Mô hình học sâu trong nhận dạng giọng nói và ứng dụng tự động ghi biên bản cuộc họp

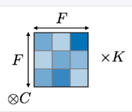
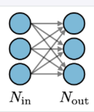
Đặng Đức Mạnh

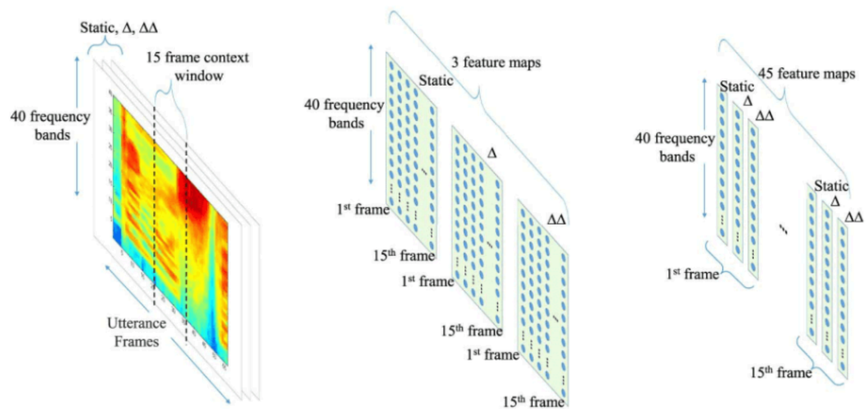
Tóm tắt

Nghiên cứu về nhận dạng giọng nói sử dụng các phương pháp học sâu đã đạt được những tiến bộ đáng kể. Ứng dụng các mô hình học sâu trong chuyển đổi tiếng nói thành văn bản đã đem lại nhiều kết quả có tính cách mạng cả về tốc độ và hiệu quả. Bài báo này đề xuất một mô hình học sâu nhận dạng giọng nói bằng cách sử dụng mạng nơ ron tích chập và mạng nơ ron hồi quy. Kết quả thử nghiệm cho thấy độ chính xác và hiệu suất của giải pháp đề xuất đã được cải thiện đáng kể.

Phương pháp

Phương pháp nghiên cứu mô hình học sâu ứng dụng ghi biên bản họp tập trung vào việc phát triển mô hình tự động chuyển đổi cuộc họp thành văn bản tóm tắt. Quá trình này thường sử dụng các mô hình ngôn ngữ tự nhiên như JASPER, BERT hoặc các biến thể của chúng để phân tích và hiểu ngữ cảnh cuộc họp. Dữ liệu từ cuộc họp được xử lý qua các bước tiền xử lý như loại bỏ tiếng ồn, chuyển đổi giọng nói thành văn bản, sau đó phân tích ngữ nghĩa và cấu trúc. Cuối cùng, hệ thống tự động tạo ra các tóm tắt chính xác, ngắn gọn, giúp người dùng dễ dàng nắm bắt nội dung cuộc họp một cách hiệu quả mà không cần xem lại toàn bộ cuộc họp.

	CONV	POOL	FC
Minh họa			
Kích thước đầu vào	$I \times I \times C$	$I \times I \times C$	N_{in}
Kích thước đầu ra	$O \times O \times K$	$O \times O \times C$	N_{out}
Số lượng tham số	$(F \times F \times C + 1) \cdot K$	0	$(N_{\text{in}} + 1) \times N_{\text{out}}$
Lưu ý	• Một tham số bias với mỗi bộ lọc • Trong đa số trường hợp, $S < F$ • Một lựa chọn phổ biến cho K là $2C$	• Phép pooling được áp dụng lên từng kênh (channel-wise) • Trong đa số trường hợp, $S = F$	• Đầu vào được làm phẳng • Mỗi neuron có một tham số bias • Số neuron trong một tầng FC phụ thuộc vào ràng buộc kết cấu



Kết quả đánh giá

Kết quả huấn luyện mô hình Jasper đạt 15,6% tỷ lệ lỗi (WER). tỷ lệ lỗi khá cao so với mô hình huấn luyện của NVIDIA. Điều này cho thấy tỷ lệ lỗi của mô hình này sẽ bị ảnh hưởng bởi hiệu năng GPU

WER (Word Error Rate) là một phép đo thường được sử dụng để đánh giá hiệu suất của mô hình nhận dạng giọng nói. Nó đo lường tỷ lệ lỗi trong việc dịch chính xác từng từ trong văn bản đúng từ giọng nói đã cho.

Kết luận

Trong đề tài này, chúng tôi đã giới thiệu mô hình học máy sử dụng mạng nơ ron tích chập và mạng nơ ron hồi quy trong nhận dạng tiếng nói. Các kết quả thực nghiệm cho thấy mô hình đề xuất tỏ ra khá hiệu quả trong chỉ số đánh giá WER, kết quả nghiên cứu cũng được ứng dụng trong xây dựng ứng dụng ghi âm biên bản cuộc họp.