

Báo báo tiến độ công việc từ 17-23/07/2023

Hồ Anh Khoa

Giảng viên hướng dẫn

PGS.TS Đinh Điền

Khoa công nghệ thông tin
Trường đại học Khoa học Tự Nhiên

Tháng 7, năm 2023






Nội dung

- 1 Tạo cơ sở dữ liệu cho meta-learning từ tập Lục Vân Tiên và Truyện Kiều
- 2 Tạo cơ sở dữ liệu từ tập Đại Việt Sử Ký Toàn Thư
- 3 Tìm hiểu cách chạy code

Nội dung

- 1 Tạo cơ sở dữ liệu cho meta-learning từ tập Lục Vân Tiên và Truyện Kiều
- 2 Tạo cơ sở dữ liệu từ tập Đại Việt Sử Ký Toàn Thư
- 3 Tìm hiểu cách chạy code

Tạo dữ liệu huấn luyện với tập Truyện Kiều và Lục Vân Tiên

This PC > DATA (D:) > Master > OCR_Nom > dataset > datasets > mono-domain-datasets >				
	Name	Date modified	Type	Size
	 luc-van-tien	7/19/2023 3:05 PM	File folder	
	 tale-of-kiieu	7/19/2023 3:05 PM	File folder	
				

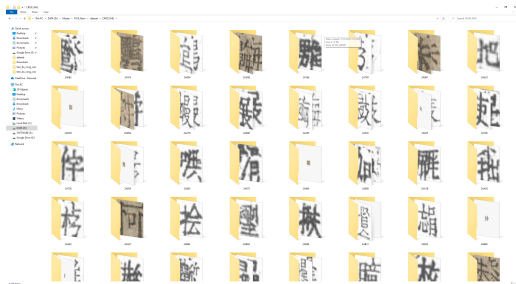
Hình: Tập dữ liệu truyện Kiều và Lục Vân Tiên

Tập dữ liệu thu được

Dữ liệu huấn luyện gồm

- 4049 chữ Nôm, tương ứng với 4049 lớp cần phân lớp.
- Mỗi lớp sẽ chứa các ảnh của chữ Nôm.
- Tên của thư mục là mã Unicode của chữ Nôm.

Tập dữ liệu thu được tiếp theo



Hình: Kết quả tạo dữ liệu huấn luyện

Nội dung

- 1 Tạo cơ sở dữ liệu cho meta-learning từ tập Lục Vân Tiên và Truyện Kiều
- 2 Tạo cơ sở dữ liệu từ tập Đại Việt Sử Ký Toàn Thư
- 3 Tìm hiểu cách chạy code

Tập dữ liệu Đại Việt Sử Ký Toàn Thư

DVSKTT_thu_l_1a	DVSKTT_thu_l_1a	DVSKTT_thu_l_1a	DVSKTT_thu_l_1a	DVSKTT_thu_l_1a	DVSKTT_thu_l_1a	DVSKTT_thu_l_1a	DVSKTT_thu_l_1b	DVSKTT_thu_l_1b	DVSKTT_thu_l_1b
_0	_1	_2	_3	_4	_5	_6	_0	_1	_2
DVSKTT_thu_l_2b	DVSKTT_thu_l_2b	DVSKTT_thu_l_2b	DVSKTT_thu_l_2b	DVSKTT_thu_l_2b	DVSKTT_thu_l_2b	DVSKTT_thu_l_3a	DVSKTT_thu_l_3a	DVSKTT_thu_l_3a	DVSKTT_thu_l_3a
_1	_2	_3	_4	_5	_6	_0	_1	_2	_3
DVSKTT_thu_l_4b	DVSKTT_thu_l_4b	DVSKTT_thu_l_4b	DVSKTT_thu_l_4b	DVSKTT_thu_l_4b	DVSKTT_thu_l_4b	DVSKTT_thu_l_4b	DVSKTT_thu_l_4b	DVSKTT_thu_l_5a	DVSKTT_thu_l_5a
_1	_2	_3	_4	_5	_6	_7	_8	_0	_1
DVSKTT_thu_l_5b	DVSKTT_thu_l_5b	DVSKTT_thu_l_5b	DVSKTT_thu_l_5b	DVSKTT_thu_l_6a	DVSKTT_thu_l_6a	DVSKTT_thu_l_6a	DVSKTT_thu_l_6a	DVSKTT_thu_l_6a	DVSKTT_thu_l_6a
_5	_6	_7	_8	_0	_1	_2	_3	_4	_5

Hình: Dữ liệu đầu vào của tập Đại Việt Sử Ký Toàn Thư

Khó khăn khi tạo dữ liệu

Khó khăn

Dữ liệu nằm **đọc** và khó phân tách một hình ảnh của câu thành các ảnh riêng lẻ của mỗi chữ Nôm.



