XÁC ĐỊNH SỰ KIỆN VÀ MÔ TẢ ĐỒNG THỜI SỬ DỤNG MÔ HÌNH SEQUENCE-TO-SEQUENCE HUẤN LUYỆN TRƯỚC CHO BÀI TOÁN MÔ TẢ VIDEO DÀY ĐẶC

JOINT EVENT LOCALIZATION AND CAPTIONING USING PRETRAINED SEQUENCE-TO-SEQUENCE MODELS FOR DENSE VIDEO CAPTIONING

Thông tin nhóm

Nguyễn Hữu Sang- 22521242

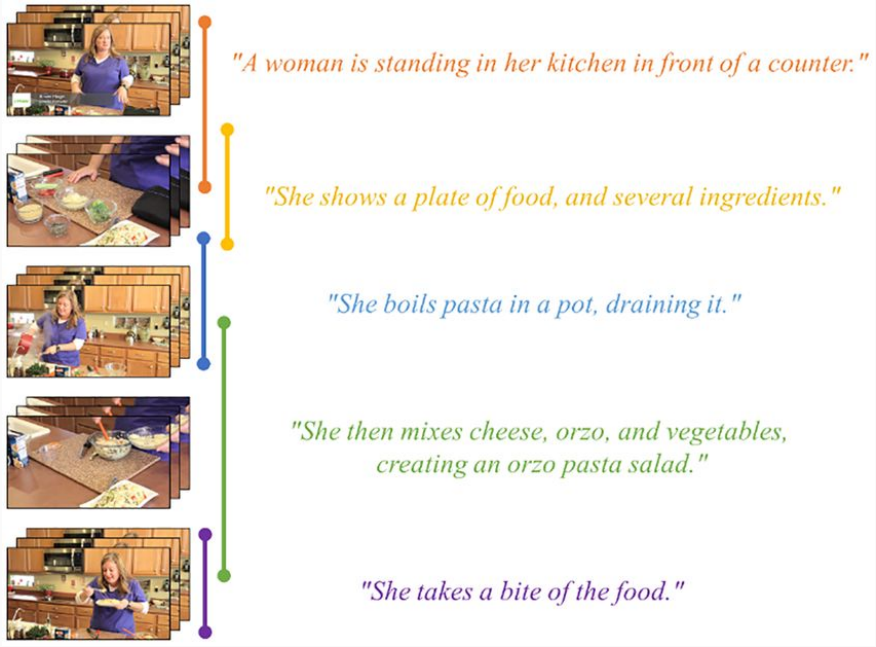
Huỳnh Ngọc Bảo Long 22520811

- Lớp: CS519.P11
- Link Github của nhóm: https://github.com/gryffin-uit/CS519.P11
- Link YouTube video: https://www.youtube.com/watch?v=n1w4DZmQbWk

Giới thiệu

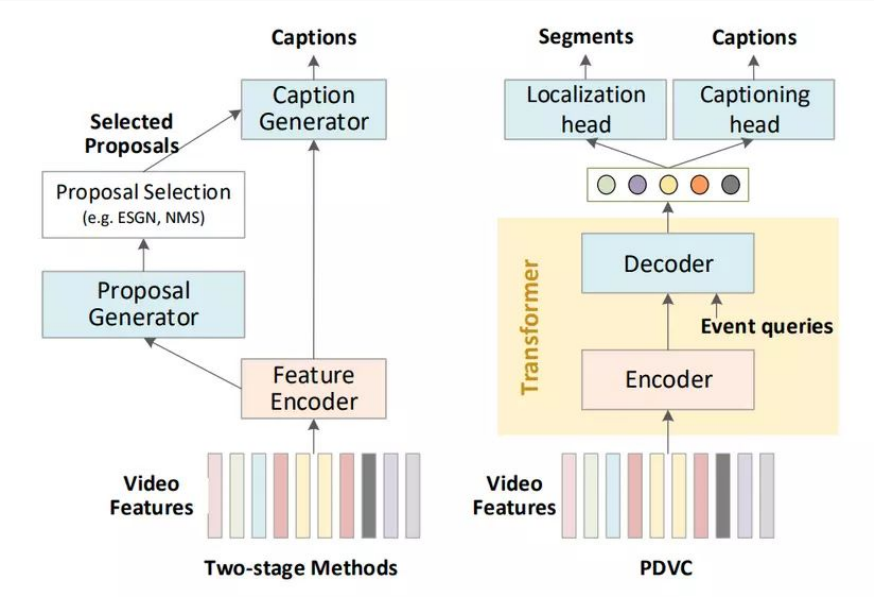
Trong bối cảnh ngày càng gia tăng lượng video được tạo ra và chia sẻ trên các nền tảng số, việc hiểu và khai thác nội dung video một cách tự động trở thành một thách thức quan trọng. Mô tả video dày đặc (dense video captioning, còn được viết là DVC) nổi lên như một bài toán trọng tâm, với mục tiêu xác định các sự kiện quan trọng trong video và tạo mô tả ngắn gọn, chính xác cho từng sự kiện.

