PROJECT REPORT : Deep Learning Term Project CS60010 Spring Semester 2024

Task: **Build encoder decoder models for Automatic Image Captioning**

Team 26 members

1. Rajanyo Paul - 23CS60R82
2. Dipan Mandal - 23CS60R04
3. Avik Pramanick - 23CS60R78
4. Soham Banerjee - 23CS60R42

Subtasks:

Part A : Design a simple CNN-based encoder for the image and RNN-based decoder model.

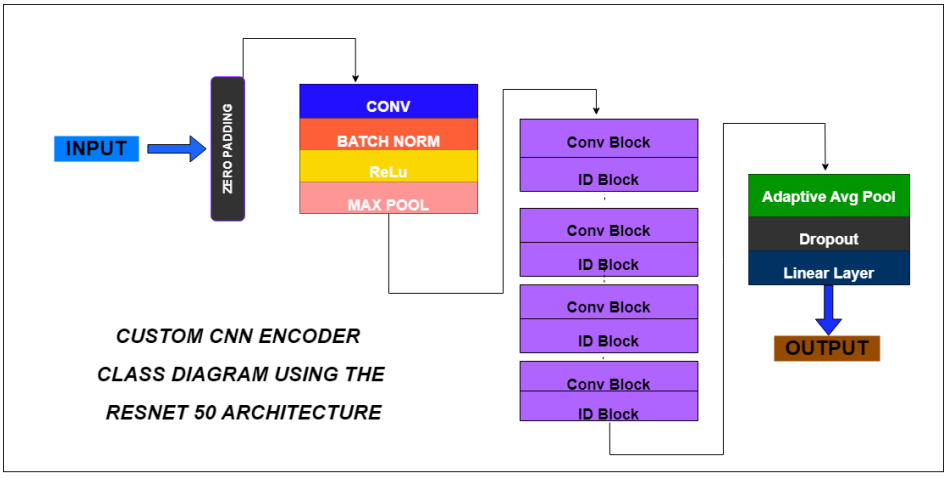
Part B : Design a transformer based encoder decoder model using a Vision Transformer (ViT) as the image encoder and a text decoder.

Part A : Methodology & Results

Data Preprocessing:
Data Collection: The dataset comprises images along with corresponding captions.
Data Parsing and Transformation: Images are resized to a uniform dimension and normalized. Captions are tokenized, and a vocabulary index is created. Special tokens for padding, start, and end of sentences are included to maintain sequence integrity.

Model Architecture:
Feature Extractor: A popular CNN (like ResNet50) is used as a feature extractor to capture visual representations of the images.
Sequence Predictor: An RNN, specifically an LSTM (Long Short-Term Memory) network, is employed to predict the sequence of words that form the caption, based on the features provided by the CNN.

Training the Model:
Loss Function: The training process involves minimizing a loss function i.e, cross-entropy loss, which measures the difference between the predicted word sequence and the actual caption.
Optimizer: An optimization algorithm, such as Adam is used to update the weights of the neural network during training.
Batch Processing: Data is processed in batches to optimize training efficiency.
Data Loaders: Custom data loaders are designed to streamline the process of loading and batching images and captions efficiently.



