

# Phân loại ký tự tiếng Anh (0-9, a-z, A-Z) dựa trên kiến trúc ResNet18 và bộ dữ liệu Chars74K

Phạm Tấn Đức  
Khoa Công Nghệ Thông Tin  
ĐH Nông Lâm TP.HCM  
23130106@st.hcmuaf.edu.vn

Trần Lê Công Hiếu  
Khoa Công Nghệ Thông Tin  
ĐH Nông Lâm TP.HCM  
23130108@st.hcmuaf.edu.vn

Trần Lợi Phát  
Khoa Công Nghệ Thông Tin  
ĐH Nông Lâm TP.HCM  
23130107@st.hcmuaf.edu.vn

## Mục lục

### Mục lục

#### Mục lục

#### I GIỚI THIỆU

#### II CÔNG VIỆC LIÊN QUAN

#### III PHƯƠNG PHÁP ĐỀ XUẤT

III-A	Tổng quan phương pháp	1
III-B	Tập dữ liệu (Dataset)	1
III-C	Tiền xử lý dữ liệu	2
III-D	Chia tập dữ liệu	2
III-E	Mô hình đề xuất – ResNet18 Fine-tuning	2
III-F	Huấn luyện mô hình	2
III-G	Suy luận (Inference)	3
III-H	Triển khai hệ thống	3
III-I	Baseline so sánh	3
III-J	Kết luận	3

#### IV THỰC NGHIỆM VÀ THẢO LUẬN

IV-A	Thiết lập thực nghiệm	3
IV-B	Chỉ số đánh giá	3
IV-C	Kết quả thực nghiệm	3
IV-D	Thảo luận	3

#### V KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

**Tóm tắt nội dung**—Trong bối cảnh chuyển đổi số ngày càng mạnh mẽ, việc nhận diện ký tự quang học (OCR) đóng vai trò then chốt trong các hệ thống tự động hóa. Nghiên cứu này tập trung vào bài toán phân loại 62 lớp ký tự tiếng Anh (bao gồm chữ số 0–9, chữ cái thường a–z và chữ cái hoa A–Z) dựa trên tập dữ liệu Chars74K. Chúng tôi đề xuất sử dụng kiến trúc mạng nơ-ron tích chập sâu ResNet18 để trích xuất đặc trưng hình thái phức tạp của các loại font chữ khác nhau. Mô hình được đánh giá toàn diện trên tập test, với Accuracy đạt 91.33%, Top-3 Accuracy đạt 97.5% và Log loss là 0.21, cho thấy khả năng nhận diện chính xác cao và tính ổn định của mô hình trên các biến thể ký tự. Để minh họa khả năng ứng dụng thực tế, nghiên cứu xây dựng một ứng dụng web bằng Flask cho phép

người dùng tải ảnh và hiển thị kết quả dự đoán kèm theo ma trận nhầm lẫn (confusion matrix) trực quan.

