Distillation-Centric Approaches in Visual Question Answering with Mixture of Experts

Huy Hoang Huynh[1,2], Tung Le[1,2], Huy Tien Nguyen[1,2,*]

[1]Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam
[2]Vietnam National University, Ho Chi Minh city, Vietnam

Correspondence: Huy Tien Nguyen. Email: ntienhuy@fit.hcmus.edu.vn
Communication: received 30/11/2024, revised 11/03/2025, accepted 04/05/2025
DOI: 10.32913/mic-ict-research.v2025.n2.1352

***Abstract*: Recent advancements in computer vision and natural language processing were applied to the Visual Question Answering task. Nonetheless, a significant proportion of models exhibiting high accuracy possess extensive architectural components. This has a significant impact on the process of bringing the technology to practical applications such as assistive devices for the blind and visually impaired, and other related fields. Our research focuses on compressing the Visual Question Answering model on the Vietnamese dataset by utilizing the knowledge distillation method. Furthermore, in order to enhance precision, we have also developed a Mixture of ViVQA Experts system that will adapt to each type of question for improving accuracy while increasing only a few parameters and not wasting time retraining the entire system from scratch. With a total of 204M parameters, this approach has reduced the size by 24.51% compared to the original model while only reducing accuracy by 6.59% on the overall test set. More specifically, we have made accuracy improvements on each question type: "number" increased by 1.35% and "color" increased by 0.48% compared to our distillation model. The code and pretrained models are available at: https://github.com/huynhhoanghuy/Distillation_ViVQA.**

Keywords: *visual question answering, knowledge distillation, mixture of expert, vision-language.*

I. Introduction

With significant progress in vision tasks and textual tasks, recent research groups have focused more on problems that combine both fields, such as Visual Question Answering (VQA) [1], Visual Captioning [2], Visual Reasoning [3], etc. To take advantage of the techniques applied in each of these fields to vision-language tasks, most existing methods require significant adjustments. Among the tasks listed, VQA is a more challenging task, requiring not only a deep understanding of images and text but also finding the correlation between them to generate answers.

Visual question answering systems aim to answer text-based questions about images [4]. The key components of modern VQA systems include image feature extractors, text/question encoders, and classifier models that predict answers based on the encoded representations. While earlier work focused heavily on recurrent models and CNN-RNN architectures, recent progress has seen the increasing adoption of attention mechanisms and transformer-based approaches. For instance, Chao Yang *et al.* [5] proposed a dual-branch model with contextual constraints to limit bias. Hangbo Bao *et al.* [6] explored unified multimodal transformer encoders for VQA. Several works have also leveraged pretrained models, such as BEiT and BARTPho, to effectively encode visual and textual semantics.

Recent studies have presented substantial advancements in the development of large-scale models [5–10] which demonstrate notable improvements in VQA performance. As the VQA task requires a wide range of knowledge (e.g., position, color, counting etc), VQA's models necessitate a significant number of parameters to effectively encapsulate these abilities. Consequently, this requirement for extensive parameterization makes traditional model distillation approaches less effective in reducing the size of these large models without compromising their performance. In addition, the heterogeneity of question types within VQA, each necessitating distinct evaluation metrics (e.g., categorical cross entropy for object detection and mean squared error for counting tasks), makes the use of a single loss function sub-optimal in the training of VQA models. To address these challenges, we propose the utilization of a distilled model as a foundation framework for the development of a **M**ixture **o**f **V**iVQA **E**xperts (MoVE). In this work, we define the teacher model as a large-scale VQA model built with complex architectures (such as ViT and ResNet [7]) that are capable of achieving high accuracy across various types of questions. However, these models contain hundreds of millions of parameters, leading to slow inference and making deployment on resource-limited devices impracti-

cal. To address this challenge, we constructed a student model, which is a more compact version of the teacher model. The student model is trained using knowledge distillation [11–13], allowing it to inherit knowledge from the teacher by minimizing divergence between their outputs and intermediate features. This technique enables the student model to preserve much of the teacher's accuracy while significantly reducing the number of parameters and computation cost. We use this distilled student model as the foundation for developing a more efficient and adaptable architecture. This system is designed to specifically accommodate the nuanced requirements of various types of questions. According to experimental results, the accuracy of the entire system has improved by 54.37% compared to 54.17% in the student model. Here, accuracy is calculated as the ratio of the number of correctly predicted answers to the total number of questions in the test set. Additionally, the accuracy on color-related questions increased by 0.48%, and the accuracy on counting-related questions increased by 1.35%. The contributions of this work are summarized as follows:

- We introduce a novel knowledge distillation methodology which entails substituting the image feature extraction modules, specifically ViT and ResNet, with smaller variants. This strategy achieves a commendable accuracy of 54.17%, closely approaching the 60.76% accuracy of the original model while concurrently achieving 24.51% reduction in the number of parameters.
- We propose a mixture network strategy that promotes a unified architecture by merging fusion features derived from both image and question inputs. At the mixture network's top layers, we establish three classifiers catering to three distinct question types: color, number, and others. This innovation marginally enhances the system's performance and serves as a preliminary validation for the efficacy of segmenting a universal classifier into multiple specialized classifiers aligned with specific question domains.
- We present detailed comparisons and related experiments on downstream tasks, including specific question types. Experimental results show that the proposed approach significantly improves accuracy.

The structure of this paper is organized as follows: Section II presents some related research, focusing on the VQA problem, knowledge distillation, and a mixture of expert techniques. Section III delves into our methodology, describing how we distill a VQA model and apply a mixture of experts to it. Section IV provides an overview of our experiments and an analysis of our findings. Finally, Section VI summarizes our study and suggests directions for future.

II. Related Works

Recently, most VQA frameworks only focus on English, lacking data sets for other languages, especially Vietnamese. Besides, the Vietnamese Visual Question Answering (ViVQA) dataset [14] introduced by Khanh *et al.* is translated from the VQA dataset that has been used by several groups. This dataset containing 10,328 images from the MSCOCO dataset and 15,000 pairs of questions and answers in Vietnamese.

