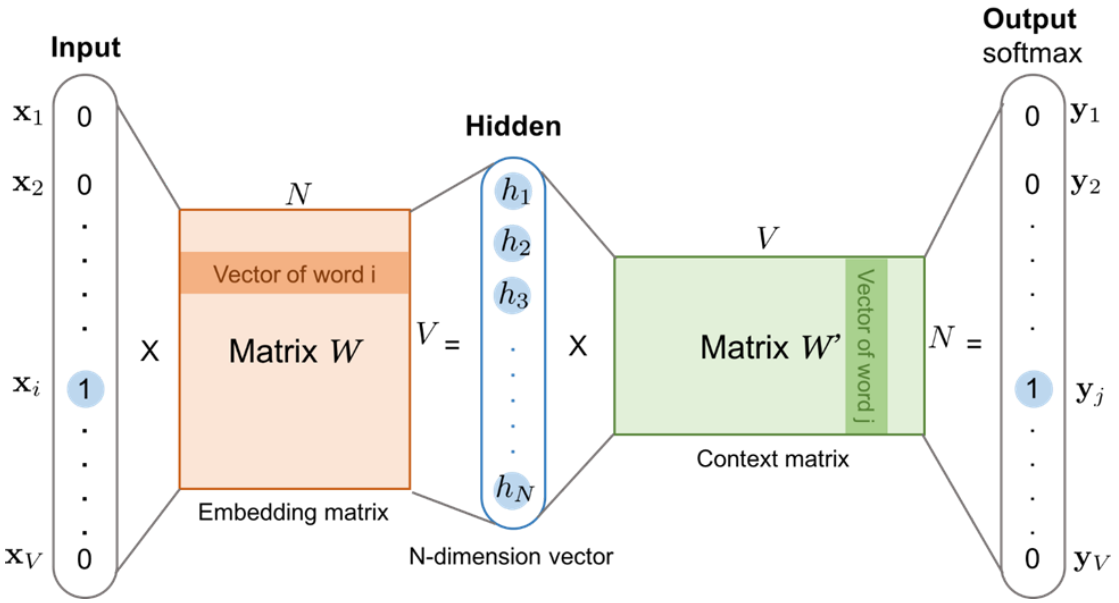


Figure 12: Training skip-gram.

- Continuous Bag of Words (CBOW): The Continuous Bag-of-Words (CBOW) model, a comparable approach for acquiring word embedding vectors, is posited as an alternative in this research. In this model, the prediction task involves estimating the target word based on the source context words, as opposed to the skip-gram model which predicts context words from a given

target word.

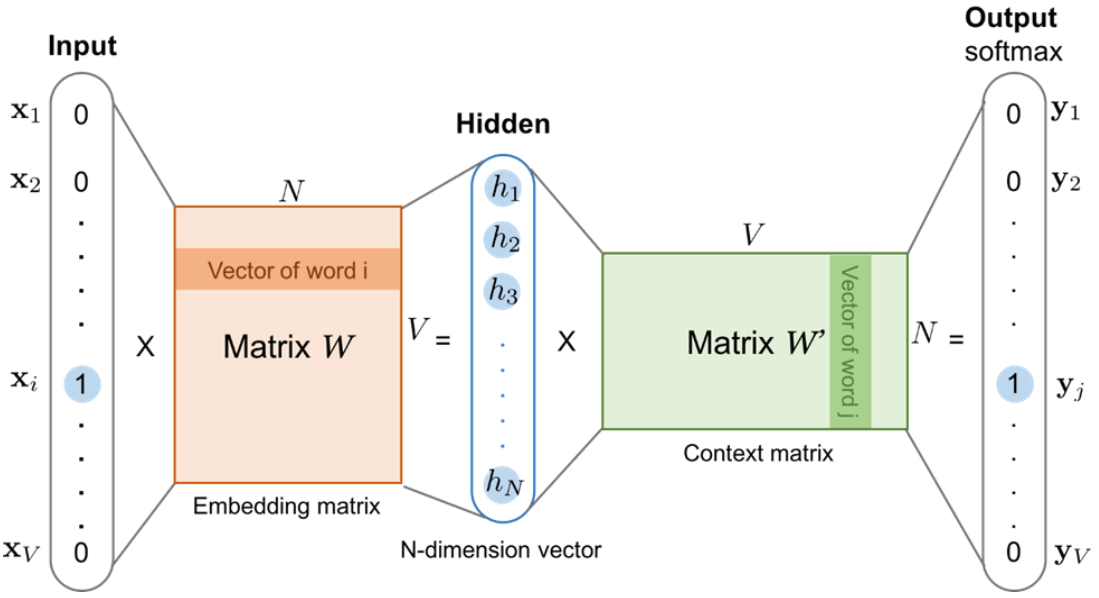


Figure 13: Training CBOW.

PhoW2V embedding was pre-trained on a 20 GB corpus of Vietnamese texts, providing word embedding vectors at syllable level and word-level with vector lengths of 100 or 300. Expressing words as semantics to enhance performance for other downstream tasks related to Vietnamese language processing.

b. PhoBERT.

The Transformer network, originally introduced by researchers from Google in their seminal paper "Attention is All You Need", has profoundly revolutionized the field of natural language processing. Acknowledging the significance of establishing a robust foundation for further research and applications in Vietnamese language processing, the development of a transformer network trained on Vietnamese data is imperative. PhoBERT is a state-of-the-art pre-trained language model that was developed by the Vietnamese AI community in 2020. It was trained on a large corpus of Vietnamese text using a Transformer architecture. The resulting model is capable of generating high-quality natural language text in Vietnamese, as well as performing a variety of downstream NLP tasks such as text classification, named entity recognition, and sentiment analysis.

One of the key innovations of PhoBERT is its use of subword tokenization, which splits words into smaller units called subwords. This allows the model to capture the morphology of Vietnamese words more effectively, as Vietnamese is a tonal language with a complex system of diacritics that can significantly alter the meaning of a word. By breaking words down into subwords, PhoBERT is able to represent these nuances more accurately, resulting in improved performance on tasks such as sentiment analysis and text classification. Another strength of

PhoBERT is its use of masked language modeling, which is a pre-training task where the model is trained to predict missing words in a sentence based on its context. This task helps the model to learn the syntactic and semantic relationships between words, as well as the overall structure of language. This in turn allows the model to generate more coherent and grammatically correct text, as well as perform better on tasks that require understanding of sentence structure such as natural language inference and machine translation. PhoBERT's embeddings have been shown to be highly effective on a variety of downstream tasks, achieving state-of-the-art performance on several benchmarks. For example, in the Sentiment Analysis in Vietnamese (SAIV) challenge, PhoBERT achieved an accuracy of 87.8%, outperforming all other models by a significant margin. In addition, PhoBERT has been used to improve the performance of other NLP models on Vietnamese text, such as named entity recognition and text classification.