**Từ khóa**—ResNet-18, Chars74K, Phân loại ký tự, Học sâu, Flask.

#### 1

#### I. GIỚI THIỆU

Nhận diện ký tự quang học (Optical Character Recognition - OCR) là một trong những bài toán kinh điển nhưng vẫn luôn giữ được tầm quan trọng cốt lõi trong lĩnh vực Thị giác máy tính và Trí tuệ nhân tạo. Với sự bùng nổ của dữ liệu số, việc chuyển đổi chính xác các ký tự từ định dạng hình ảnh sang văn bản máy tính có khả năng chính sửa là bước tiền đề không thể thiếu cho các hệ thống tự động hóa, từ việc xử lý hóa đơn, nhận diện biển số xe đến hỗ trợ người khiếm thị đọc văn bản. Thủ thách lớn nhất trong bài toán này nằm ở sự đa dạng về hình thái của các ký tự. Cùng một chữ cái nhưng có thể có hàng ngàn biến thể khác nhau về font chữ (Serif, Sans-serif, Bold, Italic) hoặc các đặc điểm viết tay riêng biệt. Đặc biệt, việc phân loại đồng thời 62 lớp ký tự (bao gồm chữ số 0–9, chữ cái hoa A–Z và chữ cái thường a–z) đòi hỏi mô hình phải có khả năng trích xuất đặc trưng cực kỳ tinh tế để phân biệt các cặp ký tự có độ tương đồng cao như '0' và 'O', '1' và 'l', hay 'p' và 'P'. Dự án này tập trung nghiên cứu và xây dựng hệ thống phân loại ký tự tiếng Anh dựa trên bộ dữ liệu Chars74K (phiên bản Digital English Font). Chúng tôi đề xuất áp dụng kiến trúc mạng nơ-ron tích chập sâu ResNet-18 (Residual Network), một kiến trúc mạnh mẽ nhờ cơ chế kết nối tắt (skip connections) giúp vượt qua rào cản về triệt tiêu đạo hàm trong các mạng sâu. Qua đó, mô hình có thể học được các đặc trưng từ mức độ đơn giản đến phức tạp một cách hiệu quả.

#### 4

#### II. CÔNG VIỆC LIÊN QUAN

Các nghiên cứu về nhận dạng ký tự quang học (OCR) và phân loại hình ảnh ký tự đã trải qua nhiều giai đoạn phát triển với nhiều hướng tiếp cận khác nhau.

Trong giai đoạn đầu, các phương pháp truyền thống thường dựa trên việc trích xuất đặc trưng thủ công như HOG (Histogram of Oriented Gradients), SIFT hoặc dùng các thuật toán học máy cổ điển như Support Vector Machines (SVM) và k-Nearest Neighbors (k-NN). Mặc dù các phương pháp này có ưu điểm về tốc độ tính toán, nhưng chúng gặp hạn chế lớn

khi đổi mặt với sự biến đa dạng của font chữ, kích thước và các biến dạng hình học phức tạp của ký tự trong bộ dữ liệu thực tế.

Sự bùng nổ của học sâu (Deep Learning) đã mở ra một kỷ nguyên mới cho bài toán phân loại ký tự. Các kiến trúc mạng no-ron tích chập (CNN) sơ khai như LeNet-5 đã chứng minh hiệu quả vượt trội trong việc tự động học các đặc trưng từ ảnh xám. Tuy nhiên, khi đổi mặt với số lượng lớp lớn (như 62 lớp ký tự trong Chars74K), các mạng nồng thường gặp hiện tượng bão hòa và không thể học được các đặc trưng ngữ cảnh sâu.

Gần đây, sự ra đời của các kiến trúc mạng sâu hơn như VGG và đặc biệt là ResNet (Residual Network) đã giải quyết được bài toán triệt tiêu đạo hàm (vanishing gradient) khi tăng độ sâu của mạng. Kiến trúc ResNet-18, với cơ chế "kết nối tắt" (Skip Connections), đã trở thành một trong những lựa chọn hàng đầu cho các tác vụ phân loại hình ảnh nhờ sự cân bằng tối ưu giữa độ chính xác và chi phí tài nguyên. Bên cạnh đó, xu hướng sử dụng mô hình tiền huấn luyện (Pre-trained models) và kỹ thuật tinh chỉnh (Fine-tuning) đã giúp các hệ thống đạt được hiệu năng cao ngay cả trên các bộ dữ liệu chuyên biệt mà không cần huấn luyện lại từ đầu, tạo nền tảng vững chắc cho việc triển khai các ứng dụng OCR thực tế.

### III. PHƯƠNG PHÁP ĐỀ XUẤT

#### A. Tổng quan phương pháp

Trong đồ án này, chúng tôi đề xuất một phương pháp phân loại ký tự tiếng Anh tự động bao gồm chữ số và chữ cái viết hoa/viết thường (0–9, A–Z, a–z), dựa trên mô hình học sâu Convolutional Neural Network (CNN) với kiến trúc ResNet18.

