Image Caption Generation with LSTM and GRU

Group 8:

Shashwat Shahi (002209773)
Foram Trivedi (002855380)
Aditya Ranjan Singh (002471528)

Introduction to Image Captioning

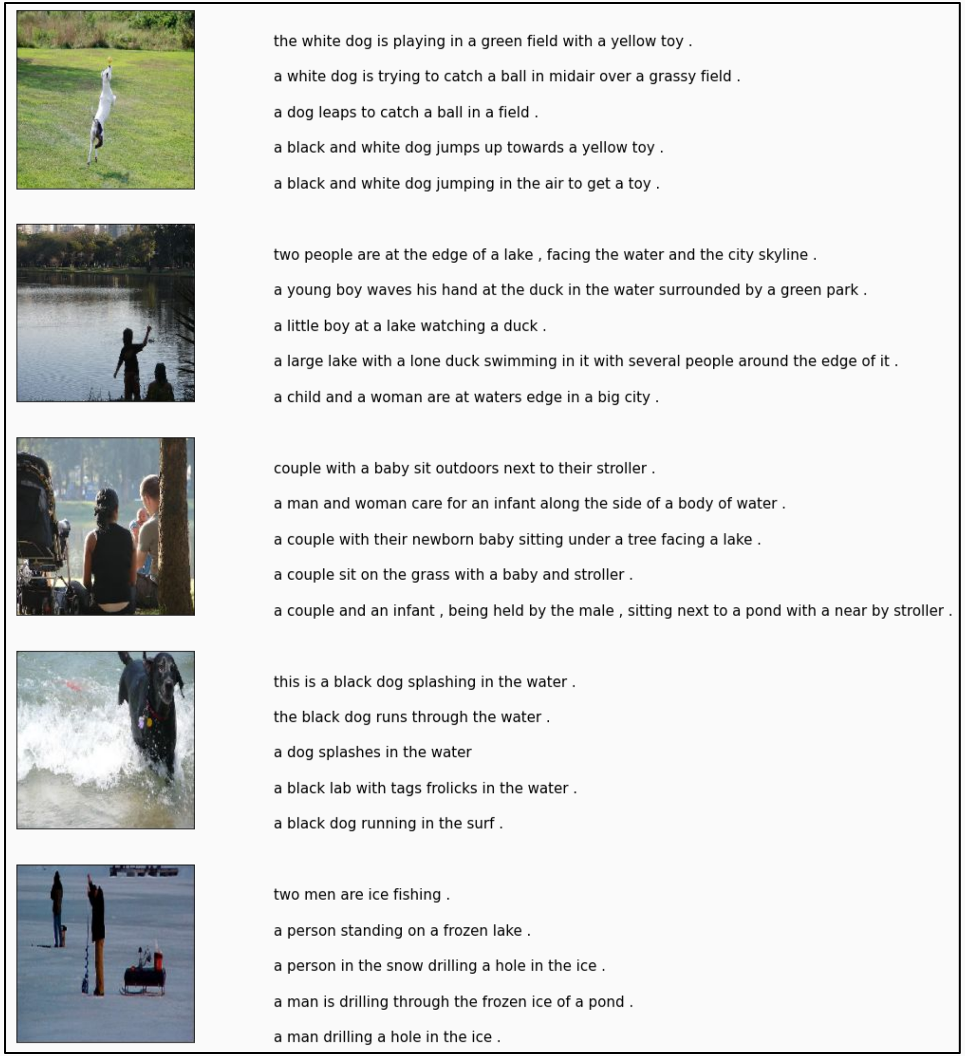
Image captioning is the process of generating textual descriptions for images.

Applications include accessibility for visually impaired users, content recommendation systems, and visual search enhancement.

- **Aim**: Generate captions for images using deep learning architectures (CNN + RNN).
- **Approach**: Combine CNNs for extracting image features with LSTM and GRU for generating descriptive captions.
- **Evaluation**: Assess caption quality using BLEU scores and semantic distance with LLMs.

Flickr_8K Dataset Overview

- Approximately 8,000 images available.
- Approximately 40,000 captions.
- Approximately 5 captions per image
- Diverse range of subjects.
- Captions include varying lengths.



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 .

this is a black dog splashing in the water .

the black dog runs through the water .

a dog splashes in the water

a black lab with tags frolicks in the water .

a black dog running in the surf .

two men are ice fishing .

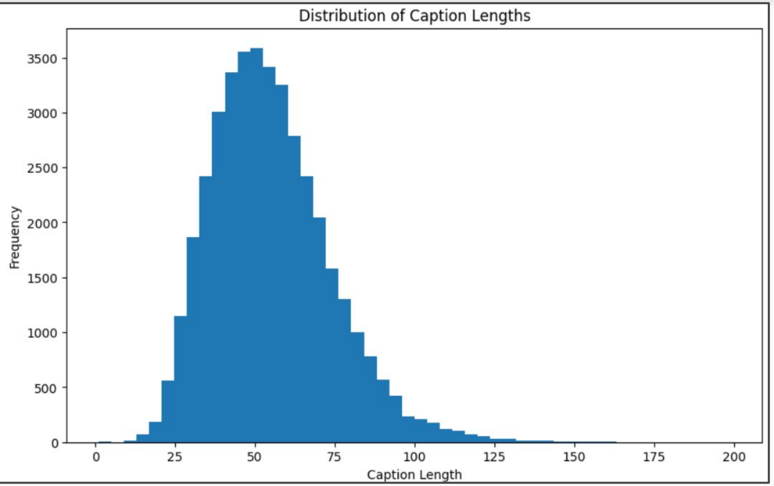
a person standing on a frozen lake .

a person in the snow drilling a hole in the ice .

a man is drilling through the frozen ice of a pond .

