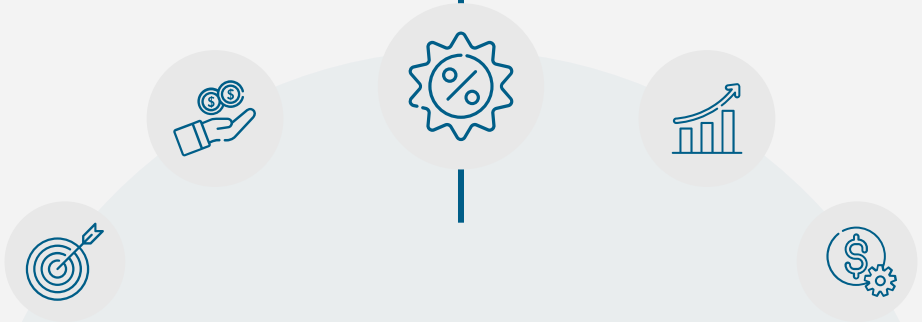


IMAGE CAPTION GENERATOR

Visionary Descriptions: Advancing Image Captions with LSTM-RNNs and Attention

NLP - 21AIE314
END SEM PROJECT

GROUP 10 BATCH A

TEAM MEMBERS

SANJANA MCS	***CB.EN.U4AIE21029***
ABINAYA N	***CB.EN.U4AIE21001***
SIDDARTH D	***CB.EN.U4AIE21064***

PROBLEM STATEMENT

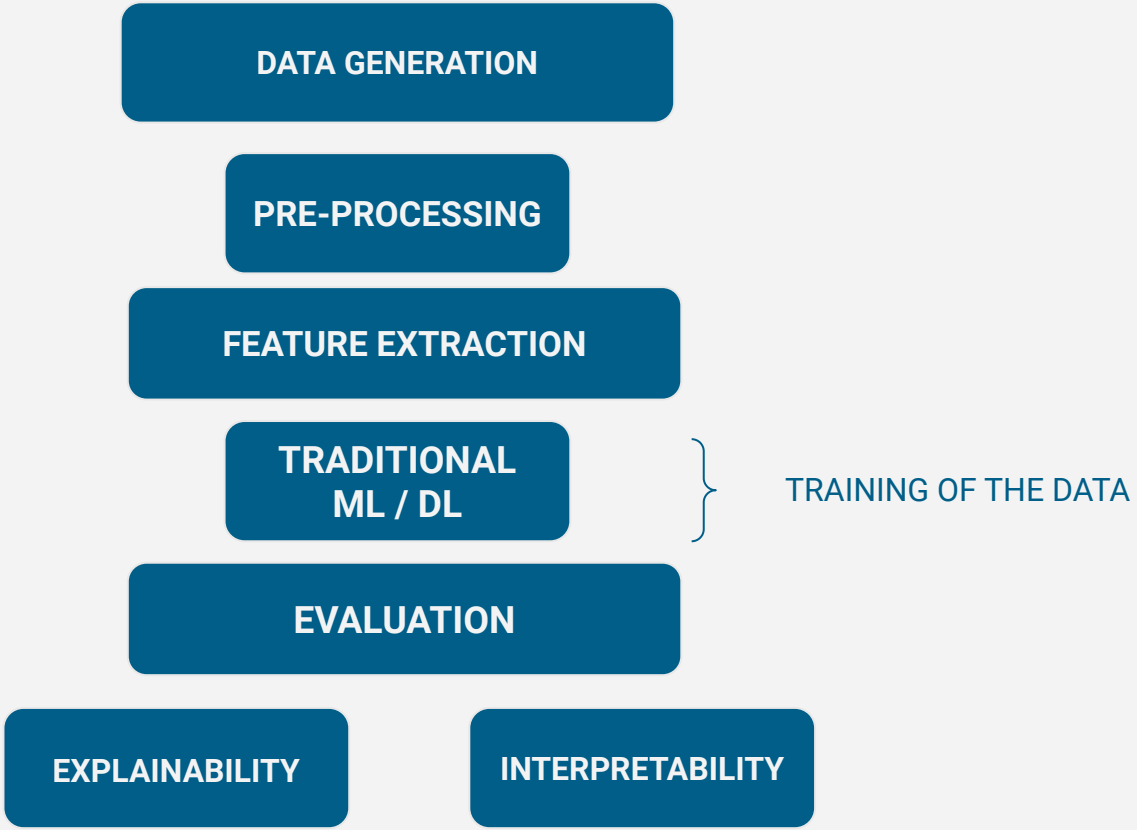
LSTM Datasets Deep Learning CNNs

Long-Short-Term-Memory IMAGE CAPTIONING **Convolutional-Neural-Networks**

ATTENTION MECHANISMS NLP Computer Vision Training

This project seeks to improve image caption generation by incorporating LSTM recurrent neural networks (RNN) with attention mechanisms. The focus is on refining the captioning system's performance by allowing it to selectively attend to important parts of the input image. By combining LSTM-CNNs and attention mechanisms, the goal is to overcome limitations in traditional models and produce more accurate and contextually relevant captions for images.

BASIC ARCHITECTURE



METHODOLOGY

Data Collection and Preprocessing:

- Gather a dataset consisting of paired images and their corresponding captions.
