CẢI TIẾN KHÔNG GIAN NHÚNG CHO MÔ TẢ ẢNH DỰA TRÊN VĂN BẢN

Bùi Cao Doanh - 19521366
Trịnh Thị Thanh Trúc - 19521059
Nguyễn Hiếu Nghĩa - 19520178

Tóm tắt

- Lớp: CS519.M11
- Link Github của nhóm: https://github.com/caodoanh2001/CS519.M11
- Link YouTube video:

Bùi Cao Doanh

Trịnh Thị Thanh Trúc

Nguyễn Hiếu Nghĩa

Giới thiệu

- Định nghĩa mô tả ảnh dựa trên văn bản: Đầu vào của bài toán là một bức ảnh, đầu ra là câu mô tả thích hợp với ngữ cảnh bức ảnh, tuy nhiên có sử dụng các thông tin về văn bản để giúp câu mô tả trở nên trọn vẹn nghĩa hơn.

Đầu vào

Mô hình

The player in the white shirt is competing for the ball with the player wearing the red number 10

Đầu ra

Giới thiệu

- Đây là một bài toán rất mới (chỉ mới được bắt đầu nghiên cứu từ năm 2020).
- Chỉ có một bộ dữ liệu được ra mắt tính tới thời điểm này (TextCaps [1]).
- Việc chọn văn bản nào trong ảnh để đề cập trong câu mô tả không phải là một vấn đề đơn giản.
- Bao gồm nhiều bài toán con: Scene text detection / recognition, Image Captioning.

Mục tiêu

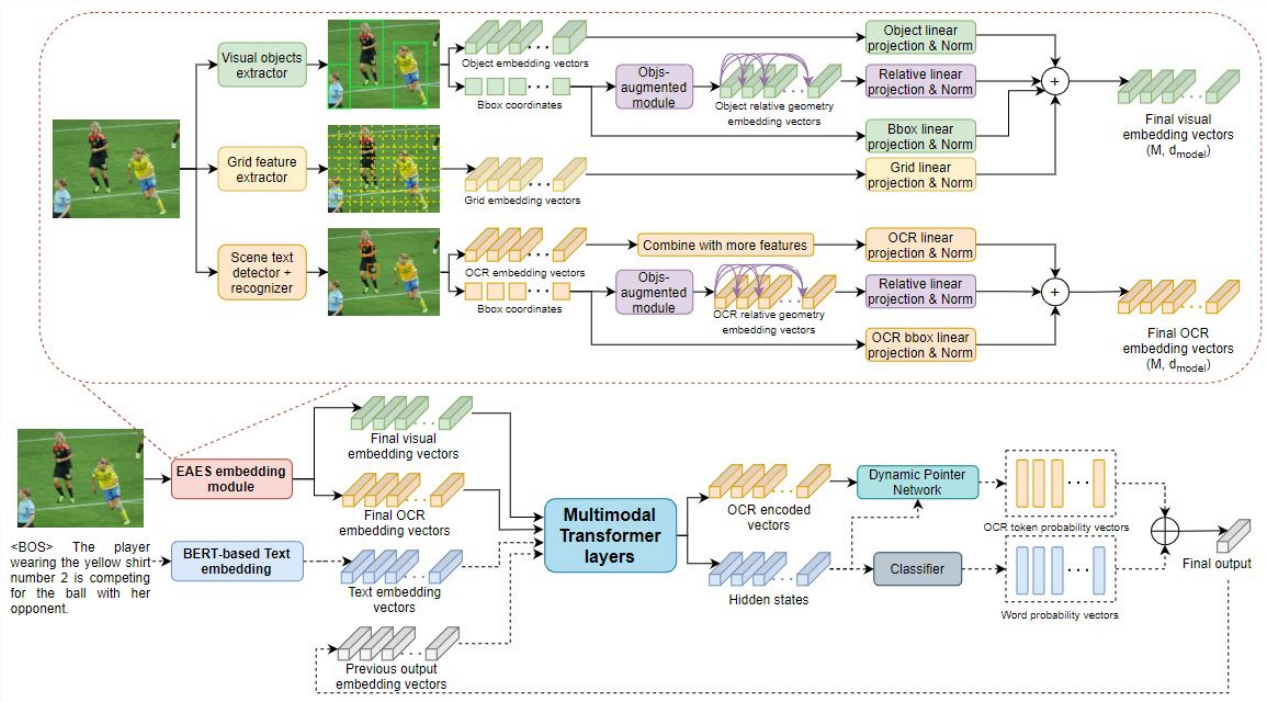
- Khảo sát các hướng tiếp cận phổ biến cho bài toán mô tả ảnh.
- Khảo sát toàn bộ các công trình hiện có cho bài toán mô tả ảnh dựa trên văn bản (5 công trình).
- Đề xuất một không gian nhúng mới cung cấp thông tin về mối quan hệ giữa các đối tượng và OCR token trên ảnh dựa trên tọa độ.
- Đề xuất kết hợp một đặc trưng lưới tăng cường ngữ cảnh toàn cục của ảnh.

