THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo: https://youtu.be/NDB7PZ1zC68
- Link slides:

• Họ và Tên: Võ Huy Khôi • MSSV: 220101042 	• Lớp: CS2205.APR2023 • Tự đánh giá (điểm tổng kết môn): 8.5/10 • Số buổi vắng: 2 • Link Github: https://github.com/HuyKhoiGrad • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: ○ Lên ý tưởng đề tài ○ Làm đề tài ○ Thuyết trình đề tài

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TỐI ƯU PHÂN ĐOẠN ĐỐI TƯỢNG TRONG VIDEO DỰA TRÊN KIẾN TRÚC BIẾN HÌNH ĐA HÌNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

DEFORMABLETRANSVOS: OPTIMAL VIDEO OBJECT SEGMENTATION WITH DEFORMABLE TRANSFORMER

TÓM TẮT *(Tối đa 400 từ)*

Phân đoạn vật thể trong video là bài toán trong lĩnh vực Thị giác máy tính có tính ứng dụng rất cao trong các lĩnh vực chỉnh sửa, phân tích video và vận hành xe tự lái. Nhiệm vụ chính là tách biệt được đối tượng chính (foreground object) và ngữ cảnh nền (background object) trong một chuỗi các khung hình của một đoạn video. Vì vậy bài toán đối mặt với rất nhiều thách thức vì đối tượng trong video sẽ biến đổi rất nhiều giữa các khung hình trong xuyên suốt video. Để có thể giải quyết bài toán, cần phải khám phá được sự phụ thuộc giữa các pixels trong cùng một khung hình và sự phụ thuộc giữa các khung hình theo thời gian. Hai sự phụ thuộc này theo thứ tự có thể được biểu diễn bằng mối quan hệ Không gian (Spatial Relationship) và mối quan hệ Thời gian (Temporal Relationship). Cơ chế Attention (Tập trung) cùng với kiến trúc Transformer [2] (kiến trúc biến hình) ra đời ban đầu nhằm mục đích khám phá các mối quan hệ Thời gian trong lĩnh vực Xử lý Ngôn ngữ tự nhiên và đã cho thấy sự thành công khi áp dụng vào lĩnh vực Nhận diện đối tượng trong ảnh với khả năng làm rõ các mỗi quan hệ Không gian giữa các pixels trong cùng một tấm hình với mô hình. Từ đó, kiến trúc Transformer đã cho thấy tiềm năng trong việc giải quyết bài toán Phân đoạn vật thể trong video và đã được áp dụng thành công với mô hình TransVOS [1]. Tuy nhiên, mô hình vẫn còn gặp một vài hạn chế trong khối lượng phép tính và tốc độ xử lý do cơ chế tính toán của kiến trúc Transformer. Trong nghiên cứu này, nhóm sẽ đề xuất một phương pháp có tiềm năng giải quyết được các vấn đề của

Transformer cho bài toán Phân đoạn vật thể trong video bằng cơ chế Derformable Attention [4] (Tập trung đa hình).

GIỚI THIỆU *(Tối đa 1 trang A4)*

Video là một loại dữ liệu đa phương tiện gồm một loạt các khung hình được sắp xếp theo thứ tự và thông thường sẽ có những đối tượng chính thể hiện nội dụng mà video muốn truyền tải. Nguồn dữ liệu video hiện nay là rất lớn với lượng thông tin mà video cung cấp là vô cùng khổng lồ. Để có thể phân tích, lưu trữ và sử dụng nguồn dữ liệu video hiệu quả, nội dung thông tin của video có thể được khai thác thông qua việc phân đoạn và theo dõi đối tượng chính trong video. Phân đoạn đối tượng trong video (Video Object Segmentation hay VOS) là một trong những bài toán cơ bản trong lĩnh vực thị giác máy tính với mục tiêu là nhận diện và phát hiện các đối tượng trong một đoạn video ở mức độ pixels. Trong đó, các pixels biểu diễn đối tượng khác nhau cũng được được phân biệt với nhau và những pixels không liên quan đến đối tượng được xếp chung vào thành vùng nền, từ đó xây dựng nên object segmentation mask.

(A) Ảnh RGB gốc

(B) Object segmentation mask

Kiến trúc Transformer với khả năng khai thác thông tin giữa các pixels trong ảnh và giữa các khung hình đã được ứng dụng thành công vào bài toán Phân đoạn vật thể trong video với mô hình TransVOS. Tuy nhiên mô hình sử dụng cơ chế Multi-head Attention [2] vẫn còn những hạn chế về mặt tính toán và tốc độ dự đoán. Vấn đề này đã được đề cập trong bài toán Nhận diện đối tượng trong ảnh với mô hình DETR [3] và đã được giải quyết bằng cơ chế Deformable Attention trong kiến trúc của Derformable Transformer [4].

Trong bài nghiên cứu này, nhóm sử dụng kiến trúc Deformable Transformer để tối ưu cho kiến trúc Transformer ứng dụng vào bài toán Phân đoạn vật thể trong video. Đồng thời sẽ ứng dụng theo hướng tiếp cận Phân đoạn đối tượng bán giám sát với việc yêu cầu object mask có sẵn ở khung hình đầu tiên để xác định đối tượng chính và tự động phân đoạn cho toàn bộ các khung hình còn lại trong chuỗi video.

Input: Chuỗi các khung hình trong một video cùng object segmentation mask của frame đầu tiên.

