OFA: UNIFYING ARCHITECTURES, TASKS, AND MODALITIES THROUGH A SIMPLE SEQUENCE-TO-SEQUENCE LEARNING FRAMEWORK

Nguyễn Vũ Dương - 20520465

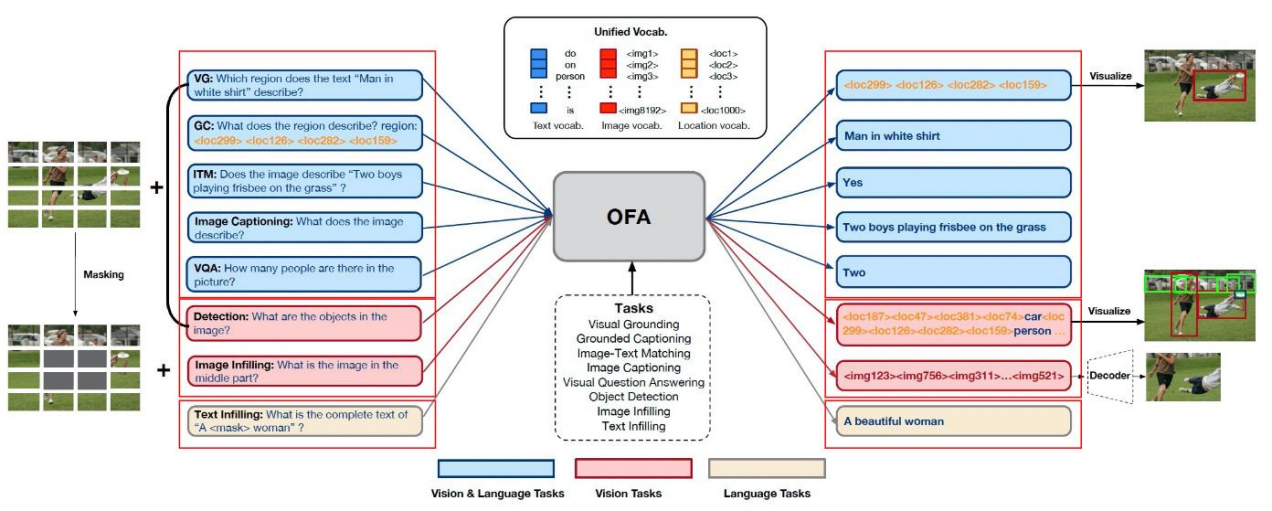
Tóm tắt

- Lớp: CS519.O11
- Link Github : https://github.com/duongve13112002/CS519.O11
- Link YouTube video: https://youtu.be/vDnrDrV8L5U
- Ảnh + Họ và Tên của các thành viên

Nguyễn Vũ Dương

Giới thiệu

Xây dựng một mô hình toàn năng có thể xử lý nhiều bài toán và dữ liệu khác nhau như con người là một mục tiêu hấp dẫn trong cộng đồng AI. Khả năng đạt được mục tiêu này có thể phụ thuộc nhiều vào việc liệu rằng các dữ liệu, bài toán và quá trình huấn luyện có thể được biểu diễn bằng chỉ một số ít hình thức, có thể được thống nhất và quản lý bởi một mô hình hoặc hệ thống duy nhất hay không.



Mục tiêu

- Nghiên cứu, khảo sát các hướng tiếp cận hiện có việc xây dựng mô hình có thể xử lý đa tác vụ.
- Đề xuất một framework không phụ thuộc vào task, modality và Task Comprehensiveness. Có thể giải quyết nhiều tác vụ bao gồm: text-to-image generation, visual grounding, visual question answering (VQA), image captioning, image classification, language modeling, ..., thông qua một framework học sequence-to-sequence đơn giản với một unified instruction-based task.
- Đánh giá mô hình với các SOTA trên các tác vụ tương ứng và xây dựng chương trình demo trực quan hóa nghiên cứu

Nội dung và Phương pháp

Nội dung 1: Nghiên cứu, khảo sát các hướng tiếp cận hiện có việc xây dựng mô hình có thể xử lý đa tác vụ.

- Đọc các bài báo liên quan đến multimodal model trong các hội nghị lớn như ICML, IJCAI,...
- Tìm các phương thức của các mô hình hiện có về cách họ xử lý nhiều loại dữ liệu khác nhau.
- Hầu hết thì các mô hình dạng này đều phải được tinh chỉnh dựa trên các mô hình có sẵn mới có thể đạt được độ hiệu quả cao.

Nội dung và Phương pháp

Nội dung 2: Xây dựng model: một framework sequence-to-sequence đơn giản mà không cần một tinh chỉnh từ một pretrain model khác

- Sử dụng mô hình Transforner encoder-decoder để xử lý các nhiệm vụ khác nhau
- Không sử dụng thêm các thành phần có thể huấn luyện được trong quá trình pretraining hoặc finetuning
- Tuân theo Normformer để thêm hai lớp chuẩn hóa (LN) và Hasdacale attention để ổn định quá trình đào tạo và tăng tốc độ hội tụ.

Nội dung và Phương pháp

Nội dung 2: Xây dựng model: một framework sequence-to-sequence đơn giản mà không cần một tinh chỉnh từ một pretrain model khác.

Với 3 đặc điểm

Task Agnostic

Biểu diễn tác vụ thống nhất

Modality Agnostic

Biểu diễn thống nhất input và ouput

Task Comprehensiveness

Đủ đa dạng về nhiệm vụ/bài toán để có khả năng tổng quát hóa một cách mạnh mẽ

Nội dung và Phương pháp

Nội dung 3: Đánh giá và so sánh với các mô hình SOTA khác trên từng lĩnh vực

- Natural Language Understanding: RoBERTa, ELECTRA và DeBERTa
- Natural Language Generation: UniLM, Pegasus và ProphetNet
- Image Classification: MoCo-v3, BEiT và MAE.
- Image Captioning: VL-T5, OSCAR, LEMON và SimVLM.
- VQA and Visual Entailment: Florence, SimVLM, VLMo và METER.
- Text-to-Image Generation: DALLE, CogView, GLIDE, Unifying.

Kết quả dự kiến

- Xây dựng mô hình hoàn chỉnh.
- Mô hình giải quyết được nhiều bài toán và dữ liệu tuy nhiên vẫn đạt được độ hiệu quả và thậm chí tốt hơn so với SOTA trên từng lĩnh vực
- Báo cáo phương pháp và kỹ thuật của mô hình đã phát triển, kết quả thực nghiệm, đánh giá
- Một demo thực hiện một vài bài toán của mô hình để trực quan

Tài liệu tham khảo

[1]. Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In International Conference on Machine Learning, PMLR, 23318–23340.

[2]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30, (2017).

[3]. Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020).