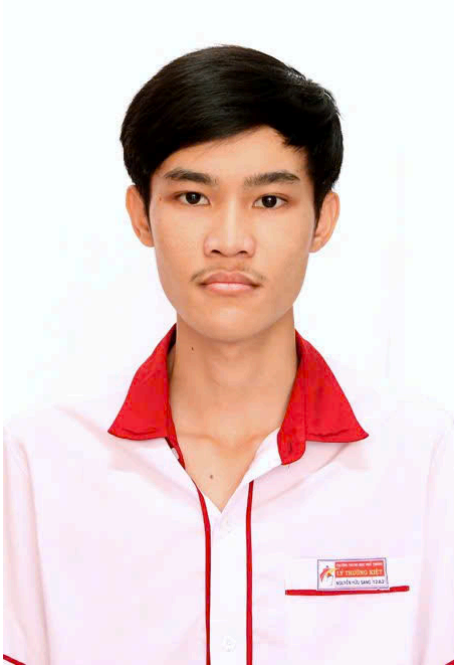
THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: https://www.youtube.com/watch?v=n1w4DZmQbWk
- Link slides: https://github.com/gryffin-uit/CS519.P11/blob/main/slides.pdf

<table>
<tr>
<td>

Họ và Tên: Nguyễn Hữu Sang
MSSV: 2252124


</td>
<td>

Lớp: CS519.P11
Tự đánh giá (điểm tổng kết môn): 9.5/10
Số buổi vắng: 0
Link Github: https://github.com/Samzaq/CS519.P11
Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:

Tìm hiểu đề tài, lên ý tưởng
Viết Đề cương nghiên cứu
Làm nội dung và thiết kế Slide

</td>
</tr>
<tr>
<td>

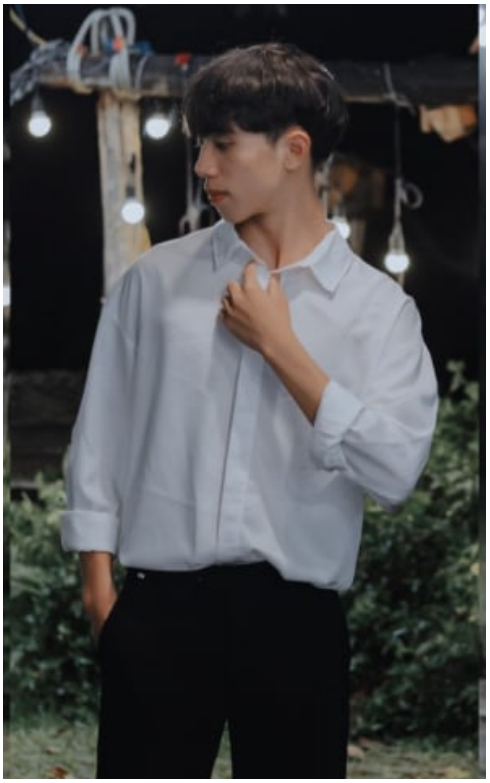
Họ và Tên: Huỳnh Ngọc Bảo Long
MSSV: 22520811

</td>
<td>

Lớp: CS519.P11
Tự đánh giá (điểm tổng kết môn): 9.5/10
Số buổi vắng: 1
Link Github: https://github.com/gryffin-uit/CS519.P11/
Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:

Tìm hiểu đề tài, lên ý tưởng

</td>
</tr>
</table>

	○ Viết Đề cương nghiên cứu ○ Làm nội dung và thiết kế Slide ○ Chỉnh sửa poster ○ Làm video thuyết trình

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI
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

TÊN ĐỀ TÀI TIẾNG ANH
JOINT EVENT LOCALIZATION AND CAPTIONING USING PRETRAINED SEQUENCE-TO-SEQUENCE MODELS FOR DENSE VIDEO CAPTIONING

TÓM TẮT

Bài toán mô tả video dày đặc (*dense video captioning*) [1] đóng vai trò quan trọng trong phân tích nội dung video, tìm kiếm thông tin đa phương tiện và các ứng dụng trí tuệ nhân tạo khác. Bài toán này yêu cầu xác định và mô tả tất cả các sự kiện diễn ra trong video, song các phương pháp hiện tại vẫn gặp nhiều hạn chế.

