**Title: Image Caption Generator Using CNN & Vision Transformers**

**Abstract**

**Project Objectives**: Image captioning is an important and challenging AI task which combines computer vision and natural language processing to generate a meaningful textual description of an image. Here, we investigate how valuable CNN-LSTM architectures and Vision Transformers produce valid and applicable captions for image datasets. CNNs are used to extract features on each spatial image, while LSTMs used to produce sequential text using extracted embeddings. Vision Transformers capture the global features of the image in order to generate better quality captions through self-attention mechanisms. MS COCO and Flickr\_8K datasets are covered within the project for the training and validation. We then compare the CNN-LSTM with Vision Transformers to identify the best method. They are assessed with the help of BLEU, METEOR, ROUGE, CIDEr and SPICE results. The accurate findings of Vision Transformers over CNN based models in discriminating fine.

**Methodology:**

The project is executed in a structured manner, starting with the acquisition of the dataset. The team uses the MS COCO and Flickr 8K datasets for the project. Initial data processing involves the conversion of images into numerical data forms. The team converts the images into feature vectors. For this process, the team uses various Convolutional Neural Networks (CNN). The team uses ResNet-50, VGG-16, and various Vision Transformers like DINO, PVT, XCIT, and SWIN to convert the images into input features. The next step in the project is the caption generation process. This process involves the preprocessing of the caption data. The team preprocesses the captions which include tokenization and embedding process. The captions are tokenized and embedding is performed to convert the captions into numerical/sequence form. The team uses a sequence model to predict the captions. The team uses the LSTM sequence generator to predict the captions based on the extracted image features. For the project, the team also includes Vision Transformers as alternative feature extractors. Vision Transformers use self-attention mechanisms and do not have the limitation of the fixed grid. The team uses Vision Transformers as an alternative to the traditional CNN in the project. Attention mechanisms are also used to make the model understand the long dependencies of the images. The model is trained using various optimization techniques, such as fine-tuning and transfer learning. Training techniques are also used to improve the model’s performance. The last section of the project is evaluation, where the team evaluates the model’s performance. For evaluation, the team uses metrics such as BLEU, METEOR, ROUGE, CIDEr, and SPICE. The team uses these metrics to evaluate the accuracy and relevance of the captions. Hyperparameter tuning is performed by adjusting the hyperparameters, e.g., learning rates, batch size, and dropout rate.

**Key Findings:**

Vision Transformers outperform CNN-based models in accuracy and efficiency, effectively capturing global image characteristics. SWIN-Transformer achieves high accuracy with its data-efficient attention mechanism, overcoming CNN-LSTM limitations in handling long-range dependencies.

**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] Jing Jie Tan, Anissa Mokraoui, Ban-Hoe Kwan, Danny Wee-Kiat Ng, Yan-Chai Hum, "Siamese-Driven Optimization for Low-Resolution Image Latent Embedding in Image Captioning", 2024 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp.79-84, 2024.

**Team:**

1.Bejgam Pranav(E22CSEU1196)

2. Varshith Reddy(E22CSEU1104)

3. Chaitanya Sai Raghav(E22CSEU1102)