Machine Learning and Predictive Analysis
Final Project

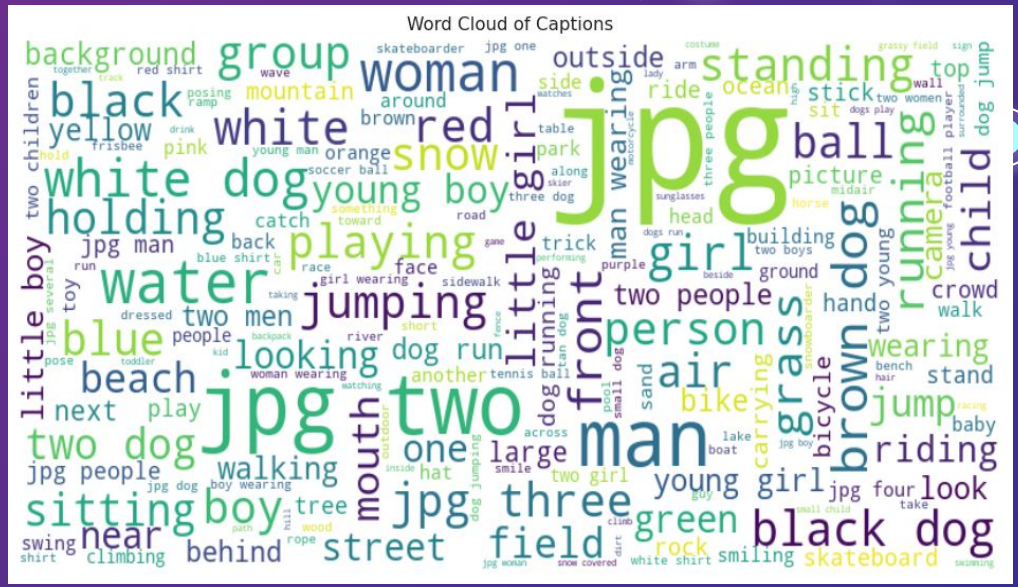
Image Caption Generation

Arushi Makraria | May 23rd 2024

Problem Statement

We aim to enhance accessibility for visually impaired individuals by enabling them to interact with visual content more effectively with the help of our image captioning model that automatically generates text-based captions for images.

The goal is to create a user-friendly solution that improves the inclusivity and usability of digital content for visually impaired users.



My Dataset

FLICKR 8K DATASET

The Flickr 8k dataset contains 8,000 images that are each paired with five captions. It is commonly used for training and evaluating image captioning models due to its diverse range of images and captions.

The dataset is widely appreciated for its variety, making it an excellent benchmark for developing models that need to generalize well across different scenes and contexts.

01

EDA

Data and Model Quality Assumption

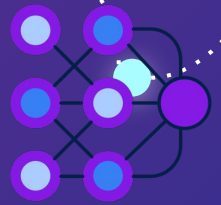
DATA

The dataset contains a diverse range of images and captions, covering various objects, scenes, and activities to help the model generalize well to new, unseen images.

Each image in the dataset is accurately paired with its corresponding captions, ensuring the text descriptions align with the visual content.

Features extracted using ResNet-50 provide meaningful and robust representations of the images.

The embedding layer in the captions model captures the semantic meaning of words and helps in generating coherent and contextually appropriate captions.



