Khoa Công Nghệ Thông tin
Ngành Khoa Học Máy Tính
Trường Đại học Xây Dựng Hà Nội

[HUCE]

ĐỀ XUẤT THUẬT TOÁN Re-CNN TRONG VIỆC HỖ TRỢ GIÁO VIÊN NHẬP ĐIỂM TỰ ĐỘNG TỪ BÀI THI TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI

Dương Công Sơn*, Hoàng Thị Thanh Tú, Phạm Hoàng Đăng Trung , Nguyễn Trần Lê Tuấn*
Email: {son167464, tuan1553564}@huce.edu.vn

---

1. NỘI DUNG ĐỀ TÀI

- Tìm hiểu tổng quan về hệ thống nhận dạng chữ viết tay và các thành phần chính.
- Tiến hành các kỹ thuật tiền xử lý ảnh
- Triển khai thuật toán phân vùng ký tự
- Xây dựng mô hình giải thuật phát hiện chữ viết và đề xuất thuật toán Re-CNN (Recurrent Convolutional Neural Network) nhằm nhận dạng chữ viết tay.
- Xây dựng phần mềm tự động nhập điểm vào danh sách lớp File Excel.
- Người sử dụng thẻ tín dụng phổ biến có trình độ học vấn là đại học
- Xây dựng giao diện người dùng.

2. Mục tiêu đề tài

Đề tài sẽ thực hiện dò văn bản và nhận diện văn bản từ mẫu giấy thi của trường Đại học Xây Dựng Hà Nội. Cụ thể, cần nhận dạng được ba thông tin chính “Họ và tên SV”, “MSSV”, “Điểm tổng kết”. Giấy thi sau khi được giáo viên chấm xong, sẽ được chụp lại và lưu vào thư mục với các điều kiện:

- Ánh sáng bình thường, tránh ngược sáng, ánh sáng điện.
- Góc ảnh: trực diện hoặc góc nghiêng không quá 10 độ.
- Không bị che khuất

Sau khi nhận dạng, đối chiếu với thông tin MSSV và họ tên sinh viên có sẵn trong danh sách lớp để tự động nhập điểm vào File Excel. Hiển thị giao diện để thầy/cô có thể làm việc, tương tác với hệ thống.

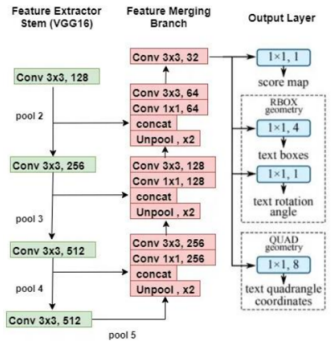
3. DATASET

Nhóm sẽ chia thành 2 bộ dữ liệu tương đương với 2 nhóm thông tin cần nhận dạng trên giấy thi:
- Bộ dữ liệu HANDS-VNOnDB2018:
HANDS – VNOnDB2018 (ICFHR2018 Competition on Vietnamese Online Handwritten Text Recognition Database) (1, 2018) là bộ dữ liệu cung cấp 1.146 đoạn văn bản viết tay bằng tiếng Việt gồm 7.296 dòng, hơn 480.000

nét và hơn 380.0d00 kí tự được viết bởi 200 người Việt Nam. Đây là bộ dữ liệu nhóm dùng để huấn luyện mô hình nhận dạng Họ và tên.
- Bộ dữ liệu MNIST:
Trong bộ dữ liệu này (2, 2010), mỗi hình là một ảnh đen trắng chứa một số được viết tay từ 0 đến 9 có kích thước là 28x28. Bộ dữ liệu vô cùng đồ sộ với khoảng 60.000 ảnh dữ liệu huấn luyện, 10.000 ảnh dữ liệu kiểm thử và được sử dụng phổ biến trong các thuật toán nhận dạng ảnh. Đây là bộ dữ liệu nhóm dùng để phục vụ cho việc huấn luyện và nhận dạng MSSV và điểm số.