Despite its strengths, PhoBERT does have some limitations. One is its reliance on pre-training on a large corpus of text, which can be computationally expensive and time-consuming. This can make it challenging for researchers with limited resources to use PhoBERT effectively. Additionally, PhoBERT's subword tokenization can lead to some loss of interpretability, as it can be difficult to map subwords back to their original words in a meaningful way. Overall, however, PhoBERT represents a significant advancement in NLP for Vietnamese, providing a powerful tool for researchers and practitioners working in this language. Its strong performance on a variety of tasks, combined with its innovative use of subword tokenization and masked language modeling, make it a highly versatile and effective language model.

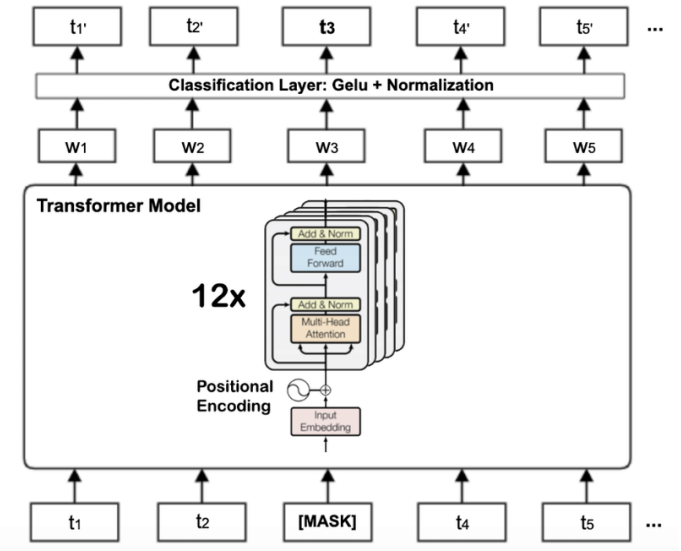


Figure 13: Training BERT with masked language modeling task.

III. Project management plan.

The purpose of this project management plan is to define the processes, procedures, and resources required to develop a Vietnamese Visual Question Answering (VQA) system. The VQA system will allow users to input natural language questions related to images and receive accurate and contextual answers in Vietnamese.

- Project Objectives: The main objectives of this project are as follows:
    - Develop a Vietnamese VQA model that can outperform the baseline model for ViVQA dataset.
    - Building a demo for visualizing the ability of the best model for answering arbitrary questions about the content of an testing image.
    - A research paper
- Project Scope: The project scope will include the following activities:
    - Research and analyze existing VQA systems and their performance in the Vietnamese language.
    - Develop a new system architecture and design the algorithm for the Vietnamese VQA system.
    - Collect and preprocess image and text data for training and testing.
    - Train and test the system using proposed algorithms.
    - Develop a user interface that allows users to upload images and input questions in Vietnamese.
    - Test and validate the system's performance through a series of evaluation metrics.
- Project Deliverables: The project deliverables will include the following:
    - A detailed project plan outlining the project scope, objectives, and timelines.
    - System architecture and design.
    - Preprocessed dataset for training and testing.
    - Trained and tested machine learning model.
    - User interface design document (Demo).
    - A report detailing the evaluation metrics and the system's performance.
- Project Timeline: The project timeline will be broken down into the following phases:
    - Research and analysis phase (2 weeks)
    - Design phase (1 weeks)
    - Data collection and preprocessing phase (1 week)
    - Training and testing phase (4 weeks)
    - User interface development phase (2 weeks)
    - Evaluation and reporting phase (2 weeks)
- Risk Management: The following risks have been identified for this project:
    - Data collection and preprocessing delays.
    - Insufficient computing resources for training and testing.
    - Algorithm performance does not meet the desired accuracy level.

The detailed plan is shown in Table 1:

Timeline	**Member's Task**		**Team's task**
	HieuLT	TuyenDC	
Week 1	Learn about VQA task	Learn about Dataset for Vietnamese VQA and baseline model	Research paper
Week 2	Learn about SOTA approach for VQA task	Download dataset and visualizing statistics	Research paper
Week 3	Finding drawbacks and proposing development directions.	Building DataLoader	Define the problems and propose the solutions
Week 4	Implementing visual encoder (4 modules)	Implementing language encoder (2 modules)	Running baseline and reporting results
Week 5	Implementing visual encoder (4 modules)	Implementing language encoder (2 modules)	Running baseline and reporting results
Week 6	Implementing fusion modules (2 modules)	Running experiments	Running baseline and reporting results
Week 7	Running experiments and writing documents	Running experiments and reporting results	Running experiments and reporting results, implementing proposed methods
Week 8	Implementing proposed methods	Running experiments and reporting results	Running experiments and reporting results, implementing proposed methods
Week 9	Running experiments and writing documents	Running experiments and writing documents	Writing paper and building demo with Streamlit
Week 10	Running experiments and writing documents	Running experiments and writing documents	Writing documents and building demo with Streamlit
Week 11	Writing documents	Writing documents	Writing documents and building demo with Streamlit
Week 12	Writing documents	Writing documents	Writing documents and building demo with Streamlit, paper

			submission
Week 13	Writing documents and build API	Writing documents and build API	Design slides for presentation, paper submission
Week 14	Writing documents and build API	Writing documents and build API	Finish slides for presentation, demo, paper rebuttal

Table 1: Project Management Plan

IV. Materials and methods.

1. Dataset.

In this project, we reuse the dataset ViVQA, a dataset for a VQA system in Vietnamese. It is translated from the COCO-QA dataset, where the images are from MS COCO, which is one of the most prestigious, vast, and diverse image datasets to date. The dataset construction procedure was divided into three stages: image collection, question generation, and dataset validation.