Hình: Dữ liệu trong Đại Việt Sử Ký Toàn Thư

Cách tiếp cận đã thử

Các tiếp cận đã thử

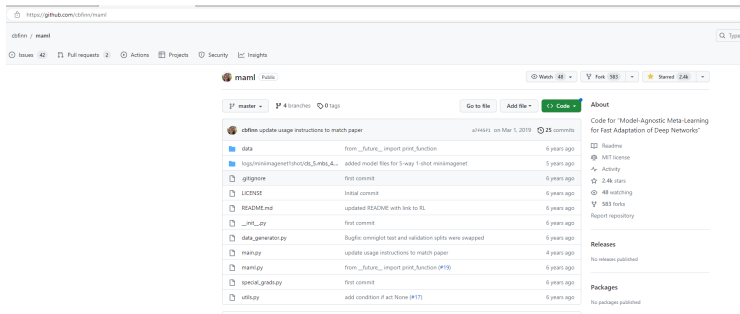
- 1 Sử dụng biên cạnh và Contour.
- 2 Sử dụng mô hình EAST cho việc phát hiện chữ Nôm.

KẾT QUẢ VẪN CHƯA TÁCH ĐƯỢC :(

Nội dung

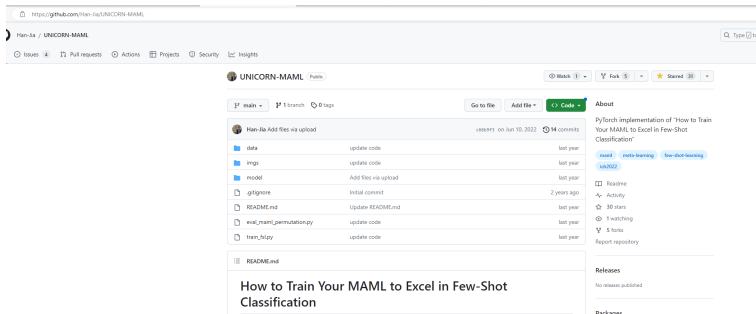
- 1 Tạo cơ sở dữ liệu cho meta-learning từ tập Lục Vân Tiên và Truyện Kiều
- 2 Tạo cơ sở dữ liệu từ tập Đại Việt Sử Ký Toàn Thư
- 3 Tìm hiểu cách chạy code

MAML



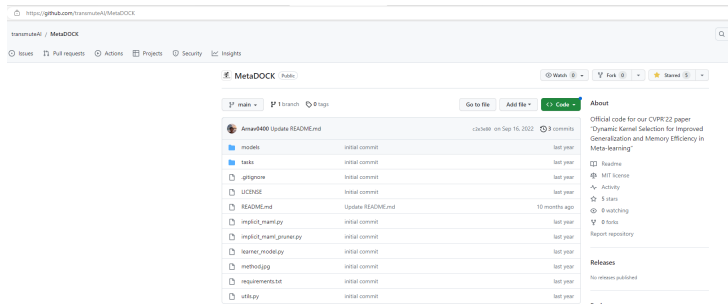
Hình: Github của phương pháp MAML

UNICORN-MAML



Hình: Github của phương pháp UNICORN-MAML

MetaDOCK



Hình: Github của phương pháp MetaDOCK

Kế hoạch tuần này (24 - 30.07)

Kế hoạch tuần này

- 1 Tiếp tục tìm cách để tách hình ảnh của câu.
- 2 Tìm kiếm phương pháp Text Detection khác cho từng từ.
- 3 Sinh tự động hình ảnh chữ Nôm ít dữ liệu bằng [TextRecognitionDataGenerator](#)
- 4 Huấn luyện và đánh giá 3 phương pháp MAML, UNICORN-MAML và METADOCK

**Cám ơn mọi người
đã lắng nghe!!!**