VQA model: The initial model [15] for this dataset employed recurrent neural networks and LSTM using GloVe word embeddings to compute image and question representations separately. These were forward to some textual/visual guided attention layers. After that, they are fused using point-wise addition and fed to a dense layer to projected into a vector for prediction. Later, [7] proposed a new approach, which using PhoBert for getting textual features and using both Vision Transformer (ViT) with ResNet for getting global and local visual features. The steps to combine them, are the same as the article mentioned before. Recently, there was a research group using BARTPho for text and BEIT for image [8]. Then, both types of features are concated together and passed through common Multi-head Self-Attention layers. This is the mechanism that allows the model to learn the correlation between visual and textual features. Finally, the model will predict the answer based on the fusion features that have been fed through the Feedforward Neural Network (FNN).

Model compression: While delivering superior results, modern deep neural networks present deployment challenges for resource-constrained platforms for deployment on devices such as mobile, jetson device, etc. due to their sheer complexity and massive computational demands. The larger the model, the longer the inference time increases, leading to a worse user experience. On the other hand, we see that recently, the field of model compression has gradually become popular, such as pruning, quantization, knowledge distillation, and so on. With pruning, pruning branches that have little influence in the model will help the compression ratio be 10 to 15 times, but there is a major weakness that is that with models with little sparseness, it is quite difficult to prune well [11]. With quantization, converting data types to a less computational and storage form and grouping calculations into clusters helps speed up execution and reduce the size of the file. However, this method still has problems when being implemented with sparse models [11]. On the other hand, the knowledge distillation method does not encounter that limitation and also has no limitation on compression ratio. In this approach, we begin with a teacher model, which is a large, high-capacity model designed with deep architectures

such as ResNet and ViT layers to handle complex tasks with strong accuracy. The teacher model typically contains hundreds of millions of parameters and is capable of learning rich multi-modal relationships. In contrast, the student model is a smaller, more lightweight version, designed with fewer parameters and simpler modules - for example, replacing large backbone networks with smaller ones like MobileNetV3. Despite its compact architecture, the student model is trained to learn from the teacher's knowledge. It simply creates a new student model which is smaller and more compact than the original teacher model [11]. Then, we use constraints between features at different levels, corresponding to each pair of layers, to transfer the knowledge that the teacher model has learned into the student model. Such as Humair Raj Khan *et al.* [12] proposed a method of knowledge distillation from the teacher monolingual VQA model to the student multilingual VQA model. They have used multiple objectives to transfer the knowledge: token embedding distillation; object attention distillation; prediction distillation. Jianfeng Wang *et al.* [13] also distill knowledge from the output and the intermediate layers on medical VQA datasets.

Knowledge Distillation: has been applied to the task of model compression. By using "knowledge" transfer technical from teacher model to student model, with the expectation that only a small student model can have the same knowledge as the large teacher model after going through the layers. Specifically, with a given sample (x_i, y_i), there is knowledge $f^T(x_i)$ in the teacher network (T) and knowledge $f^S(x_i)$ in the student network (S). In addition, the student model takes an input x_i and produces an output representation $S(x_i)$, which is trained to approximate the output of the teacher model $T(x_i)$ through the knowledge distillation process. To transfer knowledge, we should minimize the divergence between them:

$$Loss = L_S(S(x_i), y_i) + L_{KD}(f^T(x_i), f^S(x_i)) \quad (1)$$

, where $L_S(.)$ refers to the original supervised loss on the Student and label of sample. $L_{KD}(.)$ refers to a constraint that the knowledge of teacher and student should be the same. In particular, L_{KD} can be cross-entropy (CE), mean squared error (MSE), mean absolute error (MAE) function or KLdivergence (KL), ... For example, DistillBERT [16] proposed to train the small BERT model by imitating the teacher model's output probability distribution for the masked language prediction task, as well as mimicking the embedding features from the teacher. MobileBERT [17] and TinyBERT [18] distilled by utilizing the layer-wise attention distributions for comparing to the teacher model's attention distributions using MSE loss like as L_{KD}, and so on.

Mixture of experts: (MoE) [19] is an adaptive neural network architecture that allows combining multiple models (experts) to solve a problem. Each expert specializes in a certain area of the input space. Its architecture includes expert networks, a gating network to decide the combined weights of experts based on input, and a module that combines expert results according to weights [20, 21]. The strength of this model is its flexibility, which can be expanded by adding new experts without affecting old experts. MoE has been successfully applied in many fields, such as computer vision, natural language processing, speech recognition, etc. More specifically, there are a few studies on the VQA problem that have used MoE. For example, V Pahuja, et al. [22] applied MoE to optimize the CNN network using conditional computation, only partially activating the modules based on the question representation vector; G Liu, et al. [23] has proposed combining self-attention and cross-attention in cross-modal mixture experts (CMMEs) models, including visual experts and textual experts, to increase the interactivity of visual and linguistic features. In this study, we do not use MoE to optimize calculations or to increase the association between two types of features. We realize that for each type of question, there will be a separate domain. Therefore, it is necessary to have separate classifiers and a controller for navigation.

III. METHODOLOGY

In this section, we will first introduce our Mixture of ViVQA Experts system (MoVE). Given an image I and a question Q as input, this system will try to give the correct answer by classifying a list of candidate answers. The overview of our system is illustrated in Figure 1, the model consists of 3 main modules: the visual-textual representation extractor; the gating module for navigating input into expert; the classifier experts. This study will present a new approach for optimizing the Vietnamese VQA model on the Vietnamese Visual Question Answering dataset.

For the distillation phase, the NDA model [7] was designated as the teacher model, from which the distillation process commenced towards a more compact student model. This process entailed the substitution of two image extractor modules, namely ResNet and ViT, with their smaller counterparts. This strategic adjustment underscores our commitment to optimizing model's efficiency without substantially compromising its performance capabilities. This particular aspect will be elaborated upon in the specific context of Section III.1.