- Preprocess the images to a uniform size and format.
- Tokenize the captions and prepare them for training.

Model Training:

- Load the preprocessed dataset into the LSTM-RNN model with attention mechanisms.
- Train the model using the paired image-caption data, providing multiple captions for each image as input.
- Utilize optimization algorithms such as Adam or RMSprop to optimize the model parameters.
- Fine-tune hyperparameters such as learning rate and batch size to improve training efficiency.

METHODOLOGY

Caption Generation:

- Once the model is trained, provide an input image to the trained model.
- The model will utilize the learned parameters and attention mechanisms to generate a caption for the input image autonomously.
- The generated caption will reflect the model's understanding of the content and context of the input image, based on the training data.

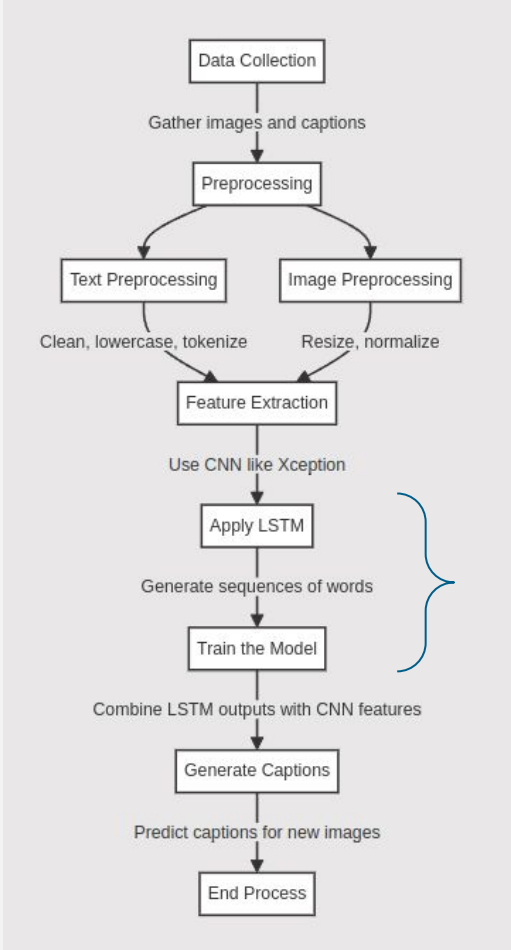
Evaluation:

- Evaluate the performance of the model-generated captions using metrics such as BLEU, METEOR, and CIDEr.
- Compare the generated captions with human-generated references to assess the quality and accuracy of the model's outputs.