MODEL

EDA – Visualizing Images with Captions

start two large tan dogs play along a sandy beach end

start two dogs playing together on a beach end

start two dogs playing in the sand at the beach end

start two dogs are making a turn on a soft sand beach end

start two different breeds of brown and white dogs play on the beach end

start climber climbing an ice wall end

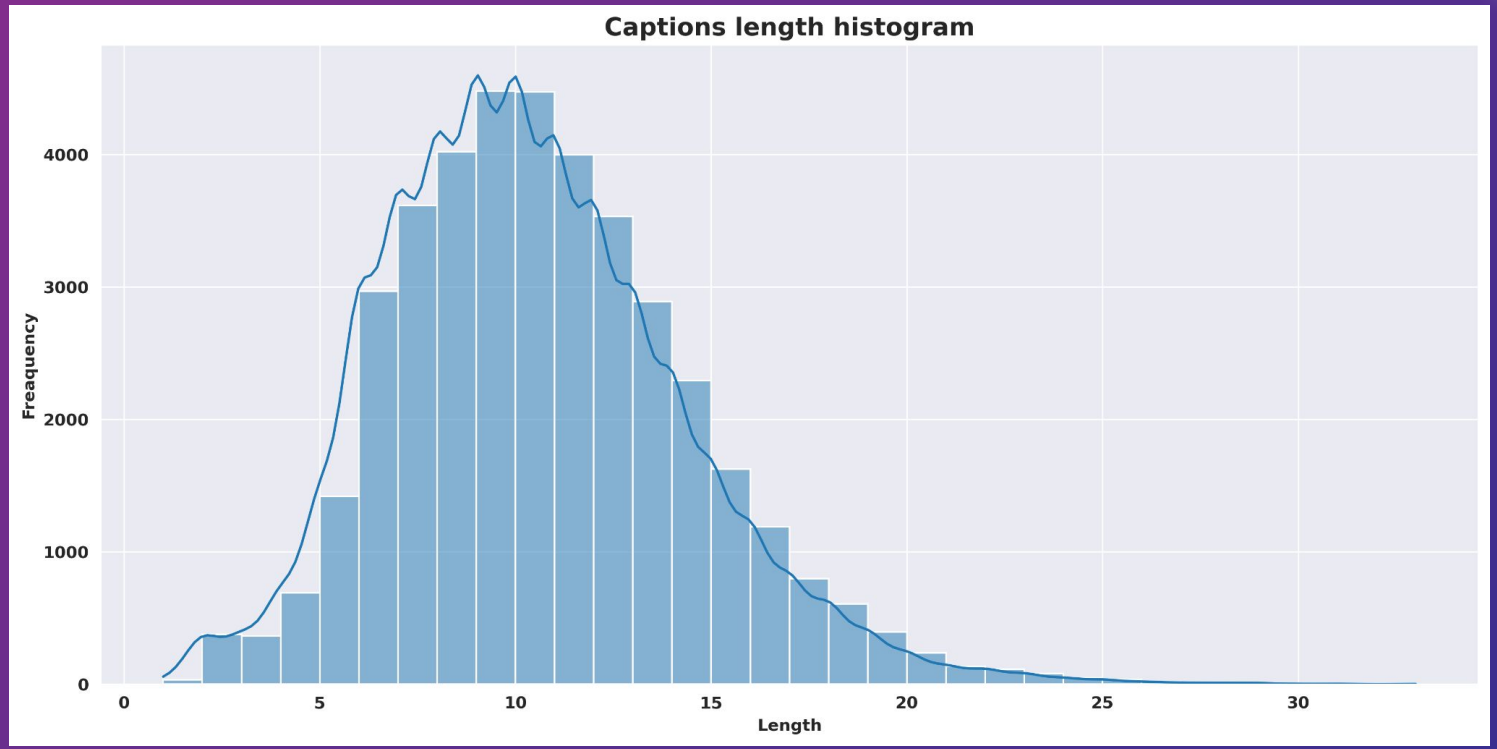
start a person in blue and red ice climbing with two picks end

start an ice climber scaling a frozen waterfall end

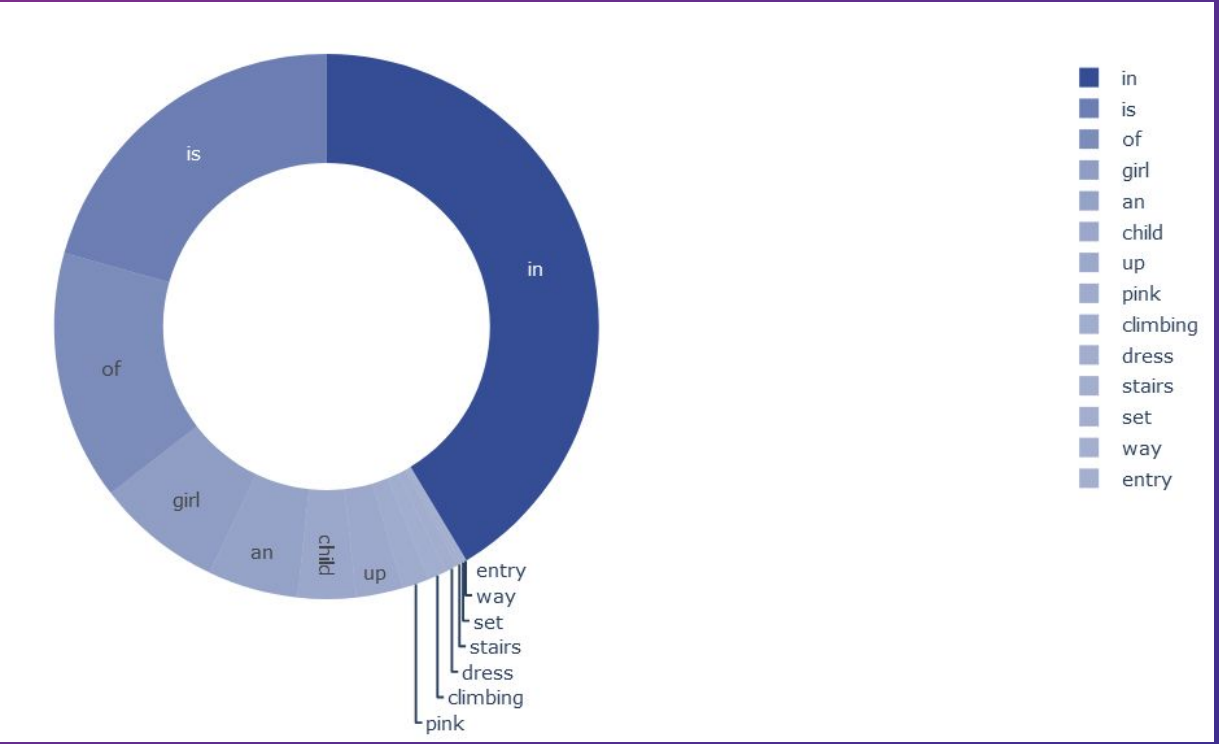
start an ice climber in a blue jacket and black pants is scaling a frozen ice wall end

