THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/CH0HIzWtNuc*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/hoangquy18/CS2205.CH190/blob/main/Qu%C3%BD%20Nguy%E1%BB%85n%20Ho%C3%A0ng%20-%20CS2205.FEB2025.DeCuong.FinalReport.Template.Slide.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Nguyễn Hoàng Quý
- MSSV: 240101066

- Lớp: CS2205.CH190
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân:
- Số câu hỏi QT của cả nhóm:
- Link Github: https://github.com/mynameuit/CS2205.xxx/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

ỨNG DỤNG MÔ HÌNH HỖN HỢP CHUYÊN GIA CHO HỌC ĐA THỂ THỨC TRÊN TIẾNG VIỆT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

MIXTURE OF EXPERTS FOR VIETNAMESE MULTIMODAL LEARNING

TÓM TẮT *(Tối đa 400 từ)*

Trong những năm gần đây, các mô hình ngôn ngữ lớn (LLM) dựa trên kiến trúc transformer đã mang lại tiến bộ vượt bậc cho xử lý ngôn ngữ tự nhiên và thị giác máy tính, nhưng đòi hỏi chi phí tính toán rất lớn, tạo thách thức tại các quốc gia hạn chế tài nguyên như Việt Nam. Kiến trúc Mô hình Hỗn hợp Chuyên gia (Mixture of Experts – MoE) nổi lên như một giải pháp nhờ chỉ kích hoạt một phần nhỏ mô hình khi suy luận, giúp giảm chi phí mà vẫn duy trì hiệu suất. Tuy đã có nhiều nghiên cứu về MoE cho mô hình lớn, việc áp dụng nó cho các mô hình thị giác-ngôn ngữ quy mô nhỏ (Small Vision-Language Model), đặc biệt với ngôn ngữ tài nguyên thấp như tiếng Việt, vẫn còn hạn chế. Nghiên cứu này đề xuất một kiến trúc MoE mới cho SVLM tiếng Việt nhằm nâng cao khả năng học đa thể thức, đồng thời thực hiện các thí nghiệm so sánh với kiến trúc dense truyền thống và phân tích tác động của việc mở rộng mô hình. Kỳ vọng kết quả sẽ chứng minh ưu thế của MoE trên các tác vụ downstream, góp phần thúc đẩy nghiên cứu và ứng dụng đa thể thức hiệu quả trong bối cảnh hạn chế tài nguyên.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong những năm gần đây, sự phát triển của trí tuệ nhân tạo (AI) đã được thúc đẩy mạnh mẽ nhờ các mô hình ngôn ngữ lớn (LLM) dựa trên kiến trúc transformer. Những mô hình này đã đạt được những thành công lớn nhờ lượng dữ liệu khổng lồ và tài nguyên tính toán dồi dào, các định luật mở rộng (scaling laws) cho thấy sự hiệu quả khi tăng kích thước mô hình và dữ liệu [1]. Tuy nhiên, tốc độ phát triển nhanh chóng kéo theo nhiều thách thức lớn về chi phí tính toán và thời gian suy luận, khiến các kiến trúc dày đặc (dense) trở nên khó áp dụng trong môi trường hạn chế tài nguyên như ở Việt Nam.

Để giải quyết bài toán này, các nhà nghiên cứu đã tìm đến những kiến trúc thay thế giúp cân bằng giữa hiệu suất và hiệu quả, trong số đó nổi bật nhất là kiến trúc Mô hình Hỗn hợp Chuyên gia (Mixture of Experts – MoE) [2]. MoE gồm tập hợp các mạng con chuyên biệt (“expert”) và cơ chế định tuyến (gating) để lựa chọn một số ít chuyên gia phù hợp ($n < E$) cho từng input đầu vào. Đây là hướng tiếp cận đầy hứa hẹn để mở rộng quy mô cho cả xử lý ngôn ngữ tự nhiên (NLP) và thị giác máy tính (CV).

