CAPTION IT!

"HARNESSING THE POWER OF CNN/RESNET MODELS AND FLICKR 8K DATASET FOR IMAGE DESCRIPTION"

AGENDA

- Problem Statement
- Technical Approach
- Data Description
- Data Pre-processing
- Deep Learning Approaches : VGG16+LSTM and RESNET50 +GRU
- Accuracy Comparison via BELU scores
- Conclusion

PROBLEM STATEMENT

Our goal is to develop a model that can automatically generate captions for images using the Flickr 8k Dataset and pretrained CNN/ResNet models

To achieve this, we will be exploring two different approaches, CNN+LSTM and ResNet+GRU. The performance of these approaches will be compared using the BLEU score, a widely used metric for evaluating the quality of machine-generated text

TECHNICAL APPROACH

- Import required libraries and modules
- Load and preprocess the dataset with images and captions
- **Split** the dataset into training and testing subset.
- Extract **image features** using a pre-trained model like **VGG16 or ResNet50**
- **Tokenize** captions using the Keras tokenizer
- Create training data with image features and tokenized captions
- Build the caption generation model, including:

  Encoder: Use extracted image features

  Decoder: Use **LSTM or GRU layers** for sequential text data processing
- Train the caption generation model on the training data
- Generate captions for test images using the trained model
- Calculate **BLEU scores** to evaluate the quality of generated captions
- Identify and visualize the best predictions based on BLEU scores

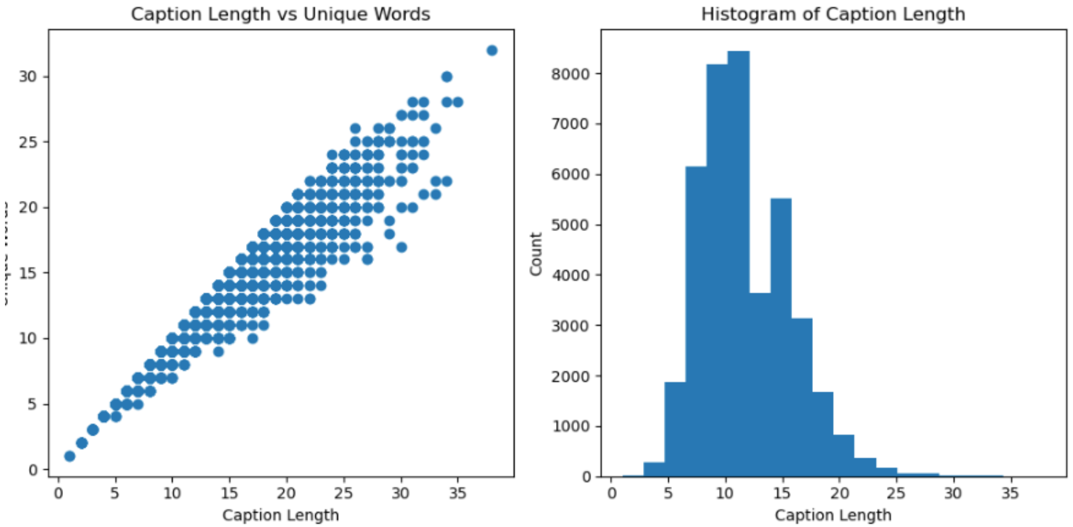
DATASET ANALYSIS

Benchmark: 8,000 images, each with 5 captions, from 6 Flickr groups; diverse scenes, no well-known people/locations

Dataset: manually curated, high-quality captions; ideal for image captioning/search model training and evaluation

Dataset: diverse images and captions for various computer vision and NLP model development and testing

- **EXPLORATORY DATA ANALYSIS**



- Scatter plot:

  X-axis: Caption length

  Y-axis: Unique words

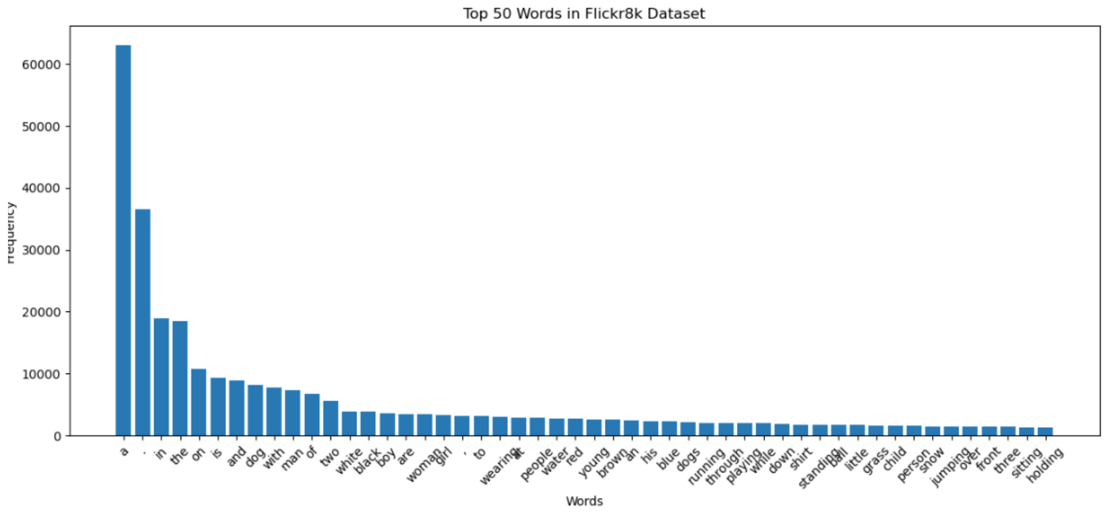
Observation: Longer captions tend to have more unique words

- Histogram:

  X-axis: Caption length range

  Y-axis: Caption count in each bin

Observation: Majority of captions have lengths between 0 and 100 characters, peaking around 50 characters



1. X-axis: 50 most frequent words
2. Y-axis: Word frequency
3. Purpose: Understand vocabulary and language patterns in captions
4. Usage: Illustrate common words and frequency distribution in presentations or reports

- the associated human-annotated captions are displayed alongside, illustrating the natural language descriptions that the image captioning model will be trained to generate.

the white dog is playing in a green field with a yellow toy .

a white dog is trying to catch a ball in midair over a grassy field .

a dog leaps to catch a ball in a field .

a black and white dog jumps up towards a yellow toy .

a black and white dog jumping in the air to get a toy .

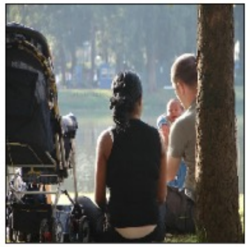
two people are at the edge of a lake , facing the water and the city skyline .

a young boy waves his hand at the duck in the water surrounded by a green park .

a little boy at a lake watching a duck .

a large lake with a lone duck swimming in it with several people around the edge of it .

a child and a woman are at waters edge in a big city .



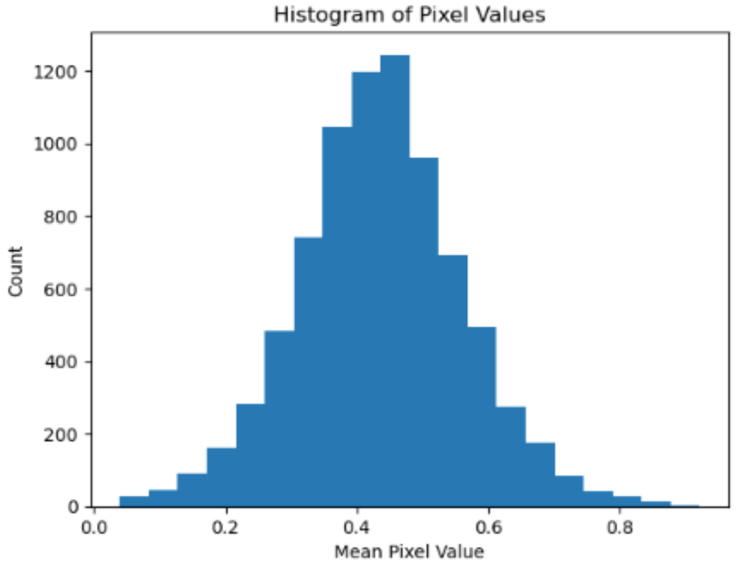
couple with a baby sit outdoors next to their stroller .