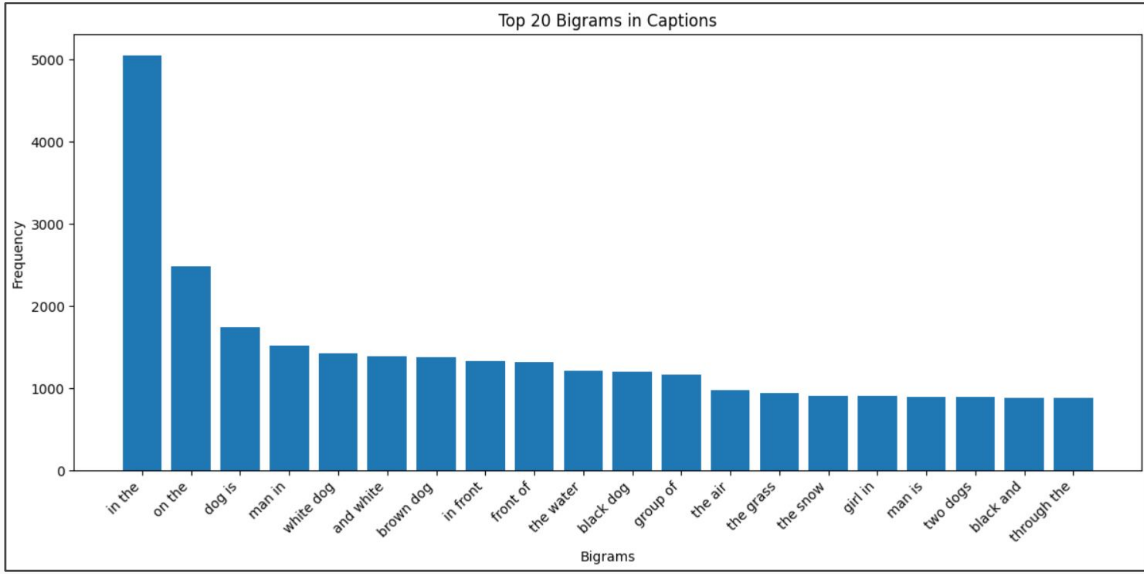
a man drilling a hole in the ice .

Caption Length Distribution & N-gram Analysis



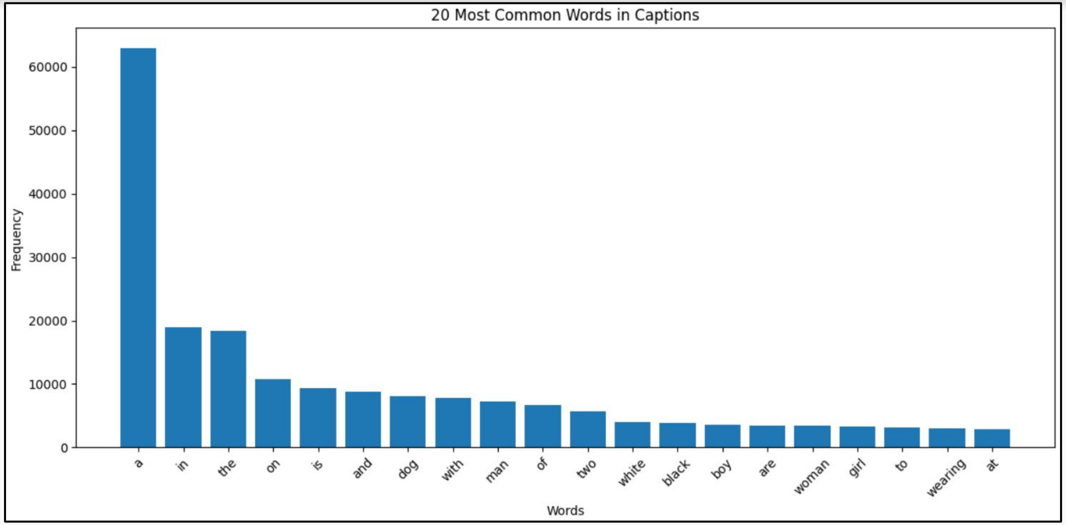
- Most captions are 40-70 characters
- Normal distribution with peak at ~50
- Few captions exceed 100 characters

