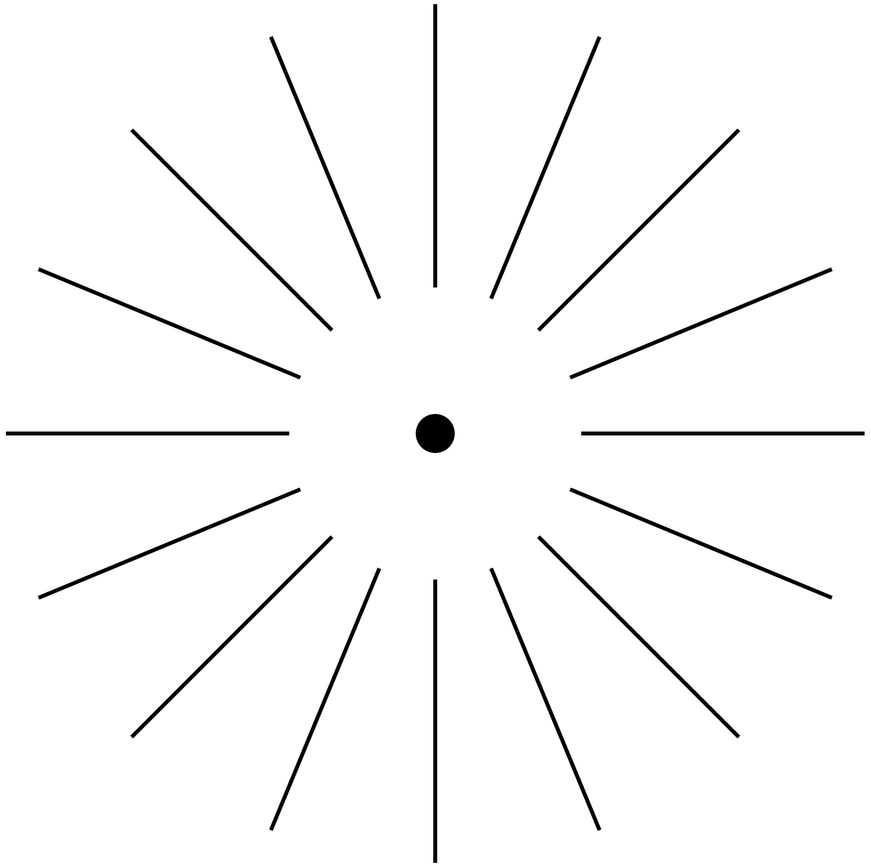
# Chú thích nội dung, hoạt động từ hình ảnh

GVHD: Võ Thanh Hùng



# Thành viên

1. Đoàn Trần Cao Trí - 2010733
2. Bùi Quốc Khánh - 2010325
3. Bùi Hoàng Minh - 2010410

Liên hệ nhóm: doantrancaotri1108@gmail.com

# Giới thiệu

- Image captioning là quá trình tạo ra mô tả văn bản tự động cho hình ảnh nhờ các mô hình máy học được huấn luyện để nhận diện và mô tả nội dung của hình ảnh một cách tự động.

- Mục tiêu của image captioning là kết hợp khả năng nhận diện hình ảnh và hiểu ngữ cảnh để tạo ra các mô tả tự nhiên và chính xác cho hình ảnh.

- Các ứng dụng:
  - hỗ trợ truy vấn hình ảnh
  - tăng cường trải nghiệm người dùng trong các ứng dụng như truyền thông xã hội, thương mại điện tử và nhiều lĩnh vực khác.

# Mục lục

# Pokemon

Size: 833 images

Train-test ratio:  70% - 30%

A large, orange dragon Pokémon with a long tail and wings.

A small, blue turtle Pokémon with a long tail.

A small, yellow cat Pokémon with a coin-shaped forehead.

# Flickr8k

Size: 8,092 images

Train-test ratio:  6,000 - 2,092

A big grey dog wearing a chain collar has a smaller brown dog pinned down.

A woman wearing a straw hat and floral skirt is sitting on cement stairs with her head in her hands.