As mentioned above, teacher models employ architectures with many parameters to deal with various domains of questions, while small student models have the limited

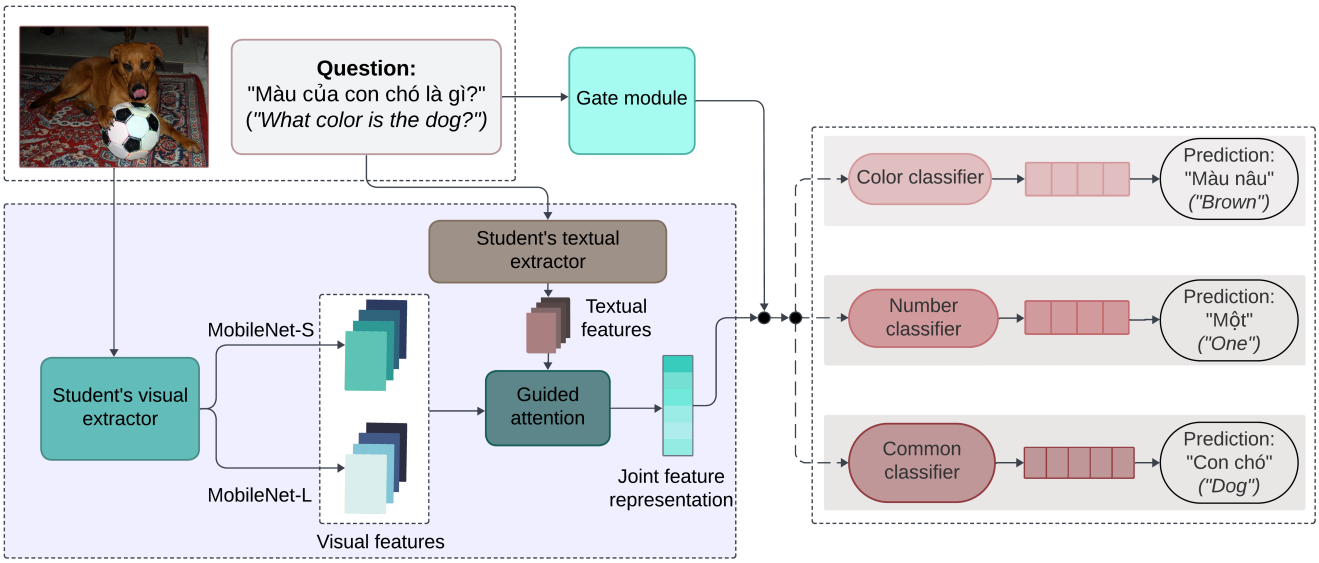

Figure 1. Our Mixture of ViVQA Experts system (MoVE). MobileNet-S and MobileNet-L stand for the MobileNetV3 model in small and large sizes, respectively. The content in parentheses is the English translation.

number of parameters. To overcome this challenge, we design separate top layers for modeling different types of tasks and mixing them together in one system to save parameters while maintaining good performance. In recent the Mixture-of-Expert paradigms [6, 19, 24], these methods all use a certain input signal to pass through a baseline model to predict which experts will receive the signal. To avoid increasing the number of parameters and computations by a large amount, we propose MoVE, an efficient rule-based pipeline for navigating inputs into corresponding experts. We analysed results and decided to build a mixture network with multiple classifiers corresponding to three specialized question types: *color*, *number*, and *others*. The *others* category includes the *object* and *location* questions. Section III.2 will provide an in-depth exploration of this particular detail.

1. Distill visual features

Given local visual features $f_{v_1} \in \mathbb{R}^{m_1 \times n_1}$, global visual features $f_{v_2} \in \mathbb{R}^{m_2 \times n_2}$, and textual features $f_t \in \mathbb{R}^{max\_len \times n_2}$, the original model [7] can combine f_{v_1} and f_{v_2} with f_t into fusion features $f_{v_1 \leftarrow t}, f_{v_2 \leftarrow t} \in \mathbb{R}^{1 \times n_2}$ by guided attention mechanism followed up textual features [7, 25] and forwarding to some reducing layers. Then the authors combined them into multi-context visual features $f_{joint\_visual} \in \mathbb{R}^{1 \times (n_1 + n_2)}$. They apply guided attention followed up f_{joint_visual} on f_t to get better textual features $f_{t \leftarrow v_j} \in \mathbb{R}^{1 \times n_2}$. Finally, they fusion $f_{t \leftarrow v_j}$ and f_{joint_visual} into joint features $f_{joint} \in \mathbb{R}^{1 \times d}$ and forward into classifier.

We assume there are other visual features of different smaller dimensions that can fuse with textual features. Our target is finding 2 smaller visual feature extraction modules and fuse that features with textual features without significant reduction in accuracy. On the other hand, we found a good candidate model: MobileNetV3 [26] which is both so light and is applied in parallel to the Resnet layers. Then we use 2 models: Large MobileNetV3 and Small MobileNetV3 for replacing ViT and ResNet-34 extractors (Fig. 2) and we call the new model is student model, which will learn knowledge from teacher model (the original model).

a) Problem formulation for distillation

Given an image I and a question Q as input, the target is finding the anwser A from a set of candidate anwsers O. The objective function is:

$$A = \arg\max_{A \in O} P(A|I, Q, O). \tag{2}$$

With traditional loss function for classification task, it only requires a constraint between prediction A and ground truth *label*:

$$Loss = L_S(A, label). \tag{3}$$

However, the student model needs additional knowledge from the teacher model to improve its precision. So that, the general loss function should be:

$$Loss = L_S(A, label) + L_{KD}(f^T(I, Q), f^S(I, Q)). \tag{4}$$

b) Our distillation method

With $L_S(.)$, we use cross-entropy loss (CE) as a traditional option.