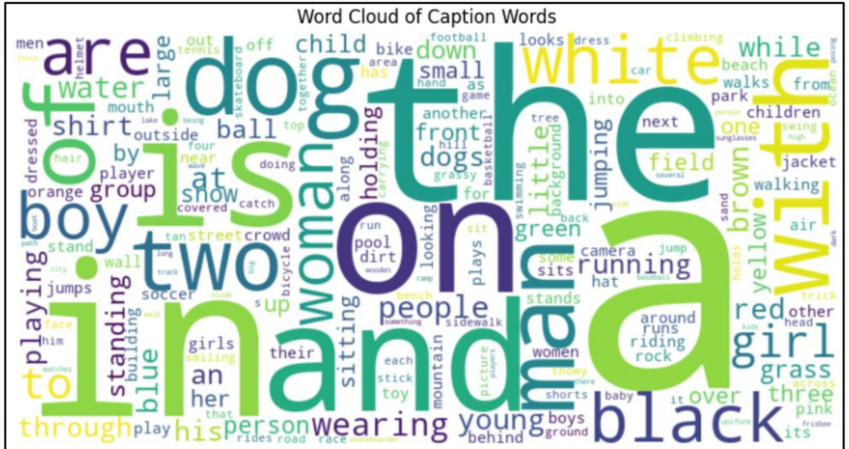
- "in the" is most frequent bigram
- Common bigrams describe objects
- Top 20 bigrams show clear patterns



Most Common Words and Word Cloud

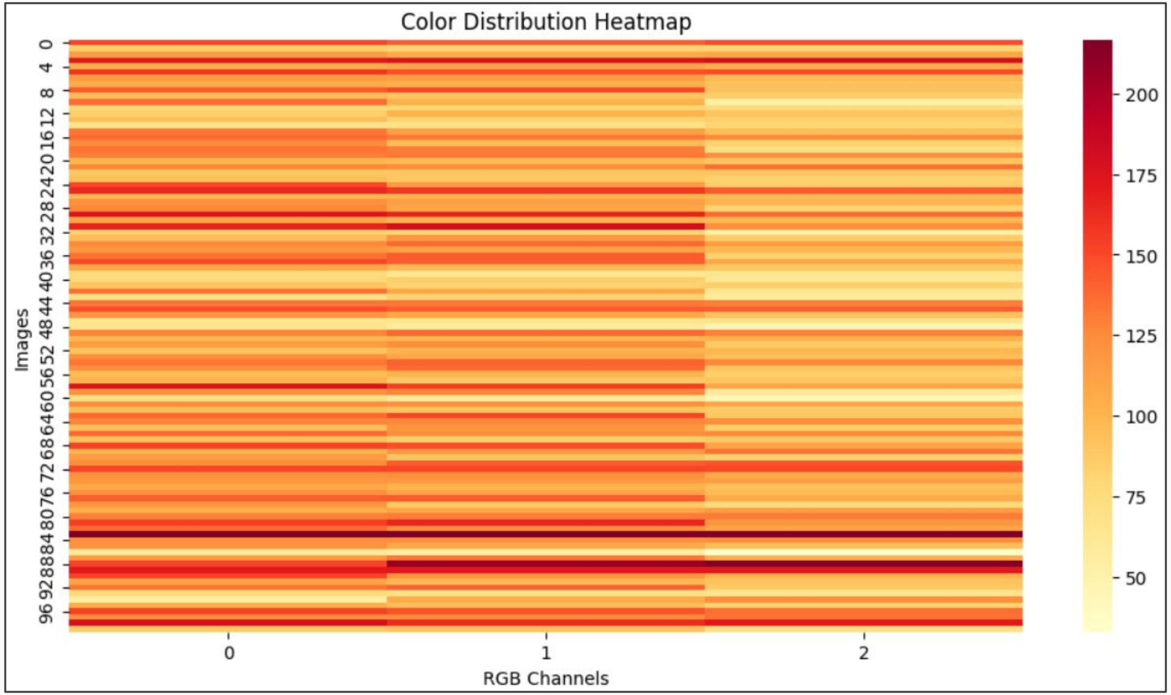
- Larger words are more frequent
- Common verbs: "is", "are", "playing"





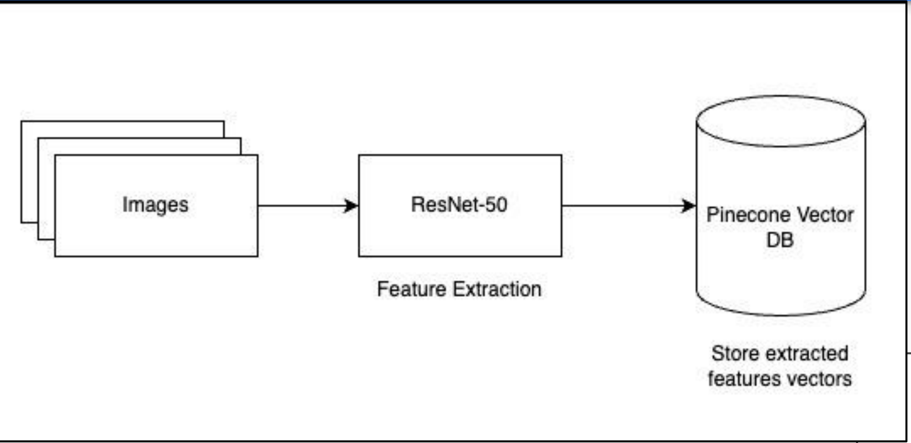
- "a" is overwhelmingly common
- Prepositions and articles dominate
- Frequent nouns: "dog", "cat", "man"

Color Distribution Heatmap



- RGB channels visualized for 10 images
- Red channel has highest intensity
- Lowest density near 0 (light yellow)

