


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo:

<https://youtu.be/vDnrDrV8L5U>

- Link slides:

<https://github.com/duongve13112002/CS519.O11>

<ul style="list-style-type: none">● Họ và Tên: Nguyễn Vũ Dương● MSSV: 20520465 	<ul style="list-style-type: none">● Lớp: CS519.O11● Tự đánh giá (điểm tổng kết môn): 8.5/10● Số buổi vắng: 1● Số câu hỏi QT cá nhân: 15● Link Github: https://github.com/duongve13112002/● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng cho bài đề xuất○ Viết bài đề xuất, poster và thiết kế slide thuyết trình○ Làm video YouTube
--	--

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI

OFA: THỐNG NHẤT KIẾN TRÚC, NHIỆM VỤ VÀ MÔ HÌNH THÔNG QUA MỘT FRAMEWORK HỌC TẬP SEQUENCE-TO-SEQUENCE ĐƠN GIẢN

TÊN ĐỀ TÀI TIẾNG ANH

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

TÓM TẮT

Đề tài, đề xuất một framework mới để thống nhất các kiến trúc, nhiệm vụ và mô thức khác nhau trong lĩnh vực AI, đặc biệt là Thị giác máy tính và Xử lý ngôn ngữ tự nhiên. OFA (One-For-All) được giới thiệu như một **sequence-to-sequence learning framework** đơn giản để giải quyết nhiều dạng bài toán và các kiểu dữ liệu khác nhau. Mô hình sử dụng kiến trúc transformer đơn giản như encoder-decoder để học các biểu diễn chung (general representation). Sau đó, so sánh độ hiệu quả của mô hình này trên từng các bài toán khác nhau (VD: Natural Language Understanding, Image Classification, Image Captioning, VQA and Visual Entailment, Natural Language Generation, Text-to-Image Generation, ...) với các mô hình SOTA và phổ biến trên từng bài toán đó. Mục tiêu, nội dung và phương pháp để hiện thực hóa mục tiêu và kết quả mong đợi đã được vạch ra trong bài đề xuất này.

GIỚI THIỆU

Xây dựng một mô hình đa năng có thể xử lý nhiều bài toán và loại dữ liệu như con người là một mục tiêu hấp dẫn trong cộng đồng AI. Khả năng đạt được mục tiêu này có thể phụ thuộc phần lớn vào việc liệu đa dạng về dữ liệu, bài toán và quá trình huấn luyện có thể được biểu diễn chỉ bằng một vài dạng có thể được hợp nhất và quản lý bởi một mô hình hoặc hệ thống duy nhất hay không.

Những phát triển gần đây của kiến trúc Transformer đã cho thấy tiềm năng của nó để trở thành một công cụ tính toán đa năng. Trong các thiết lập của học có giám sát, mô

hình " pretraining - finetuning" đạt được thành công xuất sắc trong nhiều lĩnh vực. Trong few-/zero-shot learning, các mô hình ngôn ngữ với điều chỉnh prompt/instruction chứng tỏ là có khả năng học mạnh mẽ. Những tiến bộ này đã mang đến những cơ hội quan trọng hơn bao giờ hết cho sự xuất hiện của một mô hình toàn năng (omni-model).

Để hỗ trợ khái quát hóa tốt hơn cho các vấn đề mở trong khi vẫn duy trì hiệu suất và dễ sử dụng, chúng tôi cho rằng một mô hình toàn năng nên có ba đặc điểm sau:

- **Không phụ thuộc vào bài toán (Task-Agnostic):** Biểu diễn bài toán thống nhất để hỗ trợ các loại bài toán khác nhau.
- **Không phụ thuộc vào dữ liệu (Modality-Agnostic):** Biểu diễn đầu vào và đầu ra thống nhất được chia sẻ giữa tất cả các tác vụ để xử lý các dữ liệu khác nhau.
- **Bao quát bài toán (Task Comprehensiveness):** Đa dạng về bài toán để tích lũy khả năng khái quát hóa một cách mạnh mẽ.

Do đó, chúng tôi đề xuất mô hình OFA có đầy đủ 3 đặc điểm.

