NHẬN DIỆN CẢM XÚC QUA GIỌNG NÓI SỬ DỤNG KOLMOGOROV–ARNOLD NETWORKS

Dương Thanh Nguyên - 230201048

Tóm tắt

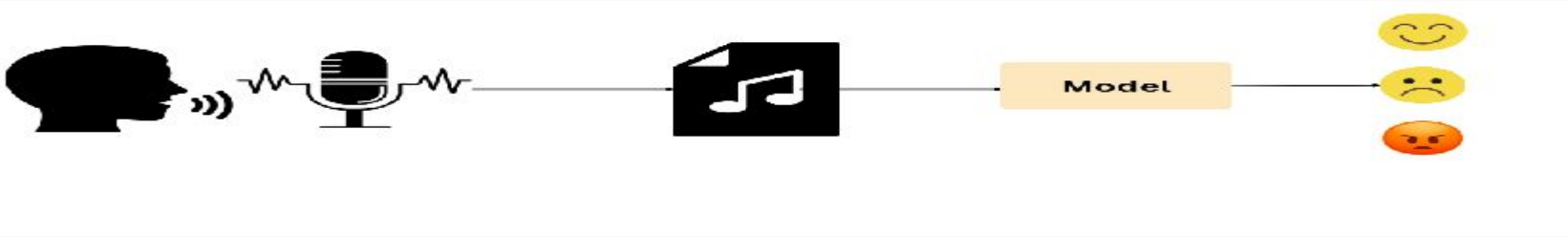
- Lớp: CS2205.APR2024
- Link Github: https://github.com/nguyendt-tn/CS2205.APR2024
- Link YouTube video: https://youtu.be/7BQmKw8UQfE
- Ảnh + Họ và Tên: Dương Thanh Nguyên
- Tổng số slides không vượt quá 10

Giới thiệu

- Tình Hình Tổng Quan
  - Cần cải thiện tương tác giữa máy tính và con người để tạo trải nghiệm người dùng tốt hơn.
  - Gia tăng vấn đề tâm lý sau đại dịch COVID-19. Thiếu hụt chuyên gia hỗ trợ tâm lý.
  - Công nghệ nhận diện cảm xúc chưa đạt hiệu quả cao với phương pháp học máy truyền thống.
- Nghiên Cứu Giải Quyết
  - Áp dụng kỹ thuật học sâu để cải thiện nhận diện cảm xúc.
  - Giới thiệu và ứng dụng Kolmogorov–Arnold Networks (KANs)[1] để nâng cao độ chính xác.
  - Xây dựng một mô hình tối ưu cho nhận diện cảm xúc dựa trên KANS.

Giới thiệu

- Input:
  - Tín hiệu âm thanh: Tín hiệu âm thanh ghi lại giọng nói của người nói.
  - Dữ liệu ngữ cảnh: Dữ liệu ngữ cảnh có thể bao gồm văn bản lời nói, hình ảnh video, hoặc thông tin về người nói và chủ đề giao tiếp.
- Output:
  - Cảm xúc: Cảm xúc của người nói được xác định dưới dạng một hoặc nhiều nhãn cảm xúc như vui, buồn, tức giận, sợ hãi, v.v.
  - Mức độ tin cậy: Mức độ tin cậy của kết quả nhận diện cảm xúc.



Mục tiêu

- Cải thiện độ chính xác trong nhận diện cảm xúc qua giọng nói sử dụng KANs
- Khám phá hiệu quả của KANs trong ứng dụng thực tế
- Xây dựng mô hình tối ưu cho nhận diện cảm xúc qua giọng nói từ KANs

Nội dung và Phương pháp

❏ Nội dung:

- Nghiên cứu và phân tích đặc trưng giọng nói để phân biệt cảm xúc.
- Xây dựng và so sánh hiệu suất của mô hình KANs với các mô hình truyền thống (LSTM, CNN, RNN).
- Đánh giá khả năng và hiệu quả của KANs trong các tình huống thực tế và môi trường âm thanh đa dạng.
- Tối ưu hóa kiến trúc và các tham số của mô hình KANs để đạt được độ chính xác tốt nhất trong nhận diện cảm xúc qua giọng nói.

Nội dung và Phương pháp

❑ Phương Pháp:

- Sử dụng các tập dữ liệu giọng nói đã được gán nhãn cảm xúc như RAVDESS, IEMOCAP
- Tiền xử lý dữ liệu bao gồm loại bỏ nhiễu và trích xuất các đặc trưng như MFCCs, formants, pitch.
- Xây dựng mô hình sử dụng KANs và huấn luyện trên dữ liệu đã tiền xử lý.
- Đánh giá hiệu suất bằng các chỉ số như accuracy, precision, recall, F1-score và so sánh với các mô hình truyền thống (LSTM, CNN, RNN).
- Xác định các tình huống ứng dụng thực tế và thực hiện thử nghiệm trên các dữ liệu giọng nói từ các môi trường âm thanh khác nhau.
- Tối ưu hóa kiến trúc và tham số của mô hình sử dụng KANs bằng cách thử nghiệm và điều chỉnh các yếu tố như số lớp, số nơ-ron, hàm kích hoạt.
- Sử dụng các kỹ thuật huấn luyện tiên tiến như learning rate annealing, early stopping, và k-fold cross-validation và giảm thiểu tài nguyên tính toán để tối ưu hiệu suất và thời gian huấn luyện của mô hình.

Kết quả dự kiến

- KANs mang lại độ chính xác cao hơn và hiệu suất tốt hơn so với các phương pháp truyền thống (LSTM, CNN, RNN) trong nhận diện cảm xúc qua giọng nói, độ chính xác hơn 85% trên bộ dữ liệu RAVDESS và hơn 75% trên bộ dữ liệu IEMOCAP.
- KANs có khả năng ứng dụng hiệu quả trong các tình huống thực tế, duy trì hiệu suất tốt trong các môi trường âm thanh phức tạp.
- Xây dựng được mô hình KANs tối ưu với hiệu suất tính toán cao và tốc độ huấn luyện nhanh, dễ dàng triển khai và ứng dụng.

Tài liệu tham khảo

- [1].Ziming Liu and Yixuan Wang and Sachin Vaidya and Fabian Ruehle and James Halverson and Marin Soljačić and Thomas Y. Hou and Max Tegmark. (2024). KANs: Kolmogorov-Arnold Networks. arXiv preprint arXiv:2404.19756.
- [2].Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.
- [3]. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.
- [4]. Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J.M., Fernández-Martínez, F. (2021). A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. Appl. Sci. 2022, 12, 327, doi: 10.3390/app12010327.

Tài liệu tham khảo

- [5]. Boigne, J., Liyanage, B., Östrem, T. (2020). Recognizing more emotions with less data using self-supervised transfer learning. arXiv preprint arXiv:2011.05585.
- [6]. Pepino, L., Riera, P., Ferrer, L. (2021). Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. Proc. Interspeech 2021, 3400-3404, doi:10.21437/Interspeech.2021-703.
- [7]. Chen, L. W., Rudnicky, A. (2021). Exploring Wav2vec 2.0 fine-tuning for improved speech emotion recognition. arXiv preprint arXiv:2110.06309.