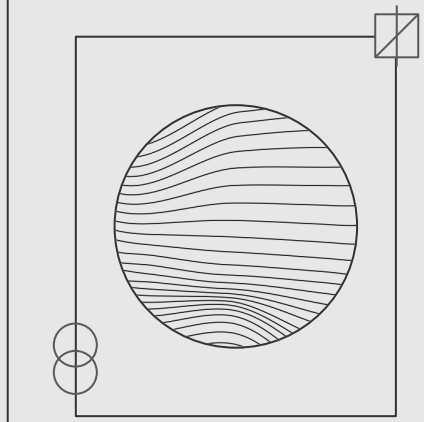
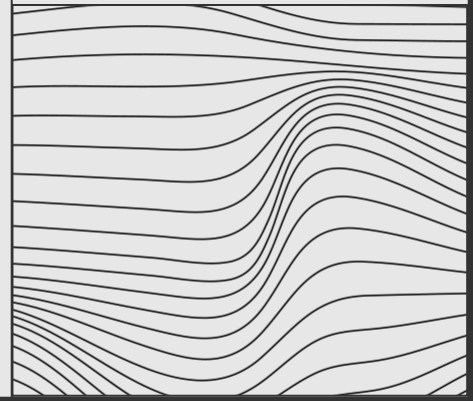
Pictophrases

Teaching Machines to describe images.

Darshan Rajopadhye
Kevin Heleodoro
Poornima Jaykumar Dharamdasani
Saumya Gupta





Introduction

Primary Objective: to create multiple image captioning models capable of generating rich and descriptive captions for images.

Secondary Objective: compare the performance of these models on a dataset to get a better understanding of how well they work.

Why?:

– Medical imaging, social media content filtering, assisting visually impaired, etc.
– Performance progress through different techniques.
– Pros and cons to creating a model from scratch.

InceptionV3

InceptionV3 was used during the pre-processing stage to extract image features and create our training and testing split.

Features:

- Loads and prepares data
- Data splitting and processing

Summary:

InceptionV3 prepares and structures the data necessary for caption generation.

Dataset

Flickr 8k Dataset

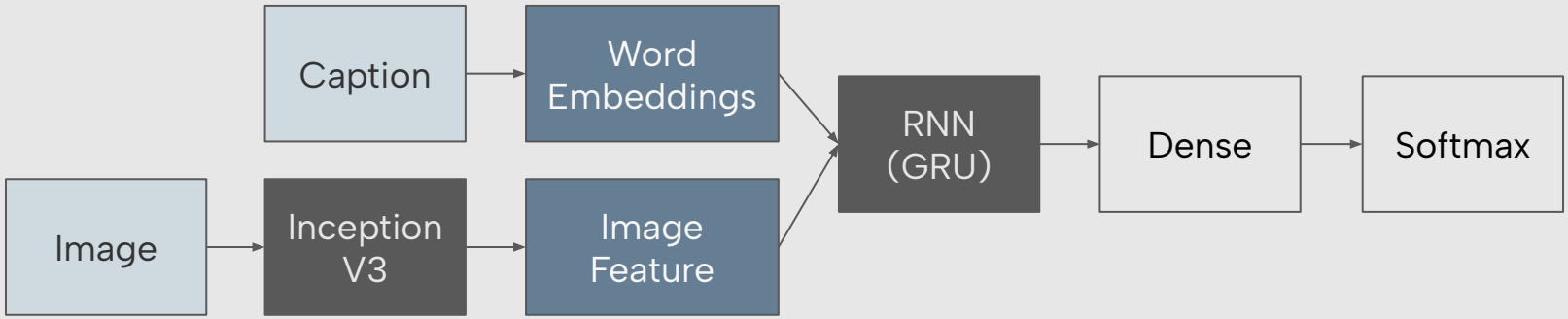
- 8,000 images
- Multiple captions per image
- Wide range of everyday scenes
- More manageable

Metrics

BLEU (Bilingual Evaluation Understudy)

- Score range of 0 to 1
- Compares machine generation vs human translated
- Widely used

RNN - Architecture



- We employed a GRU layer with 256 units to capture sequential dependencies in the caption data.
- Unlike traditional RNNs, GRUs have mechanisms like update and reset gates that allow them to capture long-term dependencies more effectively.

RNN Examples

Predicted Caption:
man is skiing down a snowy mountain

True Caption:
A skier wearing a blue jacket and helmet is skiing down a hill.

