THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://youtu.be/_p5t0YEh90A
- Link slides (dạng .pdf đặt trên Github của nhóm):
  https://github.com/nddat1811/CS2205.CH201/blob/main/SLIDE-DatNguyenDuc-Flashformer.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*

- Họ và Tên: Nguyễn Đức Đạt
- MSSV: 250201006

- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 2
- Link Github:
  https://github.com/nddat1811/CS2205.CH201/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

FLASHFORMER: MÔ HÌNH TÁCH GIỌNG NÓI ĐƠN KÊNH TỐI ƯU TỐC ĐỘ VÀ HIỆU SUẤT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

FLASHFORMER: DESIGN AND EVALUATION OF A FAST AND EFFICIENT SINGLE-CHANNEL SPEECH SEPARATION MODEL

TÓM TẮT *(Tối đa 400 từ)*

Đề tài **"FLASHFORMER: Mô hình tách giọng nói đơn kênh tối ưu tốc độ và hiệu suất"** tập trung giải quyết bài toán tách giọng nói từ một nguồn tín hiệu duy nhất. Đây là thách thức lớn trong xử lý tiếng nói, đóng vai trò quan trọng trong việc nâng cao chất lượng trợ lý ảo, thiết bị nghe nhìn và hội nghị trực tuyến. Hiện nay, các kiến trúc tiên tiến như Conv-TasNet [1] hay MossFormer2 [3] dù đạt hiệu suất cao nhưng vẫn gặp rào cản về chi phí tính toán. Cụ thể, cơ chế Transformer truyền thống có độ phức tạp tăng theo bình phương độ dài chuỗi, gây khó khăn khi triển khai trên các thiết bị tài nguyên hạn chế hoặc yêu cầu phản hồi thời gian thực.

Để khắc phục, nghiên cứu đề xuất kiến trúc lai **FlashFormer**, kết hợp khả năng mô hình hóa quan hệ dài hạn của Transformer [5] và tính hiệu quả trong ghi nhớ chuỗi của mạng FSMN [4]. Đột phá then chốt là việc tích hợp cơ chế **Flash Attention** [6] nhằm tối ưu hóa bộ nhớ ở cấp độ phần cứng. Kỹ thuật này giúp chuyển đổi độ phức tạp tính toán từ mức bình phương xuống tuyến tính, từ đó giảm đáng kể dung lượng bộ nhớ tiêu thụ và tăng tốc độ xử lý mà không làm giảm độ chính xác. Giải pháp này tạo ra sự cân bằng tối ưu giữa hiệu suất tách giọng (quality) và tốc độ thực thi (latency).

Về phương pháp thực hiện, nghiên cứu được triển khai bài bản từ khảo sát lý thuyết, xây dựng mô hình trên PyTorch đến huấn luyện và kiểm chứng trên các bộ dữ liệu chuẩn như WSJ0-2mix, Libri2Mix [7] hoặc LibriheavyMix [8]. Hiệu quả mô hình được đánh giá qua các chỉ số SI-SDR và SDR, đồng thời so sánh trực tiếp với các mô hình baseline để chứng minh sự cải thiện về tốc độ suy luận.

