

# Report on Streamlit Application for Fine-tuning CLIP Model

Name: Yuvraj Varma

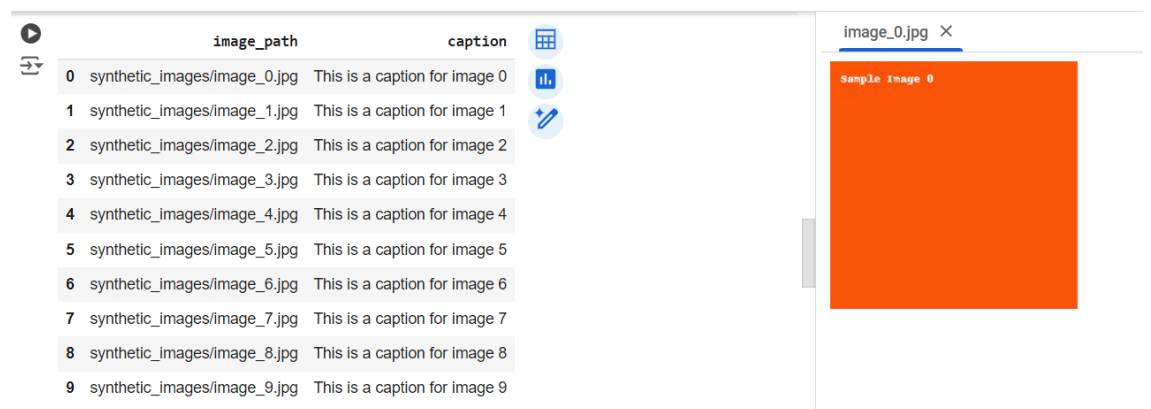
Topic: Open Source multi-modal fine-tuning on a particular domain.

## Introduction:

The goal of this project is to develop a simple Streamlit application that fine-tunes a CLIP (Contrastive Language-Image Pre-Training) model using a synthetic dataset and allows users to interactively test the fine-tuned model by uploading images and entering captions. The application will demonstrate the model's capability to associate images with textual descriptions and vice versa.

## Methodology:

1. Primary Libraries used:
  - a. Transformers
  - b. Torch
  - c. Streamlit
  - d. Pillow
  - e. Datasets
  - f. Gradio
2. Data Preparation
  - a. Generating Synthetic Images
  - b. Captioning the images



The screenshot displays a Streamlit application interface. On the left, a table lists synthetic images and their captions. On the right, an image viewer shows a sample image.

|   | image_path                   | caption                       |
|---|------------------------------|-------------------------------|
| 0 | synthetic_images/image_0.jpg | This is a caption for image 0 |
| 1 | synthetic_images/image_1.jpg | This is a caption for image 1 |
| 2 | synthetic_images/image_2.jpg | This is a caption for image 2 |
| 3 | synthetic_images/image_3.jpg | This is a caption for image 3 |
| 4 | synthetic_images/image_4.jpg | This is a caption for image 4 |
| 5 | synthetic_images/image_5.jpg | This is a caption for image 5 |
| 6 | synthetic_images/image_6.jpg | This is a caption for image 6 |
| 7 | synthetic_images/image_7.jpg | This is a caption for image 7 |
| 8 | synthetic_images/image_8.jpg | This is a caption for image 8 |
| 9 | synthetic_images/image_9.jpg | This is a caption for image 9 |

image\_0.jpg ×

Sample Image 0

- c. Forming the dataset structure

### 3. Model and Processor Initialization

The CLIP model (openai/clip-vit-base-patch32) from the Hugging Face library. The model and processor are initialized here using Transformers Library.

## 4. Data Preprocessing

Preprocess the dataset by tokenizing the text and preparing the images using the CLIP processor. This is done using the map function of the Hugging Face Dataset.

Map: 100%  8/8 [00:00<00:00, 37.63 examples/s]

Map: 100%  2/2 [00:00<00:00, 18.27 examples/s]

## 5. Training

A custom Trainer class is defined to handle the training process with a custom loss function for the CLIP model. The training arguments specify the batch size, number of epochs, and evaluation strategy.

## 6. Evaluation

The model is evaluated on the validation dataset after training.

## 7. Streamlit Application

A Streamlit application is created to allow users to interactively test the fine-tuned model. Users can upload an image and enter a caption, and the model predicts the association between the image and the text.

### Conclusion:

This project successfully demonstrates the fine-tuning of a CLIP model on a synthetic dataset and provides an interactive Streamlit application for testing the model. Users can upload images and enter captions to see the model's predictions. The steps outlined in this report provide a comprehensive guide to setting up, training, and deploying a multi-modal machine learning model using state-of-the-art tools and libraries.