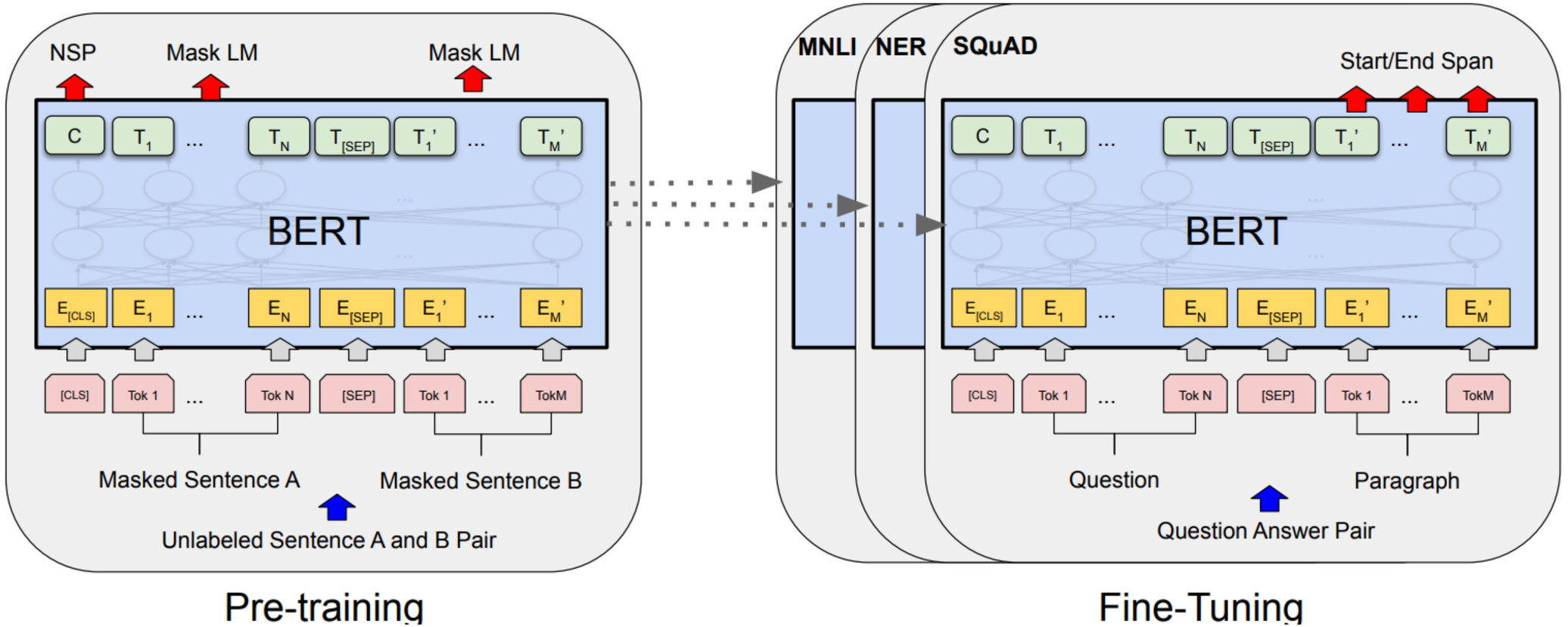
An older woman and a young girl standing in front of a large bush covered in pink flowers.

# BERT

BERT là một mô hình ngôn ngữ (Language Model) được phát triển bởi Google AI.

Kiến trúc hai chiều của BERT được hỗ trợ bởi một kỹ thuật gọi là self-attention. Self-attention cho phép BERT học các phụ thuộc dài hạn trong văn bản, điều này là cần thiết để hiểu ý nghĩa của các câu phức tạp.

