‘IMAGE CAPTION GENARATOR’

Abstract

Project Objectives: The aim of this project is to build an image caption generator by combining computer vision and natural language processing The objective of the project is to produce descriptive captions by utilizing the CNN-LSTM architecture and the Vision Transformer architecture in a contextually sensitive manner. It compares CNN-LSTM network approaches with vision transformers to elicit what works best.

Methodology: The project follows a pipeline approach where the images are first converted into feature vectors using CNNs or Vision Transformers and then given as input to sequence generators like LSTM to generate textual subtitles. The implementation involves:

Load MS COCO Dataset (the dataset is prepared until October 2023.)

CNNs (ResNet50, VGG16), Vision Transformers (DINO, PVT, XCIT, SWIN) formatted input data

Caption generation using LSTM

BLEU, METEOR, ROUGE, CIDEr and SPICE Metric to evaluate performance

Key Findings:

While CNN-based feature extraction has been proven to perform well, vision transformers have been proven to outperform such in both accuracy and efficiency.

On SWIN-Transformer, the highest accuracies are obtained thanks to its data-efficient attention mechanism.

Long range dependencies are hard for CNN-LSTM models, while global characteristics of an image are effectively captured by vision transformers.

Solution Approach:

Step1: Load data (Flickr\_8K, MS COCO), Preprocess data

Step2: Extract image features from image (CNNs, Vision Transformers)

Step3: Tokenize and preprocess captions

Step4: Model training (LSTM-based sequence generation)

Step5: Apply attention mechanisms

Step6: Evaluate metrics to measure the performance of the model

Step7: Evaluate with CNN-LSTM vs Vision Transformers & comprise Results

Reference:

[1] Efficient Image Captioning Based on Vision Transformer Models

SamarElbedwehy1,\*, T. Medhat2, TaherHamza3 and Mohammed F. Alrahmawy3

[2] Automatic image captioning combining natural language processing and

deep neural networks

Antonio M.Rinaldi, CristianoRusso, CristianTommasino

[3] Image Captioning Based on Deep Neural Networks

Shuang Liu1, Liang Bai1,a, Yanli Hu1 and Haoran Wang1

Team:

1. Bejgam Pranav E22CSEU1196

2. Varshith Reddy E22CSEU1104

3. Chaitanya Sai Raghava E22CSEU1102