THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://www.youtube.com/watch?v=MFYhRVZKsIY*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/dnsang1611/CS519.O11/blob/master/Dictionary-guidedDeepSolo_presentation.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Đoàn Nhật Sang
• MSSV: 21522542
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 9/10
	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân: 12/12
	• Số câu hỏi QT của cả nhóm: 12/12
	• Link Github: https://github.com/dnsang1611/CS519.O11/
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng đồ án
	  ○ Viết proposal, slide, poster
	  ○ Làm video youtube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

DEEPSOLO ĐƯỢC HƯỚNG DẪN BỞI TỪ ĐIỂN CHO VIỆC PHÁT HIỆN VÀ NHẬN DẠNG VĂN BẢN TIẾNG VIỆT TRONG CẢNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

DICTIONARY-GUIDED DEEPSOLO FOR VIETNAMESE SCENE TEXT DETECTION AND RECOGNITION

TÓM TẮT *(Tối đa 400 từ)*

Các thông tin phong phú, chính xác được thể hiện trong văn bản rất hữu ích trong nhiều ứng dụng dựa trên thị giác, do đó bài toán phát hiện và nhận dạng văn bản trong cảnh đã trở thành chủ đề nghiên cứu tích cực, quan trọng trong thị giác máy tính và phân tích tài liệu. End-to-end text spotting là một hướng tiếp cận của bài toán này, nhằm kết hợp 2 nhiệm vụ phát hiện và nhận dạng thành một mô hình thống nhất. Để thiết kế một mô hình hiệu quả theo hướng này, giải quyết mối quan hệ giữa 2 nhiệm vụ con là quan trọng. Mặc dù các phương pháp dựa trên Transformer loại bỏ được bộ kết hợp và hậu xử lý heuristic, chúng vẫn gặp phải vấn đề tổng hợp giữa 2 nhiệm vụ con và hiệu quả huấn luyện thấp. Để giải quyết vấn đề này, DeepSolo đã được đề xuất, với kiến trúc như DETR đơn giản, cho phép một decoder với những điểm rõ ràng phát hiện và nhận dạng văn bản đồng thời. Bên cạnh đó, DeepSolo cũng tương thích với nhãn đường, loại nhãn yêu cầu ít chi phí gán nhãn hơn so với polygon. Trong đề tài này, nhóm sẽ tập trung nghiên cứu sử dụng DeepSolo để đánh giá và chứng minh tính hiệu quả của nó đối với bài toán phát hiện và nhận dạng văn bản tiếng Việt trong cảnh. Đồng thời, nhóm cũng sẽ cải tiến DeepSolo để giúp mô hình tận dụng hiệu quả thông tin ngôn ngữ biết trước bằng việc kết hợp bộ từ điển vào quá trình huấn luyện và dự đoán.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Phát hiện và nhận dạng văn bản trong cảnh ngày càng được quan tâm vì những ứng

dụng rộng rãi của nó như: Rút trích thông tin từ ảnh, xe tự lái, hỗ trợ người khiếm thị.

Cụ thể, với tiếng Việt, bài toán này sẽ có đầu vào là hình ảnh, hình ảnh này có thể chứa một hoặc nhiều văn bản với màu sắc, độ cong, font chữ,… đa dạng; đầu ra là các polygon xác định vùng văn bản tiếng Việt trên ảnh cùng với chuỗi ký tự tương ứng.

Hình 1: Input, output của bài toán phát hiện và nhận dạng văn bản tiếng Việt

End-to-end text spotting là một hướng tiếp cận đang được chú ý để giải quyết bài toán này vì nó khai thác thác được mối quan hệ nội tại giữa detection và recognition. Trong hướng tiếp cận này, các phương pháp dựa trên DETR [1] tuy giải quyết được hạn chế của các phương pháp Region of Interest [2], Segmentation [3] nhưng vẫn còn thiếu những đại diện chung hiệu quả để phát hiện và nhận dạng văn bản. Ví dụ, TTS [4] yêu cầu thêm RNN, và thiết kế query không xem xét những đặc tính độc nhất của văn bản trong cảnh. Trong khi đó, TESTR [5] sử dụng từng decoder với query khác nhau cho từng nhiệm vụ con, dẫn đến sự không đồng nhất ngoài mong muốn.

Để khắc phục hạn chế này, DeepSolo [6] đã được thiết kế như một mô hình DETR với một Transformer decoder và một vài đầu dự đoán để giải quyết bài toán hiệu quả. Mô hình dựa trên một dạng query mới có thể mã hoá hiệu quả vị trí, hình dạng, ngữ cảnh của văn bản. Tuy nhiên, DeepSolo chỉ mới được đánh giá trên các bộ dữ liệu tiếng Anh, tiếng Trung và chưa tận dụng hiệu quả thông tin ngôn ngữ biết trước.