BERT có thể được sử dụng cho nhiều loại tác vụ xử lý ngôn ngữ tự nhiên (NLP): **Phân loại văn bản, Trả lời câu hỏi, Kết luận ngôn ngữ tự nhiên, Dịch máy**

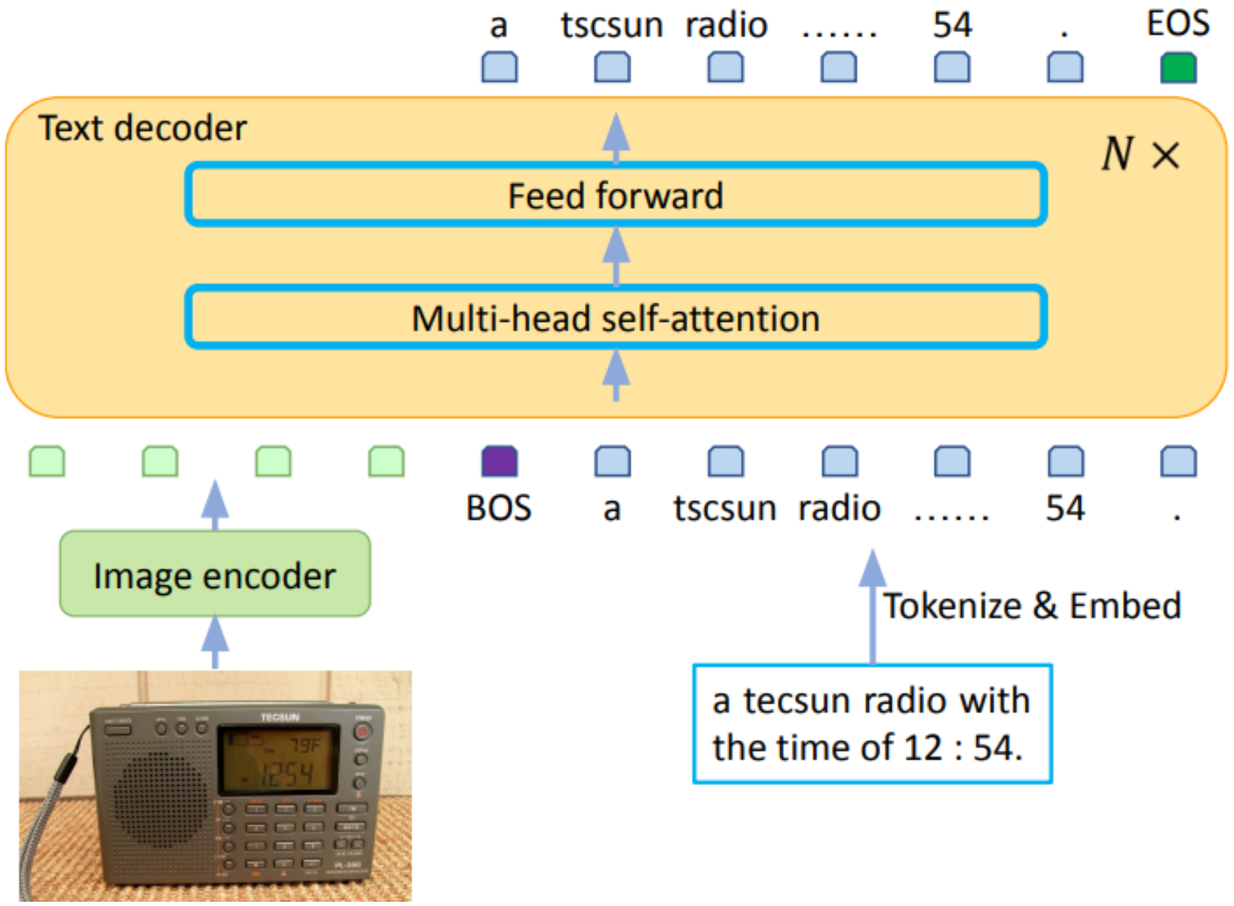


# Microsoft - GIT

GenerativeImage2Text (GIT) là một mô hình ngôn ngữ (Language Model) được phát triển bởi Microsoft AI.

GIT là một mô hình chuyển đổi (Transformer) hai chiều, được đào tạo trên một tập dữ liệu khổng lồ gồm hình ảnh và văn bản mô tả hình ảnh.