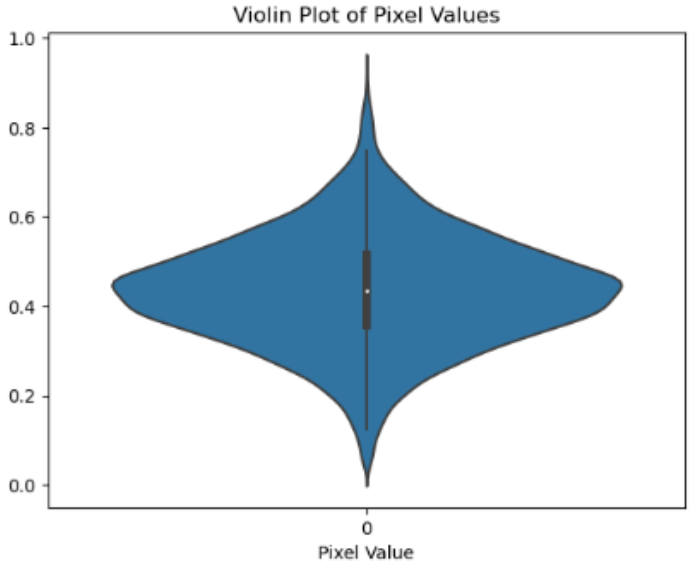
a man and woman care for an infant along the side of a body of water .

a couple with their newborn baby sitting under a tree facing a lake .

a couple sit on the grass with a baby and stroller .

a couple and an infant , being held by the male , sitting next to a pond with a near by stroller .





- The histogram shows the count of images that have a particular mean pixel value range (x-axis)
- The violin plot shows the distribution of the mean pixel values, where the wider areas represent regions with more images having that pixel value

Deep Learning Approaches

To compare the performances of the two deep learning approaches, CNN + LSTM and ResNet + GRU, for generating captions for images using the Flickr 8k Dataset and pretrained models, we can compute their BLEU scores

By comparing the BLEU scores of the two approaches, we can determine which approach performs better at generating captions for images. However, it's worth noting that BLEU scores are just one way to evaluate the performance of a captioning model

APPROACH : VGG16 & LSTM

VGG16: The VGG16 model is a pre-trained convolutional neural network (CNN) designed for image classification

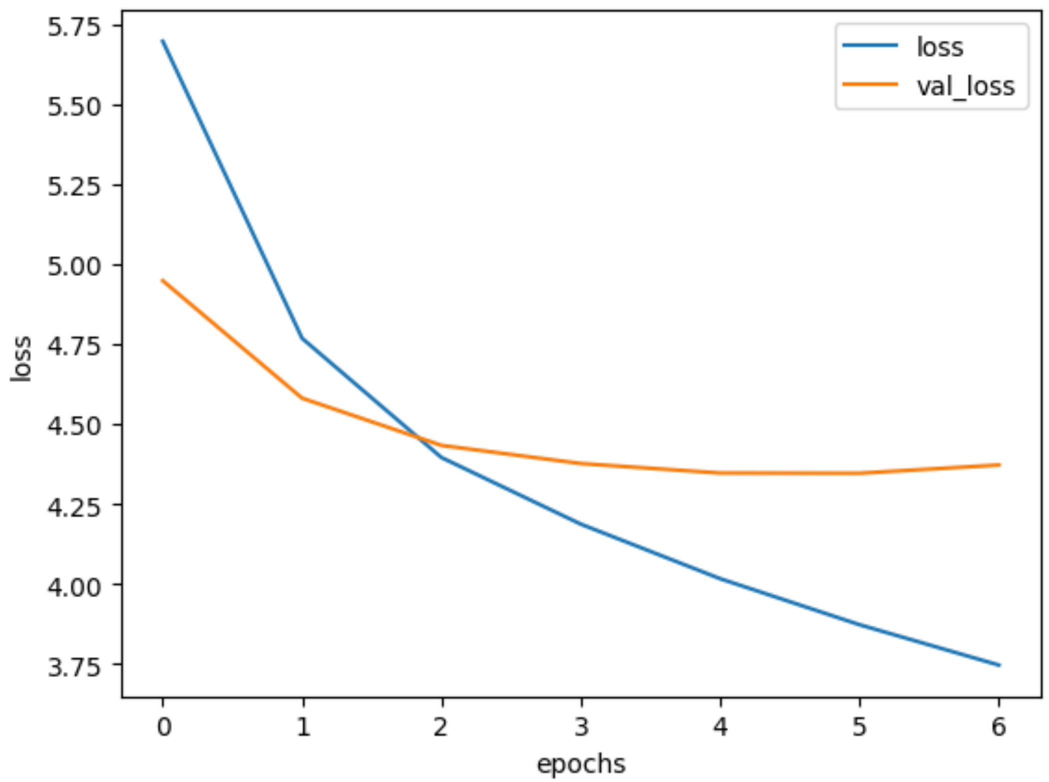
VGG16: 16-layer architecture; 13 for feature extraction, 3 for classification

