Mô hình nhẹ cho bài toán phát hiện người đang nói

Abstract

Phát hiện người đang nói là một bài toán đầy thử thách trong việc hiểu kịch bản có cả âm thanh và hình hảnh, với mục đích là phát hiện được ai đang nói trong kịch bản có một hoặc nhiều người nói. Bài toán này nhận được mối quan tâm rộng rãi bởi tính quan trọng trong việc ứng dụng như tách kênh nói, theo dõi người nói và chỉnh sửa video tự động. Những cuộc nghiên cứu hiện nay đã nỗ lực cải tiến hiệu suất bằng nhiều thông tin ứng viên đầu vào và thiết kế mô hình phức tạp. Dù những phương pháp này đạt được hiệu suất vượt trội, nhưng lượng tiêu tốn về bộ nhớ và khả năng tính toán làm cho việc ứng dụng trở nên khó khăn với nguồn tài nguyên kịch bản hạn chế. Vì vậy, chúng tôi xây dựng một kiến trúc nhẹ cho bài toán phát hiện người đang nói bằng giảm ứng viên đầu vào, tách kết cấu 2D và 3D cho tính năng chiết xuất nghe nhìn, và áp dụng đơn vị lặp lại có kiểm soát với độ phức tạp tính toán thấp cho mô hình đa phương thức. Những kết quả thí nghiệm trên bộ dữ liệu AVA-ActiveSpeaker cho rằng bộ khung làm việc của chúng tôi đạt được hiệu suất mAP tốt (94.1% với 94.2%, khoảng 23 lần) và FLOPs (0.6G với 2.6G, khoảng 4 lần). Ngoài ra, bộ khung làm việc của chúng tôi cũng thực hiện tốt trên bộ dữ liệu Columbia cho thấy sức mạnh của phương pháp. Code và trọng lượng mô hình có trên: https://github.com/Junhua-Liao/Light-ASD