Phương pháp được xây dựng theo hướng *Transfer Learning*, trong đó mô hình ResNet18 đã được huấn luyện trước trên tập dữ liệu ImageNet sẽ được tinh chỉnh (fine-tuning) để phù hợp với bài toán nhận dạng ký tự. Cách tiếp cận này giúp tận dụng tri thức thị giác tổng quát đã học, giảm thời gian huấn luyện và nâng cao độ chính xác.

Quy trình tổng quát của hệ thống được mô tả như sau:

Dataset → Preprocessing → Fine-tuning ResNet18 → Inference → Deployment (Flask)

#### B. Tập dữ liệu (Dataset)

Dataset được sử dụng trong đồ án là **Chars74K – Digital English Font**, bao gồm các ảnh ký tự tiếng Anh dạng in. Tập dữ liệu có tổng cộng 62 lớp, tương ứng với:

- 10 chữ số: 0–9
- 26 chữ cái viết hoa: A–Z
- 26 chữ cái viết thường: a–z

Mỗi lớp chứa khoảng 1016 ảnh, đảm bảo tính cân bằng dữ liệu. Ảnh có định dạng PNG, kênh màu RGB (8-bit), kích thước gốc  $128 \times 128$ . Ký tự được thể hiện bằng màu đen trên nền trắng, nên ảnh sạch và ít nhiễu.

Nhận được ánh xạ theo cấu trúc thư mục SampleXX, trong đó nhãn được xác định theo công thức:

$$\text{label} = \text{Sample\_index} - 1 \in [0, 61] \quad (1)$$

#### C. Tiền xử lý dữ liệu

Để đảm bảo dữ liệu đầu vào phù hợp với mô hình ResNet18 đã được huấn luyện trên ImageNet, chúng tôi áp dụng chuỗi biến đổi ảnh (transform) như sau:

- 1) **Resize**: Đưa tất cả ảnh về kích thước  $64 \times 64$ .
- 2) **ToTensor**: Chuyển ảnh từ dạng PIL Image sang Tensor PyTorch và chuẩn hóa giá trị pixel về khoảng  $[0, 1]$ .
- 3) **Normalize**: Chuẩn hóa ảnh theo bộ tham số Mean và Std của ImageNet: **Normalize**: Chuẩn hóa ảnh theo bộ tham số Mean và Std của ImageNet, với Mean =  $[0.485, 0.456, 0.406]$  và Std =  $[0.229, 0.224, 0.225]$ .

Việc chuẩn hóa theo ImageNet giúp mô hình tận dụng tốt hơn các trọng số đã được học sẵn, từ đó cải thiện tốc độ hội tụ và hiệu suất phân loại.

#### D. Chia tập dữ liệu

Tập dữ liệu được chia ngẫu nhiên thành ba phần nhằm đảm bảo tính khách quan trong quá trình đánh giá:

- Tập huấn luyện (Training): 70%
- Tập kiểm tra (Validation): 20%
- Tập kiểm thử (Test): 10%

Việc chia dữ liệu được thực hiện bằng phương pháp xáo trộn ngẫu nhiên chỉ mục và sử dụng Subset trong PyTorch, nhằm tránh hiện tượng rò rỉ dữ liệu.

#### E. Mô hình đề xuất – ResNet18 Fine-tuning

ResNet18 là một mạng no-ron tích chập sâu gồm 18 lớp học, sử dụng cơ chế *Residual Block* với kết nối tắt (skip connection) để giải quyết vấn đề suy giảm độ chính xác khi mạng trở nên sâu.

Ý tưởng cốt lõi của ResNet được mô tả bởi công thức:

$$H(x) = F(x) + x \quad (2)$$

Trong đồ án này, chúng tôi sử dụng ResNet18 được huấn luyện trước trên ImageNet và thực hiện các điều chỉnh sau:

- Thay thế lớp Fully Connected cuối cùng từ 1000 đầu ra thành 62 đầu ra, tương ứng với số lớp ký tự.
- Cho phép cập nhật toàn bộ trọng số của mô hình trong quá trình fine-tuning.

