Syracuse University
School of Information Studies

IMAGE CAPTION GENERATOR

Instructor: Prof. Lu Xiao

Choijamts Bataa, Gina Kuszajewski, Hendi Kushta

IST 664
NATURAL LANGUAGE PROCESSING

INTRODUCTION

Image Captioning is an intriguing field at the intersection of computer vision and natural language processing. In this project, we delve into the fascinating realm of teaching computers to describe images using human-like language.

Our goal is to empower computers with the ability to glance at a picture and effortlessly comprehend its contents. This allows people with vision issues to understand the image content clearly.

IMPORTANT DEFINITIONS

Image Captioning: Generating descriptive text for images using artificial intelligence.

CNN (Convolutional Neural Network): Deep neural network for visual data analysis.

LSTM (Long Short-Term Memory): Recurrent neural network for learning long-term dependencies.

Encoder-Decoder Architecture: Neural network pattern for transforming input to output.

Attention Mechanism: Neural network mechanism for focusing on specific input elements.

BLEU (Bilingual Evaluation Understudy): Metric for evaluating machine-generated text quality.

Transfer Learning: Technique of adapting a pre-trained model to a related task.

Image Embeddings: Vector representations capturing essential features of images.

Epoch: represents one complete pass through the entire training dataset.

LITERATURE REVIEW

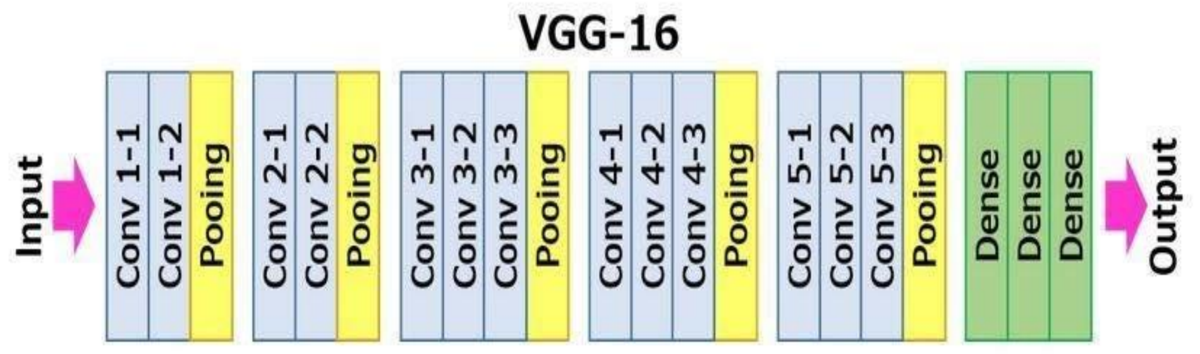
Kawshik B et.al., [1] presented that CNN and LSTM worked well together. CNN is applied to extract features and LSTM is used to create relevant captions. Jha et.al., [2] concluded that deep learning is useful to generate image caption. Sudeep V.P et.al., [3] compared CNN encoders such as VGG-16, VGG-19, ResNet-50, ResNet-101, Inception-v3, and Inception-ResNet-v2 and RNN decoders such as LSTM and GRU. They presented that ResNet-101 is better than other CNN and LSTM and GRU has almost the same results. Sri Neha et. al [5] also tested VGG-16 and ResNet-50, a residual neural network, with Flickr8k, similarly finding that ResNet-50 received higher BLEU scores overall Chang et. al [4] proposed that VGG16-LSTM-Attention paired with color analysis with the OpenCV vision library can provide a more thorough caption than a purely CNN-LSTM one. Finally, Ganesan et. al [6] utilized a CNN-LSTM model to train voice messages to detect and caption images on files, allowing the visually impaired to understand the file in its entirety.

DATA DESCRIPTION

We used the Flickr8k dataset to teach our computer how to talk about pictures. It has 8,091 images, and each picture has five captions. This diverse dataset helps train our computer to generate captions for different images. You can get the dataset from Kaggle and organize it with an "Images" folder and a file named "captions.txt" in a folder named "flickr8k."