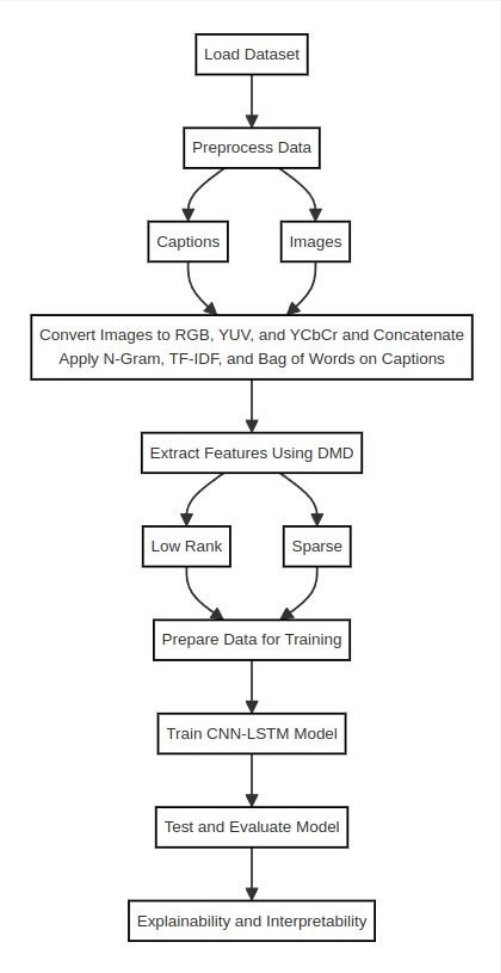
METHODOLOGY

Step	Description
Data Collection	Gather dataset of images paired with captions
Preprocessing	Preprocess images and tokenize captions for training
Model Training	Train LSTM-RNN model with attention mechanisms using dataset
Caption Generation	Generate captions for new input images using trained model
Evaluation	Evaluate generated captions using metrics such as BLEU, METEOR, and CIDEr

BASIC WORKFLOW



DMD APPROACH



LITERATURE SURVEY

https://docs.google.com/document/d/14aZChlV8hdm8Cqx0bkTLnWhoIGD7FYOFEHMYN07W8y4/edit?usp=sharing

DRAW BACKS

- We tried to combine extract features from DMD modes then put it in the CNN and provided this as an input to LSTM, since DMD is used for sequential data and the images are stationary it couldn't capture dynamics in case of images.
- Even though we are extracting features the low rank and sparse outputs are like this
- As DMD is used to find the linear dynamics from the high dimensional data (like videos and other time series movements) given a low dimensional (2D data such as image does not generate any useful information whereas model like CNN extracts features which can be fed into RNN and further enhanced by adding Attention to the RNN model

DRAW BACKS

- We tried to combine extract features from DMD modes then put it in the CNN and provided this as an input to LSTM, since DMD is used for sequential data and the images are stationary it couldn't capture dynamics in case of images.
- Even though we are extracting features the low rank and sparse outputs are like this
- As DMD is used to find the linear dynamics from the high dimensional data (like videos and other time series movements) given a low dimensional (2D data such as image does not generate any useful information whereas model like CNN extracts features which can be fed into RNN and further enhanced by adding Attention to the RNN model

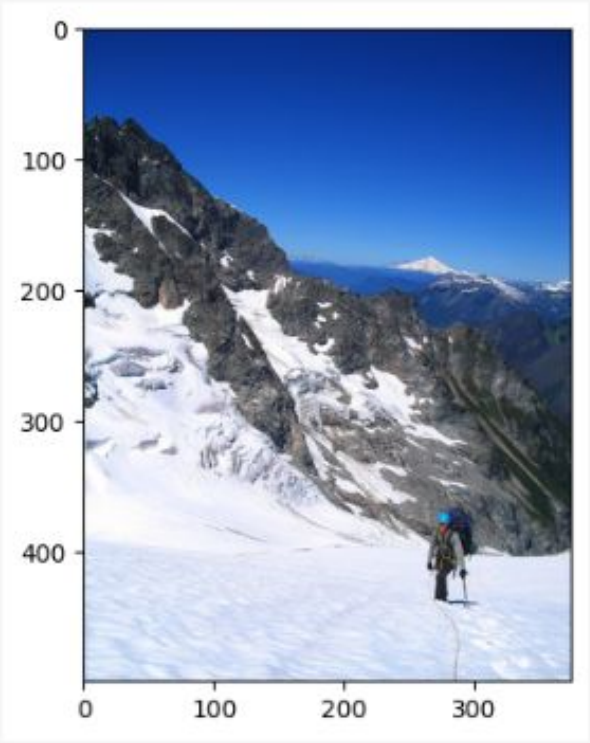
OUTPUT INFERENCES

- Successfully generated optimal captions using a CNN + LSTM model.
- Experimented with Dynamic Mode Decomposition (DMD) to improve the feature extraction process.
- Initial implementation of DMD post-CNN feature extraction resulted in poor, gibberish predictions.
- Adjusted the workflow to apply DMD during feature extraction.
- Faced training challenges due to memory constraints, preventing effective model training.