$$L_S(A, label) = CE(A, label). \tag{5}$$

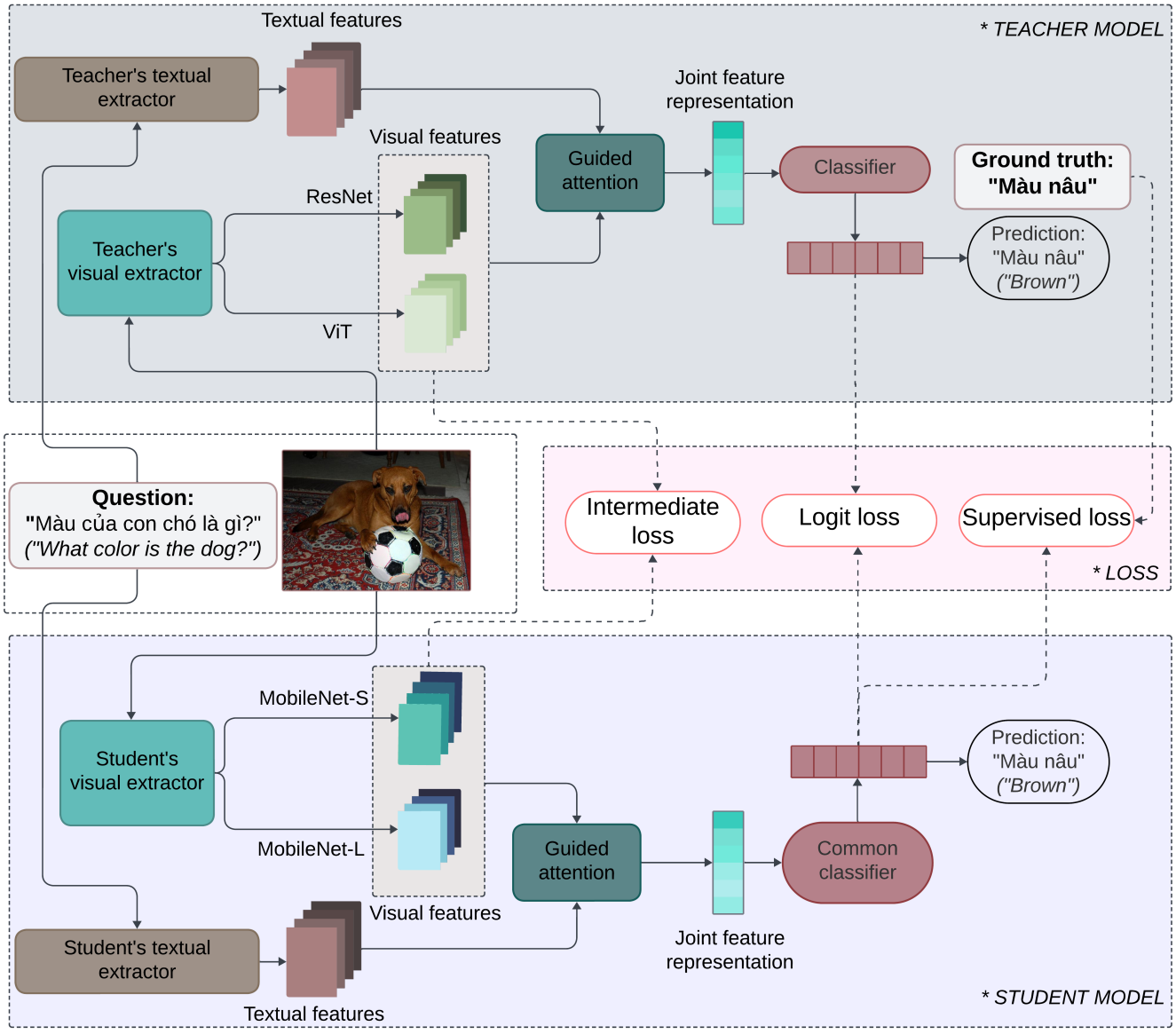

Figure 2. Illustration of the VQA model knowledge distillation flow. The content in parentheses is the English translation.

Beside that, our $L_{KD}(.)$ formulation is quite complex. We have tried to tie in student knowledge and teacher knowledge by making more constraints in intermediate features and last features (logit outputs). For intermediate features, the teacher model contains two layers: one before applying the attention mechanism and one after. In this study, we use the features at the beginning (before applying attention) to constrain the training process. Functionally, we use Kullback-Leibler divergence (KL) as the loss function for these two pairs of intermediate features. In detail:

$$\begin{aligned} L_{KD}(f^T(I,Q), f^S(I,Q)) &= L_{logit}(.) + L_{intermediate}(.) \\ &= CE(S(I,Q), T(I,Q)) \\ &+ KL(f^T_{v_1}, f^S_{v_1}) + KL(f^T_{v_2}, f^S_{v_2}) \end{aligned} \tag{6}$$

Each component of the loss function also has an associated weight factor. In addition, we are also using augmentation methods and temperature for the teacher's output and student's output to improve training progress. The formula 7 is our final loss function:

$$Loss = \alpha \cdot CE_{sup} + \beta \cdot \tau^2 \cdot CE_{logit} + \gamma \cdot KL_{f_1} + \delta \cdot KL_{f_2}. \tag{7}$$

where:

- CE_{sup}, CE_{logit}, KL_{f_1}, and KL_{f_2} is shortened form of $CE(A, label)$, $CE(S(I,Q), T(I,Q))$, $KL(f^T_{v_1}, f^S_{v_1})$, and $KL(f^T_{v_2}, f^S_{v_2})$.

- α, β, γ, and δ are the coefficients of CE_{sup}, CE_{logit}, KL_{f_1}, and KL_{f_2}.
- τ is temperature for distillation.