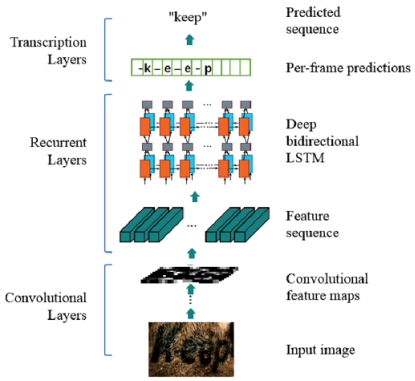
3. MÔ HÌNH DỰ KIẾN CÀI ĐẶT

Trong bài toán này, nhóm dự kiến cần 2 mô hình: Mô hình dành cho việc nhận phát hiện văn bản trên giấy thi (TEXT DETECTION) và mô hình nhận dạng dạng văn bản (TEXT RECOGNITION).
TEXT DETECTION:
Để phát hiện văn bản, nhóm chúng tôi sử dụng thuật toán EAST (Efficient and Accurate Scene Text Detection Pipeline) (3, 2017)(4, 2021)
Thuật toán bao gồm 2 giai đoạn chính. Giai đoạn một là một mạng FCN (Fully – Convolutional Neural Network) để trực tiếp nhận biết vùng chứa văn bản, cho ra kết quả là xác suất chứa văn bản và vị trí của Text Box. Tiếp theo là giai đoạn NMS (Non – Max Suppression) để gộp các Text Box thành một Bounding Box quanh văn bản
Mạng FCN gồm 3 phần chính: Lớp trích tách đặc trưng (Feature Extractor), lớp ghép các đặc trưng (Feature Merging) và lớp đầu ra (Output Layer). Dưới đây là kiến trúc tổng quan của EAST.



Hình 1: ROC curve.

Lớp tách đặc trưng: có thể sử dụng các mô hình mạng tích chập (Convolutional Network) được huấn luyện sẵn, mô hình PVANet hoặc mô hình VGG16. Cho ra các Feature Map với kích thước nhỏ hơn ảnh đầu vào.
Lớp ghép các đặc trưng: các Feature Map từ lớp trước được đưa vào lớp Unpooling để tăng gấp đôi kích thước. Sau đó được đưa vào lớp Concat (Concatenate) để ghép với Feature Map hiện có. Lớp Conv 1x1 nhằm giảm bớt chi phí tính toán, lớp Conv 3x3 để làm ngưỡng loại bỏ các Feature Map không đủ yêu cầu.
Lớp đầu ra nhận Feature Map của lớp trước. Thông qua nhiều lớp Conv 1x1 sẽ biến đổi Feature Map 32 kênh thành một kênh Score Map mang thông tin xác suất chứa văn bản và các kênh Geometry Map chứa vị trí và góc quay của văn bản.
Giai đoạn NMS: Ghép lần lượt các lớp Geometry Map lại với nhau theo từng hàng vì thường các Pixel nằm cạnh nhau theo hàng ngang có sự kết nối lớn hơn.
TEXT RECOGNITION Ta có 2 thông tin cần nhận dạng chính là "Họ và tên SV" (tức là chữ viết tay) và "MSSV", "điểm số SV" (tức là chữ số viết tay). Vì vậy mà mô hình của nhóm sẽ được huấn luyện để nhận dạng 2 loại thông tin này (chữ viết tay và số viết tay).
Nhóm chúng tôi đề xuất mô hình Re-CNN (Recurrent Convolutional Neural Network). Re-CNN là sự kết hợp của mạng CNN, RNN và các tác vụ nhận dạng chuỗi với độ dài bất kì như nhận dạng văn bản, phân loại Video, phân loại hành động… Kiến trúc mô hình Re-CNN gồm 3 giai đoạn chính bao gồm Convolutional Layer, Recurrent

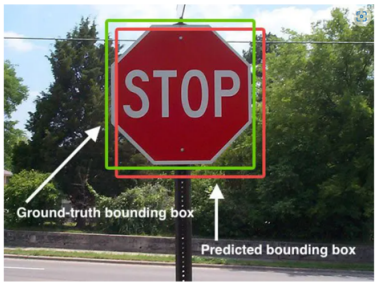


Hình 2: ROC curve.