Trained on ImageNet: 1.2M+ images, 1,000 categories; used for deep learning image classification

LSTM is a recurrent neural network that excels at capturing temporal dependencies via its memory cell, improving predictions in time-series problems

MODEL-VGG16+LSTM

```
Model: "model_1"
__________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 input_3 (InputLayer)           [(None, 30)]         0           []

 embedding (Embedding)          (None, 30, 128)      572928      ['input_3[0][0]']

 input_2 (InputLayer)           [(None, 4096)]       0           []

 CaptionFeature (LSTM)          (None, 512)          1312768     ['embedding[0][0]']

 ImageFeature (Dense)           (None, 512)          2097664     ['input_2[0][0]']

 add (Add)                      (None, 512)          0           ['CaptionFeature[0][0]
                                                                  'ImageFeature[0][0]']

 dense (Dense)                  (None, 512)          262656      ['add[0][0]']

 dropout (Dropout)              (None, 512)          0           ['dense[0][0]']

 dense_1 (Dense)                (None, 4476)         2296188     ['dropout[0][0]']

==================================================================================================
Total params: 6,542,204
Trainable params: 6,542,204
Non-trainable params: 0
__________________________________________________________________________________________________
```



PERFORMANCE EVALUATION

startseq black dog is running in the grass endseq

startseq black dog is running in the grass endseq

BLEU SCORES

```
The Mean BLEU-1 Score for the Test Set is 0.180
The Mean BLEU-2 Score for the Test Set is 0.086
The Mean BLEU-3 Score for the Test Set is 0.052
The Mean BLEU-4 Score for the Test Set is 0.037
```

Predicted: black dog is running in the snow

True: black dog is running in the water

BLEU: 0.8091067115702212

APPROACH : ResNet50 and GRU

ResNet is a widely used deep neural network architecture for image recognition and computer vision tasks, which often uses the pre-training on the ImageNet dataset comprising over 1 million labeled images across 1,000 categories.