start a man uses ice picks and crampons to scale ice end

EDA – Caption Length Histogram



EDA - Word Occurance in Captions



02

Modelling

Feature Engineering and Augmentation

1. **Text Cleaning and Pairing:**
   - The textual data had image IDs associated with it so the text was cleaned first and a separate list of image IDs was extracted then image caption pairs were generated.

2. **Image Augmentation:**
   - Applied Keras' ImageDataGenerator for random operations such as zooming, brightness manipulation, shear, histogram manipulation, and blurring, while keeping the original captions.
   - Augmented images were paired with their respective captions to enhance dataset diversity.

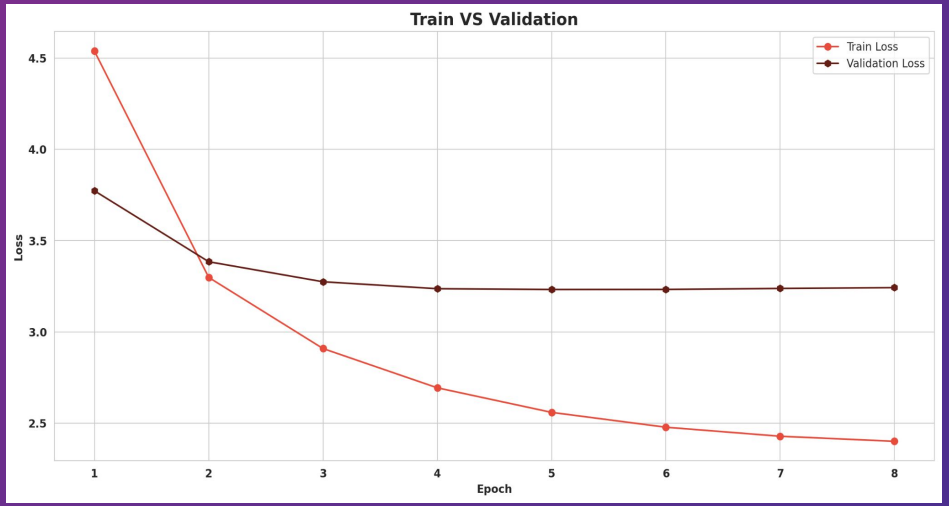
3. **Text Augmentation:**
   - Tested synonym replacement using embeddings but found it hindered model performance and was not used in the final model so decided against it.

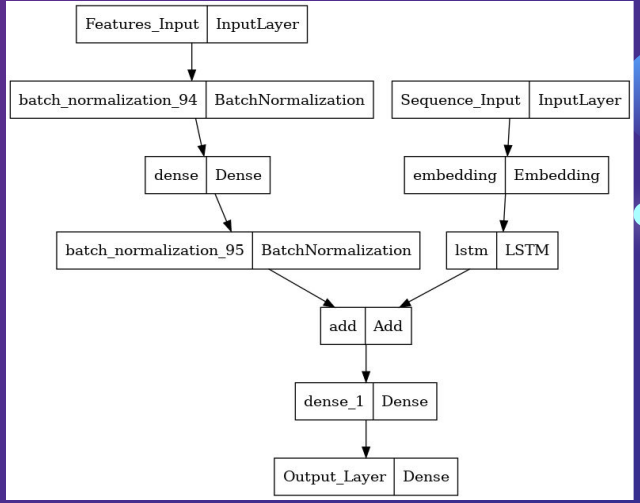
These augmentations aimed to enrich the dataset, improving the model's generalization and performance.