MỤC TIÊU

- **Thống nhất một kiến trúc:** OFA cho phép sử dụng một kiến trúc mô hình duy nhất cho nhiều loại tác vụ khác nhau, bao gồm cả uni-modal và cross-modal. Điều này giúp giảm bớt thời gian và công sức cần thiết để phát triển các mô hình mới.
- **Giải quyết được nhiều bài toán:** OFA có thể xử lý một loạt các bài toán khác nhau bằng cách chỉ cần thay đổi định dạng đầu vào và đầu ra của mô hình. Điều này làm cho OFA trở nên linh hoạt và có thể được áp dụng cho nhiều ứng dụng khác nhau.
- **Xử lý được nhiều dữ liệu:** OFA có thể xử lý các loại dữ liệu khác nhau. Điều này làm cho OFA trở nên hữu ích cho các ứng dụng multimodal.

NỘI DUNG VÀ PHƯƠNG PHÁP

NỘI DUNG

Việc đáp ứng các đặc điểm đề cập ở phần giới thiệu trong khi vẫn duy trì hiệu suất vượt trội trong các tác vụ là một thách thức. Các mô hình ngôn ngữ và đa mô hình được huấn luyện trước hiện tại dễ dàng thất bại ở một phần của các đặc điểm này, do các yếu tố sau đây:

- **Các thành phần cụ thể cho từng bài toán:** Các thành phần này khiến mô hình trở nên kém linh hoạt và không tương thích với các bài toán chưa từng thấy.
- **Sự không nhất quán giữa các giai đoạn huấn luyện:** Các công thức và mục tiêu khác nhau được sử dụng cho retraining, fine-tuning, và zero-shot, dẫn đến hiệu suất không tối ưu.
- **Biểu diễn đan xen:** Việc tích hợp phát hiện đối tượng với các bài có dữ liệu đóng cản trở việc khái quát hóa sang dữ liệu mở.

OFA có thể giải quyết được vì:

- OFA sử dụng một handcrafted instructions để thống nhất các retraining và fine-tuning tasks theo dạng sequence-to-sequence
- Transformer được sử dụng như một công cụ tính toán chung cho tất cả các dữ liệu, không có thành phần cụ thể cho từng bài toán hoặc dữ liệu nào được thêm vào.
- OFA được pretrain trên nhiều uni-modal and cross-modal tasks để đảm bảo tính bao quát nhiệm vụ.

PHƯƠNG PHÁP

Trong huấn luyện tiền đa phương thức, việc kết hợp thông tin hình ảnh và ngôn ngữ là một thách thức. Để đơn giản và tiết kiệm tài nguyên, thay vì trích xuất các đặc trưng phức tạp cho từng đối tượng, có thể sử dụng trực tiếp các mô đun ResNet để biến đổi ảnh x_v kích thước $R^{H \times W \times C}$ thành P đặc trưng patch có kích thước ẩn. Về xử lý thông tin ngôn ngữ, có thể áp dụng mã hóa byte-pair (BPE) cho chuỗi văn bản thành chuỗi các đơn vị con, sau đó nhúng chúng thành các đặc trưng.

Để xử lý các phương thức khác nhau mà không cần lược đồ đầu ra cụ thể cho từng tác vụ, việc biểu diễn dữ liệu của các phương thức khác nhau trong một không gian thống

nhất là rất quan trọng. Một giải pháp khả thi là làm rời rạc văn bản, hình ảnh và đối tượng, sau đó biểu diễn chúng bằng các token trong một bộ từ vựng thống nhất. Về biểu diễn hình ảnh, có thể sử dụng mã hóa thưa thớt để giảm độ dài chuỗi biểu diễn hình ảnh. Ví dụ, một hình ảnh có độ phân giải 256x256 được biểu diễn thành một chuỗi mã có độ dài 16x16. Mỗi mã rời rạc có tương quan mạnh với patch tương ứng.

Về biểu diễn đối tượng trong hình ảnh, có thể biểu diễn các đối tượng như một chuỗi các token rời rạc. Cụ thể, đối với mỗi đối tượng, trích xuất nhãn của nó và vùng giới hạn của nó. Các tọa độ góc liên tục (góc trên bên trái và góc dưới bên phải) của vùng giới hạn được rời rạc thành các số nguyên làm token vị trí $\langle x_1, y_1, x_2, y_2 \rangle$. Đối với nhãn đối tượng, chúng là các từ và do đó có thể được biểu diễn bằng token BPE.