Kết quả dự kiến của luận văn là một mô hình FlashFormer có tính ứng dụng cao, sẵn sàng triển khai trên các hệ thống IoT và nền tảng truyền thông hiện đại. Nghiên cứu không chỉ đóng góp về mặt học thuật mà còn thúc đẩy sự phát triển của công nghệ xử lý tiếng nói trong các điều kiện thực tế khắc nghiệt.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Bài toán tách giọng đơn kênh đóng vai trò là một thách thức quan trọng và cốt lõi trong lĩnh vực xử lý tiếng nói và học máy hiện nay. Khác với các hệ thống đa kênh, việc tách giọng đơn kênh hoàn toàn không có lợi thế từ thông tin không gian, điều này buộc các mô hình phải dựa hoàn toàn vào việc khai thác đặc trưng của tín hiệu từ một nguồn duy nhất, đặt ra yêu cầu rất cao về khả năng mô hình hóa cả đặc trưng ngắn hạn lẫn các mối quan hệ dài hạn. Trong những năm gần đây, mặc dù các phương pháp học sâu như Conv-TasNet [1], DPRNN [2] hay Moss-Former2 [3] đã đạt được những tiến bộ vượt bậc, nhưng thực tế cho thấy các mô hình này vẫn gặp nhiều hạn chế khi áp dụng vào những tình huống có độ dài tín hiệu lớn hoặc yêu cầu triển khai trên các thiết bị có tài nguyên hạn chế. Đặc biệt, các kiến trúc dựa trên Transformer truyền thống dù cho thấy khả năng mô hình hóa quan hệ dài hạn rất tốt nhưng lại phải đánh đổi bằng chi phí tính toán và bộ nhớ tăng theo bình phương độ dài chuỗi, khiến việc xử lý tín hiệu tiếng nói dài trở nên kém hiệu quả trong các ứng dụng thời gian thực

Trong khi đó, mô hình FSMN [4] sở hữu ưu điểm về tính gọn nhẹ và khả năng nắm bắt ngữ cảnh dài với chi phí thấp, nhưng lại thiếu cơ chế tập trung linh hoạt để học các quan hệ động phức tạp trong tín hiệu. Do đó, việc kết hợp khả năng biểu diễn mạnh mẽ của Transformer với tính hiệu quả của FSMN là một giải pháp tiềm năng để cân bằng giữa hiệu năng và tốc độ xử lý. Đặc biệt, sự ra đời của Flash Attention [6] giúp tối ưu hóa tính toán ở mức phần cứng, giảm chi phí bộ nhớ và tăng tốc độ suy luận khi xử lý chuỗi tín hiệu lớn. Việc tích hợp Flash Attention [6] vào kiến trúc lai Transformer–FSMN không chỉ duy trì độ chính xác cao mà còn cải thiện khả năng vận hành thực tế trên các thiết bị IoT, trợ lý ảo hay hệ thống hội nghị truyền hình.

Từ những thực tiễn trên, đề tài tập trung vào việc phát triển và đánh giá mô hình tách giọng nói đơn kênh dựa trên kiến trúc lai mang tên FlashFormer nhằm cải thiện đồng thời tốc độ suy luận và chất lượng tách giọng trong nhiều điều kiện tín hiệu khác nhau. Ý nghĩa của nghiên cứu không chỉ nằm ở việc hoàn thiện các mô hình học thuật mà còn đáp ứng nhu cầu cấp thiết về một hệ thống nhanh hơn, nhẹ hơn nhưng vẫn đảm bảo chất lượng cao để ứng dụng rộng rãi trong đời sống. Mục tiêu cụ thể của nghiên cứu bao gồm việc phân tích các hạn chế về chi phí tính toán của các mô hình baseline, xây dựng kiến trúc FlashFormer tối ưu và thực hiện huấn luyện, đánh giá toàn diện trên các bộ dữ liệu chuẩn như WSJ0-2mix, Libri2Mix [7] hay LibriheavyMix [8]. Kết quả nghiên cứu kỳ vọng sẽ khẳng định tính khả thi của việc kết hợp giữa attention tối ưu và mạng bộ nhớ tuần tự, tạo tiền đề cho các ứng dụng nhận dạng tiếng nói và truyền thông thông minh trong môi trường nhiễu

phức tạp.