Model - LSTM + CNN

Model Architecture

- This was the first model - This image captioning model has an **Encoder** that processes image features through layers of batch normalization and dense transformations, and a **Decoder** that embeds caption inputs and processes them with an LSTM layer.
- The outputs from both the Encoder and Decoder are batch normalized, combined, and passed through dense layers with ReLU and Softmax activations to generate the final caption.
- The train and test validation loss plot and the model architecture can be seen below.





Model - LSTM + CNN

Model Performance/ Result Visualization

- For evaluating model performance, we employ a combination of Greedy algorithm and Beam Search for generating captions on the test set, with BLEU score used for assessment.
- The model only had a test accuracy of 38% and from the generated images we can see that the captions were not able to properly describe what was happening in the picture.

Train Accuracy **44%**

Test Accuracy **38%**

Beam Search: a brown dog is running on a brown dog on a grassy field

BLEU-1 Beam Search: 0.66446

BLEU-2 Beam Search: 0.71131

Greedy: a man is running on a brown dog on a grassy field

BLEU-1 Greedy: 0.63017

BLEU-2 Greedy: 0.68058

Beam Search: a baseball player in a baseball

BLEU-1 Beam Search: 0.67516

BLEU-2 Beam Search: 0.72084

Greedy: a baseball player in a baseball uniform

BLEU-1 Greedy: 0.68539

BLEU-2 Greedy: 0.72993

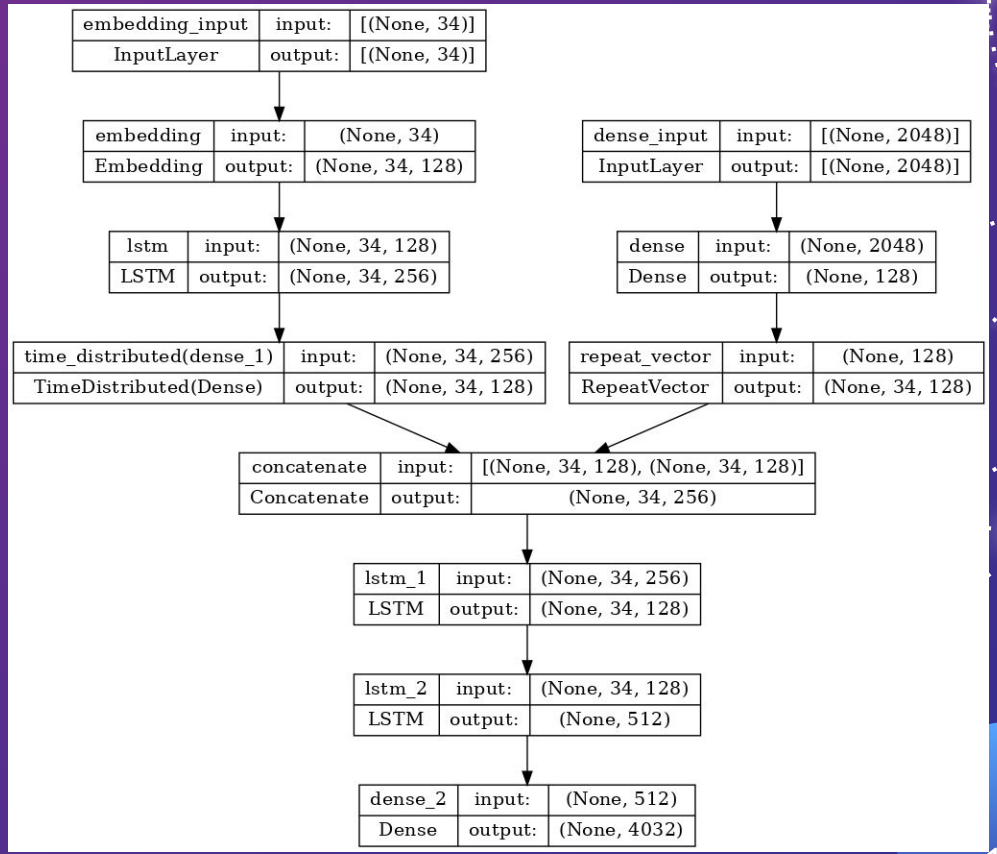
Model - LSTM + RESNET50

Model Architecture

This model architecture is a sequence-to-sequence neural network combining LSTM layers and dense layers.

It processes two inputs: one embedding input that goes through LSTM layers and another dense input that is repeated to match the LSTM sequence length.