Cách tiếp cận hai giai đoạn [2, 3], 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 [4, 5], 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. Hơn nữa, cả hai cách tiếp cận đều dựa vào dữ liệu chú thích thủ công với quy mô hạn chế, như ActivityNet Captions hoặc YouCook2, khiến việc áp dụng trên các tập dữ liệu lớn và chưa chú thích gặp nhiều khó khăn.

Các mô hình *sequence-to-sequence* huấn luyện trước trên dữ liệu Web [6] 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. Ngoài ra, chúng giúp đơn giản hóa kiến trúc bằng cách tích hợp đồng thời hai nhiệm vụ định vị và mô tả sự kiện trong một hệ thống duy nhất. Ví dụ, các mô hình như Flamingo [7] hoặc BLIP-2 [8] đã thể hiện khả năng vượt trội trong các bài toán kết hợp giữa thị giác và ngôn ngữ, mở ra tiềm năng lớn cho *dense video captioning*.

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 dựa trên các mô hình này là cần thiết. Giải pháp mới không chỉ giải quyết các vấn đề hiện có mà còn hứa hẹn nâng cao hiệu quả và khả năng ứng dụng thực tế của bài toán *dense video captioning*.

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) 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.

Tuy nhiên, việc giải quyết bài toán này không hề đơn giản. Một số câu hỏi cơ bản đặt ra là: Làm thế nào để xác định ranh giới của các sự kiện trong một video không cắt? Làm sao để mô tả các sự kiện này bằng ngôn ngữ tự nhiên một cách chính xác và gắn kết? Và quan trọng hơn, làm sao để đạt được hiệu suất cao mà vẫn duy trì tính khả thi trong các tình huống thực tế, nơi dữ liệu được chú thích thủ công rất hạn chế?

Phần lớn các phương pháp hiện tại sử dụng cách tiếp cận hai giai đoạn, tách biệt giữa việc định vị sự kiện và tạo chú thích. Điều này dẫn đến việc thiếu sự tương tác chặt chẽ giữa hai nhiệm vụ, gây giảm hiệu quả tổng thể. Một số nghiên cứu gần đây đã đề xuất tích hợp hai nhiệm vụ này trong cùng một mô hình. Tuy nhiên, các mô hình này thường phức tạp, phụ thuộc nhiều vào dữ liệu chú thích thủ công hoặc các bộ cơ chế chuyên biệt, dẫn đến chi phí tính toán cao và khó áp dụng trên quy mô lớn.

Từ đó, nghiên cứu này đề xuất phát triển một hệ thống tự động tích hợp cả hai nhiệm vụ định vị và mô tả sự kiện trong một kiến trúc thống nhất. Lý do chính để chọn hướng tiếp cận này xuất phát từ các tiến bộ gần đây trong lĩnh vực học sâu, đặc biệt là các mô hình sequence-to-sequence được huấn luyện trước trên dữ liệu lớn. Các mô hình này đã chứng minh hiệu quả trong việc học các mối liên hệ phức tạp giữa các đầu vào và đầu ra đa phương thức mà không phụ thuộc mạnh vào dữ liệu chú thích.

Đầu vào và đầu ra:

- **Đầu vào**: Video không cắt (untrimmed video), bao gồm các khung hình (frames) và lời thoại (nếu có), được trích xuất dưới dạng dữ liệu trực quan và ngôn ngữ.
- **Đầu ra**: Một chuỗi token liên kết giữa các mô tả sự kiện và các timestamp (thời gian bắt đầu và kết thúc sự kiện), tạo thành một bản tóm tắt chi tiết của video.

Cách tiếp cận này không chỉ đơn giản hóa quy trình xử lý mà còn mở rộng khả năng khai thác dữ liệu video chưa chú thích, làm tăng tính khả thi và ứng dụng thực tiễn. Nghiên cứu không chỉ kỳ vọng cải thiện hiệu quả bài toán mà còn mở ra các hướng nghiên cứu mới trong lĩnh vực xử lý video và ngôn ngữ tự nhiên.

MỤC TIÊU

Chúng tôi đề ra mục tiêu nghiên cứu này gồm:

- 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

Nội dung 1: Tìm hiểu các pre-trained sequence-to-sequence models cho Dense Video Captioning.

