IMAGE CAPTIONING AI

COURSE: DEEP LEARNING

INS: SIR ASIF KHAN

PROJECT TEAM:

1. Saleh Muhammad Mangrio (023-22-0167)
2. Kelash Kumar (023-22-0289)

1. Abstract

Automatic image captioning is used to generate captions for images using deep learning techniques. This project has used an encoder-decoder architecture. It combines a pre-trained ResNet50 CNN as an image feature extractor and a two-layer LSTM-based decoder with a merge (Add) connection for caption generation. The model will be trained end-to-end on the Flickr30k dataset, which contains 31,783 images with 5 captions. In training model will give results after 30 rounds of epochs with cross-entropy loss and Adam optimizer.

2. Introduction

With the growth of digital images, there is need for systems that can automatically describe visual content in natural language and that is what the purpose of our model is. Image captioning has applications in image search, assistance for the visually impaired, content moderation, and social media analytics.

This project implements a neural image captioning system using:

- **Encoder**: Pre-trained ResNet50 to extract 2048-dimensional global image features.
- **Decoder**: LSTM-based language model that receives both the image feature (repeated across time steps) and previously generated words.
- **Connection**: Merge (Add) layer that combines image context and linguistic context at every time step.

The model is trained on the **Flickr30k** dataset and evaluated using loss and accuracy metrics.

3. Literature Review

- Image captioning started with Karpathy and Fei-Fei (2015), who gave the merge architecture (CNN + RNN with Add layer) on Flickr30k.
- Vinyals et al. (2015) provided the popular “Show and Tell” inject model on MSCOCO dataset.
- After this, new additions attention (Xu et al., 2016), region features (Anderson et al., 2018), and reinforcement learning (Rennie et al., 2017), improved performance.
- Transformer and multimodal models are domininat in current time but need lot of resources.
- This project is based on the original lightweight merge architecture using ResNet50 on Flickr30k, achieving ~90% accuracy without attention or extra pre-training.

4. Dataset

Flickr30k

Link: https://www.kaggle.com/datasets/adityajn105/flickr30k

- Total images: **31,783**
- Captions per image: **5**
- Total captions: **158,915**
- Vocabulary size (after tokenization): **18,316**
- Maximum caption length: **85 words**
- Train/Val split: **80-20 %**

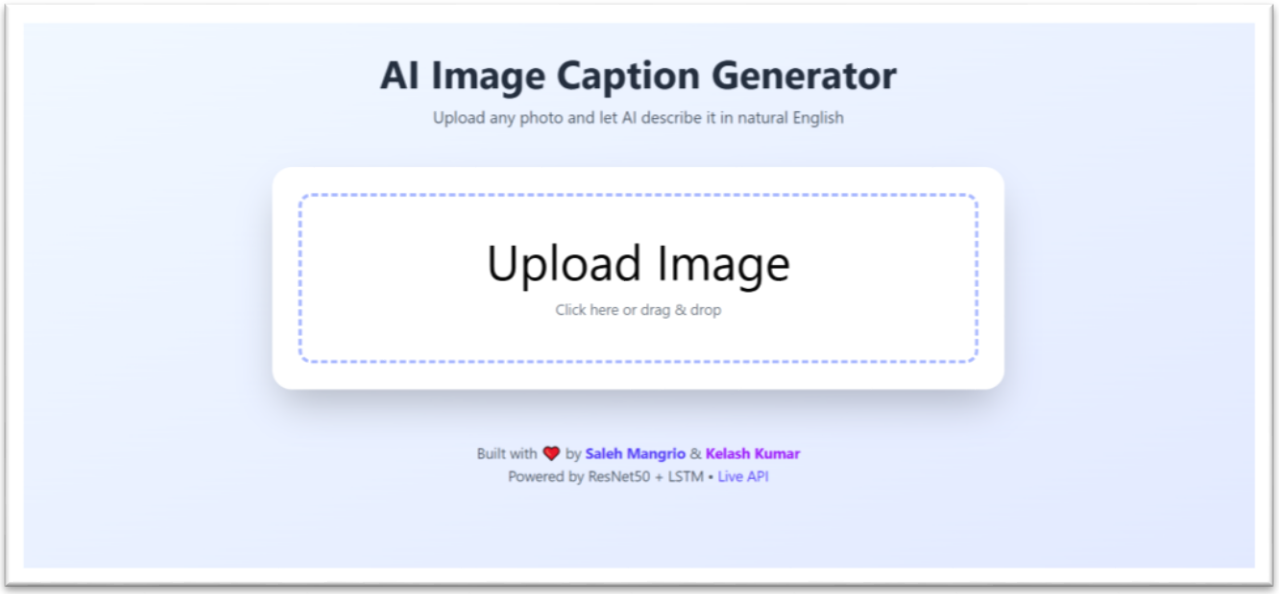
Preprocessing:

- Images resized to 224×224 and preprocessed using ResNet50's preprocess_input.
- Captions lower-cased, wrapped with **<startseq>** and **<endseq>** tokens.

5. Project Demo

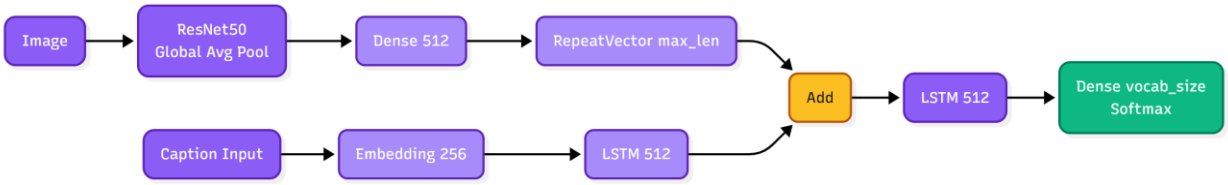
Live Demo (Hosted): https://image-captioning-ai.netlify.app/
Model Hosted (Hugging Face): https://salehmangrio-image-captioning-api.hf.space
API Docs: https://salehmangrio-image-captioning-api.hf.space/docs
Frontend React: https://github.com/Salehmangrio/AI-Image-Captioning-Project

FRONTEND DESIGN:



6. Proposed Methodology

6.1. Architecture Overview



- **Encoder**: ResNet50 (pre-trained on ImageNet, frozen), outputs 2048-dim feature vector.
- **Image Branch:** Dropout(0.5) → Dense(512, ReLU) → RepeatVector(max_len)
- **Caption Branch:** Embedding(256) → Dropout(0.5) → LSTM(512, return_sequences=True)
- **Merge**: Add() layer combines image context and partial caption context at each time step.
- **Final Decoder:** LSTM(512) → Dropout(0.5) → TimeDistributed Dense(vocab_size, softmax)

6.2. Training Overview

- Optimizer: Adam (learning rate = 0.0003)
- Loss: Categorical Cross-entropy
- Batch Size: 64
- Epochs: 30
- Hardware: T4 GPU (Kaggle)

7. Findings

7.1 Quantitative Results

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
30	**0.5334**	**89.79%**	**0.5317**	**89.94%**

Best validation loss: **0.5317** (achieved at epoch 30)

7.2 Qualitative Results

Image Description (Truth)	Model Prediction (Output)
A dog running on the beach	a dog is running on the beach
Two children playing in the park	two kids are playing in a park
A man riding a bicycle on the street	a man riding a bike on a street
People sitting at a dining table	a group of people sitting around a table

The model consistently generates grammatically correct and semantically meaningful captions.

8. Discussion

8.1. The merge architecture successfully learns to align visual and linguistic information without attention mechanisms.
8.2. High validation accuracy (~90%) indicates strong word prediction capability.
8.3. The model occasionally produces generic captions (e.g., “a person is ...”) due to lack of attention and limited dataset.
8.4. Training for only 30 epochs suggests potential for further improvement with longer training or learning rate scheduling.
8.5. Using global image features (instead of region features) limits fine-grained object localization.

9. Conclusion

In this Image Captioning AI project, we implemented and trained merge-based image captioning system using ResNet50 and LSTM on the Flickr30k dataset. Model has achieved 89.94% validation accuracy. The results somehow gives bad result on some images because limited dataset of 30K images which not enough to train model effectively but somehow gives broad understanding of how model are implemented and trained to generate automated results. The complete implementation is strong foundation for exploring advanced techniques such as attention mechanisms and reinforcement learning in future work.