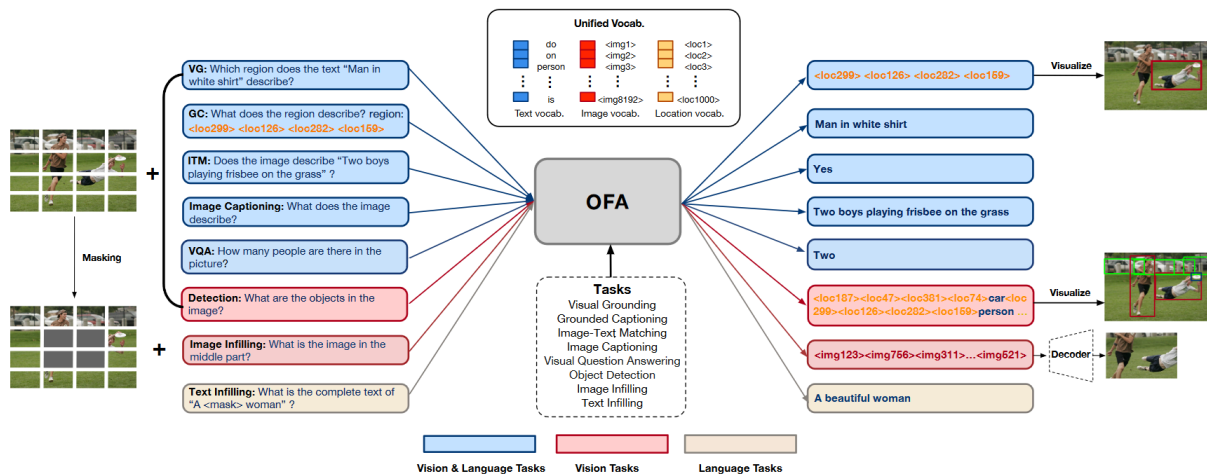
Cuối cùng, chúng tôi sử dụng một bộ từ vựng thống nhất cho tất cả các token ngôn ngữ và hình ảnh, bao gồm các đơn vị con, mã hình ảnh và token vị trí.

Về kiến trúc mô hình, chúng tôi có thể chọn Transformer làm kiến trúc chính và sử dụng khung encoder-decoder làm kiến trúc thống nhất cho tất cả các tác vụ huấn luyện tiền, tinh chỉnh và zero-shot. Cụ thể, cả encoder và decoder đều là các chồng các lớp Transformer.

Mục tiêu của chúng tôi là thiết kế một khung mô hình thống nhất tương thích với nhiều loại dữ liệu và bài toán khác nhau, cho phép mô hình tổng hợp và xử lý các bài toán mới chưa từng thấy trong cùng một mô hình. Để đạt được điều này, việc biểu diễn các bài toán tiềm năng liên quan đến các dữ liệu khác nhau trong một khuôn mẫu thống nhất là rất quan trọng. Do đó, thiết kế các pretraining task phải đảm bảo tính multitask và multimodality.

Để thống nhất các bài toán và dữ liệu, chúng tôi thiết kế một khung học tập Seq2Seq thống nhất cho quá trình tiền pretraining, fine-tuning và suy luận trên tất cả các bài toán liên quan đến các dữ liệu khác nhau. Cả các pretraining task và các downstream tasks của việc hiểu và tạo sinh cross-modal và uni-modal đều được hình thành dưới dạng Seq2Seq. Bằng cách này, chúng tôi có thể thực hiện pretraining trên dữ liệu multimodal và uni-modal để mang lại cho mô hình khả năng toàn diện. Cụ thể, chúng

tôi sử dụng cùng một lược đồ cho tất cả các tác vụ, đồng thời chỉ định các hướng dẫn được tạo thủ công để phân biệt.



KẾT QUẢ MONG ĐỢI

- Xây dựng một mô hình hoàn chỉnh dựa trên các ý tưởng trên.
- Mô hình giải quyết nhiều bài toán nhưng vẫn đạt được hiệu suất hiệu quả, thậm chí vượt trội hơn cả trạng thái nghệ thuật trong từng lĩnh vực.
- Báo cáo về các phương pháp và kỹ thuật của mô hình phát triển, kết quả thực nghiệm và đánh giá.
- Một bản demo để thực hiện một số bài toán của mô hình để trực quan hóa.

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).