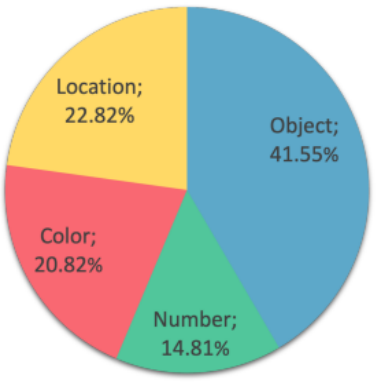
- Extract 10,328 images randomly from the MS COCO dataset.
- Translate question-answer pairs from English to Vietnamese using the COCO-QA dataset in English by employing two well-known machine translation technologies, Google Translate2 and Microsoft Translator3. However, numerous translated questions are strange and difficult to grasp.
- Before reviewing and correcting errors in translated data, annotation rules to improve the dataset's quality are provided, where question-and-answer pairs must follow the rules outlined in Table 1. Questions and answers that do not match the rules are eliminated and replaced with new ones.

No.	Descriptions
1	Each image must contain 1 - 3 questions.
2	Each question must have one corresponding and unique answer.
3	Each answer must only contain one word.
4	Q&A only about the activities and objects visible in the image.
5	Familiar English words like laptop, TV, ok, etc. are allowed.
6	Each question must be a single sentence.
7	While annotating, personal opinion and emotion must be avoided.

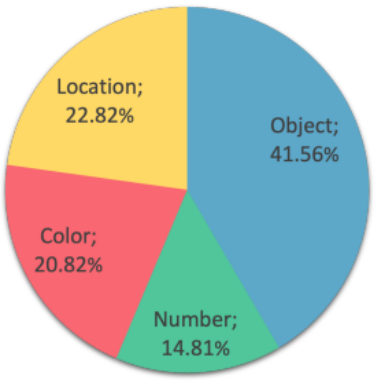
8	Questions can include a variety of activities and objectives from various perspectives.

Table 2: ViVQA dataset annotation rules.

The ViVQA dataset contains 10,328 photos and 15,000 pairs of questions and answers related to the images' content. The dataset is splitted into training and test sets in an 8:2 ratio. Because the question and answer pairings in our dataset are based on question and answer pairs from the COCO-QA dataset, the question types follow the definitions of this one, including four categories: object, number, color, and location.



Training set.



Test set.

Figure 12: The distribution of the question types on the ViVQA dataset.

Category 0: Object

- Question: những gì chứa nhiều rau?
- Answer: đĩa ăn

Figure 13: Training sample of category 0 (Object) in ViVQA dataset.

Category 1: Number

- Question: có bao nhiêu con ngựa vằn gặm cỏ trên cây bụi trong tự nhiên?
- Answer: một

Figure 14: Training sample of category 1 (Number) in ViVQA dataset.

Category 2: Color

- Question: màu của chiếc bình là gì?
- Answer: màu xanh lá

Figure 15: Training sample of category 2 (Color) in ViVQA dataset.

Figure 16: Training sample of category 3 (Location) in ViVQA dataset.

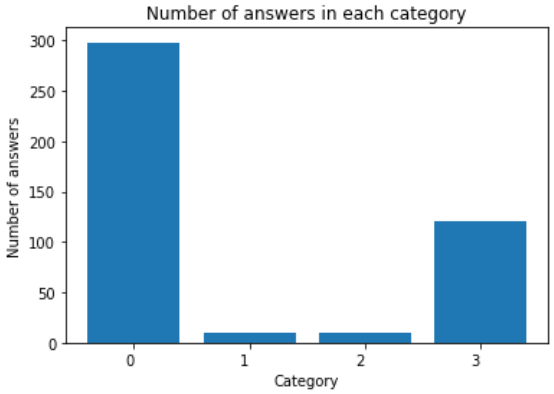


Figure 17: Number of answers in each type of question.

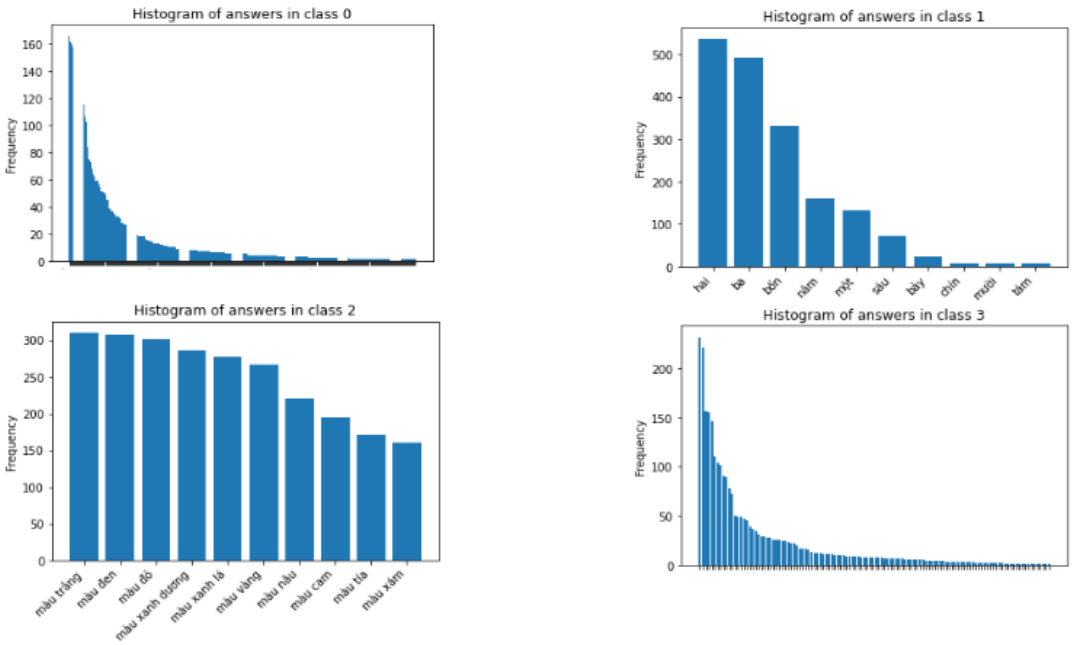


Figure 18: The distribution of each answer in each type of question.

Based on the dataset annotation rules, we can also limit the scope of the problem we are working on which is an open-end VQA problem with a fixed-size output list of words and our contribution will be to try to adapt the best models in image processing and natural language to produce a good result on this benchmark.

2. **Proposed method.**
    a. Our proposed model in Figure 19 is a novel approach to visual question answering (VQA) that combines three state-of-the-art architectures: Convolution Embedding layer in CvT, the window and shifted-window multihead attention layer in Swin Transformer, and the Bert architecture for the question encoder. This unique combination of architectures provides several advantages over traditional VQA models. We do fine-tuning pretrained weights of visual models trained on Imagenet and the whole PhoBert architecture.