Nhiệm vụ này không chỉ có giá trị trong lĩnh vực tìm kiếm và phân tích nội dung đa phương tiện mà còn hỗ trợ mạnh mẽ cho các ứng dụng như hệ thống hỗ trợ người khiếm thị, giám sát an ninh, và quản lý dữ liệu video trong các lĩnh vực như giáo dục hoặc y tế.



Giới thiệu

Tuy nhiên, các phương pháp hiện tại vẫn gặp nhiều hạn chế.



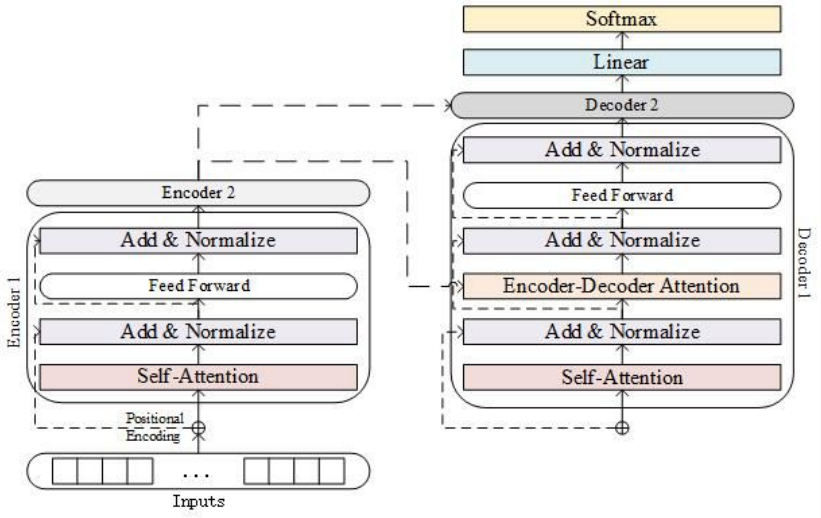
Cách tiếp cận hai giai đoạn (two-stage methods), trong đó các sự kiện được định vị trước khi mô tả, thiếu sự tương tác chặt chẽ giữa hai nhiệm vụ, dẫn đến hiệu suất chưa cao.

Ngay cả các phương pháp kết hợp (PDVC), vốn cố gắng tích hợp hai nhiệm vụ trong cùng một mô hình, vẫn phụ thuộc vào các thành phần đặc thù như bộ đếm sự kiện hoặc cơ chế phức tạp, làm tăng chi phí tính toán và giảm khả năng mở rộng

Vậy làm sao để khắc phục những nhược điểm hiện có?