Output: Các object segmentation masks cho các khung hình còn lại trong video.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Ứng dụng kiến trúc Deformable Transformer cho bài toán Phân đoạn vật thể trong video.
- Phân tích và đánh giá hiệu suất và khả năng tối ưu của mô hình khi sử dụng kiến trúc Deformabel Transformer cho bài toán Phân đoạn vật thể trong video và so sánh hiệu năng với mô hình TransVOS sử dụng kiến trúc Transfomer.
- Xây dựng ứng dụng mình họa cho phép người dùng phân tách được vật thể khỏi cảnh nền trong video được truyền vào.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Nội dung:

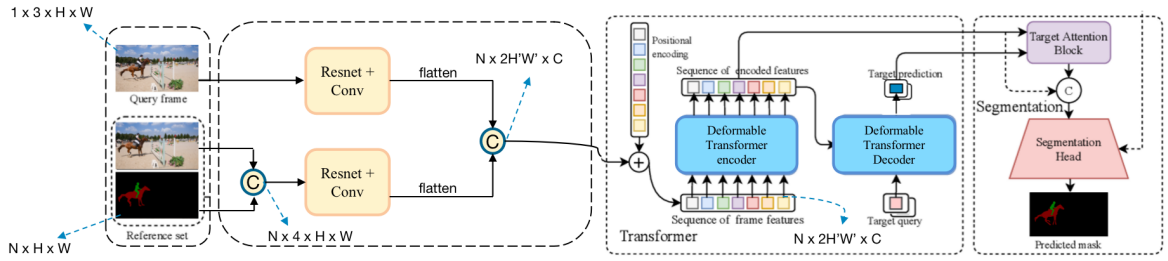
- Nghiên cứu tính chất, nguyên lí hoạt động của Multi-head Self-attention [2] và Deformable Attention [4].
- Nghiên cứu cơ chế phát hiện vật thể trong các bài toán Nhận diện vật thể trong ảnh của hai kiến trúc Transformer và Deformable Transformer.
- Thay thế kiến trúc của Transformer trong mô hình TransVOS [1] bằng Deformable Transformer.
- Huấn luyện và đánh giá các mô hình trên bộ dữ liệu đã được benchmark dành cho bài toán Phân đoạn đa vật thể là DAVIS-2017 [5] gồm: 4219 ảnh khung

hình cho huấn luyện và 2023 ảnh khung hình cho việc đánh giá.

- Xây dựng ứng dụng web cho phép người dùng đăng tải video và thực hiện Phân tách vật thể bằng mô hình DeformableTransVOS và TransVOS [1].

Phương pháp:

- Nghiên cứu bản chất của Multi-head Self-attention và ứng dụng của kiến trúc Transformer vào bài toán Nhận diện vật thể của mô hình DETR.
- Tìm hiểu cơ chế hoạt động của Deformable Attention trong kiến trúc của mô hình Deformable DETR [4] giải quyết các điểm yếu về mặt tính toán và hiệu suất của mô hình DETR [3].
- Xây dựng mô hình Deformable TransVOS bằng cách thay thế kiến trúc của Transformer trong mô hình TransVOS bằng Deformable Transformer.



- Thực hiện huấn luyện hai mô hình TransVOS và DeformableVOS theo hai giai đoạn:
  - Pretrain: Huấn luyện trên các bộ dữ liệu ảnh tĩnh có các object mask.
  - Maintrain: Huấn luyện trên tập trainset của bộ dữ liệu đã được benchmark dành cho bài toán Phân đoạn vật thể trong video là DAVIS-2017
- Đánh giá và so sánh hiệu suất của hai mô hình dựa trên tập valset của DAVIS-2017 khi:
  - Chỉ sử dụng pretrain
  - Chỉ huấn luyện maintrain
  - Kết hợp pretrain và maintrain
  - Chỉ số đánh giá:
    - J&F mean: Trung bình cộng của 2 giá trị Jaccard Coefficient và

Contour precision F-mesure.

- FPS: Số lượng frame mà mô hình xử lý trong một giây.

- Xây dựng ứng dụng web cho phép người dùng đăng tải video, thực hiện vẽ object mask cho đối tượng chính cho frame đầu, chọn mô hình và thực hiện trích xuất đối tượng trong video.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- *Báo cáo phương pháp và cơ chế hoạt động của hai kiến trúc Transformer và Deformable Transformer cho bài toán Phân đoạn đối tượng trong video. Kết quả thực nghiệm và so sánh đánh giá giữa hai phương kiến trúc.*
- *Deformable Transformer có khả năng tối ưu, giải quyết về nhược điểm tính toán lớn và tốc độ dự đoán chậm của mô hình TransVOS mà vẫn giữ được độ chính xác. Nếu thành công, nhóm sẽ đặt tên cho mô hình này là DeformbleTransVOS.*
- *Ứng dụng web hoàn chỉnh dễ sử dụng, cho phép thực hiện đăng tải video, gán nhãn đối tượng và tự động phân tách vật thể.*

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Jianbiao Mei, Mengmeng Wang, Yeneng Lin, Yong Liu:
TransVOS: Video Object Segmentation with Transformers. CoRR abs/2106.00588 (2021)

[2].Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin:
Attention is All you Need. NIPS 2017: 5998-6008

[3].Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko:
End-to-End Object Detection with Transformers. ECCV (1) 2020: 213-229

[4].Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai:
Deformable DETR: Deformable Transformers for End-to-End Object Detection.

ICLR 2021

[5].Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alexander Sorkine-Hornung, Luc Van Gool: The 2017 DAVIS Challenge on Video Object Segmentation. CoRR abs/1704.00675 (2017)