Layer và Transcription Layer.
Từ ảnh gốc khi qua lớp tích chập (Convolutional Layer) sẽ tạo ra được các Feature Map và từ đó một chuỗi các Feature Vector sẽ được tạo ra gọi là Feature Sequence để đưa vào Recurrent Layer. Mỗi Feature Vector có thể tương ứng với một vùng (region) ở ảnh gốc.
Tiếp theo các Feature Vector được đưa vào lớp Recurrent Layer. Trong lớp này, chúng tôi sử dụng mạng cnhớ dài hạn hai chiều (Bidirectional-LSTM) (5, 2022) (6, 2019) để dự đoán các kí tự từ quan hệ của các ký tự từ lớp tích chập trước đó.
Cuối là lớp phiên mã (Transcription Layer) giúp chuyển đổi mỗi khung hình (Per – Frame Prediction) được tạo bởi mạng Bidirectional-LSTM ở trước thành chuỗi kết quả cuối cùng. Cụ thể hơn, trong lớp này sẽ dự đoán từng ký tự đầu của 1 chuỗi bằng cách lựa chọn ký tự có xác suất dự đoán cao nhất trên toàn bộ từ điển có sẵn.

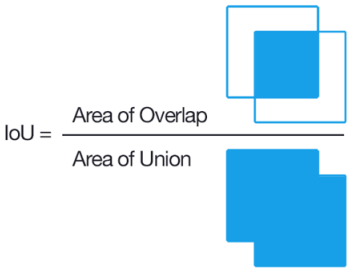
3. ĐÁNH GIÁ MÔ HÌNH

Do bài toán được triển được dựa trên 2 mô hình là TEXT DETECTION VÀ TEXT RECOGNITION, mỗi mô hình sẽ được sử dụng một phương pháp đánh giá riêng như sau:
ĐÁNH GIÁ TEXT DETECTION: Đo độ khớp giữa ground truth của bounding box với bounding box mà mô hình dự đoán, hay còn gọi là IoU (Intersection over Union) [4] [7].



Hình 3: ROC curve.

IoU được tính theo công thức sau: $IoU = (Dintchgiao)/(Dintchhp)$ Rõ ràng IoU càng tiến tới 1 thì mô hình có khả năng phát hiện TEXT càng tốt.
ĐÁNH GIÁ TEXT RECOGNITION Chúng tôi chia tập huấn luyện thành 62 nhãn bao gồm các chữ cái từ A-Z, a-z và chữ số 0-9. Tiếp theo để đánh giá khả năng nhận biết văn bản của mô hình, chúng tôi chia thành 2 bước: phân loại ký tự - character classification (bao gồm chữ cái bất kỳ hoặc chữ số bất kỳ) và phân loại từ - word classification. Phân loại ký tự sẽ đánh giá xem mô hình có phân loại từng ký tự trên vùng được phát hiện có thuộc đúng nhãn của nó hay không. Trong khi đó, phân loại từ sẽ đánh giá xem độ tốt của mô hình khi nhận dạng toàn bộ văn bản trong



Hình 4: ROC curve.

vùng ảnh.
Độ đo được chọn cho phân loại ký tự là precision, recall và F1-score [8]. Trong đó precision sẽ cho biết độ chuẩn xác của mô hình khi phân loại ký tự i vào lớp j. Recall cho biết trong tất cả ký tự thuộc lớp j thì mô hình phát hiện được bao nhiêu phần trăm. Dễ thấy, ta đều muốn cả precision và recall càng cao càng tốt, nên chỉ số F1-score sẽ được sử dụng để dung hòa 2 chỉ số này.

Tài liệu

1 2018, p. Competition on Vietnamese Online

2 2010, p. THE MNIST DATABASE of handwritten

3 2017, p. IEEE Computer Vision and Pattern Recognition

4 2021, p. Giới thiệu bài toán Scene Text Detection

5 2022, p. Deep Learning Architectures for Sequence Processing

6 2019, p. Lý thuyết về mạng LSTM