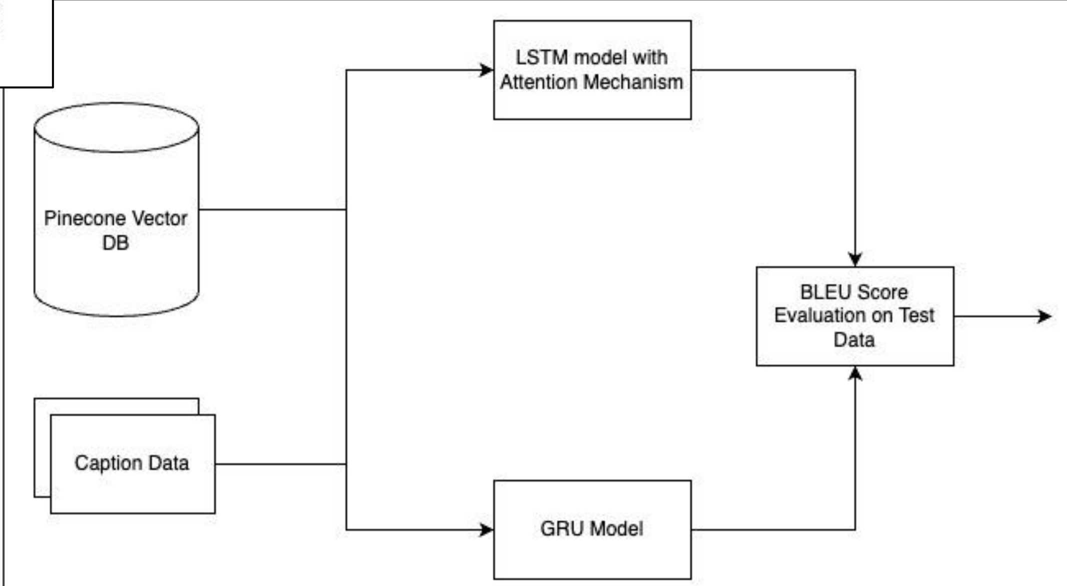
Technical Approach/Workflow



- **Feature Extraction:** ResNet-50
- **Feature Storage:** PineCone VectorDB

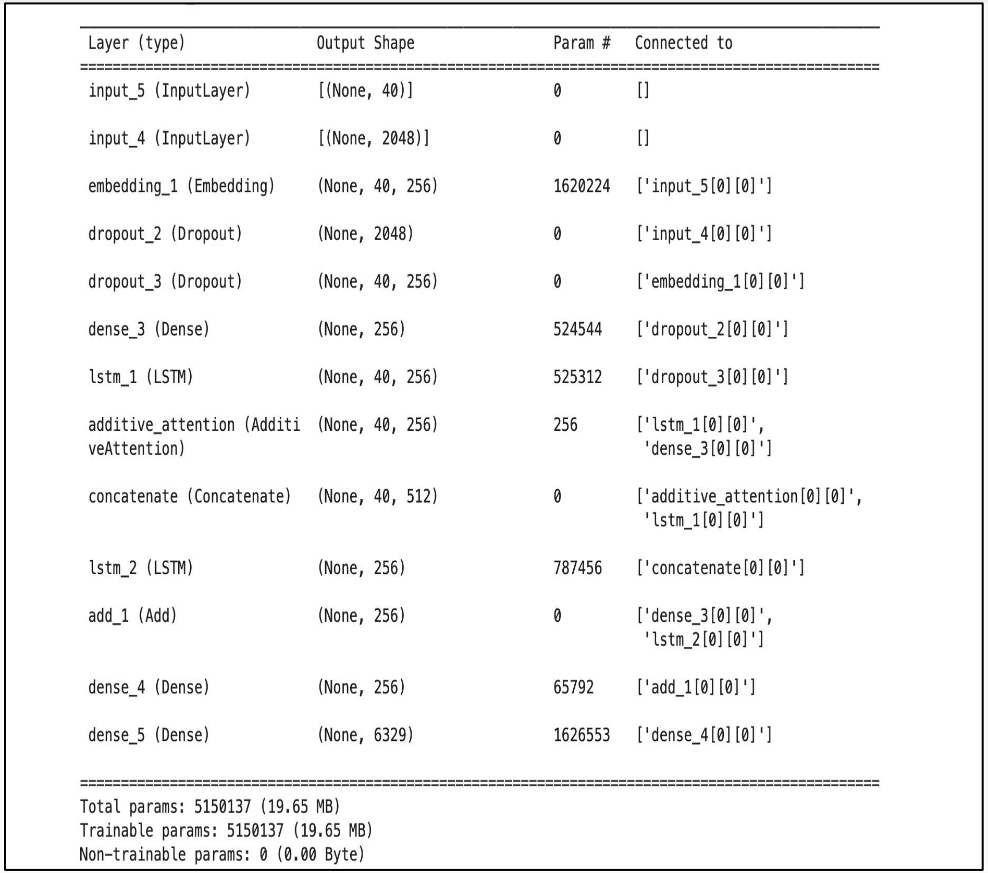
Model Architectures:

- **LSTM with Additive Attention:** Skip Connections and Contextual Information
- GRU Model


LSTM model with Attention Mechanism
Pinecone Vector DB
BLEU Score Evaluation on Test Data
Caption Data
GRU Model