2. Adapting models to question categories through mixture of experts

In the dataset under investigation, we identify four distinct categories of questions: "what", "where", "number", and "color", each characterized by its own domain of values. Notably, the categories "number" and "color" exhibit constrained value ranges and possess inherent ordinal relationships. For instance, within the "number" category, there exists a sequential correlation such that the prediction of a specific value (e.g., "seven") implies elevated likelihood estimations for adjacent numerical classes and diminished probabilities for numerically distant classes. Conversely, while the "color" category also demonstrates ordinality, particularly when analyzed in the Hue, Saturation, Value (HSV) color space, it lacks a linear, countable progression.

Acknowledging these nuances, we advocate for a tailored approach in which distinct objective functions are designated for each question type. Consequently, our proposed system incorporates three specialized classifiers: a color classifier with ten classes, a number classifier also with ten classes, and a general classifier encompassing 353 classes. This methodology aims to optimize the accuracy and efficacy of the system by acknowledging and adapting to the specific characteristics of each question type.

We built a gate module, shown in Fig. 3, that acts as a bridge between input features and a classifier. This module receives a raw question as input. It will detect important keywords to identify what type of question the question is. For misleading keywords, we will consider a few other keywords to determine more precisely. In cases where there are no keywords, the question will be classified as a "what" question. In addition, we will train the classifier module's color and number for each data type while keeping the weights of the previous feature extractors intact because our system (MoVE) uses the same visual-textual feature representation extractor. This saves a large amount of computations for the system.

Regarding the question about "number", there is an ordering relationship between classes. For the "color" classifier training process, instead of applying cross entropy (CE) as a loss function, we expect squared Earth Mover's Distance (squared EMD) as mentioned by Le Hou et al. [27], will able to force classes that are far away from candidate classes to have lower prediction values than nearby classes. For the question about "color", we arrange the color classes by HSV, with H first, V next, and S last, in ascending order. To speed up the training process, we

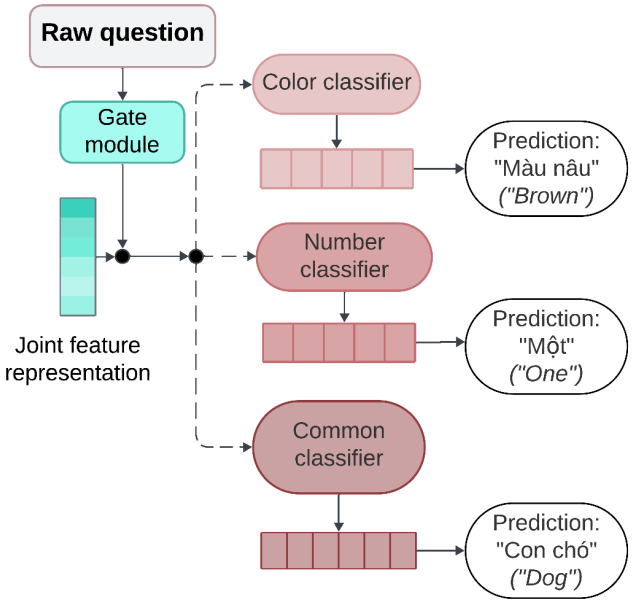

Figure 3. The application of a mixture of experts in the VQA system: Gate module with multiple classifier. The content in parentheses is the English translation.

use both squared EMD and CE because the relationships between color classes are less obvious than between number classes. Figure 1 illustrates our VQA system using a mixture of experts.

$$EMD^2 = \sum_{i=1}^{C} (CDF_i(p) - CDF_i(t))^2 \tag{8}$$

where:

- p and t are probability vector that sums to 1 of prediction and target.
- $CDF_i(.)$ is the i-th element of the $CDF(.)$ which is a function that returns the cumulative density function of its input.

In terms of architecture, we use a multi-layer perceptron (MLP) as specific classifier to encode the visual-textual features. Equation 9 and Equation 10 represent the color classifier and the number classifier, respectively.

$$\begin{aligned} h_1 &= dropout(tanh(x)) \\ h_2 &= dropout(tanh(W_2 h_1 + b_2)) \\ h_3 &= dropout(tanh(W_3 h_2 + b_3)) \\ p &= softmax(W_4 h_3 + b_4) \end{aligned} \tag{9}$$

where W_2, b_2, W_3, b_3, W_4, b_4 are the weight parameters. x is visual-textual features. p is the probability of the final answer.

$$h_1 = dropout(LeakyReLU(W_1 x + b_1))$$
$$p = softmax(W_2 h_1 + b_2) \quad (10)$$

where W_1, b_1, W_2, b_2 are the weight parameters. x is visual-textual features. p is the probability of the final answer.

IV. Experiments

In this section, first, we will demonstrate that our distillation method outperform on datasets with some strong constraints: intermediate features pairs, logit features pairs, and supervised loss between student output and label. Second, we will show our improvement in a mixture system constructed by that distilling model.

1. Experiment settings

We conduct experiments on Pytorch with the widely used Transformers library. All experiments are running on two GeForce RTX 3080 GPUs. The Gaussian noise (kernel size = (5, 9), sigma = (0.1, 5)) and random rotation (angle = 15°) are adopted as data augmentation. Regarding data preprocessing, we follow the same process as training the teacher model [7]. Table I shows the best hyperparameters we found. In addition, for knowledge distillation processing, we have decided to choose Eq. 11 is a final loss function :

$$Loss = 0.1 \cdot CE_{sup} + 0.3 \cdot 5^2 \cdot CE_{logit} + 0.3 \cdot KL_{f_1} + 0.3 \cdot KL_{f_2} \quad (11)$$

Table I
Hyperarameters settings.

Hyperparameters	Value
Temperature (τ)	5
Drop out	0.15
Learning rate	5e-4
Batch size	16

Furthermore, the number of parameters and the accuracy of the teacher model are 269,223,078 and 60.76%, respectively. This is considered a milestone to compare with the student model.