Vì vậy, trong đề tài này, nhóm sẽ tập trung nghiên cứu sử dụng DeepSolo để giải quyết bài toán phát hiện và nhận dạng văn bản tiếng Việt. Bên cạnh đó, nhóm cũng sẽ kết hợp bộ từ điển vào cả quá trình huấn luyện và suy luận [7] để cải tiến mô hình.

MỤC TIÊU

- Xây dựng bộ dữ liệu UITText chứa nhiều văn bản cong cho bài toán.
- Tìm hiểu point queries và kiến trúc của mô hình DeepSolo. Từ đó, huấn luyện, đánh giá và so sánh DeepSolo trên dữ liệu VinText [7], UITText.
- Cải tiến DeepSolo thành Dictionary-guided DeepSolo với kĩ thuật kết hợp bộ từ điển vào quá trình huấn luyện và suy luận [7]. Từ đó, huấn luyện, đánh giá và so sánh phiên bản cải tiến trên dữ liệu VinText, UITText.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Tìm hiểu bài toán phát hiện và nhận dạng văn bản tiếng Việt trong cảnh, cách xây dựng bộ dữ liệu huấn luyện và kiểm thử. Phương pháp:

- Tìm hiểu về bài toán phát hiện và nhận dạng văn bản tiếng Việt trong cảnh.
- Tìm hiểu các độ đo dùng để đánh giá trong bài toán: precision, recall, hmean cho detection và end-to-end (cả detection và recognition).
- Tìm hiểu cách xây dựng bộ dữ liệu CTW1500 [8] để xây dựng bộ dữ liệu UITText trên tiếng Việt, chứa nhiều văn bản cong hơn so với bộ dữ liệu trên tiếng Việt hiện có, VinText [7] . UITText gồm 2000 ảnh và được đánh nhãn ở mức độ dòng thay vì mức độ từ. Các ảnh được thu thập từ Internet hoặc được chụp bởi 5 người thu thập. Người thu thập được yêu cầu chụp ảnh có chứa văn bản cong ở mức độ từ hoặc mức độ dòng bằng điện thoại hoặc camera của họ. Quy trình đánh nhãn sẽ được thực hiện như CTW1500.

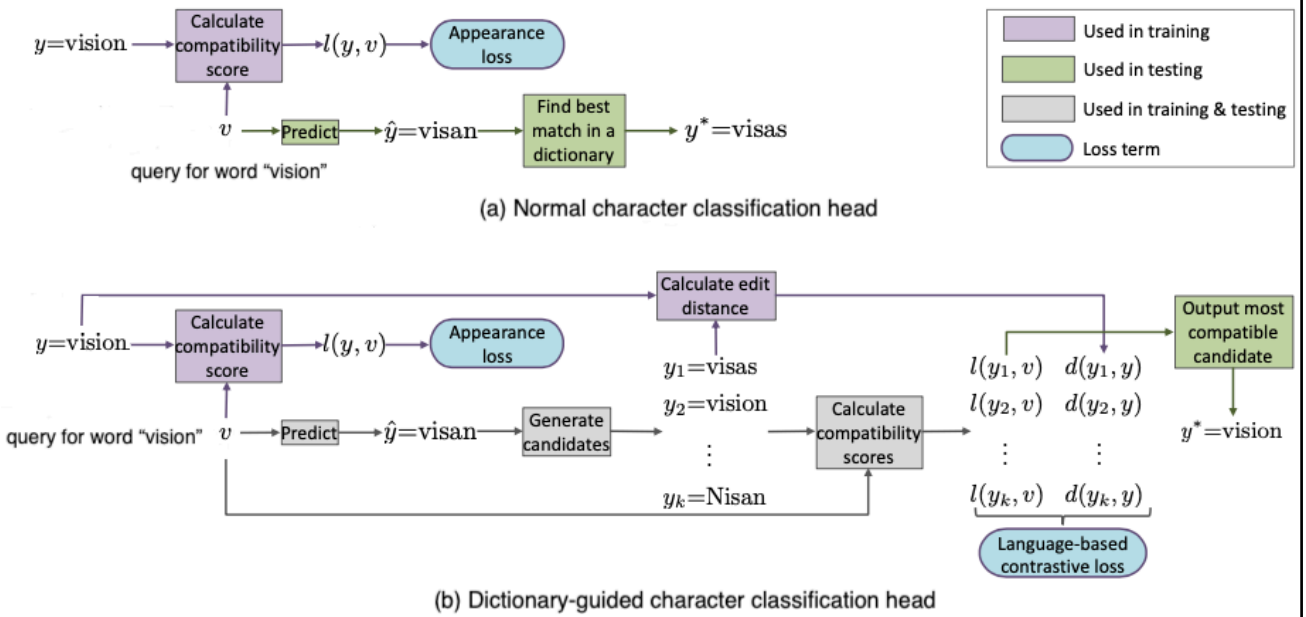
Nội dung 2: Tìm hiểu cách thiết kế point queries và kiến trúc của mô hình DeepSolo. Phương pháp là trả lời các câu hỏi sau:

- Backbone, encoder, decoder và 4 đầu phân loại có kiến trúc như thế nào? Tại sao lại sử dụng 2 kĩ thuật Multi-scale Deformable Attention và Factorized Self-Attention? Trong kĩ thuật Factorized Self-Attention, tại sao lại sử dụng intra-group attention rồi mới tới inter-group attention mà không phải ngược lại?
- Query vị trí được tạo bằng hàm mã hoá vị trí sinusoid (được định nghĩa trước) theo sau bởi MLP (Multi-layer Perceptron). Tại sao lại không sử dụng 1 lớp embedding có thể huấn luyện để thay cho cách biểu diễn query vị trí ở trên?

- Tiêu chí so khớp văn bản hoạt động như thế nào trong quá trình tối ưu?

Nội dung 3: Cải tiến DeepSolo bằng cách kết hợp bộ từ điển trong cả quá trình huấn luyện và suy luận. Áp dụng DeepSolo và phiên bản cải tiến trên dữ liệu tiếng Việt. Phương pháp:

- Tìm hiểu cách kết hợp bộ từ điển vào quá trình huấn luyện và suy luận [7]. Cụ thể, nhóm sẽ cải tiến ở character classification head (đầu dự đoán từ) như hình 2. Từ điển được dùng để khởi tạo một danh sách các từ ứng cử, và các ứng cử sẽ được đánh giá bởi compatibility scoring module. Contrastive loss được bổ sung để tối đa khả năng của những ứng cử gần với ground truth, trong khi tối thiểu khả năng của những ứng cử khác xa với ground truth.



Hình 2: Character classification head của DeepSolo thông thường và sau khi được cải tiến.

- Cài đặt, huấn luyện DeepSolo và phiên bản cải tiến trên bộ dữ liệu VinText, UITText và đánh giá kết quả với các độ đo precision, recall, hmean.
- So sánh kết quả của 2 mô hình trên với các mô hình đã có (Mask TextSpotter v3 [2], CharNet [3], TTS [4], TESTR [5]) trên VinText, UITText.
- Chuyển nhãn polygon của VinText, UITText thành nhãn đường như [6] để kiểm tra khả năng của DeepSolo và phiên bản cải tiến trên loại nhãn này.

KẾT QUẢ MONG ĐỢI

- Tài liệu về bài toán phát hiện và nhận dạng văn bản tiếng Việt, tài liệu về các độ đo precision, recall, hmean cho detection và end-to-end.
- Bộ dữ liệu UITText chứa nhiều văn bản cong.
- Tài liệu, sourcecode của mô hình DeepSolo và DeepSolo cải tiến đã được tìm hiểu và chú thích. Kết quả thực nghiệm, so sánh và đánh giá của các mô hình.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

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

[2]. Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, Xiang Bai: Mask TextSpotter v3: Segmentation Proposal Network for Robust Scene Text Spotting. ECCV (11) 2020: 706-722

[3]. Linjie Xing, Zhi Tian, Weilin Huang, Matthew R. Scott: Convolutional Character Networks. ICCV 2019: 9125-9135

[4]. Yair Kittenplon, Inbal Lavi, Sharon Fogel, Yarin Bar, R. Manmatha, Pietro Perona: Towards Weakly-Supervised Text Spotting using a Multi-Task Transformer. CVPR 2022: 4594-4603

[5]. Xiang Zhang, Yongwen Su, Subarna Tripathi, Zhuowen Tu: Text Spotting Transformers. CVPR 2022: 9509-9518

[6]. Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, Dacheng Tao: DeepSolo: Let Transformer Decoder with Explicit Points Solo for Text Spotting. CVPR 2023: 19348-19357

[7]. Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, Minh Hoai: Dictionary-Guided Scene Text Recognition. CVPR 2021: 7383-7392

[8]. Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Sheng Zhang: Detecting Curve Text in the Wild: New Dataset and New Solution. CoRR abs/1712.02170 (2017)