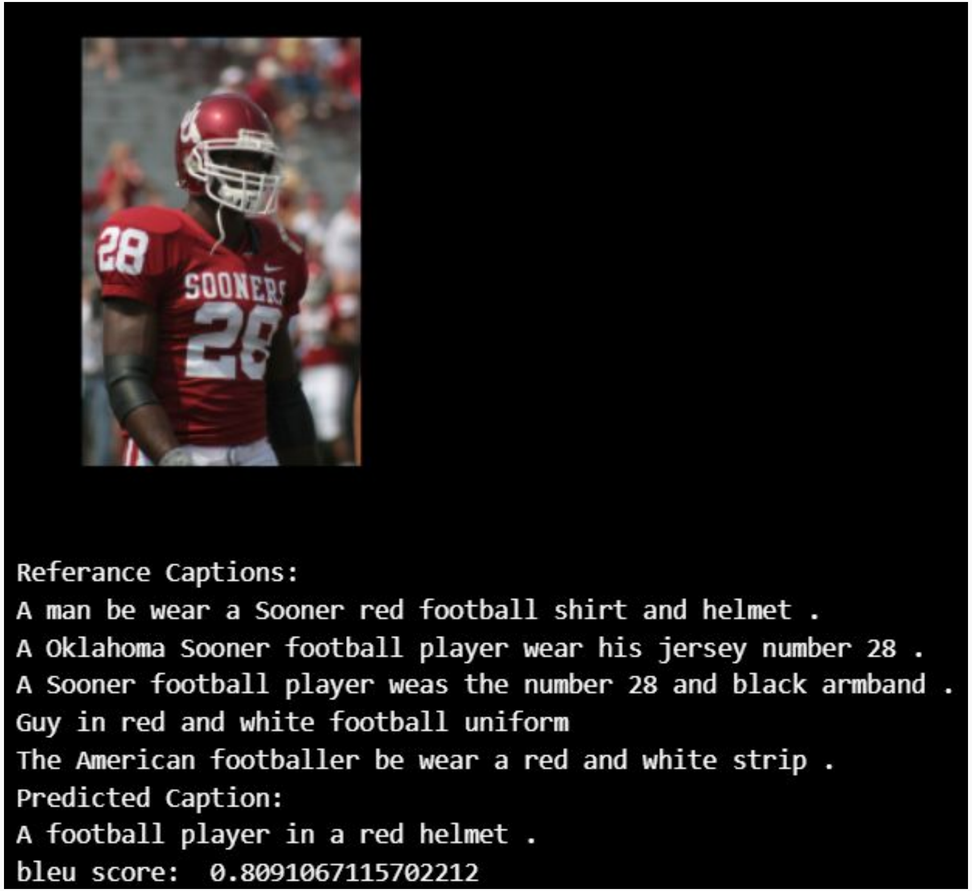
LSTM Model Implementation

- **Input Layers:** Two inputs – one for sequential data and one for non-sequential data
- **Embedding & Dropout:** The sequential input is passed through an embedding layer and a dropout layer is applied to both inputs.
- **LSTM and Attention:** The embedded input is processed by an LSTM layer, followed by an attention mechanism. The output is concatenated with the LSTM output.
- **Feature Fusion:** Another LSTM processes the concatenated output. A dense layer processes the non-sequential input, and the two are combined using element-wise addition.
- **Output Layers:** The combined result is passed through two dense layers, with the final output size being 6329.
- **Total parameters:** 5.15 million.



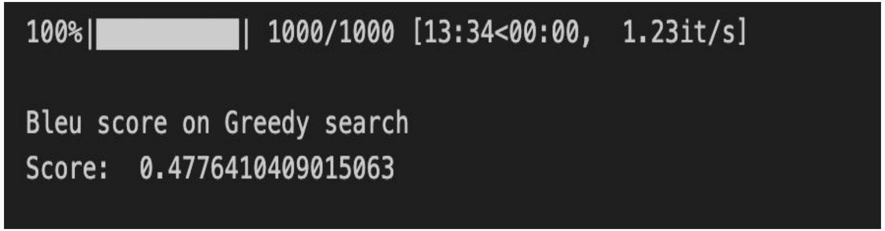
```
 Layer (type)                Output Shape                 Param #   Connected to
==================================================================================================
 input_5 (InputLayer)        [(None, 40)]                 0         []

 input_4 (InputLayer)        [(None, 2048)]               0         []

 embedding_1 (Embedding)     (None, 40, 256)              1620224   ['input_5[0][0]']

 dropout_2 (Dropout)         (None, 2048)                 0         ['input_4[0][0]']

 dropout_3 (Dropout)         (None, 40, 256)              0         ['embedding_1[0][0]']

 dense_3 (Dense)             (None, 256)                  524544    ['dropout_2[0][0]']

 lstm_1 (LSTM)               (None, 40, 256)              525312    ['dropout_3[0][0]']

 additive_attention (Additi  (None, 40, 256)              256       ['lstm_1[0][0]',
 veAttention)                                                        'dense_3[0][0]']

 concatenate (Concatenate)   (None, 40, 512)              0         ['additive_attention[0][0]',
                                                                     'lstm_1[0][0]']

 lstm_2 (LSTM)               (None, 256)                  787456    ['concatenate[0][0]']

 add_1 (Add)                 (None, 256)                  0         ['dense_3[0][0]',
                                                                     'lstm_2[0][0]']

 dense_4 (Dense)             (None, 256)                  65792     ['add_1[0][0]']

 dense_5 (Dense)             (None, 6329)                 1626553   ['dense_4[0][0]']

==================================================================================================
Total params: 5150137 (19.65 MB)
Trainable params: 5150137 (19.65 MB)
Non-trainable params: 0 (0.00 Byte)
```