- **Phương pháp thực hiện:**
  - Nghiên cứu các mô hình pre-trained như T5, BART, hoặc CLIP ViT với tiềm năng áp dụng cho dense video captioning.
  - Phân tích khả năng của các mô hình này trong việc xử lý dữ liệu video và kết nối giữa ngữ cảnh thời gian (temporal context) với các mô tả ngôn ngữ.
  - Đánh giá các mô hình dựa trên khả năng sinh caption chi tiết và phù hợp với các đoạn video đã được định vị.
- Kết quả dự kiến:
  - Đánh giá sơ bộ các ưu, nhược điểm của từng mô hình trong tác vụ dense video captioning.
  - Xác định mô hình phù hợp để thử nghiệm với các bước tiếp theo.

Nội dung 2: Phát triển module event localization chính xác cho video.

- **Phương pháp thực hiện:**
  - Thiết kế module định vị sự kiện trong video, dựa trên các phương pháp như sliding window, attention mechanism, hoặc transformer-based temporal models.
  - Kết hợp dữ liệu video (visual features) và lời thoại (audio/text features) để cải thiện độ chính xác khi định vị thời gian bắt đầu và kết thúc của sự kiện.
  - Tối ưu hóa module localization bằng cách huấn luyện trên các bộ dữ liệu có timestamp và caption đi kèm.
- Kết quả dự kiến:
  - Xây dựng được module localization có khả năng định vị chính xác các sự kiện trong video.
  - Báo cáo độ chính xác dựa trên các thước đo như IoU (Intersection over Union) và mAP (mean Average Precision).

Nội dung 3: Tích hợp localization và captioning vào một mô hình chung.

- **Phương pháp thực hiện:**
  - Tích hợp module event localization vào pipeline sequence-to-sequence để thực hiện đồng thời việc định vị sự kiện và sinh caption.
  - Sử dụng attention mechanism để kết nối các đoạn video đã được định vị với ngữ cảnh ngôn ngữ, đảm bảo tính mạch lạc của caption.
  - Tinh chỉnh pipeline bằng cách sử dụng dữ liệu video thực tế có đa dạng sự kiện và nội dung phức tạp.
- Kết quả dự kiến:
  - Phát triển mô hình tích hợp hoạt động hiệu quả trên cả hai tác vụ localization và captioning.
  - Đo lường hiệu suất bằng cách đánh giá độ chính xác trong localization và chất lượng caption thông qua BLEU, METEOR, CIDEr.

Nội dung 4: Thử nghiệm và đánh giá hiệu năng mô hình

- **Phương pháp thực hiện:**
  - 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.

- Phân tích lỗi (error analysis) để tìm ra các yếu tố ảnh hưởng đến hiệu năng, đặc biệt là ở các trường hợp video phức tạp hoặc caption không rõ ràng.

- Kết quả dự kiến:
  - Xác định hiệu suất tổng thể của mô hình so với các phương pháp khác.
  - Đề xuất hướng cải thiện dựa trên các kết quả thử nghiệm.

Nội dung 5: Khả năng ứng dụng và mở rộng mô hình.

- **Phương pháp thực hiện:**
  - 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.
  - Tìm hiểu tiềm năng mở rộng mô hình để xử lý các tác vụ liên quan như video summarization hoặc multi-modal translation.
  - Xem xét các thách thức trong việc triển khai thực tế, như xử lý video thời gian thực hoặc giảm chi phí tính toán.
- Kết quả dự kiến:
  - Đề xuất ứng dụng tiềm năng của mô hình vào các lĩnh vực khác nhau.
  - Báo cáo các thách thức và giải pháp nhằm triển khai mô hình trong thực tế.

KẾT QUẢ MONG ĐỢI

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.
- **Đóng góp cho cộng đồng nghiên cứu:**
    - Mô hình và phương pháp luận mới sẽ góp phần thúc đẩy các nghiên cứu tiếp theo trong lĩnh vực xử lý ngôn ngữ và video đa phương thức.
    - Kết quả nghiên cứu sẽ được công bố, kèm theo bộ mã nguồn và tài liệu để hỗ trợ cộng đồng.

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]. Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In CVPR, 2018.

[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.