Group 11

ACCURACY ACHIEVERS

Team Members:

Pola Gnana Shekar (21CS10052)
Simma Pavan Kumar (21CS10060)
Chityala Raviteja (21CS30016)
Sumanth Tadigoppala (21CS30053)

Deep Learning Term Project(CS60010)
Spring Semester 2024
AUTOMATIC IMAGE CAPTIONING

Introduction:

The Goal of this project is to build encoder-decoder models for **Automatic Image Captioning**. It requires designing and Implementing two different architectures, an CNN based encoder with a RNN based decoder for Part A, and transformer based encoder-decoder using a Vision Transformer (ViT) for Part B. These models will be trained and evaluated on a provided dataset, with a focus on achieving high performance metrics such as CIDEr, ROUGE-L and SPICE.

Part A: CNN-RNN Model for Image Captioning

Methodology:

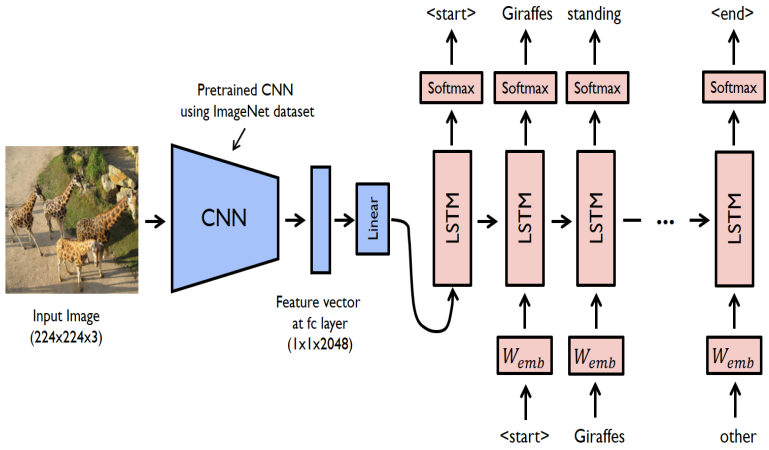
➢ **Preprocessing:**
- Images undergo standard preprocessing steps, including resizing, center cropping, tensor conversion and normalization.
- Captions are tokenized and converted into numerical indices using a vocabulary mapping.

➢ **Model Creation:**
- **Encoder:** Utilizes a pre-trained ResNet-50 CNN to extract high-level image features.

- **Decoder:** Employs an LSTM-based RNN to generate captions. It integrates image features and tokenized captions, leveraging embeddings for context-rich captioning.

➢ **Model Training and Validation:**

- The model is trained using Adam optimizer and cross-entropy loss, ensuring efficient convergence.
- Validation ensures model robustness and prevents overfitting through continuous monitoring.



➢ **Evaluation:**

- Quality assessment is conducted using CIDEr, ROUGE-L and SPICE scores, providing insights into caption generation performance against ground truth.

Results and Analysis:

Metrics	Score	Remarks
CIDEr	0.313	The CIDEr score, at 0.313, indicates a moderate level of agreement between the generated captions and the reference captions
ROUGE-L	0.270	A ROUGE-L score of 0.270 suggests that the generated captions have a moderate overlap with the reference captions in terms of longest common subsequences
SPICE	0.086	A SPICE score of 0.086 indicates a relatively low level of semantic similarity between the generated and reference captions

The model performs moderately well, with a CIDEr score of 0.313 indicating a moderate level of agreement with the reference captions, a ROUGE-L score of 0.270 indicating moderate overlap with the reference captions, and a SPICE score of 0.086 indicating a relatively low level of semantic similarity with the reference captions.

Part B: Transformer Based Image Captioning using ViT

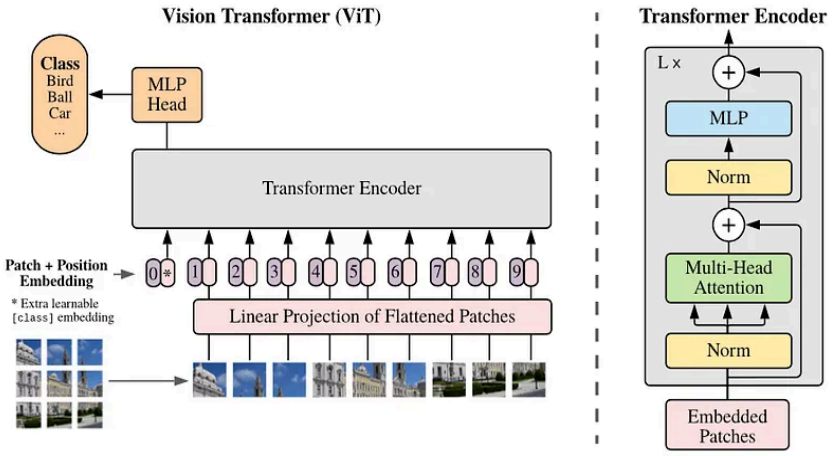
Methodology:

➢ **Preprocessing:**

- The database is structured into train,test and validation sets, with captions truncated to 30 words for uniformity.
- Images are processed using a ViT-based feature extractor, while captions are tokenized using a BERT tokenizer for encoding.

➢ **Model Creation:**

- The VisionEncoderDecoderModel is configured with a pre-trained ViT encoder and a BERT based decoder, facilitating image captioning.
- Special tokens and beam search parameters are set to optimize the model's caption generation capabilities.



➢ **Model Training and Validation:**

- Training is conducted using the Seq2Seq Trainer, incorporating specific training arguments, evaluation metrics, and a subset of the dataset for efficiency.
- To prevent overfitting and improve generalization, early stopping is employed based on the validation loss, halting training when the model's performance on the validation set starts to degrade.

➢ **Evaluation:**

- The model's performance is assessed using CIDEr, ROUGE-L, and SPICE scores. These metrics measure the similarity and relevance of generated captions to the ground truth references.
- CIDEr evaluates diversity and uniqueness, ROUGE-L focuses on n-gram overlap, and SPICE assesses semantic similarity.

Results and Analysis:

Metrics	Score	Remarks
CIDEr	0.214	The CIDEr score of 0.214 indicates a basic level of caption quality, suggesting room for enhancement in generating more descriptive and relevant captions.
ROUGE-L	0.262	The ROUGE-L score of 0.262 indicates that the model's generated captions have a moderate overlap with the ground truth references at the token level, showing some ability to capture the essence of the reference captions.
SPICE	0.097	The SPICE score of 0.097 suggests that the model's generated captions have limited semantic similarity to the ground truth references

The model shows potential and it also has room for improvement. It can generate captions (CIDEr: 0.214) with some overlap (ROUGE-L: 0.262) with the references but lacks semantic similarity (SPICE: 0.097). Further refinement is needed for better quality and relevance.

Conclusion:

In comparing the two models for image captioning, Model A, using a CNN encoder and RNN decoder, demonstrates slightly better performance than Model B, which utilizes a ViT encoder and BERT decoder. Model A achieves a CIDEr score of 0.313, a ROUGE-L score of 0.270, and a SPICE score of 0.086, indicating moderate agreement, overlap, and semantic similarity with the reference captions, respectively. On the other hand, Model B achieves a CIDEr score of 0.214, a ROUGE-L score of 0.262, and a SPICE score of 0.097, suggesting slightly lower performance in terms of agreement, overlap, and semantic similarity. However, both models show potential for improvement to enhance the quality and relevance of the generated captions.