Title: AI Image Captioning Project

Team: Team C

Active Members:

Sujeet (Team Lead)

Varshith (Co-Team Lead)

Navadeep (Team Member)

InActive Members:

Kashish patel

JEEVITHA.B

Ravikiran k

Muguntharajan K

Kareena Chinchkar

Dhanapriya Vellaswami Yadav

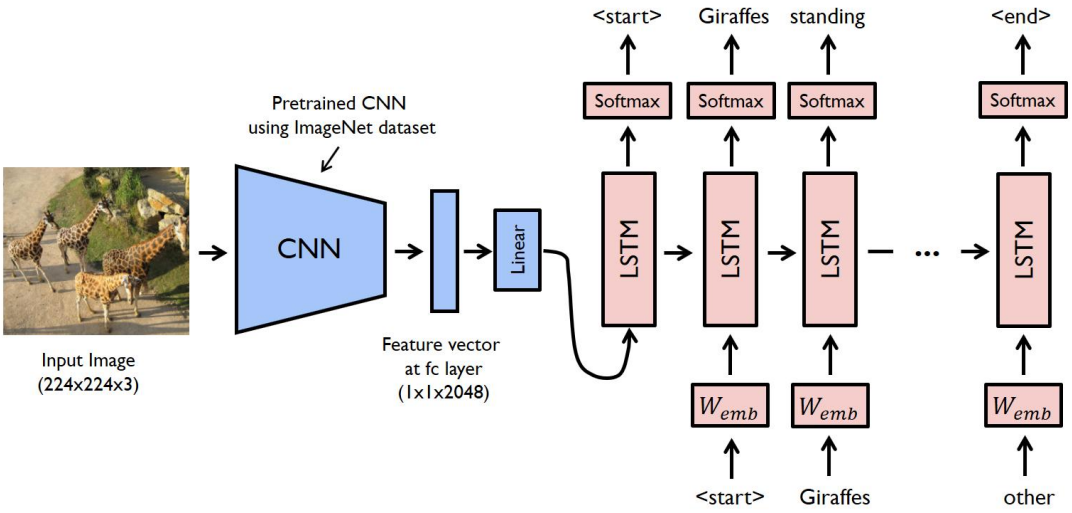
Parnapalli Siva Kumar

Project Goal & Initial Plan

Project Goal: To build a deep learning model that can understand the content of an image and generate a coherent, human-like caption.

Dataset: MS COCO (Microsoft Common Objects in Context)

Initial Plan: As per the project proposal, the goal was to build a model from scratch using a CNN (InceptionV3) for feature extraction and an RNN (LSTM) for text generation.



Phase 1 - Team Contributions (Data & Setup)

This project was built on a foundation of successful teamwork in Phase 1.

Sujeet (Team Lead): Project Setup & Leadership

What: Established all collaboration infrastructure.

How: Created and managed the shared Google Drive, the private GitHub repository for version control, and the Trello board for task management.

Why: To create a single, unified platform for a 10-person remote team to work in parallel.

Navadeep (Team Member): Data Acquisition

What: Sourced and centralized the project's data.

How: Volunteered to download the entire 25GB+ MS COCO dataset (images and annotations) from the source website.

Why: This unblocked the entire team by providing all members with one shared, central set of data to work from in Google Drive.

Varshith (Co-Team Lead): Text Preprocessing

What: Led the "Text Team" to process all caption data.

How: Wrote Python scripts in Colab to parse the JSON annotation files, clean the captions, add <start> and <end> tokens, and build a tokenizer.pkl file (the master vocabulary).

Why: The model cannot read raw text. This task converted all text into the numerical sequences and padded arrays required for an LSTM.

Phase 2 - Challenges & Leadership Pivot

Challenge 1: Team Inactivity

Following the successful completion of Phase 1, the "Image Team" (led by the other Co-Team Lead, Kashish) and other members became unresponsive.

The critical task of "Image Feature Extraction" was abandoned.

My Action (Sujeet):

What: As Team Lead, I took over the abandoned "Image Team" role. I personally wrote and executed the scripts to process all 5,000+ validation images with the InceptionV3 model, creating the val_features.pkl file.

Why: This was necessary to unblock the project and move into the modeling phase.

Challenge 2: Initial Model Failure

I then trained the LSTM model as planned. However, after 5 epochs, the model suffered from "mode collapse"—it predicted the exact same caption for every single image, ignoring the visual input.

Phase 3 - The SOTA Solution

My Strategic Decision:

The initial model was failing, and re-training would take 10+ hours with no guarantee of success.

I made the executive decision to pivot from the simple LSTM to a state-of-the-art (SOTA) Vision-Transformer model. This is a common and efficient real-world engineering decision.