- One advantage of this model is its ability to handle complex and diverse inputs. The Convolution Embedding layer in CvT is a powerful image encoder that can handle different image sizes and resolutions. This allows the model to process images of varying quality and complexity, making it more versatile and robust than other VQA models that are limited to specific image sizes or resolutions.
- The window and shifted-window multihead attention layer in Swin Transformer is another key component of the proposed model that provides several advantages. This layer is able to attend to different regions of the image at multiple scales, allowing the model to extract more detailed and fine-grained features from the image. Additionally, the shifted-window approach helps to reduce computational overhead and memory usage, making the model more efficient and scalable.
- The Bert architecture for the question encoder is another important aspect of the proposed model. By leveraging the power of pre-trained language models, the Bert architecture is able to capture the complex semantic and syntactic relationships in natural language questions. This allows the model to better understand the meaning of the question and extract the relevant information from the image.
- Another advantage of the proposed model is its ability to perform joint attention at the fusion module. This allows the model to combine the information from both the image and the question in a more effective and efficient way. By learning to distill what information is needed to answer the question, the model is able to focus on the most important and relevant features of the image and question, resulting in more accurate and precise answers.

  However, there are some drawbacks for training our model:

- While the use of pretrained models is a common practice in deep learning, fine-tuning these models on datasets with different languages and cultures can be challenging. In our case, the pretrained models were trained on the English language and culture, while the Vietnamese VQA dataset is in Vietnamese. This presents several challenges, such as the model's ability to generalize to a different language and culture, potential biases in the model's predictions, and the need for large amounts of labeled data in Vietnamese. However, we don't have that much data as doing experiments in a benchmark.

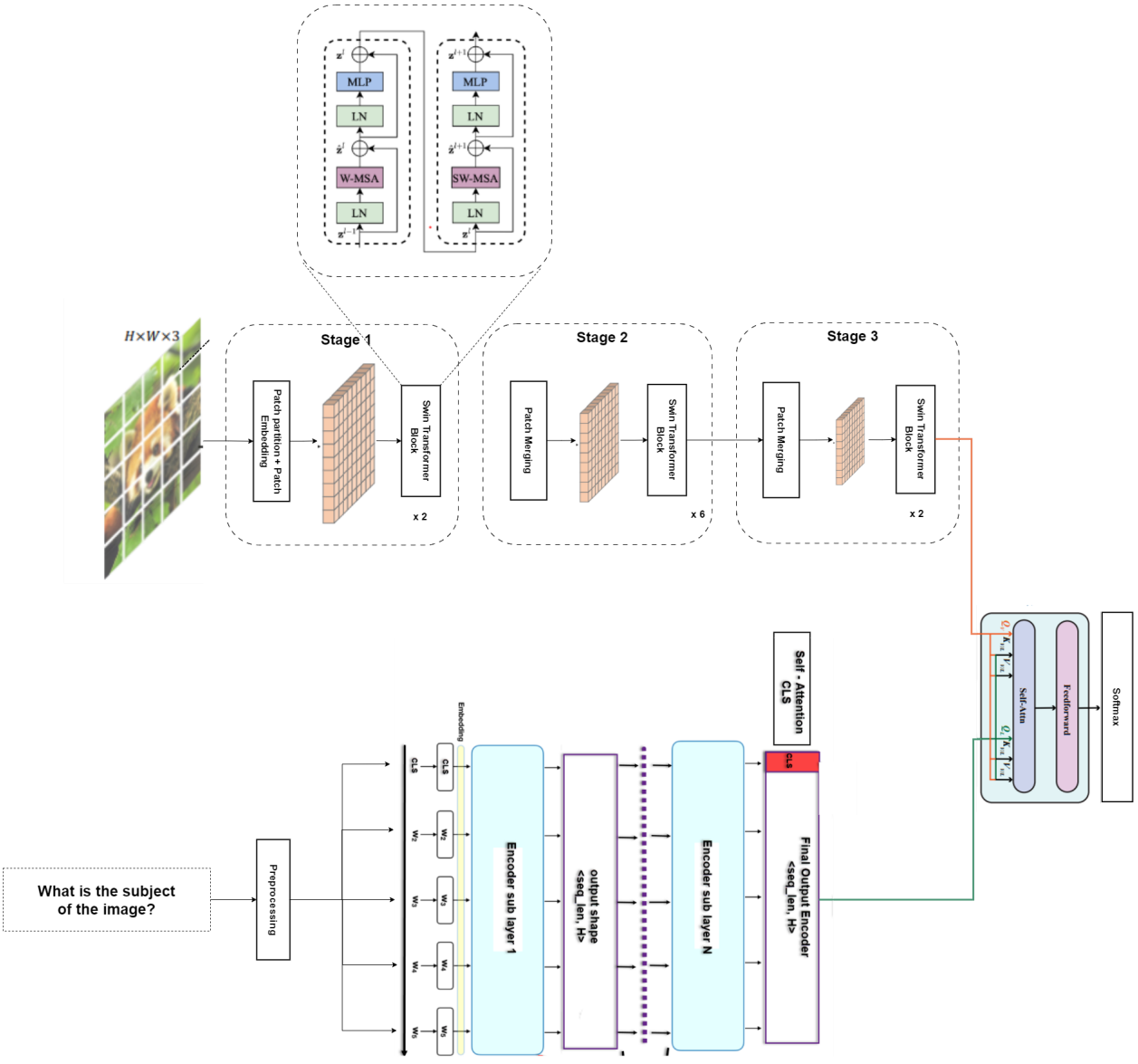


Figure 19: Our proposed model architecture for the ViVQA task.

b. We conduct our experiments with the CLIP-ViT and PhoBert model as a hybrid model that combines the Clip-Vit visual encoder with the PhoBert language model.