Giới thiệu

Các mô hình *sequence-to-sequence* huấn luyện trước trên dữ liệu Web đã chứng minh hiệu quả trong việc xử lý đồng thời các chuỗi sự kiện và mô tả, với khả năng học từ dữ liệu không chú thích và giảm sự phụ thuộc vào các bộ dữ liệu thủ công



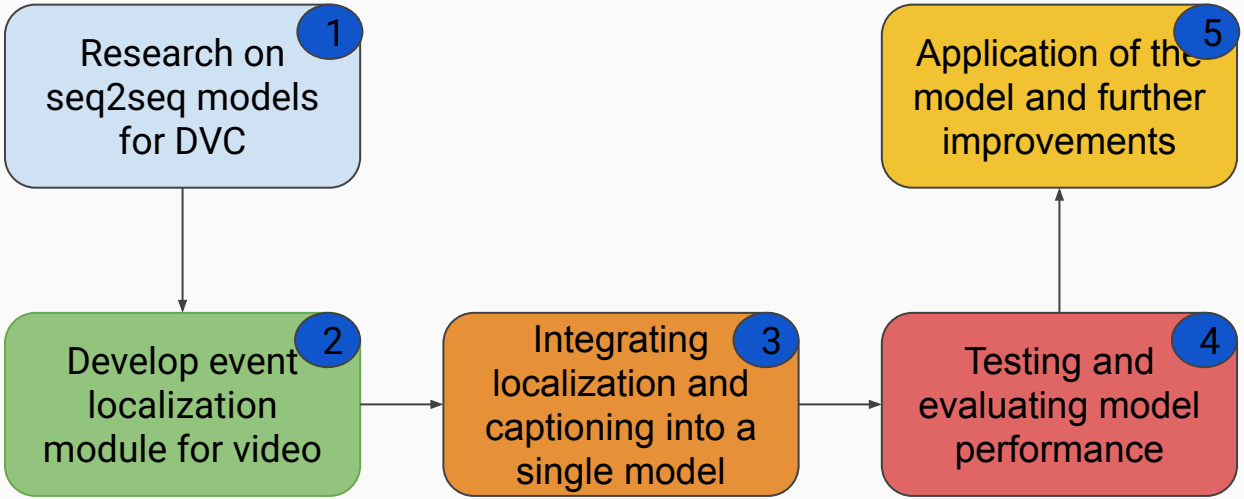
Model T5

Dựa trên các hạn chế hiện tại và những tiến bộ từ các mô hình *sequence-to-sequence*, việc đề xuất một phương pháp cải tiến ứng dụng các model này vào bài toán dense video captioning là một lựa chọn tiềm năng cao

Mục tiêu

- **Phát triển mô hình sequence-to-sequence tích hợp**: Xây dựng mô hình hiện đại có khả năng đồng thời định vị và mô tả chi tiết các sự kiện trong video, đảm bảo tính chính xác và mạch lạc theo thời gian.
- **Tận dụng dữ liệu video có lời thoại**: Sử dụng dữ liệu video tự nhiên, không được chú thích sẵn để giảm chi phí, đồng thời khai thác hiệu quả mối liên hệ giữa hình ảnh, âm thanh và ngôn ngữ.
- **Đánh giá được quá trình fine-tune**: Đo lường hiệu quả mô hình thông qua log-likelihood, nhằm đánh giá khả năng dự đoán chính xác và sự phù hợp của các mô tả được sinh ra so với dữ liệu thực tế.