2. Experimental results

a) Distillation results

1) *Determine the feasibility of knowledge distillation:* Before focusing on this research, we need to check the feasibility of distillation. We assume that if the compact student model is capable of learning knowledge from the complex teacher model, then it is at least capable of learning knowledge when both models have the same architecture. Our loss function is constructed using CE_{sup} and KL_{logit}, with a learning rate of 10^{-4} and a dropout rate of 0.3. This configuration gave approximately the same results as the teacher achieved. Specifically, we achieve an accuracy of 59.66%, compared to 60.76% of the original model. The accuracy is calculated as the percentage of correctly predicted answers out of the total number of questions in the test set. This result has reinforced our confidence in the distillation pipeline and the performance of the network output so that we can continue to develop the following parts.
2) *Determine which module to distill:* We sequentially explored distillation processes for the ResNet branch alone and subsequently for both the ResNet and ViT branches. Based on the results Table II, the distillation of a singular branch resulted in a model size reduction by only 9.47%, and a decrease in the number of parameters by 7.5%. Conversely, the distillation of two branches significantly enhanced the model's compactness, achieving 25.71% reduction in size, and 24.51% decrease in parameter count. However, such substantial reductions in parameter volume were accompanied by a decline in model accuracy. As the relatively minor reductions of the single-branch distillation process did not meet our expectations for model efficiency improvements, we elected to proceed with the dual-branch distillation strategy. Concurrently, we initiated efforts to identify and implement measures aimed at mitigating the adverse impact on accuracy, thereby balancing the trade-off between model compactness and performance efficacy.

Table II
Experiments of distillation on 1 and both 2 visual branches of the teacher model. Green arrows (↑/↓) signify positive trends of the evaluation metrics, while red arrows (↑/↓) signify negative trends.

Distillation	Gain
ResNet → MobileNetv3 (small)	Size: 995.8 MB (9.47%↓) Num of params: 248,859,142 (7.5%↓) Best training acc: 98.3 % (2.18%↑) Best testing acc: 58.5 % (3.72%↓)
ResNet → MobileNetv3 (small), ViT → MobileNetv3 (large)	Size: 816.0 MB (25.82%↓) Num of params: 203,945,157 (24.51%↓) Best training acc: 98.3 % (2.18%↑) Best testing acc: 38.1 % (37.29%↓)

3) *Determine which knowledge constraints are feasible:* As mentioned, in the distillation process, in addition to two objective function: the alignments of the student model's predictions with the teacher model's predictions and with the true labels.Our investigation

reveals that the incorporation of additional constraints at intermediate stages of the model could further enhance the efficacy of the distillation process. In the original model, there are two main intermediate features: before applying the guided attention method and after applying the guided attention method. According to Table III, utilizing features prior to the application of attention mechanisms offers substantial benefits within the context of model distillation.

Table III
EXPERIMENTS OF DISTILLATION USING INTERMEDIATE FEATURES BEFORE/AFTER APPLYING GUIDED ATTENTION.

Loss function	*Before*	*After*
$0.90 \cdot KL_{logit} + 0.10 \cdot CE_{sup} + 0.00 \cdot KL_{f_1} + 0.00 \cdot KL_{f_2}$	46.35	43.72
$0.30 \cdot KL_{logit} + 0.10 \cdot CE_{sup} + 0.30 \cdot KL_{f_1} + 0.30 \cdot KL_{f_2}$	**49.78**	**45.32**
$0.36 \cdot KL_{logit} + 0.10 \cdot CE_{sup} + 0.36 \cdot KL_{f_1} + 0.18 \cdot KL_{f_2}$	47.59	43.87
$0.18 \cdot KL_{logit} + 0.10 \cdot CE_{sup} + 0.36 \cdot KL_{f_1} + 0.36 \cdot KL_{f_2}$	49.77	43.55

In our investigation, we conduct a comprehensive evaluation of several loss functions, including Cross-Entropy (CE), Kullback-Leibler (KL) divergence, Mean Squared Error (MSE), and Mean Absolute Error (MAE), to ascertain the optimal configuration for our model. Furthermore, we investigated the coefficients of these components to determine their respective impacts on model performance. As a default, we use supervised loss as the CE and only change the remaining loss functions. Finally, with the config mentioned in Section IV.1, our best accuracy is 54.17%. Specifically, accuracy for "color" question is 53.12%, for "number" question is 42.79%, for "where" question is 44.96%, and for "what" question is 63.80%.

Table IV
THE COMPARISON BETWEEN THE MOVE AND THE STUDENT MODEL ON THE NUMBER QUESTION.

Class	Student (%)	MoVE (%)	Class Distribution (%)
One	23.33	**26.67**	6.76
Two	**53.73**	52.24	30.18
Three	**45.76**	44.07	26.58
Four	43.75	**51.04**	21.62
Five	30.30	30.30	7.43
Six	**25.00**	18.75	3.60
Seven	0.00	**40.00**	1.13
Eight	16.67	16.67	1.35
Nine	0.00	**50.00**	0.45
Ten	0.00	0.00	0.90
Total	42.79	**44.14**	100

b) Mixture of experts results

We implement a rule-based gate module and three classifiers, each tailored to a specific category of question types: (1) number, (2) color, and (3) a combined category for "what" and "where" questions. This architectural configuration is designed to optimize the model's performance across the diverse spectrum of question types encountered within the dataset.

In the ViVQA dataset, the rule-based gate classifier reaches 99.25%. Employing a rule-based module ensures both high accuracy and rapid execution speeds, thereby reducing the necessity for deploying deep learning models in the gate module.

Regarding accuracy improvements, the number classifier's accuracy increased by 1.35%, and the color classifier's accuracy increased by 0.48%. The entire system reached 54.37% on the whole dataset, while the total number of parameters increased by only 3.1% compared to the student model.