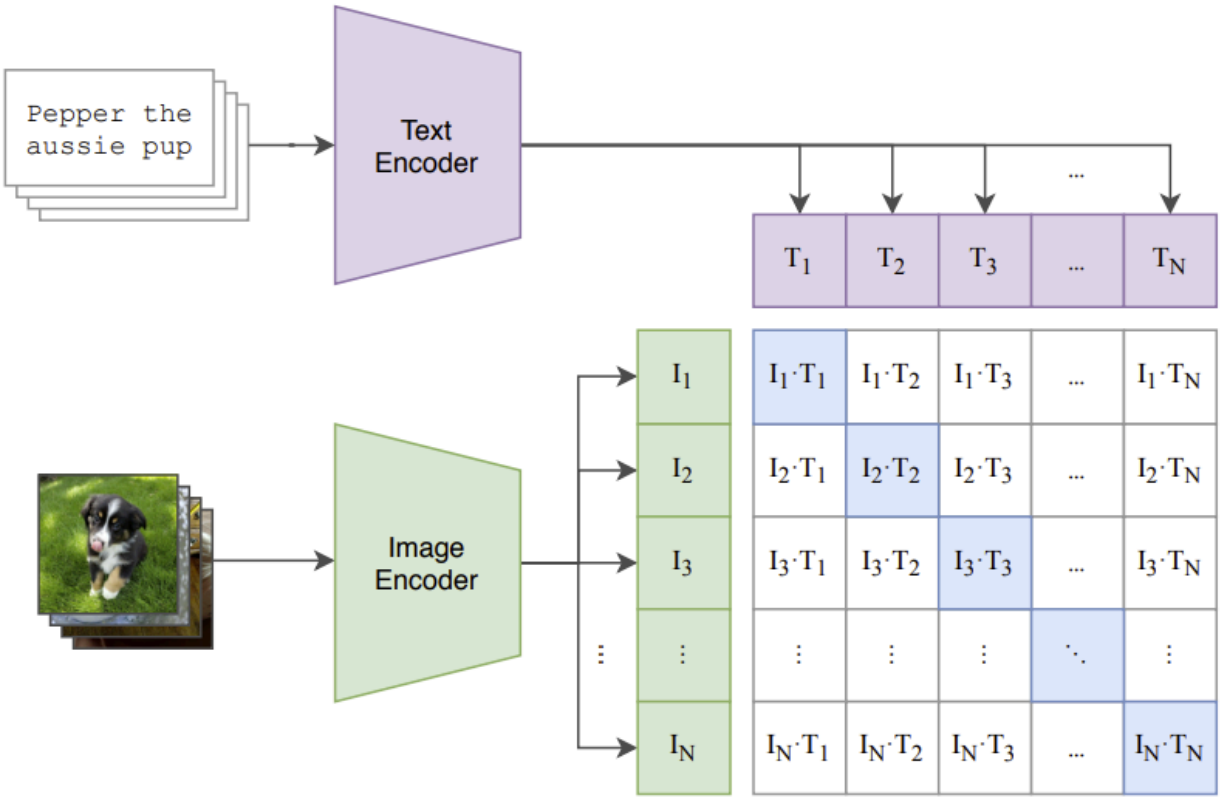
Năm 2024, các mô hình MoE nổi bật như Mixtral-8x7B, DeepSeekMoE, Switch Transformer và GLaM đã chứng minh thành công trong NLP. Lấy cảm hứng từ đó, các nhà nghiên cứu đã điều chỉnh MoE cho thị giác máy tính, tận dụng cơ chế chọn chuyên gia để xử lý thông tin hình ảnh hiệu quả hơn [3]. Điều này cũng thúc đẩy sự phát triển của các mô hình thị giác-ngôn ngữ (Vision-Language Model), phục vụ các nhiệm vụ như tạo chú thích ảnh, trả lời câu hỏi trực quan (VQA) và suy luận đa thể thức.

Dù phần lớn những nghiên cứu về MoE tập trung vào các mô hình lớn, việc tối ưu

hóa cho các mô hình nhỏ cũng rất quan trọng. Các nhà nghiên cứu đã thành công áp dụng MoE vào các mô hình ngôn ngữ nhỏ (SLM) như BERT hay RoBERTa, giúp giảm chi phí tính toán mà vẫn giữ kết quả tốt[4]. Tuy nhiên, việc ứng dụng MoE cho các mô hình thị giác-ngôn ngữ nhỏ (SVLM), đặc biệt ở các ngôn ngữ tài nguyên thấp như tiếng Việt, vẫn chưa được nghiên cứu rộng rãi.

