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

Nguyễn Khắc Thái - 250101062

Tóm tắt

Họ và Tên: Nguyễn Khắc Thái

MSHV: 250101062

YouTube [Youtube]

GitHub [Github]

Giới thiệu

- **Bối cảnh:** Nhận diện văn bản ngoại cảnh (STR) có vai trò quan trọng trong:
  - Hỗ trợ người khiếm thị.
  - Xe tự hành, robot điều hướng.
  - Số hóa và trích xuất thông tin tự động.
- **Thách thức hiện tại:**
  - Văn bản nghệ thuật (Art-Text), phông chữ phức tạp.
  - Điều kiện ánh sáng kém, bị che khuất.
  - Văn bản nằm ngoài từ điển (Out-of-Vocabulary - OOV).
  - **Vấn đề cốt lõi:** Sự nhầm lẫn giữa các ký tự có hình dạng tương đồng (Visual Ambiguity).

Giới thiệu

- **Input:** Một bức ảnh bất kỳ + Danh sách bounding box (vị trí văn bản).
- **Output:** Chuỗi ký tự văn bản tương ứng được nhận diện.
- **Ràng buộc:**
  - Hỗ trợ tiếng Việt (có dấu) và tiếng Anh.
  - Bao gồm chữ số và ký tự đặc biệt.
  - Không giới hạn góc chụp, chất lượng ảnh.

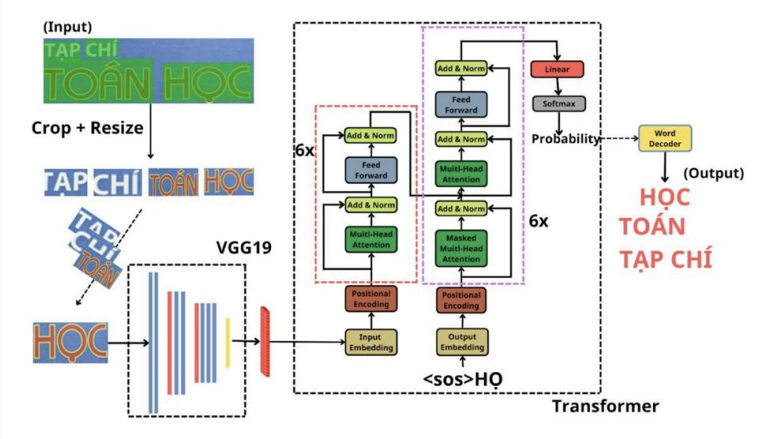
(Box ID)	(Text Prediction)
1	“HOT”
2	“&”
3	“COLD”
4	“TRÀ”
5	“SỮA”
6	“&”
7	“XIÊN”
8	“QUE”

Mục tiêu

- **Xây dựng mô hình STR:** Kết hợp mạng CNN (trích xuất đặc trưng) và Transformer (dự đoán chuỗi) để xử lý văn bản trong ảnh tự nhiên.
- **Đề xuất Hàm mất mát mới:** Phát triển hàm **Cluster Character Loss (CCL)** nhằm giải quyết vấn đề nhầm lẫn giữa các ký tự có hình dạng giống nhau (Ví dụ: 0 và O, l và I).
- **Đánh giá hiệu quả:**
  - Cải thiện độ chính xác (Accuracy).
  - Giảm sai số chỉnh sửa (Levenshtein distance).
  - Kiểm thử trên các tập dữ liệu: VinText, ICDAR 2013 và dữ liệu tự thu thập.

Nội dung và Phương pháp

- Mô hình được thiết kế theo các bước:
  - **Cropping:** Cắt vùng ảnh chứa văn bản từ bounding box.
  - **Feature Extractor (VGG19):** Trích xuất đặc trưng hình ảnh.
  - **Context Modeling (Transformer Encoder):** Mã hóa ngữ cảnh.
  - **Prediction (Transformer Decoder):** Giải mã và dự đoán ký tự.



Kết quả dự kiến

Về mặt mô hình:

- **Hoàn thiện hệ thống STR:** Xây dựng thành công pipeline nhận diện văn bản từ ảnh đầu vào đến văn bản đầu ra.
- **Chứng minh tính đúng đắn của CCL:**
  - Kỳ vọng hàm mất mát CCL sẽ giúp mô hình hội tụ tốt hơn.
  - Mô hình có khả năng "hiểu" và phân biệt được các ký tự trong nhóm Cluster (ví dụ: phân biệt được số 0 và chữ O trong ngữ cảnh cụ thể).

Kết quả dự kiến

Về mặt hiệu năng

- **Cải thiện độ chính xác (Accuracy):**
  - Dự kiến độ chính xác trên tập VinText đạt khoảng **70%**.
  - Dự kiến cải thiện độ chính xác trên tập ICDAR 2013 từ **1% trở lên** so với mô hình gốc.
- **Giảm sai số (Levenshtein Distance):**
  - Kỳ vọng giảm thiểu khoảng cách chỉnh sửa chuỗi ký tự.
  - Giảm tỷ lệ sai sót trên các tập dữ liệu khó (Art-text, Out-of-Vocabulary).

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.