#### F. Huấn luyện mô hình

Quá trình huấn luyện được thực hiện với các thiết lập sau:

- Hàm mất mát: CrossEntropyLoss
- Thuật toán tối ưu: Adam
- Learning rate:  $1 \times 10^{-3}$
- Batch size: 64
- Số epoch: 20
- Thiết bị: GPU (CUDA)

Trong mỗi epoch, mô hình thực hiện lan truyền xuôi để dự đoán, tính loss, lan truyền ngược để cập nhật trọng số và đánh giá độ chính xác trên tập validation. Sau khi huấn luyện hoàn tất, mô hình được lưu dưới dạng file .pth để phục vụ suy luận.

## G. Suy luận (Inference)

Trong giai đoạn suy luận, ảnh đầu vào được xử lý với cùng pipeline preprocessing như tập test. Mô hình ResNet18 đã huấn luyện sẽ dự đoán xác suất cho 62 lớp, từ đó lựa chọn:

- Ký tự có xác suất cao nhất (Top-1)
- Danh sách Top-K kết quả có xác suất cao

Chi số dự đoán được ánh xạ ngược lại thành ký tự tương ứng trong bộ ký tự chuẩn.

## H. Triển khai hệ thống

Hệ thống được triển khai dưới dạng một ứng dụng web sử dụng framework Flask. Người dùng có thể tải lên ảnh ký tự, hệ thống sẽ thực hiện suy luận và trả về kết quả dự đoán gần như tức thời.

Việc triển khai bằng Flask giúp hệ thống dễ dàng mở rộng thành API hoặc tích hợp vào các ứng dụng thực tế như OCR và nhận dạng văn bản tự động.

## I. Baseline so sánh

Để đánh giá hiệu quả của mô hình học sâu, chúng tôi xây dựng một mô hình baseline sử dụng Logistic Regression. Ảnh được chuyển sang grayscale, resize về  $32 \times 32$ , sau đó làm phẳng thành vector đặc trưng.

Mô hình baseline đạt độ chính xác khoảng 85%, trong khi ResNet18 đạt 91.33%, cho thấy mô hình học sâu có khả năng học các đặc trưng hình ảnh phức tạp như nét chữ và đường cong tốt hơn.

## J. Kết luận

Phương pháp đề xuất dựa trên fine-tuning ResNet18 cho thấy hiệu quả cao trong bài toán phân loại ký tự tiếng Anh, đạt độ chính xác cao và có khả năng triển khai thực tế. Đây là giải pháp cân bằng giữa độ chính xác, tốc độ xử lý và chi phí tính toán, phù hợp cho các hệ thống nhận dạng ký tự tự động.

# IV. THỰC NGHIỆM VÀ THẢO LUẬN

## A. Thiết lập thực nghiệm

Thực nghiệm được tiến hành trên tập dữ liệu **Chars74K – Digital English Font**, bao gồm các ảnh ký tự tiếng Anh dạng in, được gán nhãn tương ứng với 62 lớp ký tự gồm chữ số (0–9), chữ cái viết hoa (A–Z) và chữ cái viết thường (a–z).

Dữ liệu được tiền xử lý thông qua các bước chuẩn hóa ảnh, bao gồm resize ảnh về kích thước  $64 \times 64$ , chuyển đổi sang tensor và chuẩn hóa theo bộ tham số Mean và Std của ImageNet. Sau đó, tập dữ liệu được chia thành ba phần: tập huấn luyện (70%), tập validation (20%) và tập kiểm thử (10%). Việc chia dữ liệu được thực hiện ngẫu nhiên nhằm đảm bảo tính khách quan trong đánh giá.

Ba mô hình được sử dụng để so sánh trong thực nghiệm bao gồm:

- Logistic Regression với đặc trưng ảnh được làm phẳng (flatten) từ ảnh grayscale.
- Mô hình CNN ResNet18 được fine-tuning trên tập Chars74K.

- Mô hình ResNet18 pretrained được sử dụng làm mô hình chính để xuất.

Tất cả các mô hình được huấn luyện và đánh giá trên cùng tập dữ liệu để đảm bảo tính công bằng trong so sánh.