Do đó, việc phát triển kiến trúc MoE cho SVLM mở ra một hướng đi đầy tiềm năng để nâng cao nhiệm vụ học đa thể thức. Vì vậy, trong đề tài này chúng tôi sẽ cung cấp một cái nhìn tổng quan toàn diện về vai trò của kiến trúc MoE trong các mô hình thị giác-ngôn ngữ quy mô nhỏ (Small scale MoE vision-language model), cụ thể như sau:
- Xây dựng bộ dữ liệu tổng hợp (synthetic data) phục vụ các tác vụ học đa thể thức ở Việt Nam
- Đề xuất kiến trúc mới dựa trên MoE nhằm nâng cao khả năng học sự tương tác giữa văn bản và hình ảnh (dự kiến đạt hiệu suất cao hơn so với kiến trúc dense truyền thống)
- Thực hiện các thí nghiệm toàn diện để so sánh hiệu suất giữa kiến trúc MoE đề xuất và kiến trúc dense trên các tác vụ đa thể thức (downstream tasks).
- Phân tích tác động của khả năng mở rộng mô hình (small, base, large) đến hiệu suất học đa thể thức

Input: 1 văn bản và hình ảnh
Output: Similarity giữa văn bản và hình ảnh đó



Ảnh 1: Ảnh minh họa contrastive learning

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

- Nghiên cứu và đề xuất kiến trúc mới dựa trên MoE nhằm nâng cao khả năng học sự tương tác giữa văn bản và hình ảnh (dự kiến đạt hiệu suất cao hơn so với kiến trúc dense truyền thống).
- Thực hiện các thí nghiệm toàn diện để so sánh hiệu suất giữa kiến trúc MoE đề xuất và kiến trúc dense trên các tác vụ đa thể thức (downstream tasks: VQA, image captioning).
- Phân tích tác động của khả năng mở rộng mô hình (small, base, large) đến hiệu suất học đa thể thức.

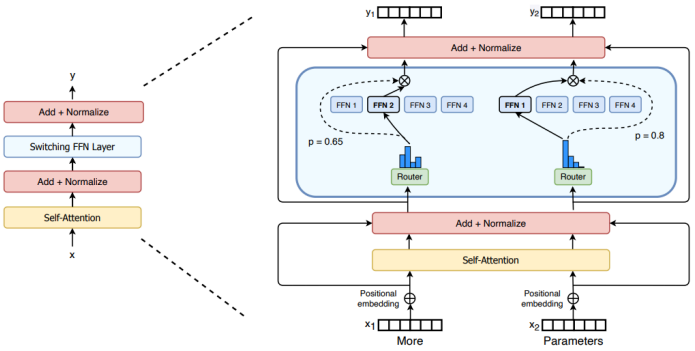
NỘI DUNG VÀ PHƯƠNG PHÁP

Bài nghiên cứu này bao gồm 4 nội dung chính, bao gồm:

- Bộ dữ liệu: Thu nhập và khám phá các bộ dữ liệu về multimodal trên tiếng Việt như:

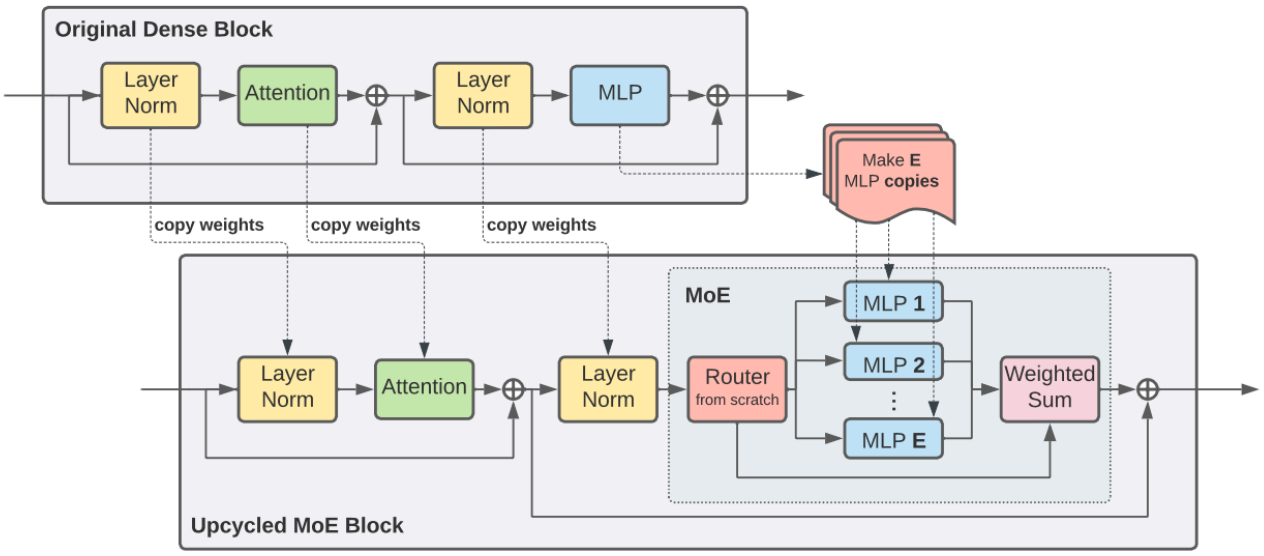
UIT-ViLC, ViOCRVQA, ViVQA, ...