Results and Evaluation (BLEU scores)



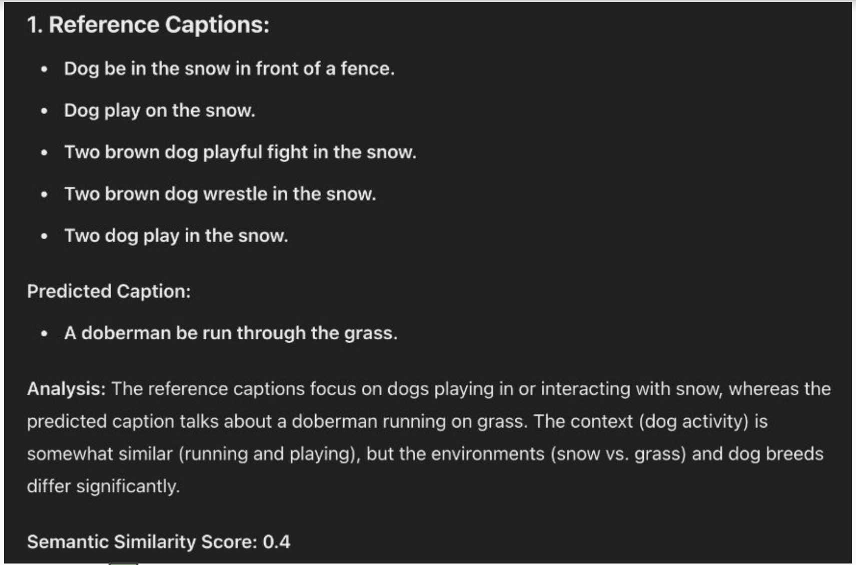
```
Referance Captions:
A man be wear a Sooner red football shirt and helmet .
A Oklahoma Sooner football player wear his jersey number 28 .
A Sooner football player weas the number 28 and black armband .
Guy in red and white football uniform
The American footballer be wear a red and white strip .
Predicted Caption:
A football player in a red helmet .
bleu score:  0.8091067115702212
```

Mean BLEU Score on Test Data



```
100%|██████████| 1000/1000 [13:34<00:00,  1.23it/s]

Bleu score on Greedy search
Score:  0.4776410409015063
```

Semantic Similarity Computation with Large Language Model (LLM)



1. Reference Captions:

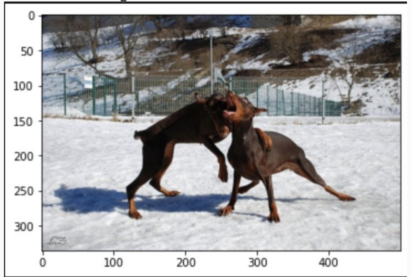
- Dog be in the snow in front of a fence.
- Dog play on the snow.
- Two brown dog playful fight in the snow.
- Two brown dog wrestle in the snow.
- Two dog play in the snow.

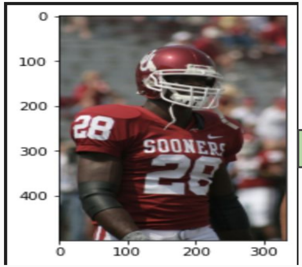
Predicted Caption:

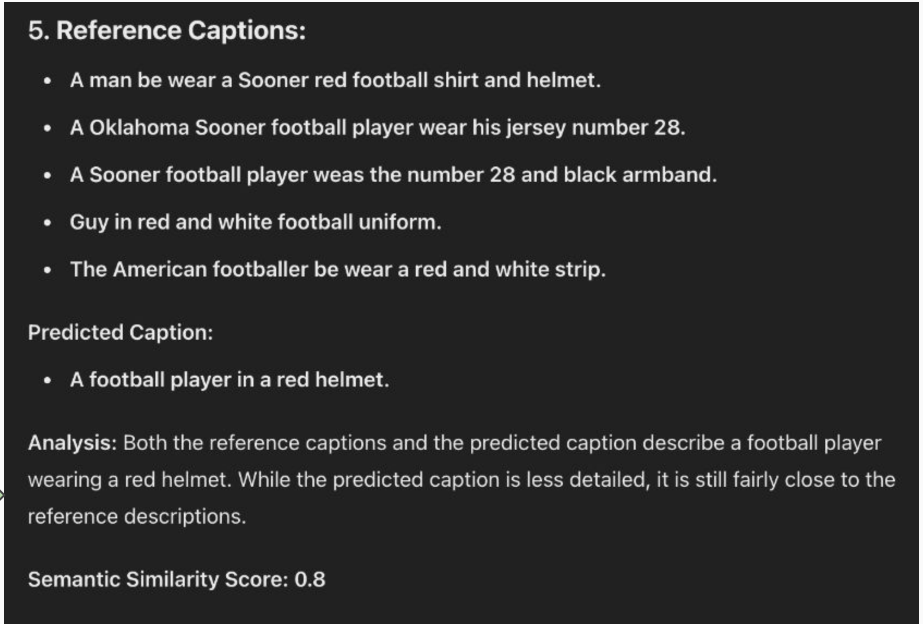
- A doberman be run through the grass.