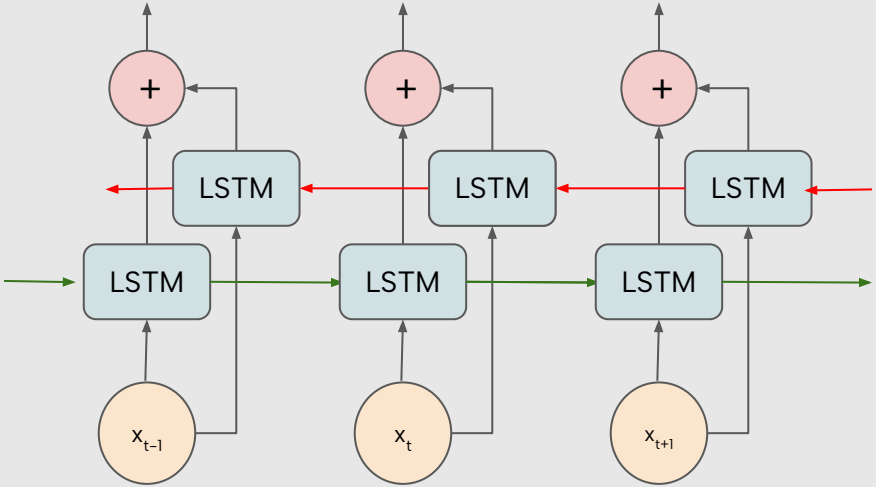
Predicted Caption:
man in red shirt is standing outside building

True Caption:
A group of people looking at sound equipment.

Predicted Caption:
man is sitting on rock by the mountain

True Caption:
A little girl balances on rocks on the beach.

Bi-LSTM Architecture



- Bi-LSTM processes input sequences in both forward and backward directions, capturing information from both past and future contexts.
- Bi-LSTM helps capture long-range dependencies in the data, allowing the model to better understand the relationships between different elements in a sequence.
- Solves the vanishing gradient problem.

BiLSTM Examples

Predicted Caption:
a man in a yellow shirt is airborne on a motorcycle

True Caption: A cyclist in a helmet is driving down a slope on his bike .

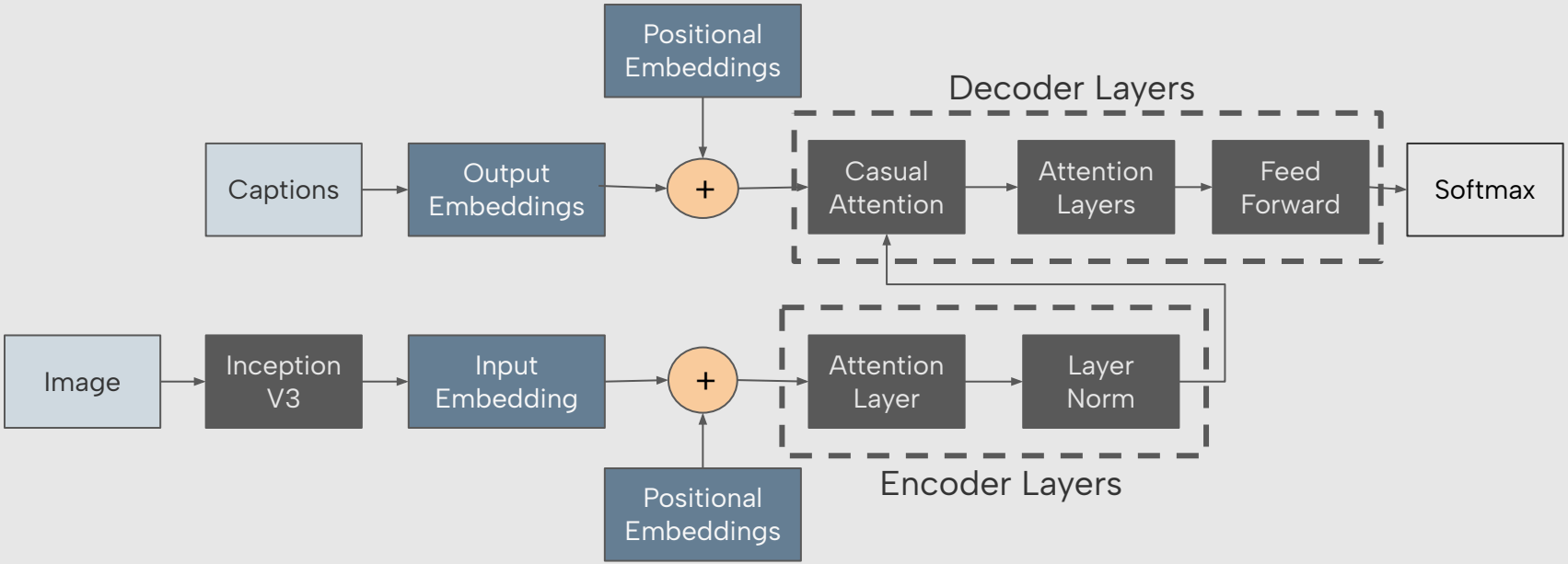
Predicted Caption:
a girl in a purple jacket is riding a blue and red and yellow coat across a field

True Caption : A man in street racer armor is examining the tire of another racer 's motorbike .