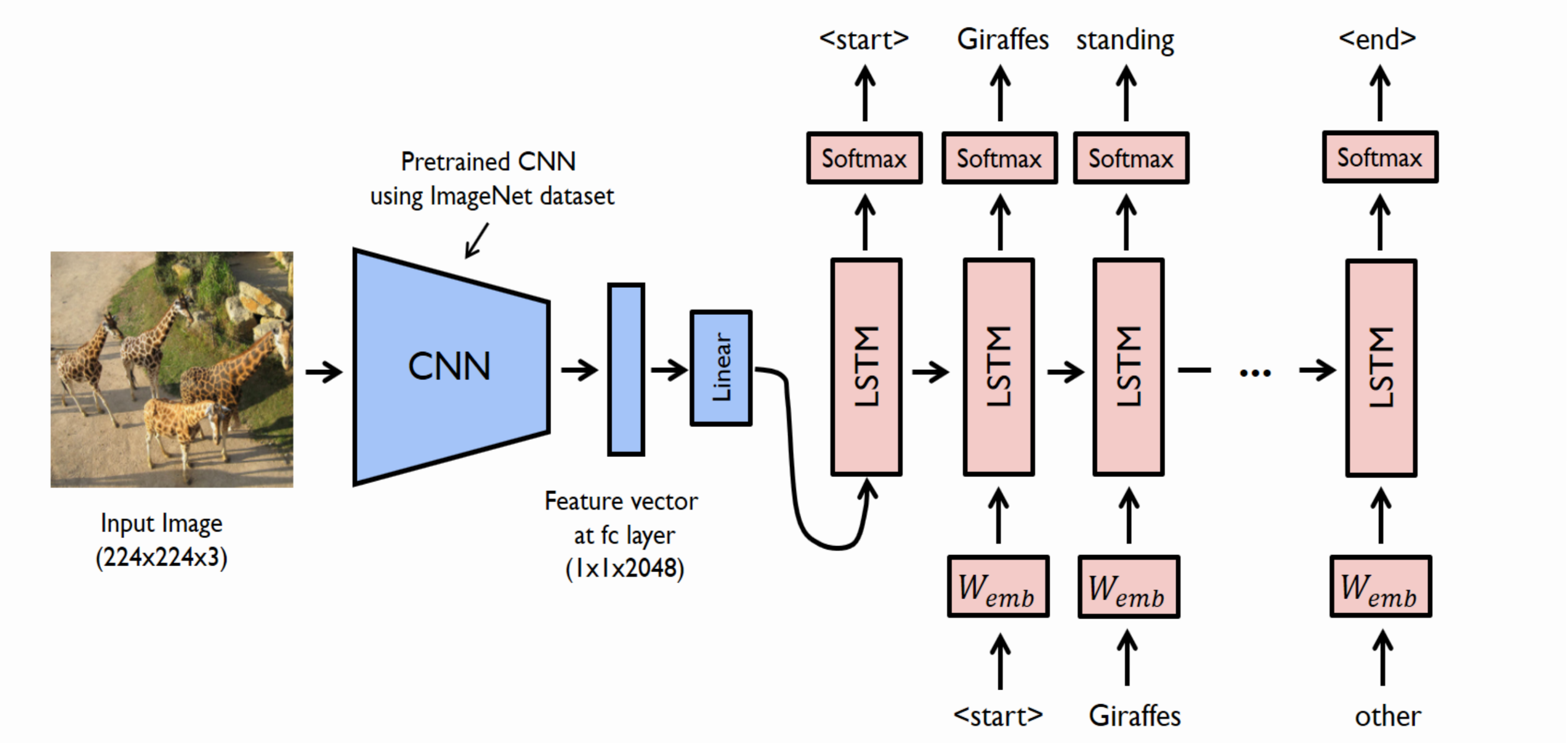
MODEL

VGG16 Model



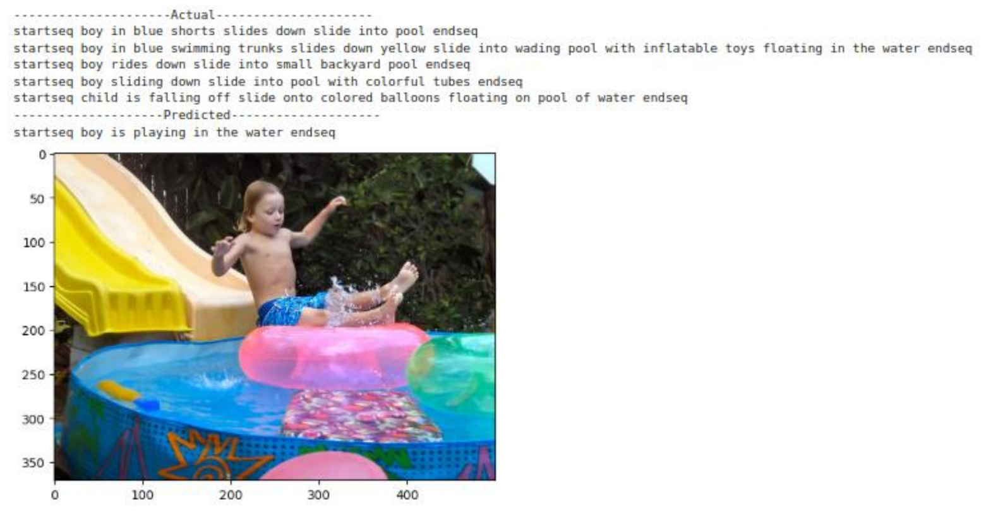
The VGG-16 model is a CNN with 16 layers before the output. It's part of a family of models used for image recognition. After seeing a picture, it breaks the image down into different parts, like shapes and colors to recognize what is in the image. People use VGG-16 to help computers understand and talk about pictures in a way that's almost like how humans do!