We conducted a comparative analysis between the student model and the MoVE system, focusing on two specific domains:

- **Number Questions**: As evidenced in Table IV, the MoVE system demonstrated capability in resolving certain questions labeled as "seven" and "nine", which were previously unsolvable by the student model. However, a notable decrement in MoVE's performance is observed within some classes, which constitute a significant portion of the dataset for number-related questions. This discrepancy has led to a marginal improvement in the overall accuracy of the system. Upon examining the distribution of classes, it is observed that the MoVE model achieves substantial advancements in performance for classes characterized by a lower frequency of samples.
- **Color Questions**: According to Table V, MoVE's performance was found to be inferior to that of the student model in merely three out of ten classes ("black", "purple", and "green"). The performance in the remaining classes exhibited a consistent improvement, barring the unchanged outcomes for "blue" and "white." Despite these gains, the improved classes do not represent a substantial fraction of the color-question dataset, resulting in only a slight enhancement in accuracy. Among the eight classes demonstrating varied performance levels, the MoVE model outperformed the student model in five classes, as opposed to the three classes where the student model excelled.

These findings provide a foundational basis for directing future efforts towards refining the MoVE system, with an aim to address the identified limitations and bolster the accuracy of the model across both the number and color question domains.

Table V
THE COMPARISON BETWEEN OUR MOVE AND THE STUDENT MODEL ON THE COLOR QUESTION.

Class	Student (%)	MoVE (%)	Class Distribution (%)
Orange	38.10	**40.48**	6.72
Black	**37.50**	32.14	8.96
Red	65.33	**68.00**	12.00
Brown	65.28	**66.67**	11.52
Purple	**58.54**	56.10	6.56
White	47.06	47.06	13.60
Yellow	59.42	**62.32**	11.04
Gray	31.43	**40.00**	5.60
Blue	52.70	52.70	11.84
Green	**57.89**	55.26	12.16
Total	53.12	**53.60**	100

3. Overall Results

Based on Table VI, training the student model with knowledge distillation improved accuracy by 7.07% compared to the student model without distillation. Furthermore, the student model with knowledge distillation reduced nearly 65 million parameters while only decreasing accuracy by 6.59% when compared to the teacher model.

These two findings demonstrate the successful application of knowledge distillation techniques to the VQA task. Moreover, implementing the mixture of experts approach helped improve the student model's predictive capabilities while only marginally increasing the number of parameters (approximately 2.9% increase compared to the student model). Further details on this improvement are provided in Section IV.2.b. However, since MoVE's improvements were primarily concentrated in classes with low distribution in the dataset, the overall results showed only modest gains.

Table VI
OVERALL RESULTS BETWEEN MOVE MODEL, STUDENT MODEL FROM SCRATCH, STUDENT MODEL WITH KNOWLEDGE DISTILLATION AND ORIGINAL MODEL.

Model	No. Parameters (M)	Accuracy (%)
Teacher model	~269	60.76
Student model (w/o knowledge distillation)	~204	47.10
Student model (w/ knowledge distillation)	~204	54.17
MoVE model	~210	54.37

Although the teacher model remains the most accurate, with an accuracy of 60.76%, its large size and resource demands make deployment challenging. Our distilled student model, and subsequently the MoVE system, aim to provide a practical trade-off between accuracy and efficiency by reducing model size. Moreover, the mixture-of-experts system further improves accuracy on certain challenging question categories with minimal additional parameters.

V. DISCUSSION

We propose a method called Distilling Step-by-Step with Multiple Rationales for extracting rationales from large language models (LLMs) to provide informative supervision in training small task-specific models. Our results demonstrate that this approach outperforms both traditional LLMs and other methods that use only rationales. Moreover, it reduces the need for extensive training datasets by replacing human-generated rationales. Distilling Step-by-Step proposes a resource-efficient training-to-deployment paradigm compared to existing methods. Further studies affirm the generalizability and the strategic design choices of our method. Below, we discuss the limitations, future directions, and ethical considerations of our work.

Our approach has several limitations. Firstly, it requires the use of the GPT-3.5 API, which is not only limited but also incurs costs. Secondly, although rationales significantly enhance the performance of our task-specific models, the evaluation of a rationale's quality whether it is effective or not remains unclear. Finally, despite the success of using LLM rationales, there is evidence that LLMs exhibit limited reasoning capabilities in more complex reasoning and planning tasks [28].

Looking ahead, there are several directions for future work with our method. First, we could explore the use of other LLMs, such as GPT-4.0 or Gemini, to potentially improve the quality of rationales, anticipating that newer models might offer enhanced capabilities. Second, we aim to investigate how the quality of rationales impacts the performance of task-specific models. Finally, while we have observed success in using LLM rationales for textual entailment tasks, we plan to extend this method to tasks like image description to enhance both the classification and description of images using similar strategies.

It is important to acknowledge that the behavior of our smaller downstream models may inherit biases from the larger teacher LLMs. We anticipate that advancements in reducing antisocial behaviors in LLMs could be adapted to enhance the ethical performance of smaller language models.

VI. CONCLUSION

In this study, we introduce a Mixture of ViVQA Experts system (MoVE) that takes in visual-textual representations from image-question pairs and routes them to specialized expert classifiers to resolve three primary question types that are frequently encountered in visual question answering datasets - namely $color$, $number$, and $other$ questions. The gating module, which performs dynamic routing between expert modules, helps our overall VQA system better adapt

to distinct data domains posed by different question types, thereby improving answer prediction accuracy compared to single module baselines.

Based on the findings of this work, we envision several promising future research directions to further advance the Mixture of ViVQA Experts system. One clear extension is to develop more fine-grained expert modules for additional question types beyond *color* and *number* that capture niche domains requiring specialized reasoning. Another useful improvement is to apply knowledge distillation to the textual branch feature extractors in a teacher-student training framework in order to reduce model size and improve inference speed while retaining accuracy gains. This approach has the potential to facilitate the porting of multi-expert systems to platforms with limited resources. Finally, expanding the ViVQA mixture-of-experts formulation proposed here to multi-lingual settings can allow leveraging non-English training data. By developing question-type prediction modules for other languages and aligning polyglot data across mixture components, such systems can answer visual questions posed in multiple global dialects. We are of the opinion that advancing research in these directions can lead to a broader and more practical viability for VQA systems.

