Tên đề tài : Image caption generate và ứng dụng .

Mục tiêu : Tạo một model Image Caption Generate bằng tiếng việt và thêm các tính năng tốt nhất có thể để tạo nên một sản phẩm hoàn chỉnh .

Các thông tin về tạo model Image Caption Generate :

Data set :

Image Data : Kaggle .

Caption tiếng việt : Sử dụng model tiếng anh có sẵn để sinh caption bằng tiếng anh sau đó sử dụng model envit5-translation để sinh ra các caption dữ liệu để có được các caption có mô tả gần mới model hướng tới nhất .

Training :

Cơ sơ vật chất :

+) google cloud platform (train image , deployment , prediction) 1.375$ / hour

+) The HuggingFace :+) Nvidia L4 VRAM 14GB MEM 30GB 0.8$/hours

+) Nvidia A10G VRAM 24GB MEM 30GB 1$/hours

+) AWS : Các instance

P4 instance : Nvidia A100 GPU (dùng train model suất cao)

[p4d.24xlarge](https://us-east-1.console.aws.amazon.com/ec2/home?region=us-east-1#InstanceTypeDetails:instanceType=p4d.24xlarge) : 32.7726 / hour

G5 instance : Nvidia A10G GPU (dùng train model với hiệu suất trung )

G5. large 1.19 $/hour

G5g.2x large 1.58$/hour

Budget : +) hiện tại 300$ trên gg cloud platform đến 9/12/2024

+) 60$ cash (hiện tại)

+) 250$ (có thể sẵn sàng chi cho đến khi hoàn thành dự án theo từng giai đoạn )

Hướng phát triển dự án :

Hướng 1 : Hoàn thiện mô hình image captioning đã tạo , sau đó train kết hợp với model dense captioning để tạo ra một caption chi tiết nhất có thể dành cho bức ảnh đầu vào , rồi kết hợp với text -to-speech model để tạo nên một mô tả về image cho thính giả .

Hướng 2 : Hoàn thiện mô hình Image Captioning hiện có, sau đó kết hợp với Dense Captioning để tạo ra các mô tả chi tiết cho từng đối tượng trong ảnh. Tiếp tục tích hợp mô hình Text Generation nhằm mở rộng các mô tả chi tiết này, từ đó tạo ra một câu chuyện liền mạch và phong phú. Hệ thống sẽ tự động phát triển nội dung từ các bức ảnh và mô tả đã có, tạo ra những câu chuyện đa dạng, chi tiết và gắn kết, mở rộng khả năng ứng dụng trong sáng tác và truyền thông tự động..

Hướng 3 : Hoàn thiện mô hình image caption bằng tiếng việt ,sau đó kết hợp với Dense Captioning để tạo ra các mô tả chi tiết cho từng đối tượng trong ảnh , sau đó kết hợp multi-model embeding để tạo ra các chiều kết hợp giữa text và image để tạo nên một thanh tìm kiếm 2 chiều với mục đích : Tìm kiếm từ text đến hình ảnh , tìm kiếm từ hình ảnh đến các text thuộc phần caption ( “tìm kiếm image đến các image liên quan thông caption kèm theo các caption liên quan “ tham khảo ).

Hướng 4 : Hoàn thiện mô hình image caption bằng tiếng việt , kết hợp với Dense Captioning để tạo ra các mô tả chi tiết cho từng đối tượng trong ảnh, sau đó kết hợp Q&A model để trả lời các câu hỏi về thông tin trong ảnh đã được mô tả qua 2 model trước đó .