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

- **Phân tích và đánh giá các hạn chế của các mô hình hiện đại:** Tập trung chỉ ra những rào cản về chi phí tính toán và hiệu năng xử lý các chuỗi tín hiệu dài của các kiến trúc như Conv-TasNet [1], DPRNN [2] và MossFormer2 [3].
- **Xây dựng kiến trúc lai Transformer–FSMN tích hợp Flash Attention:** Thiết kế mô hình FlashFormer nhằm tối ưu hóa khả năng mô hình hóa quan hệ ngữ cảnh dài hạn, đồng thời giảm mạnh chi phí bộ nhớ và tăng tốc độ suy luận trong thời gian thực
- **Huấn luyện và kiểm chứng hiệu năng mô hình:** Thực nghiệm trên các bộ dữ liệu chuẩn (WSJ0-2mix, Libri2Mix [7], LibriheavyMix [8]) để xác định mức cải thiện về chất lượng tách giọng (SI-SDR, SDR) và khả năng tiết kiệm tài nguyên so với các phương pháp tiên tiến hiện nay.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. **Nghiên cứu tổng quan các phương pháp hiện có trong lĩnh vực tách giọng nói đơn kênh**

Mục tiêu

Làm rõ cơ sở lý thuyết, các hướng tiếp cận hiện đại và những hạn chế còn tồn tại trong bài toán tách giọng nói đơn kênh, đặc biệt tập trung vào các mô hình miền thời gian. Thêm vào đó là đánh giá khả năng mô hình hóa quan hệ ngữ cảnh dài hạn, chi phí tính toán và tính phù hợp khi triển khai trên các hệ thống thực tế.

Công việc

Thu thập và phân tích các mô hình tiêu biểu như Conv-TasNet [1], DPRNN [2], MossFormer2 [3] và các biến thể dựa trên Transformer [5].

Đánh giá ưu điểm, nhược điểm, độ phức tạp tính toán và khả năng mở rộng đối với dữ liệu dài.

Khảo sát vai trò của FSMN [4] và Flash Attention [6] trong việc tối ưu hóa xử lý chuỗi.

Phương pháp

Đọc và tổng hợp tài liệu từ các công bố khoa học quốc tế từ thông qua Google Scholar, IEEE, từ GVHD.

Phân tích định tính các kiến trúc mô hình dựa trên cấu trúc mạng, độ phức tạp tính toán và yêu cầu tài nguyên.

So sánh định tính giữa các kiến trúc, phân tích mô hình theo các tiêu chí đánh giá trong

lĩnh vực tách giọng như SI-SDR, SI-SDRi.

2. **Đề xuất kiến trúc mô hình FlashFormer**

Mục tiêu

Đề xuất mô hình **FlashFormer**, một kiến trúc lai kết hợp giữa Transformer và Feedforward Sequential Memory Network (FSMN) [4], đồng thời tích hợp cơ chế Flash Attention [6] nhằm tối ưu hóa quá trình tính toán. Trong kiến trúc đề xuất, Transformer đảm nhiệm vai trò học các quan hệ phụ thuộc toàn cục và biểu diễn đặc trưng giàu ngữ cảnh, FSMN hỗ trợ mở rộng vùng ngữ cảnh dài với chi phí thấp, còn Flash Attention giúp giảm tiêu thụ bộ nhớ và tăng tốc độ xử lý attention khi làm việc với chuỗi tín hiệu dài.

Công việc

Thiết kế kiến trúc lai kết hợp Transformer [5] và FSMN [4] nhằm tận dụng ưu điểm của cả hai: khả năng biểu diễn mạnh mẽ của Transformer và tính hiệu quả của FSMN trong xử lý ngữ cảnh dài.

Tích hợp cơ chế Flash Attention [6] để tối ưu hóa bộ nhớ và rút ngắn thời gian tính toán attention khi xử lý chuỗi tín hiệu dài.

Thử nghiệm và điều chỉnh các siêu tham số quan trọng của mô hình để đạt hiệu năng tốt nhất, bao gồm số lớp, kích thước ẩn, số đầu attention và độ sâu của bộ nhớ FSMN.

Phương pháp

Xây dựng và triển khai mô hình bằng Python sử dụng các thư viện học sâu như PyTorch, cho phép linh hoạt trong thiết kế và kiểm thử.

Thực hiện kiểm thử theo từng mô-đun để đánh giá độc lập hiệu quả của từng thành phần trước khi kết hợp thành kiến trúc hoàn chỉnh.