- The Clip-Vit visual encoder, as mentioned, is a variant of the ViT (Vision Transformer) architecture that was introduced in the CLIP (Contrastive Language-Image Pre-training) model. The Clip-Vit model is pre-trained on a large dataset of images using a self-supervised learning approach, which allows it to learn to encode visual information without relying on explicit labels or annotations. The Clip-Vit model is usually then fine-tuned on a specific task, such as visual question answering, to improve its performance.
- The PhoBert language model, on the other hand, is a variant of the RoBERTa architecture that has been specifically trained on Vietnamese text. The PhoBert model has a deep understanding of the nuances and complexities of the Vietnamese language, which makes it an ideal choice for processing Vietnamese textual information.
- In the Clip-Vit combined with PhoBert model, the Clip-Vit visual encoder and the PhoBert language model are combined in a multi-modal architecture. The model takes an image and a textual question in Vietnamese as input and outputs an answer to the question in Vietnamese as output. The image is first passed through the Clip-Vit visual encoder, which encodes the visual information into a high-dimensional feature space. The textual question is then processed by the PhoBert language model, which encodes the textual information into a contextualized embedding. The two encodings are then fused together using a cross-modal attention mechanism, which allows the model to understand the relationship between the visual and textual information. The fused encoding is then passed through a series of fully connected layers to generate the final answer to the question. The Clip-Vit + PhoBert model architecture has several advantages, including its ability to understand both visual and textual information, its deep understanding of the Vietnamese language, and its ability to generate high-quality answers to open-end Vietnamese visual question answering tasks.

Again, one significant drawback of using an English pre-trained model like Clip-Vit for non-English languages is the potential for language bias. Pre-trained models are trained on large-scale datasets, which can lead to the model learning biases and stereotypes that exist in the training data. When applied to non-English languages, these biases can manifest themselves in different ways, leading to inaccuracies and errors in the model's outputs.

- Another drawback of using an English pre-trained model for non-English languages is the potential for domain mismatch. When applied to non-English languages, the model may encounter images or concepts that it has not seen before or are not well represented in the training data, leading to reduced performance and accuracy.
- Furthermore, the use of an English pre-trained model for non-English languages can also lead to challenges in fine-tuning the model for specific tasks. Fine-tuning is the process of adapting a pre-trained model to a specific

task or domain by retraining the model on a smaller dataset. However, when using an English pre-trained model for non-English languages, the training data may not contain enough examples to effectively fine-tune the model, leading to overfitting or underfitting.

c. Combining pre-trained CNN and transformer models, such as Visual Attention Network (VAN) and PhoBERT, respectively, can also leverage the strengths of each and enable the model to better understand and reason about the world around us. This combination can be particularly useful for Vietnamese VQA, where both visual and textual understanding is required in the Vietnamese language.

- One advantage of combining VAN and PhoBERT for Vietnamese VQA is the ability to handle complex language and visual information. PhoBERT, as mentioned, is a pre-trained language model that has been specifically trained on Vietnamese text, allowing it to understand and process the nuances of the language. This is important for VQA, as the questions and answers can often contain complex language, idiomatic expressions, and cultural references.Similarly, VAN is a pretrained CNN model that has been specifically designed for image captioning tasks. By attending to different regions of an image and generating a textual description, VAN can extract visual features from an image and understand its content. This is important for VQA, as the model must be able to understand the visual content of the image and generate an appropriate answer based on that understanding. By combining VAN and PhoBERT, the model can process both visual and textual information and generate more accurate answers.
- Another advantage of combining VAN and PhoBERT for Vietnamese VQA is the ability to leverage pretraining. Both models have been pre-trained on large datasets, allowing them to learn general features that can be transferred to new tasks. This is important for VQA, as it is a task that requires a large amount of training data to achieve good results. By leveraging the pretraining of both models, the model can achieve better results with less training data, making it more efficient and practical to use in real-world applications.
- Furthermore, the combination of VAN and PhoBERT can lead to more explainable results in Vietnamese VQA. VAN's attention mechanism enables the model to attend to different regions of the image while generating a textual description, providing insight into the visual features that are important for the task. This can be particularly useful in applications where it is important to understand the reasoning behind the model's decisions. Similarly, PhoBERT's transformer architecture enables it to capture the semantic meaning of the question and generate an appropriate answer based on that understanding. This can help to ensure that the model generates answers that are consistent with the meaning of the question, leading to more accurate and explainable results.

d. The CvT + PhoBert model is also considered well-suited for Vietnamese VQA because it can effectively capture both the visual and linguistic aspects of the input data. The Convolutional Vision Transformer (CvT) can extract visual features from the input image, while the Pretrained Hierarchical Bidirectional Encoder Representations from Transformers (PhoBert) can capture the contextual and semantic information from the input text in Vietnamese language. By combining these two techniques, the CvT + PhoBert model can effectively reason about the relationship between the image and the text, and generate accurate answers to the questions.

- One advantage of using the CvT + PhoBert model for Vietnamese VQA is that it has already been pretrained on large-scale datasets in Vietnamese language, such as the VnCoreNLP corpus, which contains more than 1 million sentences. This pretrained model can be fine-tuned on smaller datasets for specific tasks, such as Vietnamese VQA, to further improve its performance. This allows for efficient use of computational resources and faster training times.
- The CvT model has also shown promising results in transfer learning, where a model is pretrained on a large dataset and then fine-tuned on a smaller task-specific dataset. The CvT model can be pretrained on large-scale datasets, such as ImageNet, and then fine-tuned on smaller datasets for specific tasks, such as object detection or image segmentation. This allows for efficient use of computational resources and faster training times, while still achieving state-of-the-art performance on a wide range of tasks. With the removal of positional embedding while training, it can operate on images of arbitrary size, showing the ability to effectively capture both local and global spatial features in an image.

V. Results.

1. Experimental results.

By defining traditional models or pretrained transformers for the Vietnam VQA task, we conduct training from scratch with traditional models and fine-tuning large pretrained models, with settings that are tweaked to fit our problem. Pytorch is our main programming language. The training process is conducted on Colab, Kaggle and GPU Server, estimated at 1 hour for 50 epochs of training. Our experimental results are shown in Table 3.

System	**Metric**		
	Accuracy	WUPS 0.9	WUPS 0.0

LSTM + W2V	0.3228	0.4132	0.7389
LSTM + FastText	0.3299	0.4182	0.7464
LSTM + EMLO	0.3154	0.4114	0.7313
LSTM + PhoW2Vec	0.3385	0.4318	0.7526
Bi-LSTM + W2V	0.3125	0.4252	0.7563
Bi-LSTM + FastText	0.3348	0.4268	0.7542
Bi-LSTM + ELMO	0.3203	0.4247	0.7586
Bi-LSTM + PhoW2Vec	0.3397	0.4215	0.7616
Co-attention + PhoW2Vec	0.3496	0.4513	0.7786
CvT + PhoBert	0.3805	0.5382	0.7943
Clip-Vit + PhoBert	0.5227	0.5641	0.8308
Pretrained CNN (Visual Attention Network) + PhoBert	0.5979	0.6157	0.8623
Swin Transformer + PhoBert	**0.6201**	**0.6814**	**0.8719**

Table 3. Experimental results on ViVQA dataset.

a. The Swin Transformer and PhoBert combined models have achieved state-of-the-art results in this task, outperforming other models in terms of accuracy, from 34.96% to **62.01%** and from 45.13 to **68.14** in WUPS 0.9 score , from 77.86 to **87.19** in WUPS 0.0 score. The model has shown its strength in encoding effectively each of two types of data.