Nội dung và Phương pháp



1. Nghiên cứu các mô hình pre-trained như T5, BART, hoặc CLIP ViT với tiềm năng áp dụng cho dense video captioning.

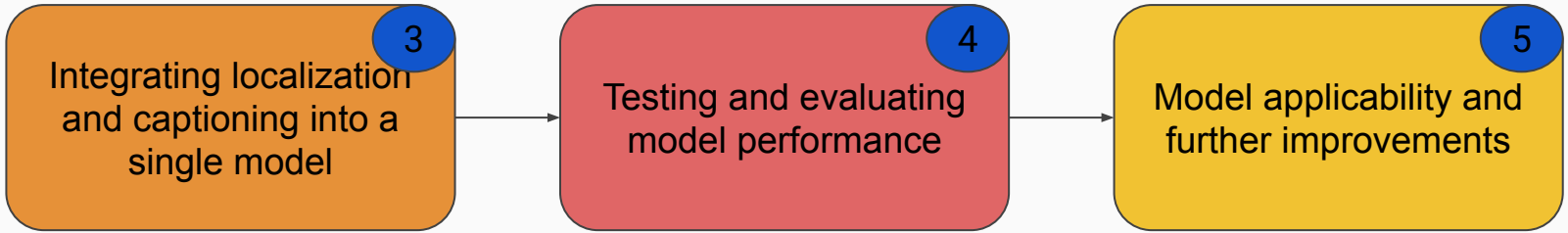
2. Thiết kế module định vị sự kiện trong video, dựa trên các phương pháp như sliding window, attention mechanism, hoặc transformer-based temporal models.

Nội dung và Phương pháp

3. Tích hợp module event localization vào pipeline sequence-to-sequence để thực hiện đồng thời việc định vị sự kiện và sinh caption.

4. Sử dụng các bộ dataset benchmark như ActivityNet Captions hoặc YouCook2 để thử nghiệm hiệu suất của mô hình. So sánh với các mô hình hiện có trong cùng tác vụ, tập trung vào cả thời gian xử lý và độ chính xác.

5. Khảo sát các ứng dụng thực tiễn như tự động tạo phụ đề cho video, phân tích nội dung video phục vụ giám sát, hoặc hỗ trợ người dùng khiếm thính.



Kết quả dự kiến

Kết quả nghiên cứu sẽ đạt được một mô hình sequence-to-sequence tiên tiến, tích hợp hiệu quả hai tác vụ **localization** và **captioning** để xử lý video một cách toàn diện. Cụ thể:

- Đối với tác vụ event localization:
  - Mô hình có khả năng định vị chính xác các sự kiện trong video, xác định thời gian bắt đầu và kết thúc của từng sự kiện với độ chính xác cao.
  - Các thuật toán tối ưu hóa được áp dụng để cải thiện chỉ số IoU (Intersection over Union) và đảm bảo hiệu quả trên các đoạn video có nội dung đa dạng.
- Đối với tác vụ captioning:
  - Mô hình có thể sinh ra các đoạn mô tả ngôn ngữ chi tiết, mạch lạc và ngữ nghĩa phong phú, phù hợp với các sự kiện đã được định vị.
  - Hiệu suất của captioning sẽ được đánh giá dựa trên các chỉ số như BLEU, METEOR, CIDEr và SPICE, với kỳ vọng đạt kết quả cạnh tranh so với các phương pháp hiện tại.
- Về tích hợp hai tác vụ:
  - Một pipeline đồng bộ được phát triển, trong đó các module localization và captioning hoạt động phối hợp hiệu quả.
  - Hệ thống đảm bảo khả năng xử lý video phức tạp, duy trì tính nhất quán giữa thời gian sự kiện và nội dung mô tả.
- Khả năng tổng quát hóa:
  - Mô hình được kỳ vọng có khả năng áp dụng rộng rãi trên các bộ dữ liệu video khác nhau, từ các video ngắn có nội dung cụ thể đến các video dài với ngữ cảnh phức tạp.
  - Hiệu quả trên các video chưa từng xuất hiện trong quá trình huấn luyện sẽ là một điểm nhấn, thể hiện khả năng tổng quát hóa và ứng dụng thực tế.
- Ứng dụng tiềm năng:
  - Cung cấp giải pháp tự động tạo phụ đề video chi tiết, hỗ trợ người dùng khiếm thính hoặc ngôn ngữ khác.
  - Hỗ trợ phân tích video trong các lĩnh vực như giám sát, giáo dục, giải trí, và truyền thông.

Tài liệu tham khảo

- [1]. Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In ICCV, 2017.
- [2]. Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In BMVC, 2020.
- [3]. Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In ICCV, 2017.
- [4]. Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In CVPR, 2021.
- [5]. Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In ICCV, 2021.
- [6]. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR, 2020.
- [7]. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In NeurIPS, 2022
- [8]. J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” arXiv preprint arXiv:2301.12597, 2023.