Tiến hành các thí nghiệm nhỏ nhằm đánh giá sơ bộ khả năng hội tụ, mức độ tiêu thụ tài nguyên và tốc độ suy luận của mô hình, làm cơ sở cho việc tối ưu tiếp theo.

3. **Huấn luyện và đánh giá mô hình trên các bộ dữ liệu chuẩn**

Mục tiêu

Xác minh hiệu năng của mô hình **FlashFormer** được đề xuất trong các điều kiện thực tế và tiến hành so sánh trực tiếp với các phương pháp hiện có trong lĩnh vực tách giọng đơn kênh. Thông qua quá trình huấn luyện và đánh giá toàn diện, mục tiêu là kiểm chứng khả năng mô hình hóa quan hệ ngữ cảnh dài hạn, chất lượng tách giọng, khả năng tổng quát hóa và mức độ tối ưu về tài nguyên.

Công việc

Thực hiện chuẩn hóa toàn bộ các bộ dữ liệu chuẩn như WSJ0-2mix, Libri2Mix [7],

LibriheavyMix [8] hoặc WHAM!, bao gồm các bước tách tập huấn luyện – kiểm tra – xác thực, xử lý mức âm lượng, tạo các mixture theo cấu hình quy định, và đảm bảo sự đồng nhất giữa các mẫu âm thanh.

Tiến hành huấn luyện mô hình với nhiều thiết lập cấu hình khác nhau nhằm đánh giá mức độ ổn định, độ nhạy với siêu tham số và khả năng hội tụ của mô hình.

Ghi lại các chỉ số loss, thời gian huấn luyện mỗi epoch, mức sử dụng GPU/CPU và tốc độ xử lý để phục vụ phân tích sau này.

Thực hiện so sánh định lượng với các mô hình baseline như Conv-TasNet [1], DPRNN [2] hoặc MossFormer2 [3] nhằm xác định mức cải thiện về chất lượng tách giọng và tốc độ xử lý.

Phương pháp

Tiền xử lý dữ liệu bao gồm chuẩn hóa mức âm lượng, chia tập huấn luyện – xác thực – kiểm tra và đảm bảo tính đồng nhất giữa các mẫu.

Sử dụng các thước đo đánh giá được cộng đồng nghiên cứu chấp nhận rộng rãi trong lĩnh vực tách giọng đơn kênh, đảm bảo kết quả thu được có tính khách quan và dễ dàng so sánh với các phương pháp khác.

Ngoài đánh giá bằng chỉ số số học, tiến hành nghe mẫu đầu ra của mô hình (subjective listening test) để đánh giá cảm nhận về chất lượng tín hiệu. Điều này giúp bổ sung góc nhìn toàn diện hơn khi các chỉ số định lượng chưa phản ánh hết chất lượng âm thanh.

Kết hợp số liệu từ nhiều thí nghiệm khác nhau, sử dụng biểu đồ, bảng tổng hợp và mô hình thống kê để phân tích sự ảnh hưởng của từng siêu tham số hoặc từng thành phần kiến trúc.

4. Phân tích, thảo luận và rút ra kết luận

Mục tiêu

Đánh giá mức độ đáp ứng của mô hình đối với các mục tiêu nghiên cứu đã đề ra, đồng thời xác định khả năng ứng dụng thực tiễn của mô hình trong các hệ thống xử lý tiếng nói, đặc biệt là các môi trường yêu cầu tốc độ cao và tài nguyên hạn chế.

Công việc

Phân tích chi tiết các kết quả thí nghiệm nhằm nhận diện điểm mạnh và những hạn chế của mô hình, bao gồm chất lượng tách giọng, tốc độ suy luận và khả năng mở rộng sang các tập dữ liệu hoặc điều kiện âm thanh khác.

Thảo luận về tính khả thi khi triển khai mô hình trong các hệ thống thời gian thực hoặc những thiết bị có tài nguyên hạn chế, đánh giá mức độ phù hợp với các yêu cầu thực tế.

