DL Term Project

Team ID: G-04

Team : Not so Deep Learners

Members:

1. 21CS10002 Adarsh Patel
2. 21CS30019 Galipelli Sai Mallikarjun
3. 21CS30042 Ritesah M
4. 21CS10022 Dhruv Agja

Part-A :

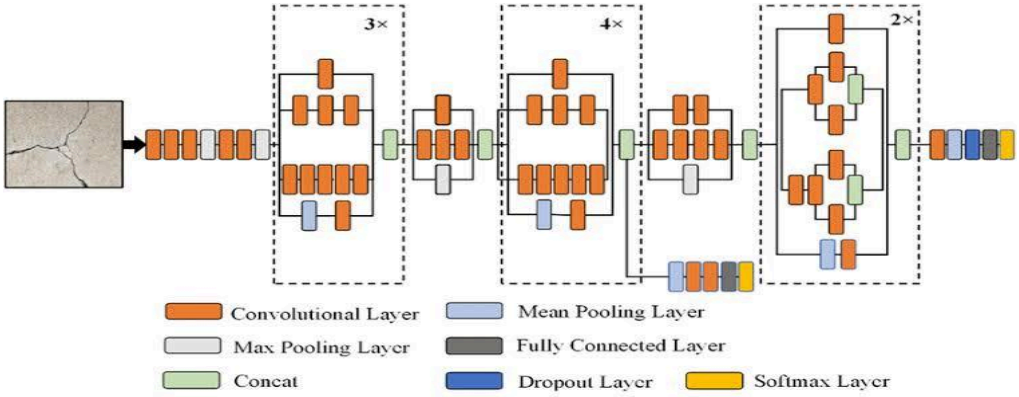
Methodology:

- Preprocessing:

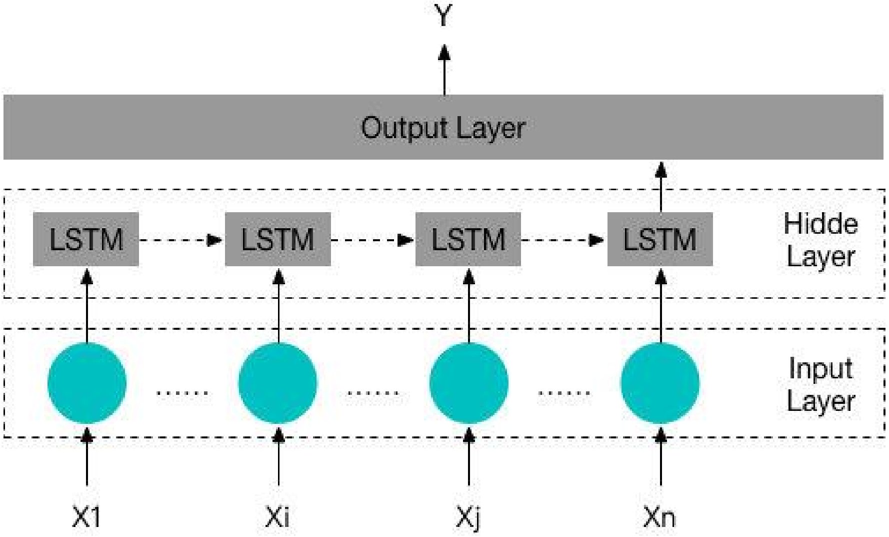
  In the image captioning project, the preprocessing begins with loading the image data and associated captions. Images are resized and normalized using transformations from `torchvision.transforms`, ensuring uniformity in input dimensions. Captions undergo tokenization using `spacy` and are then numericalized with a custom `Vocabulary` class. This step prepares the data for the CNN-RNN model, providing a structured input format.

- Model Creation:

  The Inception-v3 CNN, used in the `EncoderCNN` class, excels at extracting rich, hierarchical features from images, capturing intricate details. This model's depth and convolutional operations aid in understanding image content at various levels. On the other hand, the LSTM-based RNN in `DecoderRNN` generates captions sequentially, leveraging its recurrent nature to maintain context and coherence throughout the caption. Combining these in `CNN_RNN` merges the strengths: image features enrich captioning, while captions guide the model's understanding of visual context, creating a robust framework for image captioning tasks.