The outputs from these two paths are concatenated, passed through additional LSTM layers, and finally through a dense layer to produce the output.



Model - LSTM + RESNET50

Model Performance and Evaluation

Test Accuracy **89.02%**

Train Accuracy **86.04%**

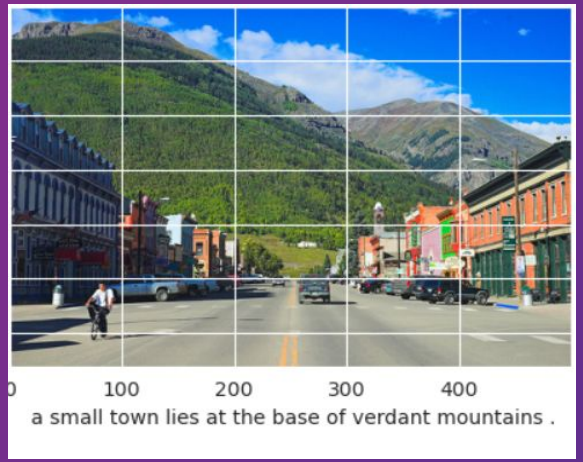
The training loss decreases steadily over 200 epochs, indicating effective learning and convergence of the model.

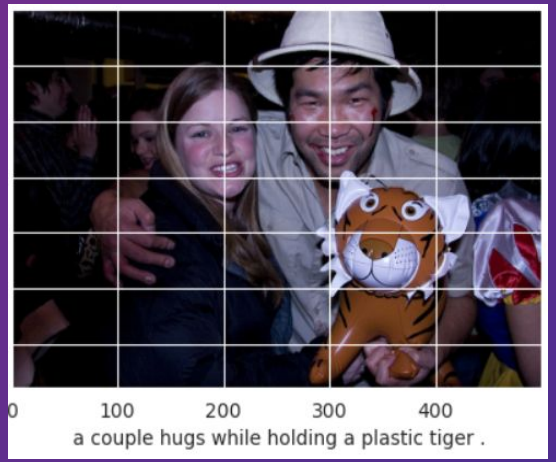
The model had a test Accuracy of 89% and train accuracy of 86% showing it generalizes well.

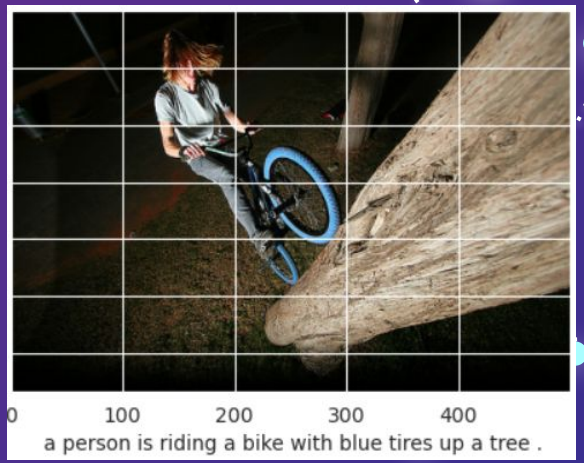
To handle overfitting, an early stopping with patience of 3 was added so as to halt training if the model's accuracy doesn't improve for 3 consecutive epochs.

Model - LSTM + RESNET50

Visualized Model Results







Based on the above results we can see that the LSTM + RESNET50 model was able to generate very accurate captions for the images that properly described the content of a picture proving this model to be a good choice.

03

Results & Future Work

Final results

Model Comparison: LSTM + CNN vs. LSTM + ResNet 50

Performance Metrics:

LSTM + CNN: Accuracy 38%

LSTM + ResNet 50: Accuracy 86%

- **Better Superior Accuracy:** The LSTM + ResNet 50 model achieved a test accuracy of 86%, which is significantly higher than the LSTM + CNN model. This means ResNet 50 is much better at generating accurate image captions.
- **Advanced Feature Extraction:** ResNet 50, being a deeper network, is capable of capturing more intricate features and patterns in images compared to a standard CNN.
- **Proven Architecture:** ResNet 50 is a well-known and trusted model in image processing. It's great at understanding complex images, making it perfect for generating accurate image captions.

Future work

Moving forward, we aim to tackle remaining challenges such as understanding context and dealing with ambiguous scenes.

Attention mechanisms and multimodal learning can be explored to enhance our model's performance.

Capturing nuanced relationships between objects and contexts for more contextually relevant captions. Exploring reinforcement learning techniques and new evaluation metrics will also be key to refining our model and ensuring it delivers high-quality, semantically rich captions in real-world scenarios.

GITHUB REPOSITORY LINK