# MIXTURE OF EXPERTS FOR VIETNAMESE MULTIMODAL LEARNING

**Nguyễn Hoàng Quý**

University of Information Technology
HCMC, Vietnam

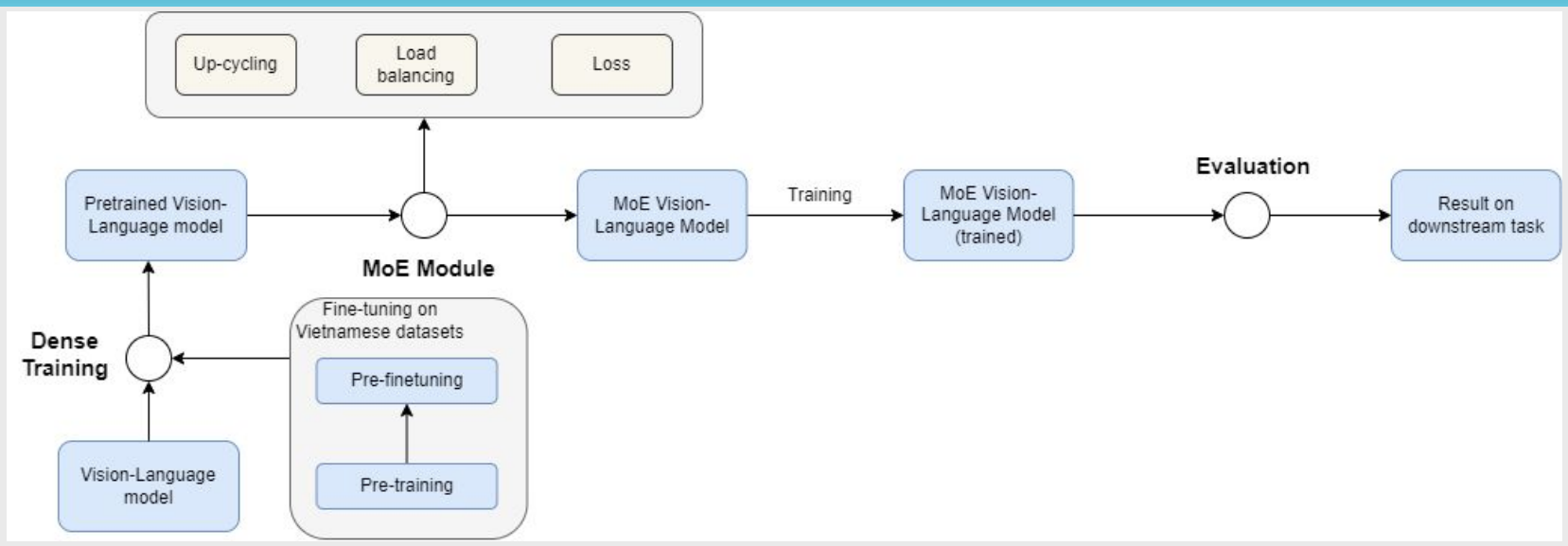
### What ?

- Propose a novel Mixture of Experts (MoE) architecture for small-scale Vietnamese vision-language models.
- Build a synthetic multimodal dataset for Vietnamese downstream tasks.
- Conduct comprehensive experiments between MoE and traditional dense architectures.
- Analyze the impact of model scaling (small, base, large) on multimodal learning performance.

### Why ?

- LLMs achieve remarkable performance thanks to abundant resources, but they are often impractical in low-resource environments like Vietnam.
- Mixture of Experts offers a promising solution to reduce computational costs while maintaining high efficiency.
- Developing MoE architectures for small-scale vision-language models,, is a crucial step toward expanding multimodal AI applications.

## Overview




## Description

### 1. Dataset Construction

- Collect and explore Vietnamese multimodal datasets (e.g., UIT-ViLC, ViOCRVQA, ViVQA).
- Build a synthetic dataset tailored for multimodal learning tasks.

### 2. Model Design

- Develop a Mixture of Experts (MoE)-based architecture to align text-image features.
- Experiment with cross-modal and inter-modal learning using language models (BERT, RoBERTa) and vision backbones (ViT, VGG), as well as pretrained vision-language models (CLIP, SigCLIP).

### 3. Training Strategies

- Explore up-cycling FFN (reusing pretrained models and duplicating experts).
- Implement load balancing and expert specialization to distribute knowledge effectively.
- Apply multimodal and MoE-specific loss functions (e.g., contrastive loss, sigclip loss, z-loss, importance loss).

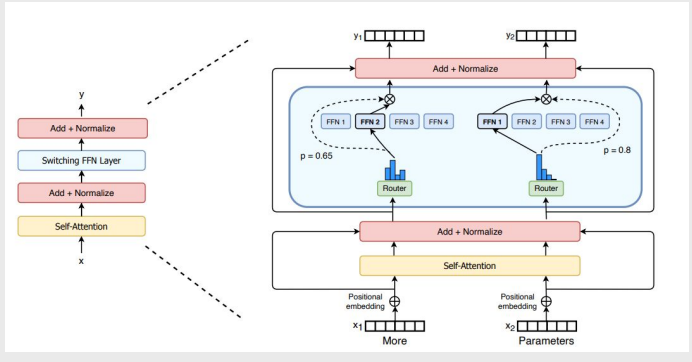



*Figure 1*. Example of Mixture of Expert Block.

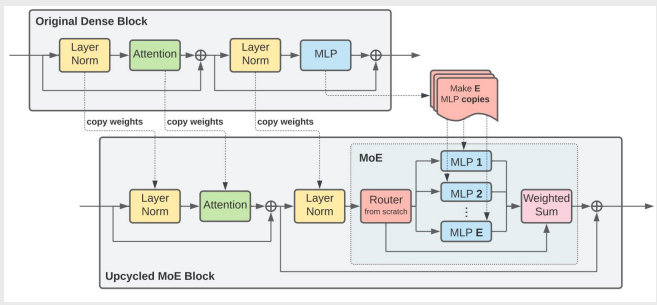



*Figure 2*. Example of Up-cycling training.

### 4. Evaluation & Analysis

- Conduct experiments to assess the effect of varying expert numbers, loss functions, and backbone models.
- Compare MoE vs. dense architectures on downstream tasks (image captioning, VQA) and analyze scaling impacts (small, base, large).

NII
Nguyễn Hoàng Quý – Trường Đại học Công Nghệ Thông Tin
TEL : 0974729224 Email : quynh.19@grad.uit.edu.vn