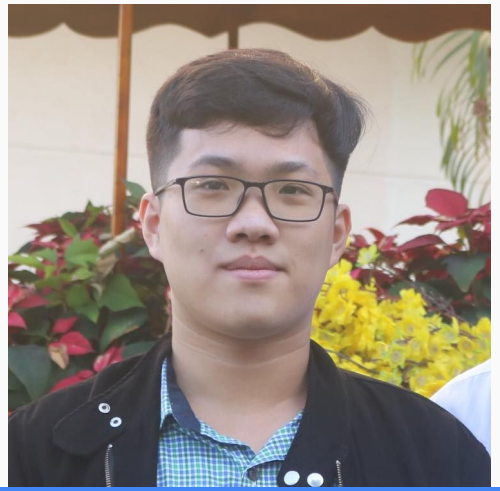
NGHIÊN CỨU MÔ HÌNH HỖN HỢP CHUYÊN GIA CHO HỌC ĐA THỂ THỨC TRÊN TIẾNG VIỆT

MIXTURE OF EXPERTS FOR VIETNAMESE MULTIMODAL LEARNING

GVHD: PGS. TS Lê Đình Duy
Học viên: Nguyễn Hoàng Quý - 240101066

Tóm tắt

- Lớp: CS2205.CH190
- Link Github của nhóm: https://github.com/hoangquy18/CS2205.CH190
- Link YouTube video: https://youtu.be/CH0HIzWtNuc
- Họ và tên: Nguyễn Hoàng Quý - 240101066



Giới thiệu

Bối cảnh:

- LLM & transformer → tiến bộ vượt bậc NLP, CV
- Thách thức: Chi phí tính toán cao → khó áp dụng ở Việt Nam
- Cần một giải pháp cân bằng hiệu suất & tài nguyên

Giới thiệu

Giải pháp: Mixture of Experts (MoE)

- Chỉ kích hoạt một nhóm nhỏ chuyên gia (expert)
- Chỉ kích hoạt một nhóm nhỏ chuyên gia khi suy luận
- Thành công ở mô hình lớn

→ Question: Có áp dụng tốt cho mô hình nhỏ, ngôn ngữ ít tài nguyên (như tiếng Việt) không?

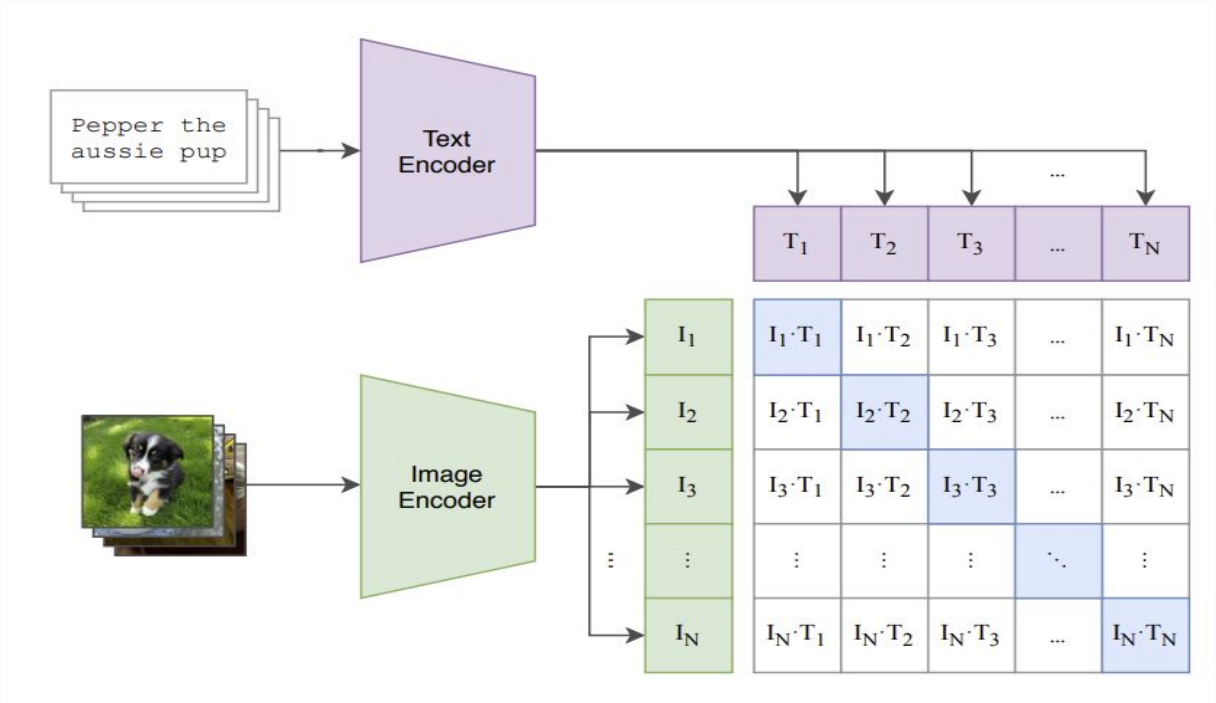
Mục tiêu

- Đề xuất kiến trúc MoE mới cho SVLM tiếng Việt
- Huấn luyện và so sánh với mô hình dense truyền thống
- Phân tích khả năng mở rộng khi tăng số lượng tham số

Nội dung và Phương pháp

- Thu nhập & khai thác dữ liệu đa thể thức tiếng Việt: UIT-ViLC, ViOCRVQA, ViVQA, …
  - Xử lí dữ liệu văn bản
  - Xử lí dữ liệu hình ảnh
- Xây dựng multimodal backbone: căn chỉnh đặc trưng văn bản – hình ảnh (cross-modal, inter-modal)

Nội dung và Phương pháp



Nội dung và Phương pháp

- Triển khai huấn luyện MoE:
  - Upcycling
  - Load balancing và expert specialization
- Đánh giá hiệu suất với các baseline khác trên downstream task:
  - Image Captioning
  - Visual Question Answering

Kết quả dự kiến

- 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.
- Mô hình có khả năng mở rộng: Hiệu suất mô hình tỉ lệ với số lượng tham số

Tài liệu tham khảo

[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).