Image Caption Generator

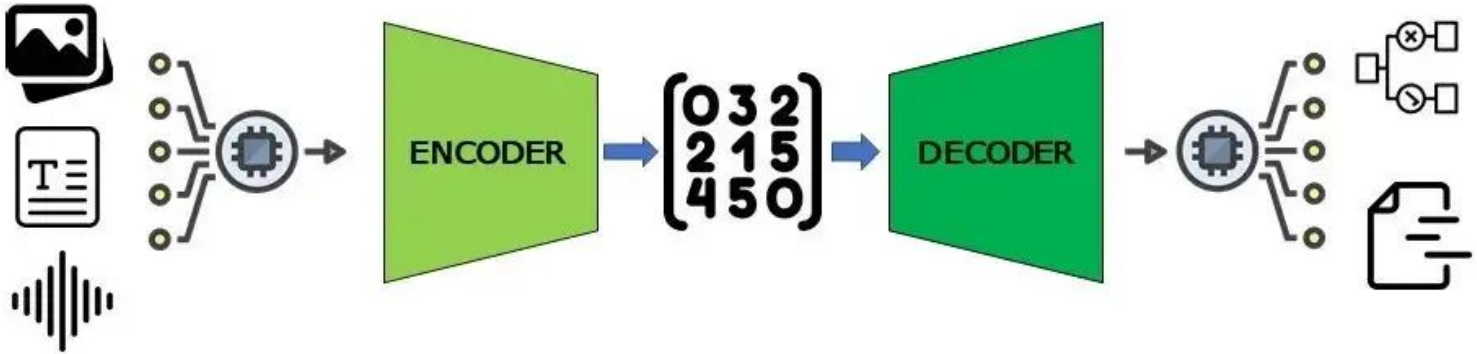
By

Jithesh KB
Poojashree NS
Tharun Mukka
Priyanka Math

Introduction

- End-to-end sequence-to-sequence embedding task
  - input sequences are the image pixels
  - output is a caption that describes the image
- Mutual exclusivity of image and text sequences
- Requires the use of two interconnected models
  - Image encoding model
  - Text decoding model

Encoder-Decoder architecture



Flickr 8K Dataset

- Collection of 8,000 images
- 5 possible captions for each image

Why?

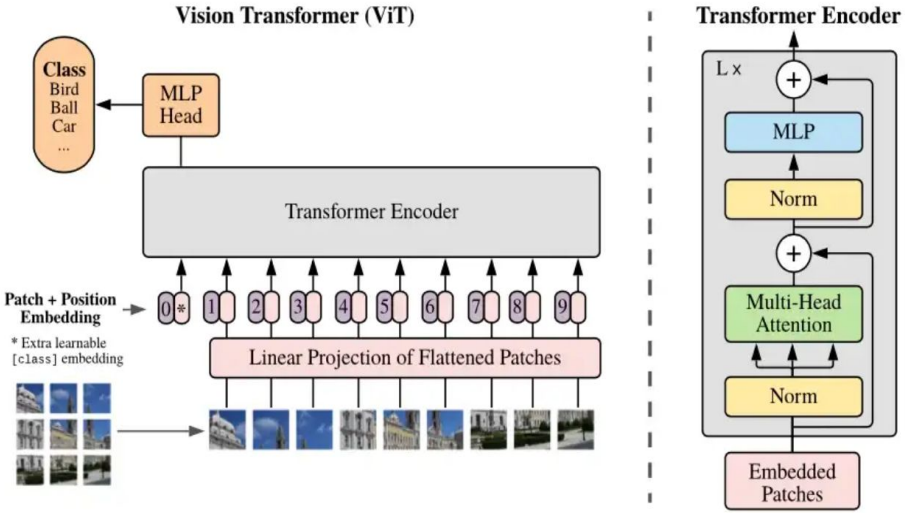
- Small in size, hence can be trained easily on low-end laptops/desktops
- Data is properly labelled
- Available for free