Đề xuất các hướng nghiên cứu tiếp theo để tiếp tục hoàn thiện mô hình, chẳng hạn như

cải thiện cấu trúc kiến trúc, mở rộng sang dữ liệu đa dạng hơn hoặc tối ưu hóa quá trình suy luận.

Phương pháp

Sử dụng các biểu đồ, bảng số liệu và công cụ phân tích trực quan để trình bày và đánh giá kết quả một cách rõ ràng và có hệ thống.

Đối chiếu kết quả thu được với các nghiên cứu liên quan nhằm đặt mô hình vào bối cảnh tổng thể của lĩnh vực tách giọng nói, từ đó làm rõ những đóng góp mới của đề tài.

Tổng hợp và trình bày các nhận định khoa học một cách logic, chặt chẽ, làm cơ sở cho phần kết luận và đánh giá chung của luận văn.

KẾT QUẢ MONG ĐỢI

Kết quả quan trọng nhất của nghiên cứu là việc xây dựng thành công FlashFormer, một kiến trúc tách giọng nói đơn kênh lai kết hợp giữa Transformer [3], FSMN [4] và cơ chế Flash Attention [6]. Mô hình này hướng đến mục tiêu tối ưu hóa đồng thời hiệu suất tách giọng và tốc độ xử lý, tạo ra sự cân bằng vượt trội giữa khả năng mô hình hóa ngữ cảnh dài hạn và chi phí tính toán. Nhờ vào cấu trúc Transformer được tinh chỉnh kết hợp với tầng FSMN [4] giúp mở rộng ngữ cảnh hiệu quả và Flash Attention [6] tối ưu hóa truy cập bộ nhớ, mô hình được kỳ vọng sẽ đạt tốc độ suy luận nhanh hơn đáng kể trên các chuỗi tín hiệu dài.

Về mặt thực nghiệm, FlashFormer dự kiến đạt chất lượng tách giọng cạnh tranh hoặc vượt trội so với các mô hình baseline hiện có khi đánh giá trên các bộ dữ liệu chuẩn cho bài toán tách giọng nói đơn kênh. Các chỉ số đánh giá phổ biến như SI-SDR, SI-SDRi hoặc SDR được kỳ vọng cải thiện rõ rệt, trong khi tốc độ suy luận được nâng cao, đặc biệt đối với các tín hiệu có độ dài lớn. Điều này cho thấy mô hình không chỉ mạnh về mặt biểu diễn mà còn phù hợp với các yêu cầu xử lý thời gian thực.

Bên cạnh các thông số kỹ thuật, mô hình còn được kỳ vọng sẽ cho thấy sự linh hoạt và khả năng tổng quát hóa tốt trong nhiều điều kiện tín hiệu khác nhau. Điều này mở ra cơ hội ứng dụng rộng rãi trong các bài toán liên quan như tăng cường tiếng nói, nhận dạng giọng nói trong môi trường nhiễu phức tạp và tích hợp vào các nền tảng giao tiếp thông minh. Với những kết quả đạt được, FlashFormer không chỉ là một giải pháp tách giọng hiệu quả mà còn đóng góp một hướng tiếp cận tiềm năng cho các nghiên cứu về mô hình lai trong lĩnh vực xử lý tiếng nói.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," 2019

[2] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," 2020

[3] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma, "MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation," 2024

[4] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu, "Feedforward Sequential Memory Networks: A New Structure to Learn Long-term Dependency," 2015

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Łukasz Kaiser, "Attention Is All You Need," 2017

[6] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra and Christopher Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," 2022

[7] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge and Emmanuel Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," 2020

[8] Zengrui Jin, Yifan Yang, Mohan Shi, Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Liyong Guo, Lingwei Meng, Long Lin, Yong Xu, Shi-Xiong Zhang, Daniel Povey, "LibriheavyMix: A 20,000-Hour Dataset for Single-Channel Reverberant Multi-Talker Speech Separation, ASR and Speaker Diarization," 2024