THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: Here
- Link slides: Here

- Họ và Tên: Nguyễn Khắc Thái
- MSSV: 250101062

- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 0
- Số câu hỏi QT của cả nhóm: 0
- Link Github: Here

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

CẢI TIẾN HÀM MẤT MÁT TRONG BÀI TOÁN NHẬN DIỆN VĂN BẢN NGOẠI CẢNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

IMPROVEMENT LOSS FUNCTION IN SCENE TEXT RECOGNITION

TÓM TẮT

Nhận diện văn bản ngoại cảnh (Scene Text Recognition - STR) là một bài toán quan trọng trong lĩnh vực thị giác máy tính, đóng vai trò then chốt trong các ứng dụng như điều hướng robot, hỗ trợ người khiếm thị và tự động hóa trích xuất thông tin. Tuy nhiên, các phương pháp hiện tại vẫn gặp nhiều thách thức khi xử lý văn bản có phong cách nghệ thuật (Art-Text), bị che khuất, điều kiện ánh sáng kém hoặc các từ nằm ngoài từ điển (Out-of-Vocabulary). Một trong những vấn đề lớn là sự nhầm lẫn giữa các ký tự có hình dạng tương đồng (ví dụ: '0' và 'O').

Nghiên cứu này đề xuất giải pháp cải tiến mô hình nhận diện văn bản bằng cách áp dụng Tư duy tính toán (Computational Thinking) và giới thiệu một hàm mất mát mới có tên là **Cluster Character Loss (CCL)**. Mô hình sử dụng kiến trúc VGG19 làm backbone để trích xuất đặc trưng và Transformer để dự đoán chuỗi văn bản. Hàm CCL được thiết kế để tăng cường hình phạt đối với các nhầm lẫn giữa các ký tự thuộc cùng một nhóm tương đồng về hình ảnh, giúp mô hình học phân biệt tốt hơn. Kết quả thực nghiệm trên các bộ dữ liệu VinText, ICDAR 2013 và bộ dữ liệu tự thu thập cho thấy phương pháp đề xuất cải thiện độ chính xác và giảm sai số Levenshtein so với mô hình cơ sở.

GIỚI THIỆU

Trong kỷ nguyên số hóa, nhu cầu trích xuất thông tin từ hình ảnh tự nhiên ngày càng tăng cao. Bài toán STR không chỉ dừng lại ở việc nhận diện văn bản trong tài liệu scan mà mở rộng ra các biển báo, biển hiệu cửa hàng và bao bì sản phẩm ngoài thực tế. Các nghiên cứu gần đây như của Nguyen et al. (2021) hay Wan et al. (2020) đã đạt được những tiến bộ đáng

kể bằng cách kết hợp thông tin ngữ nghĩa và từ điển. Tuy nhiên, các mô hình vẫn thường xuyên thất bại khi gặp các phông chữ nghệ thuật phức tạp hoặc khi văn bản không tuân theo quy tắc ngôn ngữ thông thường (ví dụ: tên riêng, mã số).

Vấn đề cốt lõi được xác định là sự mơ hồ về mặt hình ảnh giữa các ký tự. Các hàm mất mát truyền thống thường coi tất cả các sai sót là như nhau, không tập trung giải quyết triệt để sự nhầm lẫn giữa các ký tự "gần giống nhau" (Visual Ambiguity).

Dựa trên cách tiếp cận Tư duy tính toán, nghiên cứu này phân rã bài toán lớn thành các phần nhỏ (Decomposition), nhận diện các mẫu sai sót thường gặp (Pattern Recognition) để thiết kế thuật toán tối ưu. Cụ thể, chúng tôi tập trung xây dựng hàm mất mát Cluster Character Loss (CCL) tích hợp vào kiến trúc mạng nơ-ron sâu (Deep Learning) để giải quyết bài toán nhận diện văn bản tiếng Việt và tiếng Anh trong điều kiện ngoại cảnh phức tạp.

