BÀI THU HOẠCH

CHỦ ĐỀ: VIDEO SUMMARIZATION

A. Thông tin cá nhân

MSSV: 19127273
Huỳnh Thị Mỹ Thanh

B. Bài thu hoạch

I. Problem statement

Input: một video

Output: video có độ dài ngắn hơn nhưng vẫn giữ được nội dung chính

II. Challenge

- Tính chủ quan của việc xác định những phần quan trọng của video. Người khác nhau có thể có ý kiến khác nhau về những phần của video được coi là quan trọng, phụ thuộc vào sở thích và mục tiêu của họ.
- Lượng dữ liệu rất lớn trong quá trình tóm tắt video. Video có thể dài vài giờ và ngay cả một tóm tắt ngắn cũng có thể chứa một lượng dữ liệu đáng kể.
- Nhiều video có sự thay đổi nội dung theo thời gian (ví dụ như quá trình phát triển của 1 loài sinh vật nào đó), nhưng việc tóm tắt lại yêu cầu các phần quan trọng phải được giữ nguyên, như vậy có phải đồng nghĩa với việc phải giữ cả video dài?

III. Methods

Truyền thống:

1. Keyframe extraction: chọn một số frame đại diện cho các phần quan trọng của video để tạo thành một bản tóm tắt.
2. Feature extraction: trích xuất các đặc trưng quan trọng của video như màu sắc, chuyển động và âm thanh. Sau đó các đặc trưng này được kết hợp để tạo thành một bản tóm tắt.

Deep learning

1. ***Supervised learning***
    - **Bi LSTM:** để xác định các sự kiện quan trọng trong video, kết hợp hai LSTM chạy song song trên cùng một chuỗi đầu vào ở 2 chiều (xuôi, ngược), đồng thời xây dựng hai hidden state khác nhau để biểu diễn thông tin từ mỗi chiều. Sau đó, các hidden state từ hai chiều được concatenated với nhau để tạo thành một hidden state duy nhất. Quá trình này cho phép Bi LSTM học được đặc trưng

của dữ liệu từ cả hai chiều và giúp cải thiện độ chính xác của mô hình trong các tác vụ như phân loại, dự đoán và tổng hợp.

- **Seq2seq:** được sử dụng để tạo ra các câu mô tả tóm tắt cho các sự kiện trong video. Kiến trúc của Seq2Seq bao gồm một mô hình encoder và một mô hình decoder. Encoder sẽ đọc dữ liệu đầu vào (như là các frame) và biểu diễn chúng thành một vector ngữ nghĩa. Vector này sẽ được truyền vào mô hình decoder để tạo ra các câu mô tả tóm tắt.
- **Fully convolution sequence network:** nó chỉ sử dụng các lớp convolution và không sử dụng các lớp fully connected. FCSN có thể chạy trên các khung hình của video và sinh ra một heatmap để biểu diễn mức độ quan trọng đối với các frame. Heatmap này được sử dụng để tính toán các điểm đánh dấu (landmarks) trên video, là các điểm đại diện cho các sự kiện quan trọng. Sau khi các điểm đánh dấu được tính toán, một mô hình tóm tắt có thể được sử dụng để tạo ra các câu mô tả tóm tắt cho các sự kiện này. FCSN đã được chứng minh là hiệu quả trong tác vụ video summarization, và nó có thể xử lý các video với kích thước khác nhau một cách hiệu quả. Mặc dù FCSN không sử dụng các lớp fully connected, nó vẫn có thể học được các đặc trưng tốt và sinh ra các câu mô tả tóm tắt chất lượng cao cho các sự kiện trong video.

2. ***Unsupervised learning***
   - **CRNN:** dùng CNN rút trích đặc trưng => đưa vào RNN học đặc trưng (về thời gian và nội dung của các frame, âm thanh...)
   - **3D – CNN:** dùng trên dãy ảnh 3 chiều (xếp chồng các frame lên) , có khả năng học các đặc trưng không gian và thời gian của các frame.
   - **GANs:** sử dụng một mạng phân biệt và một mạng sinh (generator) để sinh ra các mẫu mới từ dữ liệu đầu vào.
3. ***Weakly supervised learning***

Chỉ một phần nhỏ các tập dữ liệu được gán nhãn. Cụ thể trong video summarization, các video có sẵn được sử dụng để training model, nhưng chỉ một phần của các tập dữ liêu được gán nhãn chính xác. Điều này có nghĩa là các sự kiện quan trọng được xác định bằng cách sử dụng các phương pháp tổng hợp và học không giám sát thay vì dựa trên các tập dữ liệu được gán nhãn.

Sử dụng các kỹ thuật như clustering, classification và regression để xác định các sự kiện quan trọng trong video. Những phương pháp này dựa trên các đặc trưng thô của video như âm thanh, hình ảnh và văn bản để xác định các sự kiện quan trọng trong video.

Phương pháp weakly supervised learning được sử dụng để giảm bớt số lượng tập dữ liệu được gán nhãn chính xác cần thiết để đào tạo mô hình và đồng thời tăng tính linh hoạt của mô hình trong việc tạo ra video summarization chính

xác và hiệu quả. Tuy nhiên, một hạn chế của phương pháp này là độ chính xác của mô hình có thể bị giảm do sự thiếu chính xác của các tập dữ liệu được gán nhãn.

IV. Conclusion

Video summarization tựa như quá trình giảm thiểu độ dài của video bằng các tạo ra một video phiên bản tóm tắt đại diện cho nội dung của video. Việc tạo ra các video tóm tắt giúp người xem có thể nắm bắt nội dung dễ dàng hơn.

Tuy nhiên video summarization vẫn là một lĩnh vực nghiên cứu đầy thách thức, đặc biệt là với các video dài và phức tạp. Để tạo ra các tiểu cảnh tóm tắt video chất lượng cao, cần phải có dữ liệu huấn luyện đầy đủ và phù hợp để các mô hình học máy và học sâu có thể học được các đặc trưng quan trọng của video. Ngoài ra, các mô hình cần được đánh giá và tinh chỉnh kỹ lưỡng để đảm bảo tính hiệu quả và độ chính xác trong quá trình tóm tắt video.

Video summarization là một lĩnh vực đầy triển vọng và hứa hẹn mang lại nhiều tiện ích cho người dùng. Sự phát triển của các phương pháp và mô hình trong lĩnh vực này sẽ giúp cho việc tạo ra các tiểu cảnh tóm tắt video trở nên dễ dàng hơn, từ đó đáp ứng nhu cầu của người dùng trong việc tiết kiệm thời gian và tối ưu hóa trải nghiệm xem video.