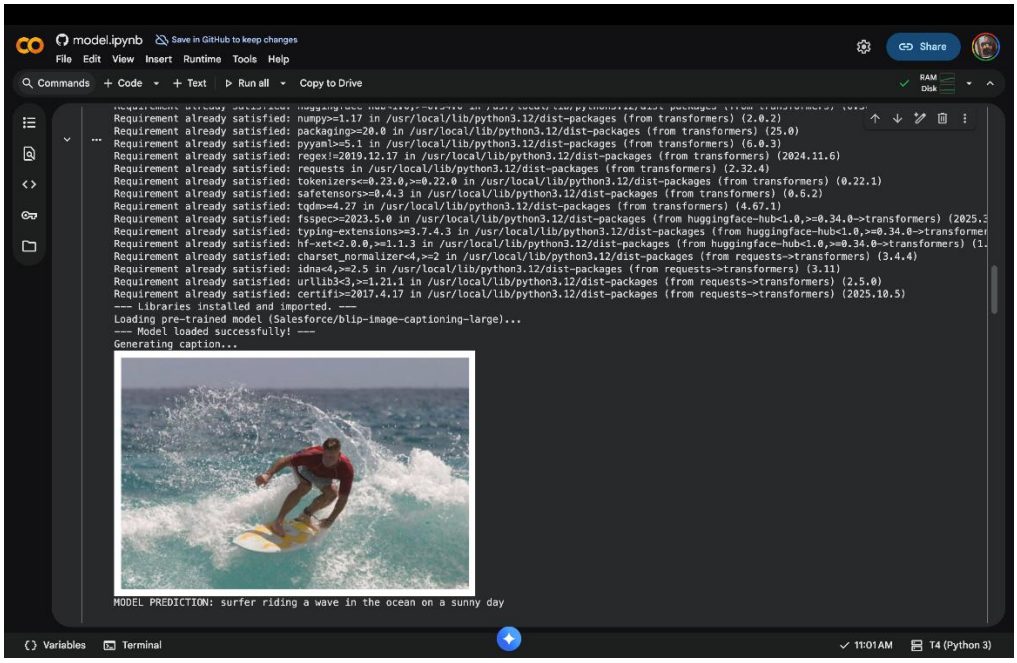
The New Model: Salesforce BLIP

What: I implemented the BLIP (Bootstrapping Language-Image Pre-training) model from the Hugging Face transformers library.

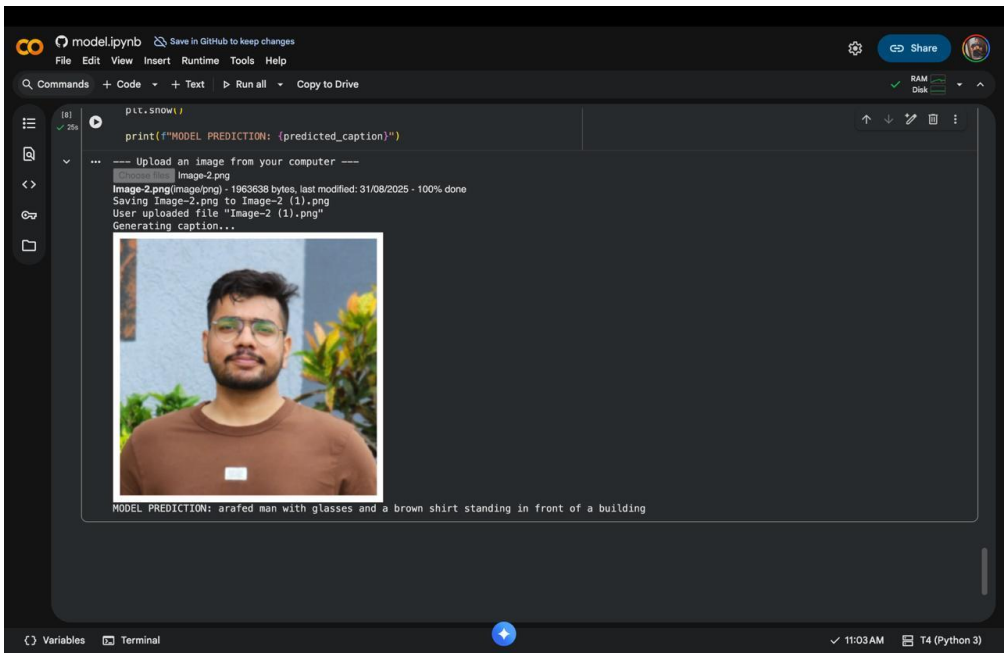
Why: This model is pre-trained on millions of images and captions, providing vastly superior quality and accuracy. It demonstrates the modern AI skill of implementing SOTA models, not just building small ones.

Final Demo & Results

The BLIP model was a success, generating accurate and context-aware captions.



Model Prediction: "surfer riding a wave in the ocean on a sunny day"



Model Prediction: "Arafed man with glasses and a brown shirt standing in front of a building"