Encoder & Decoder approach

- Image captioning can be tackled as **Encoder & Decoder Task**
  - Task 1 - **Input to the decoder is an Image**, tackled by a Vision Transformer (ViT)
  - Task 2 - **Output from the Decoder is Text describing the objects in the image**, which can be tackled by Roberta, BERT or any other state of the art Language model

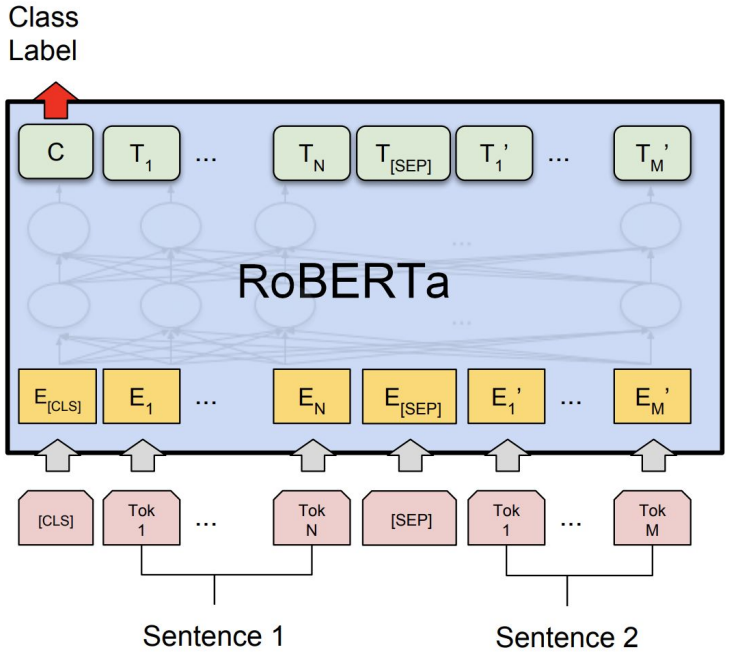
Encoder - Vision Transformers

- Conveys the idea that each image can be seen as a group of images that can be compared to tokens or words in a sentence
- Any CV task can utilize this 3D image tensor tokens as an input to a transformer
- Overcomes CNN's problem of limited association