Inception_v3 (pretrianed, Encoder)



LSTM (RNN Decoder).

- Results:

Metric	Score
CIDEr	0.03951025476649118
SPICE	0.09861253988760205
Rouge-L Precision	0.331611980959768
Rouge-L Recall	0.23380320720378248
Rouge-L F1-Score	0.26138034767078994

- Analysis:

The CIDEr score highlights room for improvement in aligning with reference captions, while the SPICE score indicates moderate semantic relevance. Fine-tuning the model's architecture and training parameters, along with exploring advanced techniques, could enhance caption quality significantly. These scores offer valuable insights for refining the image captioning model, aiming for more accurate and contextually relevant captions.

Part-B :

Methodology:

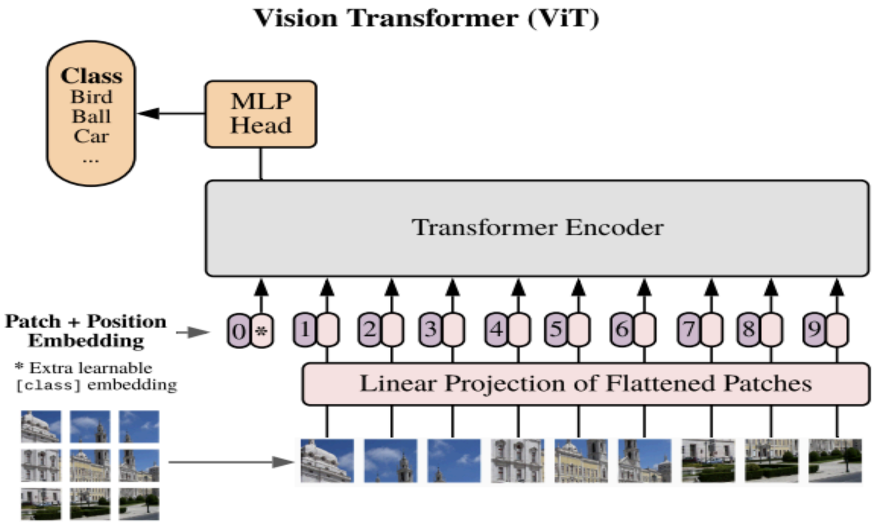
- Preprocessing:

The image captioning project begins with preprocessing steps, where the ImageCaptionDataset class is utilized to handle image and caption data. This involves loading CSV files containing image paths and captions for both training and validation datasets. Images are processed using the ViTImageProcessor for pixel value extraction, while captions undergo tokenization using the BertTokenizer. This preprocessing ensures the data is properly formatted and ready for input into the VisionEncoderDecoderModel.

- Model Creation:

The VisionEncoderDecoderModel represents a great fusion of the Vision Transformer (ViT) and BERT architectures, revolutionizing image captioning tasks. ViT excels at extracting image features, utilizing self-attention mechanisms to capture global context efficiently. BERT, renowned for its language understanding abilities, is seamlessly integrated, providing a deep comprehension of textual input. The model's configuration, with specialized start and padding tokens, ensures precise caption generation. By combining ViT's visual understanding with BERT's semantic prowess, this model achieves a holistic understanding of image-text relationships. Its robust architecture guarantees captions that not only describe image content but also align with the nuanced semantics, setting new standards in the field of multimodal AI.

The model is fine-tuned on the training dataset to adapt its parameters and weights specifically to the image-caption pairs in the training set. This process optimizes the model's ability to generate accurate and contextually relevant captions.


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

- Results:

Metric	Score
CIDEr	0.10201522379203559
SPICE	0.13907834708641043
Rouge-L Precision	0.34221915579261264
Rouge-L Recall	0.26280162830809567
Rouge-L F1-Score	0.27451505099608364

- Analysis:

The CIDEr score of 0.102 reflects moderate consensus with reference captions, while the SPICE score of 0.139 emphasizes the model's strength in producing semantically rich captions. These scores indicate the model's success in accurately describing image content, promising performance in generating contextually appropriate captions.