CNN + LSTM

```
start man is climbing up snowy mountain end

<matplotlib.image.AxesImage at 0x785a2c16ea90>
```



DMD POST CNN RESULT

Generated Caption: skier wearing blue coat is standing on snowy terrain with mountains in the background wearing blue background is skies on the ground in front of mountains behind him arms him two other

CNN + LSTM

```
start dog is jumping over barrier end
```

```
<matplotlib.image.AxesImage at 0x7cf9ac47fd90>
```

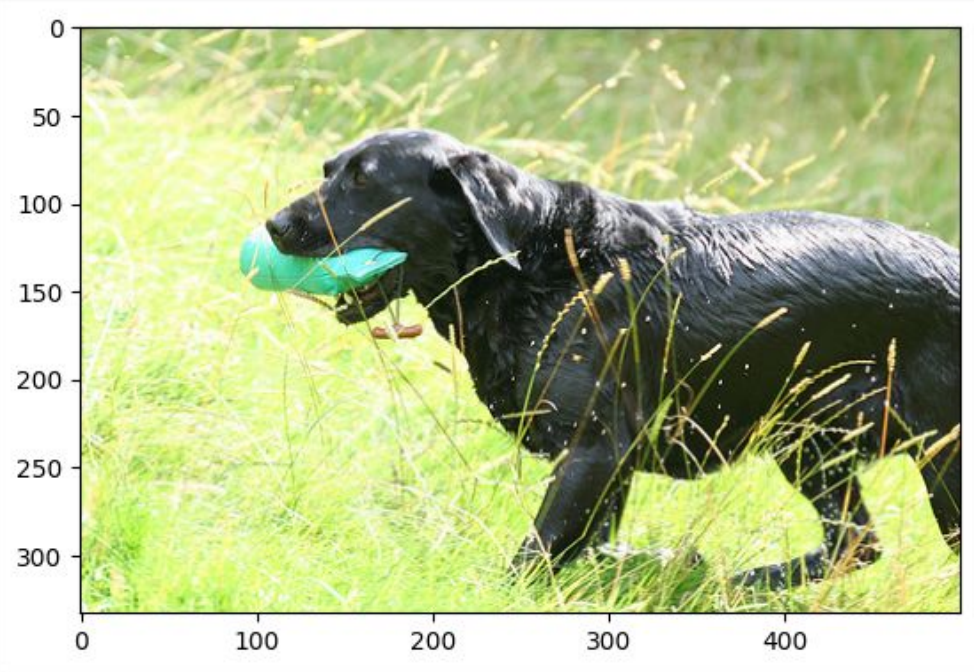
DMD POST CNN RESULT

Generated Caption: enjoys numbers on sandy field watching watched of opposing hockey players run in the air behind spectators players run on the air while another players watches is smiling behind the opposing players

CNN + LSTM

```
start black dog is running through the grass end

<matplotlib.image.AxesImage at 0x7cf98c354810>
```



DMD (feature extraction)

Original Caption: black dog is running through the grass
Predicted Caption: a a

CNN + LSTM

```
start man in wetsuit is wakeboarding on the water end

<matplotlib.image.AxesImage at 0x7cf94c66de50>
```

DMD (feature extraction)

Original Caption: man in wet suit is wakeboarding on water
Predicted Caption: in in

CONCLUSION

our project has laid the groundwork for significant advancements in image captioning technology. Moving forward, we plan to integrate attention mechanisms into our model to enhance its focus on key elements within images. Furthermore, scaling our approach to larger datasets will allow us to improve the robustness and generalization of our caption generator. By pursuing these avenues of enhancement, we are poised to push the boundaries of image captioning, bringing us closer to a future where machines can accurately and meaningfully describe visual content across diverse languages and domains.

THANK YOU