Nội dung và Phương pháp

- Nội dung:
  - Khảo sát, chỉ ra hạn chế của các nghiên cứu trước trên bài toán mô tả ảnh dựa trên văn bản.
  - Tìm hiểu phương pháp M4C-Captioner [2] - baseline phổ biến cho bài toán mô tả ảnh dựa trên văn bản.
  - Đặt giả thiết: liệu có một đặc trưng nào đó tốt hơn biểu diễn được thông tin vị trí của các đối tượng có trong ảnh và ngữ cảnh toàn cục của ảnh ?
  - Tìm hiểu các dạng đặc trưng lưới và đặc trưng vị trí của các đối tượng trong ảnh, làm cơ sở kết hợp tạo thành không gian nhúng tăng cường.

Nội dung và Phương pháp

- Phương pháp:



Nội dung và Phương pháp

- Phương pháp:
  - Để khảo sát các phương pháp cho bài toán mô tả ảnh truyền thống, chúng tôi chủ yếu tiếp cận các công trình khảo sát.
  - Để khảo sát các phương pháp cho bài toán mô tả ảnh dựa trên văn bản, chúng tôi khảo sát 5 công trình hiện có cho bài toán này [3][4][5][6][7] được đăng trên các hội nghị, tạp chí uy tín (CVPR, AAAI, MM, JIFS).
  - Để trích xuất đặc trưng vị trí, chúng tôi thử thực hiện các phép tính giống ở công trình [8].
  - Để trích xuất đặc trưng lưới, chúng tôi thử sử dụng mô hình pre-trained ở công trình [9].

Kết quả dự kiến

- Nếu thử nghiệm thành công, chúng tôi sẽ đặt tên phương pháp là EAES (Effective Augmented Embedding Spaces).
- 2 công bố quốc tế:
  - Hội nghị quốc tế (NICS, RIVF hoặc KSE)
  - Tạp chí Q3
- Cuốn báo cáo tổng hợp các nội dung nghiên cứu.
- Một ứng dụng minh họa trực quan hóa nghiên cứu.

Tài liệu tham khảo

[1] Sidorov, Oleksii, et al. "Textcaps: a dataset for image captioning with reading comprehension." European Conference on Computer Vision. Springer, Cham, 2020.
[2] Hu, Ronghang, et al. "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[3] Xu, Guanghui, et al. "Towards Accurate Text-based Image Captioning with Content Diversity Exploration." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[4] Wang, Jing, Jinhui Tang, and Jiebo Luo. "Multimodal attention with image text spatial relationship for ocr-based image captioning." Proceedings of the 28th ACM International Conference on Multimedia. 2020.
[5] García, Rafael Gallardo, et al. "Searching for memory-lighter architectures for OCR-augmented image captioning." Journal of Intelligent & Fuzzy Systems Preprint: 1-12.
[6] Wang, Zhaokai, et al. "Confidence-aware Non-repetitive Multimodal Transformers for TextCaps." arXiv preprint arXiv:2012.03662 (2020).
[7] Gallardo García, Rafael, et al. "Towards Multilingual Image Captioning Models that Can Read." Mexican International Conference on Artificial Intelligence. Springer, Cham, 2021.
[8] Guo, Longteng, et al. "Normalized and geometry-aware self-attention network for image captioning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[9] Jiang, Huaizu, et al. "In defense of grid features for visual question answering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.