GIT có thể được sử dụng cho nhiều loại tác vụ xử lý ngôn ngữ tự nhiên (NLP) liên quan tới hình ảnh: **Mô tả hình ảnh, Trả lời câu hỏi về hình ảnh, Tạo hình ảnh từ văn bản**



# Công cụ train

**Colab** cung cấp truy cập miễn phí đến **GPU T4** của Google. GPU này được phát triển bởi NVIDIA và được tối ưu hóa cho công việc học máy và tính toán thông qua mạng.
Dưới đây là một số thông số cơ bản của GPU T4:

- Kiến trúc: Turing
- GPU RAM: 15GB
- Hỗ trợ cho Tensor Cores để tăng tốc tính toán tensor

# Pokemon

- learning rate: 5e - 5
- train batch size: 8
- eval batch size: 16
- time training: 18 minutes
- early stopping: True

| Epoch | Training Loss | Validation Loss | Wer Score |
|---|---|---|---|
| 0 | 9.158200 | 7.557196 | 74.681159 |
| 1 | 6.862600 | 5.745167 | 21.407076 |
| 2 | 4.977800 | 3.866543 | 11.232310 |
| 4 | 1.366500 | 0.636763 | 4.313299 |
| 5 | 0.377500 | 0.142961 | 1.605712 |
| 6 | 0.088700 | 0.047337 | 0.849531 |
| 8 | 0.022900 | 0.030367 | 0.506394 |
| 9 | 0.017700 | 0.029864 | 0.462063 |
| 10 | 0.013300 | 0.029311 | 1.227195 |
| 12 | 0.009100 | 0.030604 | 1.867860 |

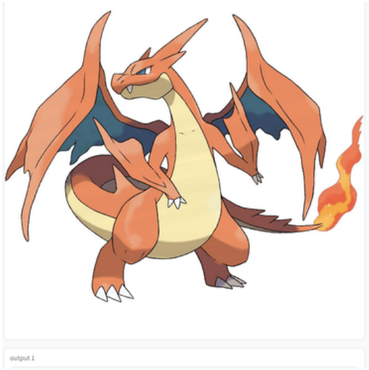
# Flickr8k

- learning rate: 5e - 5
- train batch size: 8
- eval batch size: 16
- time training: 40 minutes
- early stopping: True

| Step | Training Loss | Validation Loss | Wer Score |
|---|---|---|---|
| 50 | 7.114700 | 4.257147 | 1.803492 |
| 100 | 2.012600 | 0.283577 | 6.577424 |
| 150 | 0.113500 | 0.069973 | 8.001378 |
| 200 | 0.050100 | 0.067654 | 10.451218 |
| 250 | 0.033500 | 0.068245 | 10.812529 |
| 300 | 0.022200 | 0.071887 | 12.210752 |
| 350 | 0.014100 | 0.074315 | 9.185633 |
| 400 | 0.008400 | 0.077360 | 10.007352 |
| 450 | 0.005900 | 0.079654 | 10.657528 |
| 500 | 0.004200 | 0.081749 | 10.687548 |
| 550 | 0.003100 | 0.083451 | 10.299740 |
| 600 | 0.002400 | 0.083748 | 10.847144 |

# Kết quả dự đoán

# Pokemon



a drawing of a fire breathing dragon

a candle with a blue flame on top of it

a red and blue dragon with its mouth open

# Kết quả dự đoán

# Flickr8k

output 1

a young boy in blue jeans and a white shirt is standing by a rock wall.

a young boy in blue jeans and a white shirt is standing by a rock wall

output 1

a little girl in a pink outfit is looking at the camera.

a little girl in a pink outfit is looking at the camera

output 1

a man in a yellow vest is driving a yellow vehicle through the woods.

a man in a yellow vest is driving a yellow vehicle through the woods

# Tổng kết

- Đề tài về image captioning trên tập dataset Pokémon và Flickr là một nghiên cứu thú vị trong lĩnh vực trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên.

- Tập dataset Pokémon cung cấp một bộ sưu tập hình ảnh về các Pokémon khác nhau, trong khi tập dataset Flickr bao gồm một loạt ảnh từ cộng đồng người dùng trên toàn thế giới.

- Qua đó, nghiên cứu này không chỉ mở ra những tiềm năng trong việc ứng dụng image captioning trong lĩnh vực trò chơi và giải trí mà còn đề xuất các ứng dụng tiềm năng trong việc xử lý hình ảnh đa dạng từ các nguồn dữ liệu khác nhau.