- The key features of the Swin Transformer is the use of a hierarchical feature aggregation mechanism. The model first divides the input image into non-overlapping patches and applies a set of convolutional layers to each patch. These patches are then aggregated into a smaller set of feature maps, which are processed using a set of transformer blocks. The output of these blocks is then aggregated again to produce the final feature representation of the image. The use of hierarchical feature aggregation allows the Swin

Transformer to efficiently process large-scale images while maintaining a high level of accuracy. This is particularly useful for VQA tasks, where the model needs to attend to different parts of the image to generate a relevant answer.

- On the other hand, PhoBert is a BERT-based model that has been pre-trained on a large corpus of Vietnamese text. BERT (Bidirectional Encoder Representations from Transformers). The PhoBert model is pre-trained on a large corpus of Vietnamese text, which allows it to learn the language-specific features of Vietnamese language. The pre-training process of PhoBert involves two tasks: masked language modeling and next sentence prediction. In masked language modeling, the model is trained to predict a missing word in a sentence given the surrounding context. In next sentence prediction, the model is trained to predict whether two sentences are consecutive in a text corpus. The pre-training process allows the PhoBert model to learn the statistical patterns and relationships between the words in the Vietnamese language. This knowledge can then be transferred to downstream tasks such as VQA, where the model needs to generate a relevant answer in Vietnamese.
- The combination of the Swin Transformer and PhoBert models in the Vietnamese VQA task is particularly effective because it allows the model to leverage the strengths of both models. The Swin Transformer can efficiently process the input images and generate relevant visual features, while PhoBert can understand the questions in Vietnamese and generate relevant textual features. The combination of these two features allows the model to generate a more accurate and relevant answer to the given question.
    b. The results of the Clip-Vit + PhoBert model on the Vietnamese VQA dataset show that it performs better than the baseline model, but still has room for improvement. The model achieves an accuracy of 0.5227, a WUPS score of 0.5641 for a threshold of 0.9, and a WUPS score of 0.8308 for a threshold of 0.0. While these results are an improvement over the baseline model, they are not as high as those achieved by state-of-the-art models on other VQA datasets.
- One reason for the relatively low performance of the Clip-Vit + PhoBert model on the Vietnamese VQA dataset could be the complexity of the Vietnamese language. Vietnamese is a tonal language with a complex grammar and vocabulary, which can make it difficult for models to accurately capture the meaning of questions and generate appropriate answers. Additionally, the Vietnamese VQA dataset may not be as large or diverse as other VQA datasets, which can limit the ability of the model to generalize to new examples.
- However, the Clip-Vit + PhoBert model is a combination of a highly effective image processing model (Clip-Vit) and a state-of-the-art language model specifically designed for Vietnamese (PhoBert). This approach represents an important step forward in developing effective and accurate models for

cross-modal retrieval tasks in non-English languages. From here, it also opens up a new direction in multimodal problems for separating the training embedding problem as a contrastive learning task (foundation model) before applied to downstream ones. However, we expect that there will be more such open-source embedding for multimodal tasks in the future as we still encounter cases where reusing an individual embedding for visual features with a different encoding model leads to severe performance degradation due to data mismatch.

c. The Pretrained CNN (Visual Attention Network) + PhoBert model achieved an accuracy of 0.5979 and a WUPS score of 0.6157 for the WUPS 0.9 metric, and a WUPS score of 0.8623 for the WUPS 0.0 metric. Compared to the baseline model, Co-attention + PhoW2Vec, the Pretrained CNN (Visual Attention Network) + PhoBert model achieved significantly higher scores across all three metrics. The baseline model had an accuracy of 0.3496 and WUPS scores of 0.4513 for the WUPS 0.9 metric and 0.7786 for the WUPS 0.0 metric.

- The main advantage of the Pretrained CNN (Visual Attention Network) + PhoBert model over the baseline model is the use of a pre-trained CNN for image feature extraction and PhoBert for text processing. This combination enables the model to better understand both the image and the question, resulting in more accurate answers. The pretrained CNN is able to extract high-level visual features from the image, such as edges, shapes, and textures, which can be used to more accurately answer questions about the image. PhoBert, on the other hand, is a language model that has been specifically designed for the Vietnamese language. It has been pre-trained on a large corpus of Vietnamese text, allowing it to better understand the nuances and complexities of the language. This makes it particularly effective for answering questions in Vietnamese, as it can more accurately interpret the meaning of the question and generate a relevant answer. Because the Pretrained CNN (Visual Attention Network) is trained for Image Captioning Task, it has the advantage of a multimodal perspective model when capturing relevant local information used for language tasks, these local features are showing the ability to transfer better than global features in the transformer's model, as training transformers require a large of training data to achieve better results.

d. Although the results when training with CvT + PhoBert are not superior, the experimental results still show a fairly good generalization ability of the model. Although the accuracy is not too high compared to the baseline model (0.3805 vs 0.3496), the model achieves the same WUPS parameter as other transformers models (0.5382 for WUPS 0.9). In general, as the field of computer vision continues to evolve, models like the CvT are likely to play an increasingly important role in enabling

machines to understand and interpret visual information in a variety of contexts with adaptive size input images and the ability to learn both local and global visual features.