Analysis: The reference captions focus on dogs playing in or interacting with snow, whereas the predicted caption talks about a doberman running on grass. The context (dog activity) is somewhat similar (running and playing), but the environments (snow vs. grass) and dog breeds differ significantly.

Semantic Similarity Score: 0.4







5. Reference Captions:

- A man be wear a Sooner red football shirt and helmet.
- A Oklahoma Sooner football player wear his jersey number 28.
- A Sooner football player weas the number 28 and black armband.
- Guy in red and white football uniform.
- The American footballer be wear a red and white strip.

Predicted Caption:

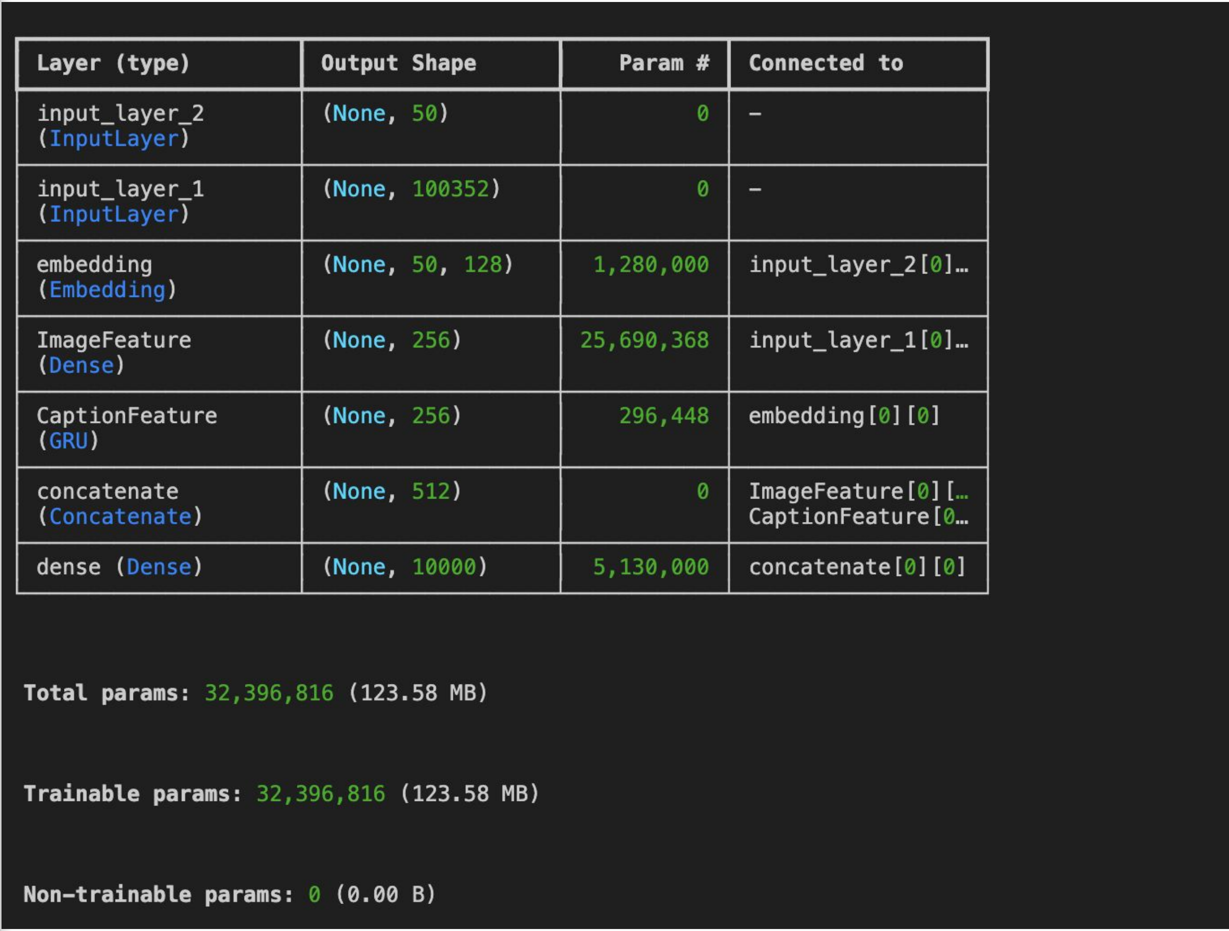
- A football player in a red helmet.

Analysis: Both the reference captions and the predicted caption describe a football player wearing a red helmet. While the predicted caption is less detailed, it is still fairly close to the reference descriptions.