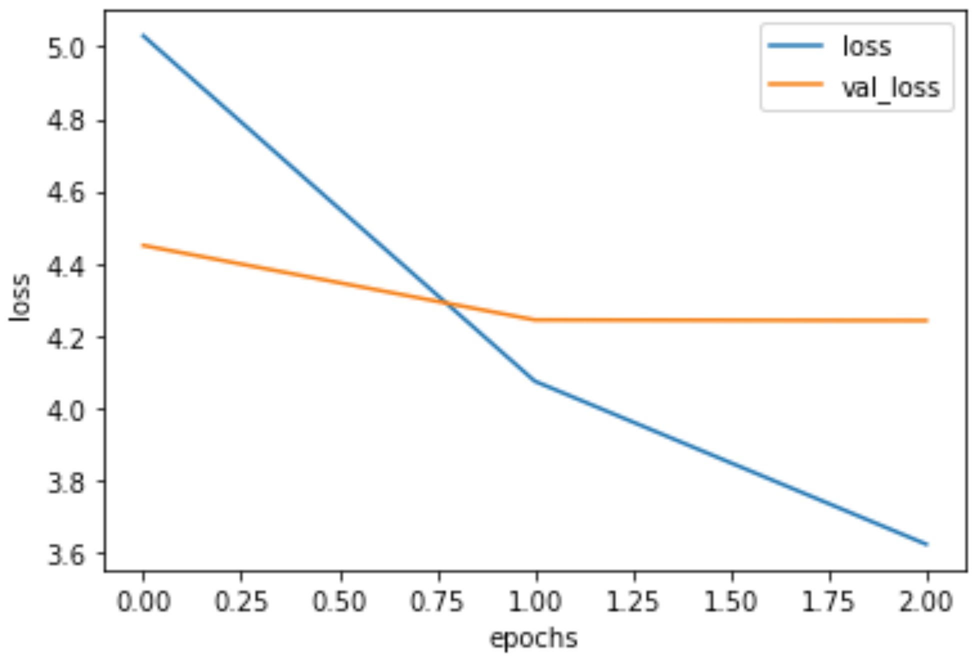
ResNet50 is a 50-layer deep residual network with convolutional, max-pooling, and fully connected layers, designed for image recognition tasks.

GRU is a recurrent neural network that efficiently captures temporal relationships using its update and reset gates, enhancing predictions in time-series tasks

MODEL:RESNET50+GRU

```
____________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
====================================================================================================
 input_5 (InputLayer)           [(None, 50)]         0           []

 input_4 (InputLayer)           [(None, 100352)]     0           []

 embedding_1 (Embedding)        (None, 50, 128)      1280000     ['input_5[0][0]']

 ImageFeature (Dense)           (None, 256)          25690368    ['input_4[0][0]']

 CaptionFeature (GRU)           (None, 256)          296448      ['embedding_1[0][0]']

 concatenate (Concatenate)      (None, 512)          0           ['ImageFeature[0][0]',
                                                                  'CaptionFeature[0][0]']

 dense_2 (Dense)                (None, 10000)        5130000     ['concatenate[0][0]']

====================================================================================================
Total params: 32,396,816
Trainable params: 32,396,816
Non-trainable params: 0
____________________________________________________________________________________________________
None
```

Plotting Loss & Validation Loss:



BLEU scores

```
The Mean BLEU-1 Score for the Test Set is 0.138
The Mean BLEU-2 Score for the Test Set is 0.050
The Mean BLEU-3 Score for the Test Set is 0.029
The Mean BLEU-4 Score for the Test Set is 0.022
```

COMPARISION

VGG16+LSTM

The Mean BLEU-1 Score for the Test Set is 0.180
The Mean BLEU-2 Score for the Test Set is 0.086
The Mean BLEU-3 Score for the Test Set is 0.052
The Mean BLEU-4 Score for the Test Set is 0.037

RESNET50+GRU

The Mean BLEU-1 Score for the Test Set is 0.138
The Mean BLEU-2 Score for the Test Set is 0.050
The Mean BLEU-3 Score for the Test Set is 0.029
The Mean BLEU-4 Score for the Test Set is 0.022

CONCLUSION

The VGG16 model with LSTM achieved higher BLEU scores across all four metrics compared to the ResNet50 model with GRU, indicating that the VGG16-LSTM combination is more effective at generating accurate and relevant image captions.

The performance difference between the two models suggests that the choice of pre-trained image feature extraction model (VGG16 vs. ResNet50) and the choice of sequence model (LSTM vs. GRU) can significantly impact the quality of generated captions.

Although the VGG16-LSTM model outperforms the ResNet50-GRU model in this comparison, there is still room for improvement in both models. Future work could explore different architectures, optimization techniques, or training strategies to further enhance the performance of image captioning models.