2. **Fail case analysis.**

Through testing and testing, we found that this benchmark data for Vietnamese VQA still has some shortcomings:

- The first issue with this Vietnamese VQA benchmark dataset is that it is not large enough for fine-tuning pretrained transformers models for English. Pretrained transformers models, such as BERT, RoBERTa, and GPT, have achieved state-of-the-art performance on a wide range of natural language processing tasks, including VQA. These models are typically pre-trained on large-scale datasets, such as Wikipedia and BookCorpus, and then fine-tuned on smaller task-specific datasets. However, the size of the Vietnamese VQA benchmark dataset is relatively small compared to other popular VQA datasets, such as VQA v2.0 and COCO-QA, which can lead to data mismatch and overfitting. Data mismatch occurs when the distribution of data in the training dataset is significantly different from the distribution of data in the fine-tuning dataset. This can lead to poor generalization performance of models, as they may not have learned to handle the specific characteristics of the new scenario
- The second issue with the Vietnamese VQA benchmark dataset is that some questions may not be very true for normal communication, as they are translated from the COCO-QA dataset. The COCO-QA dataset is a popular VQA dataset in English that contains questions and answers that are collected from real-world images. However, the translation of these questions and answers into Vietnamese may not always preserve the meaning and intent of the original questions and answers. This can distort the learning process of the model, as it may learn to associate incorrect answers with certain questions, leading to poor performance.
- In addition, there are some conundrums where the input image contains very little data for making decisions of the model. Despite the involvement of attention-based models for learning and retaining the most important information of the photo, those difficult cases need to be given more attention and appear in the training dataset so that the model can learn to distinguish and learn to make correct decisions.
- Through conducting experiments, we found that model errors often lie in the questions of categories 1 (Number) and categories 3 (Location), while we can do very well with questions of category 0 (Object) and category 2 (Color). The reason may lie in the fact that category 1 and category 3 questions often require the model to pay attention to different positions of the image, synthesize all related information to be able to make an accurate decision, this can be solved if we increase the number of heads and the number of

multihead attention layers, however, it can lead to overfitting because the training dataset is not large and comprehensive enough. Thus, what is needed now for the further development of modeling for the VQA task is to increase the number and variety of training sample questions, especially difficult questions in category 1 and category 3.

Question: có bao nhiêu kế hoạch hai chiến tranh thế giới cổ điển ngồi trên đường băng

Answer: ba

Pred_answer: ba

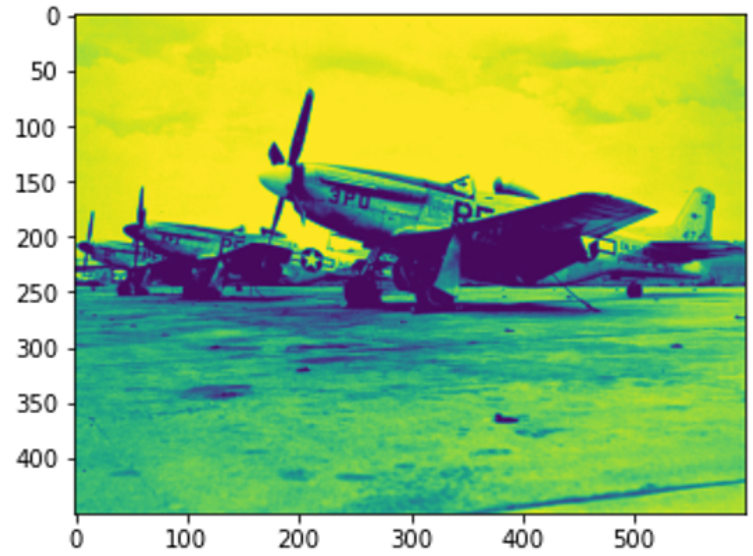


Figure 20: Translation error in dataset. However, there are cases where the model can still understand the context of the sentence and give the correct answer.

Question: người phụ nữ đang kiểm tra điện thoại di động ở đâu?

Answer: nhà vệ sinh

Pred_answer: phòng tắm

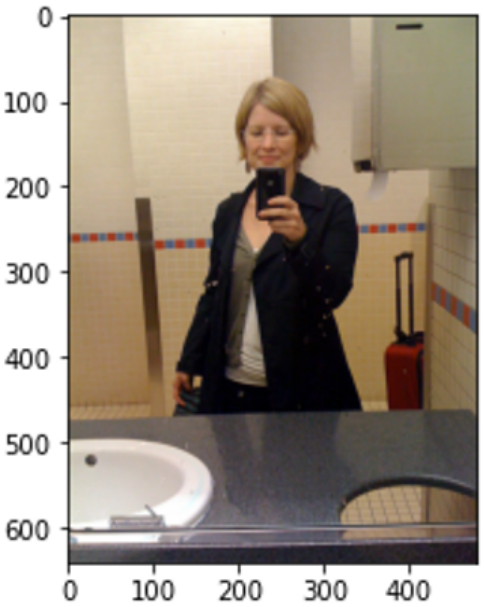


Figure 21: Failing case with conundrum. However, we can understand that these 2 results are completely close.

Question: cà rốt tươi ở đâu

Answer: phòng bếp

Pred_answer: thùng chứa

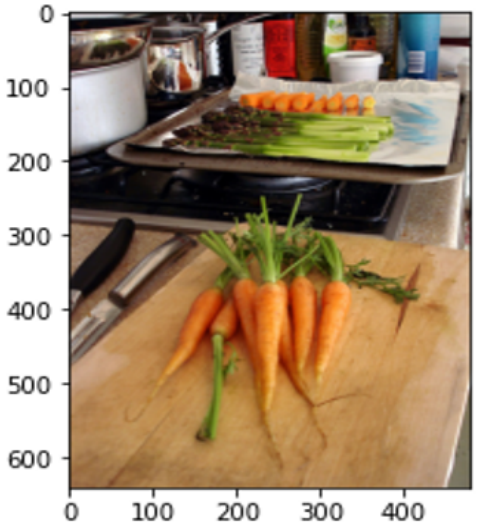


Figure 22: Failing case with conundrum. There are still carrots in the tray.

../viq_images/COCO_000000264884.jpg

You selected question: có bao nhiêu người đàn ông cầm vợt

Predict

Model predicting...

Answer: hai

Figure 23: Experimenting with our demo UI, the results showed that the ability to work on large objects, but still struggles with small objects.

../viq_images/COCO_000000219254.jpg

Enter question (Vietnamese):

cái gì trong ảnh

Predict

Answer: xe máy

Figure 24: Experimenting with our demo UI, the model gave the answer correctly.

VI. Discussion.

1. About our experimental results

In our experiment, we have used state-of-the-art models for encoding images and text in Vietnamese so far in our area of knowledge. Although the results outperform the baseline models, it will not getting more than a threshold because of the following 2 reasons:

- The first challenge is related to the size and quality of the dataset used for training VQA models. The quality of the dataset is a crucial factor in the performance of any machine learning model, including VQA models. The dataset used for training VQA models needs to be large enough to capture the variability of both images and text. This means that the dataset needs to include a diverse range of images and questions that cover various topics and scenarios. However, currently available datasets for VQA in Vietnamese are relatively small and lack diversity. This limits the ability of VQA models to learn and generalize well to new data. Moreover, the quality of the dataset is also a concern as it can have a significant impact on the performance of the model. The quality of the dataset can be affected by several factors such as incorrect labeling, missing data, and unbalanced data. This can result in biased models that are not able to perform well on new data.
- The second challenge is related to the effectiveness of the models used for encoding images and text. VQA models rely on two main components, an image encoder and a text encoder, which are used to extract features from images and text respectively. The effectiveness of these encoders plays a crucial role in the overall performance of the VQA model. The current state-of-the-art models for encoding images and text in Vietnamese are designed independently of each other, and they do not effectively learn pattern matching between the two distribution zones. This limits the ability of the model to understand the relationship between images and text, resulting in lower accuracy in answering questions. Furthermore, current models that are developed for VQA in Vietnamese are just pairing the best performing encoders in each field without being able to learn effectively pattern matching between 2 distribution zones.