REFERENCES

[1] A. C. A. M. de Faria, F. d. C. Bastos, J. V. N. A. da Silva, V. L. Fabris, V. d. S. Uchoa, D. G. d. A. Neto, and C. F. G. d. Santos, "Visual question answering: A survey on techniques and common trends in recent literature," *arXiv preprint arXiv:2305.11033*, 2023.

[2] B. Xin, N. Xu, Y. Zhai, T. Zhang, Z. Lu, J. Liu, W. Nie, X. Li, and A.-A. Liu, "A comprehensive survey on deep-learning-based visual captioning," *Multimedia Systems*, vol. 29, no. 6, pp. 3781–3804, Dec 2023.

[3] q. yue, R. Feng, F. Matsuzawa, T. Arai, and K. Iwata, "A survey of visual reasoning," 2023.

[4] N. S. Nguyen, V. S. Nguyen, and T. Le, "Advancing vietnamese visual question answering with transformer and convolutional integration," *Computers and Electrical Engineering*, vol. 119, p. 109474, 2024.

[5] C. Yang, S. Feng, D. Li, H. Shen, G. Wang, and B. Jiang, "Learning content and context with language bias for visual question answering," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[6] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, "VLMo: Unified vision-language pre-training with mixture-of-modality-experts," in *Advances in Neural Information Processing Systems*, 2022.

[7] A. D. Nguyen, T. Le, and H. T. Nguyen, "Combining multi-vision embedding in contextual attention for vietnamese visual question answering," in *Image and Video Technology*. Cham: Springer International Publishing, 2023, pp. 172–185.

[8] K. V. Tran, K. Van Nguyen, and N. L. T. Nguyen, "Bartphobeit: Pre-trained sequence-to-sequence and image transformers models for vietnamese visual question answering," in *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE, 2023, pp. 1–6.

[9] V. Joshi, P. Mitra, and S. Bose, "Multi-modal multi-head self-attention for medical vqa," *Multimedia Tools and Applications*, Oct 2023.

[10] A. A. Yusuf, C. Feng, X. Mao, R. Ally Duma, M. S. Abood, and A. H. A. Chukkol, "Graph neural networks for visual question answering: a systematic review," *Multimedia Tools and Applications*, Nov 2023.

[11] A. Berthelier, T. Chateau, S. Duffner, C. Garcia, and C. Blanc, "Deep model compression and architecture optimization for embedded systems: A survey," *Journal of Signal Processing Systems*, vol. 93, no. 8, pp. 863–878, Aug 2021.

[12] H. Raj Khan, D. Gupta, and A. Ekbal, "Towards developing a multilingual and code-mixed visual question answering system by knowledge distillation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1753–1767.

[13] J. Wang, S. Huang, H. Du, Y. Qin, H. Wang, and W. Zhang, "Mhkd-mvqa: Multimodal hierarchical knowledge distillation for medical visual question answering," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 567–574.

[14] K. Q. Tran, A. T. Nguyen, A. T.-H. Le, and K. V. Nguyen, "ViVQA: Vietnamese visual question answering," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Shanghai, China: Association for Computational Lingustics, 11 2021, pp. 683–691.

[15] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6281–6290.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[17] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: a compact task-agnostic BERT for resource-limited devices," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2158–2170.

[18] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4163–4174.

[19] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[20] M. Nguyen, H. Abbass, and R. McKay, "A novel mixture of experts model based on cooperative coevolution," *Neurocomputing*, vol. 70, pp. 155–163, 12 2006.

[21] R. Ebrahimpour, E. Kabir, H. Esteky, and M. R. Yousefi, "View-independent face recognition with mixture of experts," *Neurocomputing*, vol. 71, no. 4, pp. 1103–1107, 2008, neural Networks: Algorithms and Applications 50 Years of Artificial Intelligence: a Neuronal Approach.

[22] V. Pahuja, J. Fu, and C. J. Pal, "Learning sparse mixture of experts for visual question answering," *arXiv preprint arXiv:1909.09192*, 2019.

[23] G. Liu, J. He, P. Li, S. Zhong, H. Li, and G. He, "Unified transformer with cross-modal mixture experts for remote-sensing visual question answering," *Remote Sensing*, vol. 15, no. 19, 2023.

[24] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers:

scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 1, jan 2022.

[25] J. Feng, E. Tang, M. Zeng, Z. Gu, P. Kou, and W. Zheng, "Improving visual question answering for remote sensing via alternate-guided attention and combined loss," *International Journal of Applied Earth Observation and Geoinformation*, vol. 122, p. 103427, 2023.

[26] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[27] L. Hou, C.-P. Yu, and D. Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv preprint arXiv:1611.05916*, 2016.

[28] K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati, "Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change)," in *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

Nguyen Tien Huy received his B.S. degree in Software Technology in 2010 and his M.S. degree in Computer Science in 2015, both from the University of Science, Ho Chi Minh City, Vietnam (VNU-HCM). He earned his Ph.D. in Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 2019. He is currently a lecturer in the Department of Computer Science, Faculty of Information Technology, VNU-HCM. His research interests include Natural Language Processing and Deep Learning, with a particular focus on intelligent systems and multimodal data analysis.

Email: ntienhuy@fit.hcmus.edu.vn

Tung Le is a Lecturer of Information Technology at the University of Science, Ho Chi Minh City, Vietnam (HCMUS). He earned his B.S. degree in Computer Science from the Honour Program at the University of Science, Vietnam National University, Ho Chi Minh City, in 2012, and his M.S. degree in Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 2018. He completed his Ph.D. in Information Science at JAIST in 2021. His research interests include information retrieval, natural language processing, visual question answering, and multi-modal systems.

Email: tungle@fit.hcmus.edu.vn

Huynh Hoang Huy received the B.S. degree in Computer Science from the University of Science, Ho Chi Minh City, Vietnam (VNU-HCM), in 2021. He has completed a Master's degree at the University of Science. His research focuses on deep learning and its advancement toward real-world implementation.

Email: hhoanghuy@fit.hcmus.edu.vn