MỤC TIÊU

1. Xây dựng mô hình nhận diện văn bản ngoại cảnh (STR) dựa trên kiến trúc VGG19 kết hợp với Transformer để xử lý trích xuất đặc trưng và dự đoán chuỗi.
2. Đề xuất và cài đặt hàm mất mát mới "Cluster Character Loss" (CCL) nhằm giảm thiểu sự nhầm lẫn giữa các ký tự có hình dạng tương đồng (cụm ký tự).
3. Đánh giá hiệu quả của giải pháp trên các bộ dữ liệu tiêu chuẩn (ICDAR 2013), dữ liệu tiếng Việt (VinText) và dữ liệu Art-text tự thu thập; chứng minh sự cải thiện về độ chính xác và giảm sai số Levenshtein.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Quy trình xử lý dữ liệu và bài toán (Problem Formulation & Preprocessing):

- **Định nghĩa bài toán:** Đầu vào là ảnh chứa văn bản đã được cắt (cropped images) dựa trên bounding box; đầu ra là chuỗi ký tự tương ứng.
- **Tiền xử lý:** Chuẩn hóa kích thước ảnh đầu vào về cố định ($H \times W = 32 \times 128$) để phù hợp với mạng CNN. Xây dựng bộ từ điển ký tự bao gồm 99 ký tự (chữ cái tiếng Việt,

chữ số, ký tự đặc biệt).

2. Kiến trúc mô hình (Model Architecture):

- **Feature Extractor:** Sử dụng mạng VGG19 backbone để trích xuất các đặc trưng thị giác (feature maps) từ ảnh đầu vào.
- **Sequence Modeling:** Sử dụng kiến trúc Transformer với cơ chế Self-Attention (bao gồm Encoder và Decoder) để chuyển đổi feature maps thành chuỗi văn bản dự đoán. Mô hình được thiết lập với 6 layer và 4 head-attention.

3. Đề xuất cải tiến Hàm mất mát (Cluster Character Loss - CCL):

- **Xây dựng cụm ký tự (Cluster definition):** Phân tích thống kê các sai sót thường gặp để nhóm các ký tự tương đồng (ví dụ: nhóm {0, O, o, D, Q}, nhóm {l, 1, I}).
- **Cơ chế hoạt động:** Thiết lập công thức hàm loss có bổ sung trọng số phạt (penalty k_j). Nếu mô hình dự đoán sai nhưng ký tự dự đoán nằm trong cùng cụm tương đồng với nhãn thực (ground truth), hệ thống sẽ ghi nhận và điều chỉnh trọng số để mô hình "chú ý" hơn vào các đặc trưng tinh vi phân biệt các ký tự này trong các lần cập nhật trọng số tiếp theo.

4. Thực nghiệm và Đánh giá (Implementation & Evaluation):

- **Huấn luyện:** Thực hiện trên bộ dữ liệu VinText (tiếng Việt).
- **Đánh giá:**
    - Kiểm thử trên tập VinText (tiếng Việt) và ICDAR 2013 (tiếng Anh).
    - Đánh giá khả năng xử lý Art-text và Out-of-Vocabulary (OOV).
- **Độ đo:** Sử dụng độ chính xác (Accuracy - bao gồm case-sensitive) và khoảng cách chỉnh sửa (Levenshtein Distance) để đo lường mức độ sai lệch ký tự.

KẾT QUẢ MONG ĐỢI

1. Chứng minh được tính hiệu quả của hàm mất mát CCL: Giúp mô hình hội tụ tốt hơn và giảm thiểu các lỗi sai do nhầm lẫn hình ảnh.

2. Cải thiện hiệu năng định lượng: Tăng độ chính xác (Accuracy) trên tập dữ liệu VinText (kỳ vọng ~70%) và ICDAR 2013 (kỳ vọng tăng >1% so với baseline).
3. Giảm sai số Levenshtein: Chứng minh mô hình có khả năng dự đoán sát với thực tế hơn, đặc biệt trên các mẫu dữ liệu khó như văn bản nghệ thuật hoặc văn bản nằm ngoài từ điển huấn luyện.

TÀI LIỆU THAM KHẢO

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

[2]. Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, Hwalsuk Lee: *On Recognizing Texts of Arbitrary Shapes with 2D Self-Attention.* CVPR Workshops 2020: 2326-2335.

[3]. Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluís Gómez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernández Mota, Jon Almazán, Lluís-Pere de las Heras: *ICDAR 2013 Robust Reading Competition.* ICDAR 2013: 1484-1493.

[4]. Zhaoyi Wan, Minghang He, Haoran Chen, Xiang Bai, Cong Yao: *Text Scanner: Reading Characters in Order for Robust Scene Text Recognition.* AAAI 2020: 12120-12127.