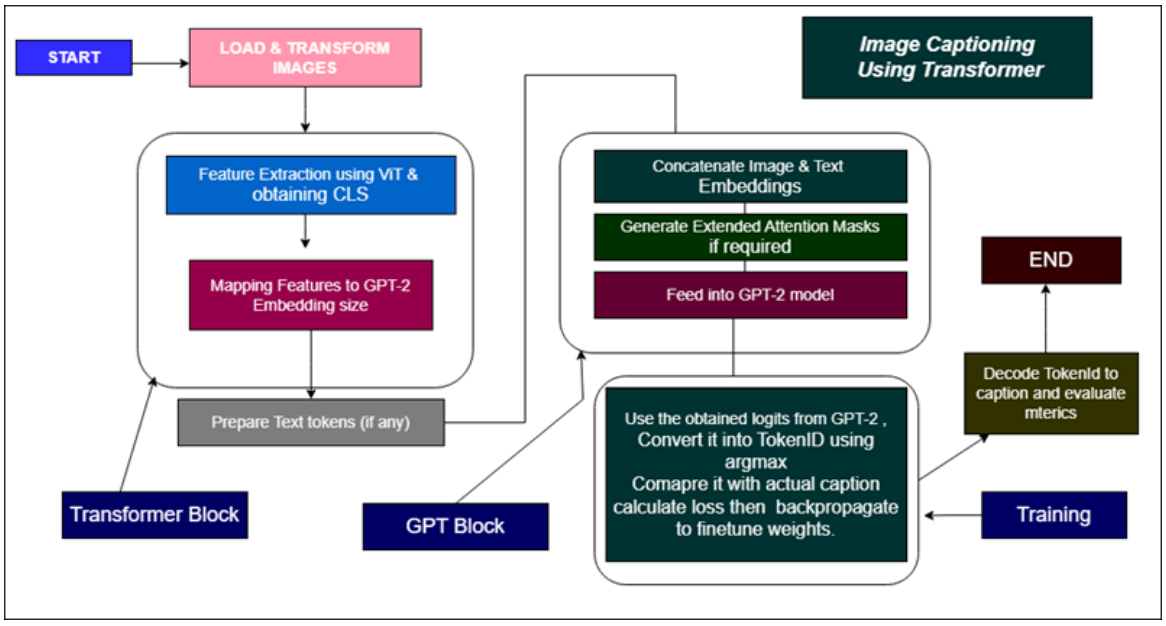
The metric scores for part A
The ROUGE-L score is : 0.269982538788438
The CIDEr score is: 0.028215051499469436
The SPICE score is: 0.09003152037129201

Part B : Methodology & Results

Data Preprocessing:
Images were resized, converted to tensors, and randomly flipped to enhance the dataset. Captions were tokenized using the GPT2Tokenizer, with adjustments made to accommodate variable-length inputs. A CustomDataset class was designed to manage the image-caption pairs, allowing batch processing through PyTorch's DataLoader.

Model Architecture:
We integrated a Vision Transformer (ViT) and GPT-2 to create a model that processes images and generates captions. The ViT model encoded images into embeddings, which were then fed into the GPT-2 model to generate captions based on the image features.

Training the Model:
We trained our model using Cross-Entropy Loss and optimized it with AdamW, focusing on fine-tuning parameters like epochs, batch size, and learning rate. The model diagram is given below.



Metrics Calculation:
We employed ROUGE-L, CIDEr, and SPICE metrics to quantitatively assess the quality of the generated captions. Functions tweak_data() and evaluate_metrics() were implemented to prepare data for metric calculation and to compute the scores. This was the same for both the parts.

The metric scores for part B
The ROUGE-L score is : 0.3971661692771957
The CIDEr score is: 0.3354663638540146
The SPICE score is: 0.13363108855120315

Finally, the model's performance was summarized by outputting the calculated metrics and displaying generated captions alongside their corresponding images for qualitative assessment.

Analysis from both parts

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation):
Scores : Part A (0.270) vs. Part B (0.397) : The higher ROUGE-L score in part B suggests that its captions are more similar to the reference captions in terms of structure and word choice. This indicates better fluency and possibly a more accurate reflection of the key elements in the images compared to part a.

CIDEr (Consensus-based Image Description Evaluation):
Scores : Part A (0.028) vs. Part B (0.335) : Part B's significantly higher CIDEr score indicates that part B's captions are more informative and relevant to the actual content of the images.

SPICE (Semantic Propositional Image Caption Evaluation):
Scores : Part A (0.090) vs. Part B (0.134) : The higher SPICE score for part B suggests that its captions are more semantically accurate, capturing more details about the objects and their interactions within the images. Part b likely provides a richer and more precise description of the image content.

Conclusion

ViT Transformer, as used in part B , has superior performance in understanding context and generating coherent, relevant text compared to the CNN + RNN architecture. This is primarily due to the nature of transformers that allows them to consider all parts of the input data simultaneously. GPT's capabilities in generating fluent and contextually appropriate text further complement the transformer's strengths.