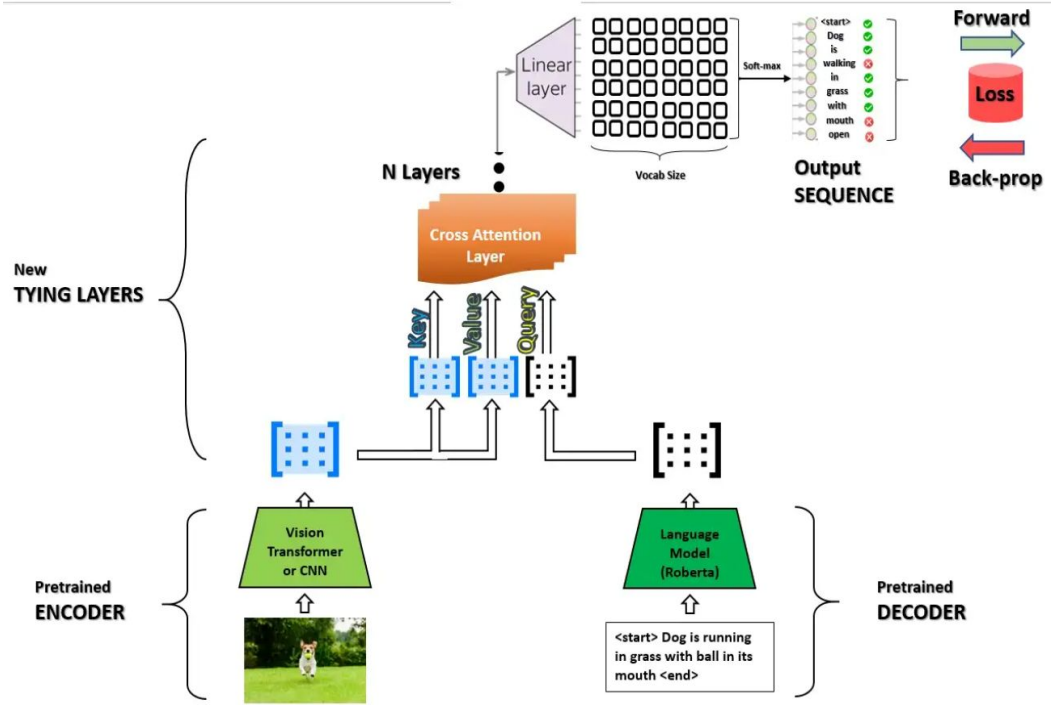
Decoder - RoBERTa

- Includes fine-tuning the original BERT model
- Trained on a vast dataset that goes over 160GB of uncompressed text
- Model learns to predict intentionally hidden sections of text within otherwise unannotated language examples
- Saves training time and complexity on the actual task of captioning



Architecture - Vision Encoder and Decoder

- Pre-trained models initialize the vision encoder decoder
- Encoder embeddings are employed as KEY & VALUE
- Decoder embeddings as QUERY
- Illustration of teacher-forcing instruction



Metrics

- **Perplexity**, it is a commonly used metric for measuring a language model's performance
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**, it measures number of matching n-grams between the sequence generated by the model and desired sequence
    - Precision
    - Recall
    - F1 score

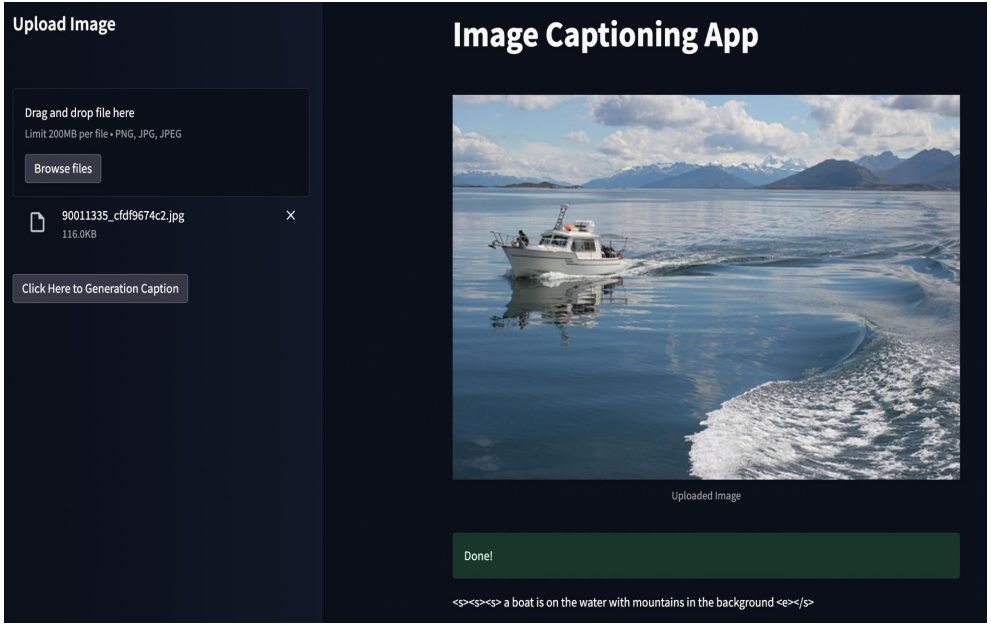
MLOps

MLflow : https://dagshub.com/poojashreeNS/CMPE_297_FinalProject.mlflow/#/experiments/3

Dagshub: https://dagshub.com/poojashreeNS/CMPE_297_FinalProject

Web Application

- Streamlit web app for the interaction between UI and the trained Encoder-Decoder model
- Deployed on Huggingface spaces
- Deployment Link



Demo

Thank You