Semantic Similarity Score: 0.8

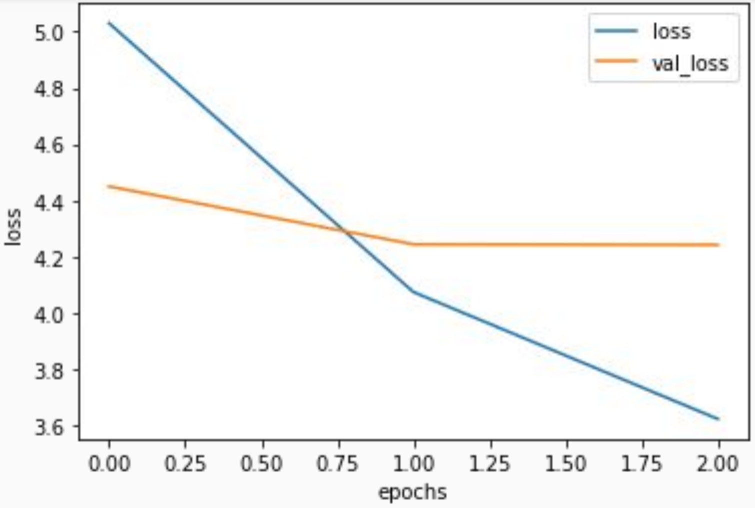
GRU Architecture

- **Input Layers:** Image features (`input_layer_1`, shape (None, 100352)) and text sequences (`input_layer_2`, shape (None, 50)).
- **Feature Extraction:** Image features reduced to (None, 256) via a dense layer. Text sequences converted to 128-dimensional embeddings, processed by a GRU to output (None, 256).
- **Fusion Layer:** Concatenates features into a vector of size (None, 512).
- **Output Layer:** Final dense layer produces output for caption generation.

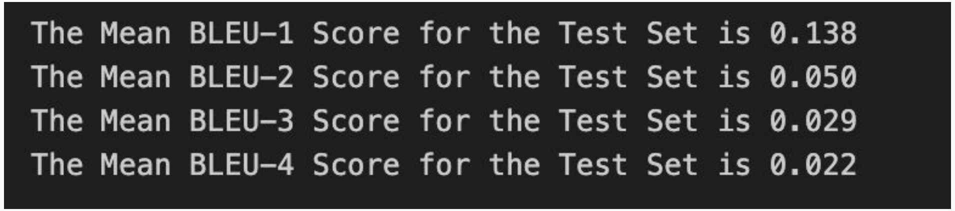


```
┌──────────────────────┬──────────────────┬──────────────┬────────────────────┐
│ Layer (type)         │ Output Shape     │      Param # │ Connected to       │
├──────────────────────┼──────────────────┼──────────────┼────────────────────┤
│ input_layer_2        │ (None, 50)       │            0 │ -                  │
│ (InputLayer)         │                  │              │                    │
├──────────────────────┼──────────────────┼──────────────┼────────────────────┤
│ input_layer_1        │ (None, 100352)   │            0 │ -                  │
│ (InputLayer)         │                  │              │                    │
├──────────────────────┼──────────────────┼──────────────┼────────────────────┤
│ embedding            │ (None, 50, 128)  │    1,280,000 │ input_layer_2[0]…  │
│ (Embedding)          │                  │              │                    │
├──────────────────────┼──────────────────┼──────────────┼────────────────────┤
│ ImageFeature         │ (None, 256)      │   25,690,368 │ input_layer_1[0]…  │
│ (Dense)              │                  │              │                    │
├──────────────────────┼──────────────────┼──────────────┼────────────────────┤
│ CaptionFeature       │ (None, 256)      │      296,448 │ embedding[0][0]    │
│ (GRU)                │                  │              │                    │
├──────────────────────┼──────────────────┼──────────────┼────────────────────┤
│ concatenate          │ (None, 512)      │            0 │ ImageFeature[0][…  │
│ (Concatenate)        │                  │              │ CaptionFeature[0…  │
├──────────────────────┼──────────────────┼──────────────┼────────────────────┤
│ dense (Dense)        │ (None, 10000)    │    5,130,000 │ concatenate[0][0]  │
└──────────────────────┴──────────────────┴──────────────┴────────────────────┘

 Total params: 32,396,816 (123.58 MB)

 Trainable params: 32,396,816 (123.58 MB)

 Non-trainable params: 0 (0.00 B)
```

GRU Results



1. **Training Loss (Blue Line):**
   - The training loss decreases steadily over the course of the 2 epochs, indicating that the model is learning and improving on the training data.
   - This consistent decline suggests that the model is fitting well to the training set.
2. **Validation Loss (Orange Line):**
   - The validation loss decreases initially but then plateaus after around 1 epoch.
   - This plateau suggests that while the model is improving on the training data, it is not significantly improving on unseen validation data after a certain point.



```
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
```

Analysis:

- **Potential Overfitting**
- **Can Include dropout and normalization layers to avoid overfitting**

Comparative Analysis of BLEU Scores

LSTM	GRU
0.47	0.23

Analysis Results:

1. LSTM achieves higher accuracy due to the Attention Mechanism.

2. GRUs are computationally more efficient, making them suitable for applications where speed is critical.

3. GRU’s scores can be increased by increasing the complexity of the architecture.

4. LSTM’s scores can be increased by training for more epochs and more data.

Future Directions

- **Train for more epochs** to improve model convergence.
- **Increase dataset size** for better generalization.
- Incorporate **attention with GRU** to enhance sequence learning.
- **Experiment with transformer architecture** for improved performance on long sequences.

Thank You