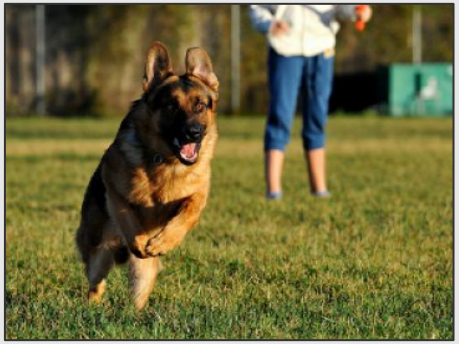
Predicted Caption:
a baseball player in a purple jersey about to kick the ball

True Caption : A baseball player slides toward a base .

Transformer Architecture



Transformer Examples



Predicted Caption:
a brown dog is running through a field

True Caption:
A brown dog is running in the grass

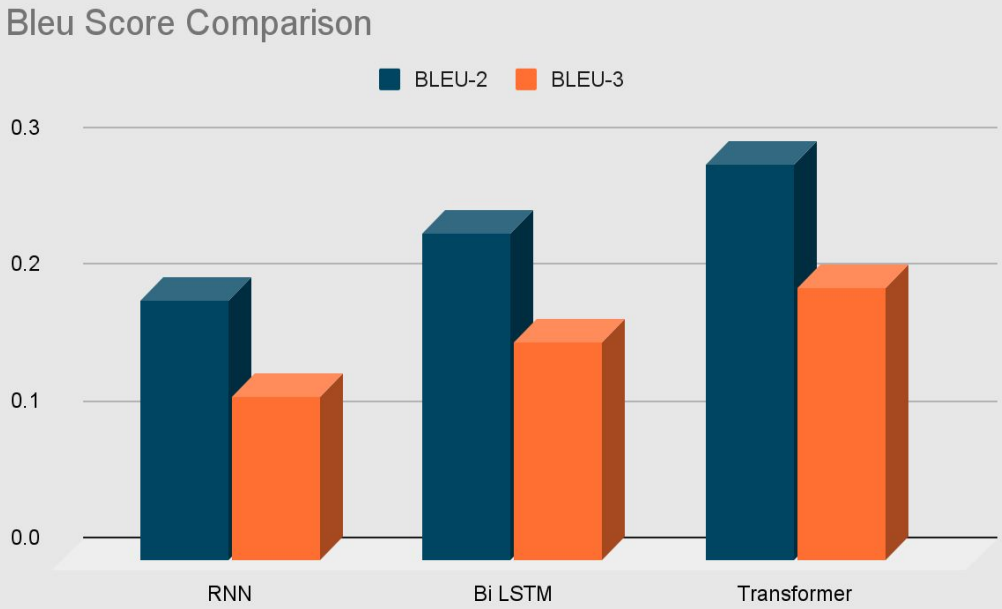
Predicted Caption:
a man in a blue shirt is riding a bike through the woods

True Caption:
A man wearing a blue helmet riding a bike in the woods

Predicted Caption:
a group of people are standing in front of a crowd

True Caption:
Girls in light blue outfits perform a choreographed dance

Results


Bleu Score Comparison
BLEU-2
BLEU-3
0.3
0.2
0.1
0.0
RNN
Bi LSTM
Transformer

Takeaways

Model Performance

The results demonstrate a clear progression in performance.

Transformers

The attention mechanisms just outperform everything.

RNN vs Bi-LSTM

All that compute for just a little improvement in performance.

Future Scope

The data is endless and so are the hyperparameters.

Comparison

Predicted Captions:

1. RNN:

A man is sitting on a bike

2. Bi LSTM :

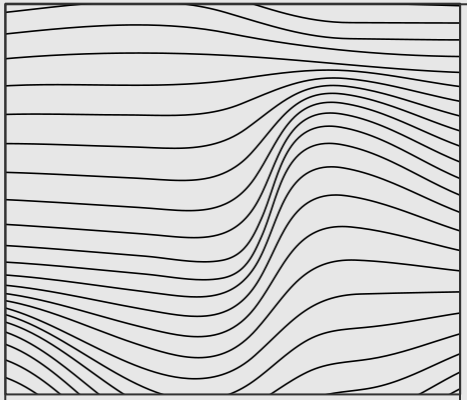
A man wearing a red helmet is climbing up a snow hill.

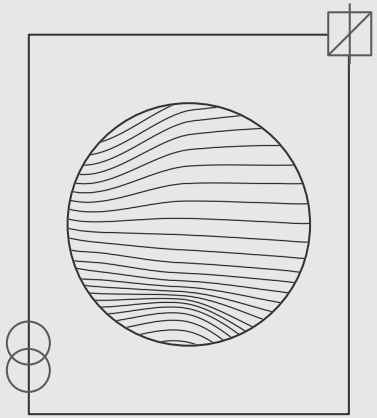
3. Transformer :

A man in a red shirt is standing on a rock.

True Caption:

A woman with her backpack sits on a large rock and looks down over the mountains .





Questions