- Nghiên cứu và xây dựng mô hình giúp căn chỉnh đặc trưng giữa văn bản-hình ảnh: sử dụng cross-modal và inter-modal learning.
    + Phần này sẽ tiến hành thử nghiệm và chọn ra các mô hình phù hợp cho bài toán. Ví dụ như text thì có các mô hình nổi bật như BERT, RoBERTa; phần vision nghiên cứu về các mô hình như ViT, VGG, ... Ngoài ra nghiên cứu về các mô hình language-vision đã tiền huấn luyện trước như CLIP, SigCLIP, ...
    + Sau đó biến đổi FFN sang Multi-FFN (MoE)



Ảnh 2: Ảnh minh họa kiến trúc MoE.

- Nghiên cứu các phương pháp huấn luyện mô hình MoE:
    + Upclyling FFN [5]: Nghĩa là sử dụng một mô hình đã được huấn luyện sẵn, sau đó sẽ duplicate lên nhiều phần khác nhau, giúp tận dụng được các đặc trưng đã được huấn luyện -> tuy nhiên có vấn đề là phải learning gate (router) từ đầu.
    + Cơ chế Load Balancing & Expert Specialization: Nghiên cứu các phương pháp giúp phân phối tới các Expert hiệu quả. Nghiên cứu cách phân bổ kiến thức đều lên mọi Expert hay mỗi Expert sẽ đảm nhiệm 1 chức năng riêng.
    + Các hàm loss về Multimodal và MoE: contrastive loss, sigclip loss, z-loss, importance-loss, ...



Ảnh 3: Ảnh minh họa cơ chế Up-cycling.

- Cơ chế đánh giá:

+ Các thử nghiệm trên mô hình MoE: tăng/giảm số lượng expert, sự ảnh hưởng của loss, các backbone model, ...

+ So sánh và đánh giá giữa việc sử dụng mô hình gốc và mô hình MoE trên các tác vụ downstream task cho tiếng Việt như: Image captioning, VQA, …

KẾT QUẢ MONG ĐỢI

- Các thí nghiệm toàn diện về việc xây dựng mô hình MoE trên tiếng Việt.
- Các mô hình sử dụng cơ chế MoE cho kết quả tốt hơn mô hình gốc trên các tác vụ downstream task cho tiếng Việt.
- Tính mở rộng: Khi tăng số lượng tham số thì hiệu suất mô hình cũng tăng theo.
- Viết thành 1 bài báo khoa học và được đăng tại 1 tạp chí uy tín.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

[2]. Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., & Huang, J. (2024). A survey on

mixture of experts. arXiv preprint arXiv:2407.06204.

[3]. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., ... & Houlsby, N. (2021). Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems, 34, 8583-8595.

[4]. Zuo, S., Zhang, Q., Liang, C., He, P., Zhao, T., & Chen, W. (2022). Moebert: from bert to mixture-of-experts via importance-guided adaptation. arXiv preprint arXiv:2204.07675.

[5]. Komatsuzaki, Aran, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. "Sparse upcycling: Training mixture-of-experts from dense checkpoints." arXiv preprint arXiv:2212.05055 (2022).