## B. Chỉ số đánh giá

Hiệu năng của các mô hình được đánh giá thông qua các chỉ số phổ biến trong bài toán phân loại đa lớp, bao gồm Accuracy, Precision, Recall và F1-score. Các chỉ số này được định nghĩa như sau:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Trong đó,  $TP$ ,  $TN$ ,  $FP$  và  $FN$  là lần lượt biểu thị số lượng mẫu dự đoán đúng dương, đúng âm, sai dương và sai âm. Đối với bài toán đa lớp, các chỉ số Precision, Recall và F1-score được tính theo trung bình macro (macro-average) để phản ánh hiệu suất tổng thể trên tất cả các lớp ký tự.

## C. Kết quả thực nghiệm

Bảng I trình bày kết quả so sánh hiệu năng của các mô hình trên tập kiểm thử.

Bảng I  
So sánh hiệu năng các mô hình trên tập test

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (Flatten)	85.00%	N/A	N/A	N/A
ResNet18 (fine-tuned)	91.33%	N/A	N/A	N/A

Kết quả thực nghiệm cho thấy mô hình **ResNet18 fine-tuned** đạt hiệu năng cao nhất, với độ chính xác trên tập kiểm thử đạt **91.33%**. Mô hình hội tụ tốt sau khoảng 20 epoch huấn luyện, với training loss giảm đều và validation accuracy đạt giá trị cao nhất ở epoch 18.

Trong khi đó, mô hình Logistic Regression sử dụng đặc trưng ảnh làm phẳng chỉ đạt độ chính xác khoảng 85%. Điều này cho thấy phương pháp truyền thống gặp hạn chế trong việc học các đặc trưng hình ảnh phức tạp như đường cong, nét chữ và cấu trúc không gian của ký tự.

## D. Thảo luận

So với mô hình baseline, ResNet18 cho thấy ưu thế rõ rệt nhờ khả năng tự động trích xuất đặc trưng phân cấp từ dữ liệu ảnh. Các lớp tích chập trong ResNet18 học được các đặc trưng cơ bản như cạnh và góc ở các lớp đầu, trong khi các lớp sâu hơn nắm bắt được hình dạng và cấu trúc tổng thể của ký tự.

Việc sử dụng kỹ thuật fine-tuning trên mô hình pretrained ImageNet giúp tăng tốc độ hội tụ và cải thiện độ chính xác, đặc

biệt trong bối cảnh tập dữ liệu có quy mô vừa như Chars74K. Mặc dù xuất hiện dấu hiệu overfitting nhẹ ở các epoch cuối, mô hình vẫn duy trì hiệu năng cao trên tập kiểm thử.

Nhìn chung, kết quả thực nghiệm khẳng định rằng các mô hình CNN sâu kết hợp với Transfer Learning, điển hình là ResNet18, là lựa chọn hiệu quả cho bài toán phân loại ký tự tiếng Anh, vượt trội so với các phương pháp truyền thống dựa trên đặc trưng thủ công.

## V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong nghiên cứu này, chúng tôi đã xây dựng một hệ thống phân loại ký tự tiếng Anh dựa trên ResNet18 fine-tuned trên tập dữ liệu Chars74K. Mô hình đạt độ chính xác 91.33% trên tập test, vượt trội so với mô hình baseline Logistic Regression. Kết quả cho thấy mô hình học sâu có khả năng trích xuất đặc trưng hình ảnh phức tạp và ổn định trên các biến thể ký tự khác nhau.

Trong tương lai, có thể cải thiện hiệu năng bằng cách áp dụng các mạng sâu hơn, kỹ thuật tăng cường dữ liệu (data augmentation), hoặc mở rộng từ bài toán phân loại ký tự đơn lẻ sang nhận diện chuỗi ký tự (OCR trên văn bản dài). Hệ thống cũng có thể được tích hợp trực tiếp vào các ứng dụng thực tế như nhận dạng biển số, xử lý hóa đơn, hay hỗ trợ người khiếm thị.