Therefore, there is a need for new approaches that can effectively learn the relationship between images and text in Vietnamese. In order to address these challenges, there are several strategies that can be employed.

- First, there is a need for larger and more diverse datasets for VQA in Vietnamese. This can be achieved by collecting data from various sources, including social media platforms, news websites, and online forums. Additionally, there is a need for improved data labeling techniques to ensure the quality of the dataset. This can be achieved by employing crowd-sourcing techniques or using semi-supervised learning approaches.
- Second, there is a need for new approaches that can effectively learn the relationship between images and text in Vietnamese. This can be achieved by developing new models that are specifically designed for VQA in Vietnamese, which can effectively learn the relationship between images and text. For instance, models that integrate multi-modal features and learning-based approaches can be employed to enhance the relationship between images and text.

In conclusion, VQA in Vietnamese is a challenging task that requires the development of models that can effectively learn the relationship between images and text. The challenges of dataset size and quality, and the effectiveness of the models used for encoding images and text, are major factors affecting the performance of VQA models for Vietnamese language. However, by employing new strategies and approaches, we can overcome these challenges and improve the performance of VQA models

2. **In a bigger view.**

In recent years, the field of multimodal learning has seen tremendous advancements with the emergence of state-of-the-art models such as CLIP, DALL-E, and GPT-3. These models have demonstrated remarkable capabilities in encoding and understanding both text and images, leading to a wide range of applications, including image captioning, visual question answering, and text-to-image synthesis. However, the majority of these models have been developed and trained on English datasets, which limits their applicability in other languages such as Vietnamese.

One of the key challenges in developing multimodal models for Vietnamese is the scarcity of high-quality datasets. Unlike English, which has a vast amount of annotated data available, Vietnamese has a relatively smaller corpus, which makes it challenging to train large-scale models. Moreover, the quality of the datasets is not always up to the mark, with noise and inconsistencies affecting the overall performance of the models. This is particularly true for complex tasks such as image captioning, where a single error in the input text can significantly affect the output. Another challenge in developing multimodal models for Vietnamese is the lack of investment in research and development. While several Vietnamese language models have been developed in recent years, such as PhoBert and ViDeBERTa, most of these models have been trained on text data only, without considering the multimodal aspect. This limits their applicability in tasks that require understanding both text and images, such as image captioning and visual question answering. Furthermore, the lack of investment in research and development has also limited the availability of computational resources required to train large-scale models.

To overcome these challenges, a stronger investment in both data and computational resources is needed. One way to address the scarcity of high-quality datasets is to encourage the creation of annotated datasets through collaborations between industry and academia. This would require investments in infrastructure, tools, and resources to support the creation and curation of large-scale datasets that cover the entire variability of text and images. Additionally, the development of transfer learning techniques that can leverage pre-trained models on English datasets and fine-tune them on Vietnamese datasets could also be explored. Furthermore, research and development in multimodal learning for Vietnamese must be given more attention to unlock its full potential. This would require investments in both human resources and infrastructure to develop state-of-the-art models that can encode and understand both text and images in Vietnamese. This includes the development of transfer learning techniques that can leverage pre-trained models on text-only datasets and fine-tune them on multimodal datasets. Additionally, investments in computational resources such as high-performance computing clusters and cloud-based infrastructure could accelerate the pace of research and development in this area.

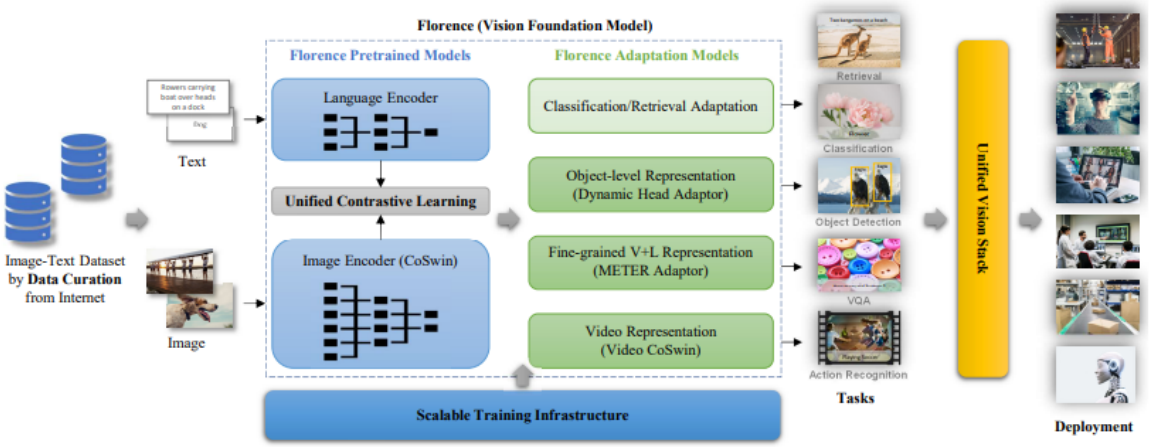


Figure 24: Training Florence using contrastive strategy, its embedding can be transferred to downstream tasks with excellent performance.

VII. Appendix

- Github: https://github.com/daocongtuyen2x/Vietnamese_Visual_Question_Answersing
- Reference:

- https://arxiv.org/abs/2202.09741
- https://arxiv.org/pdf/2111.11432
- https://arxiv.org/abs/2103.15808
- https://arxiv.org/abs/2103.14030
- https://arxiv.org/abs/2003.00744
- https://arxiv.org/abs/2103.00020
- https://arxiv.org/abs/2212.13296
- https://arxiv.org/abs/2010.11929
- https://arxiv.org/abs/1606.00061
- https://www.researchgate.net/publication/355214763_ViVQA_Vietnamese_Visual_Question_Answering