LSTM Model Training

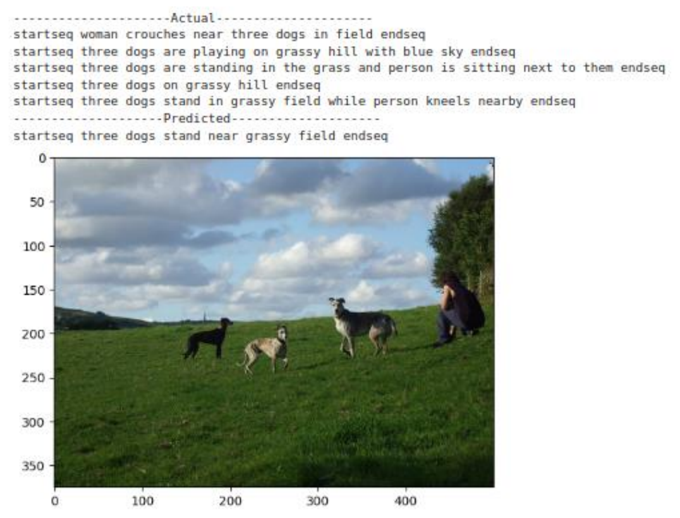


LSTM model teaches the computer to recognize patterns and make predictions in sequences of information, allowing it to understand and generate meaningful outputs.

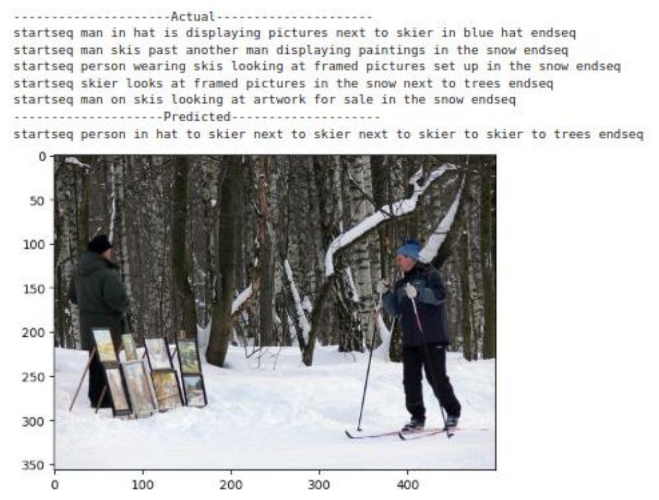
This neural network, using a pretrained CNN, extracts features from an input image. The features undergo linear processing guided by learned weights, followed by an LSTM layer to generate sequential captions. Softmax predicts words, and word embeddings capture the meaning of each word in the sequence. "Start" tokens signify the caption's beginning.



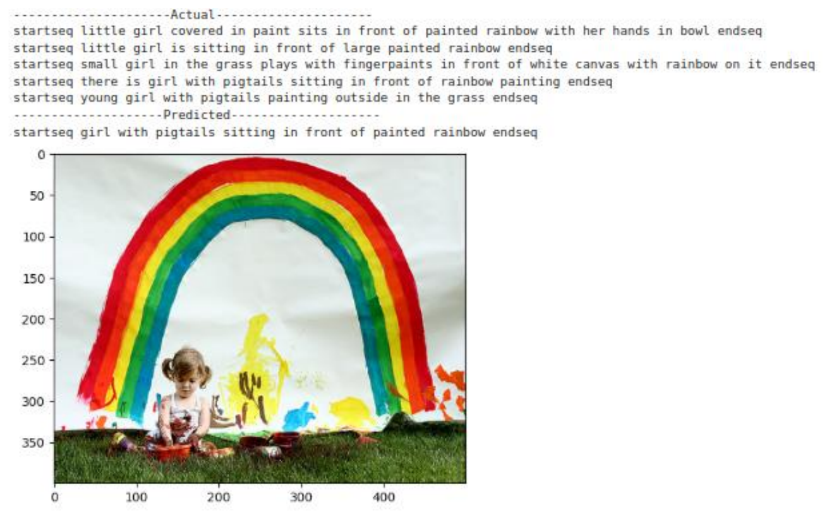
```
---------------------Actual--------------------
startseq boy in blue shorts slides down slide into pool endseq
startseq boy in blue swimming trunks slides down yellow slide into wading pool with inflatable toys floating in the water endseq
startseq boy rides down slide into small backyard pool endseq
startseq boy sliding down slide into pool with colorful tubes endseq
startseq child is falling off slide onto colored balloons floating on pool of water endseq
--------------------Predicted-------------------
startseq boy is playing in the water endseq
```



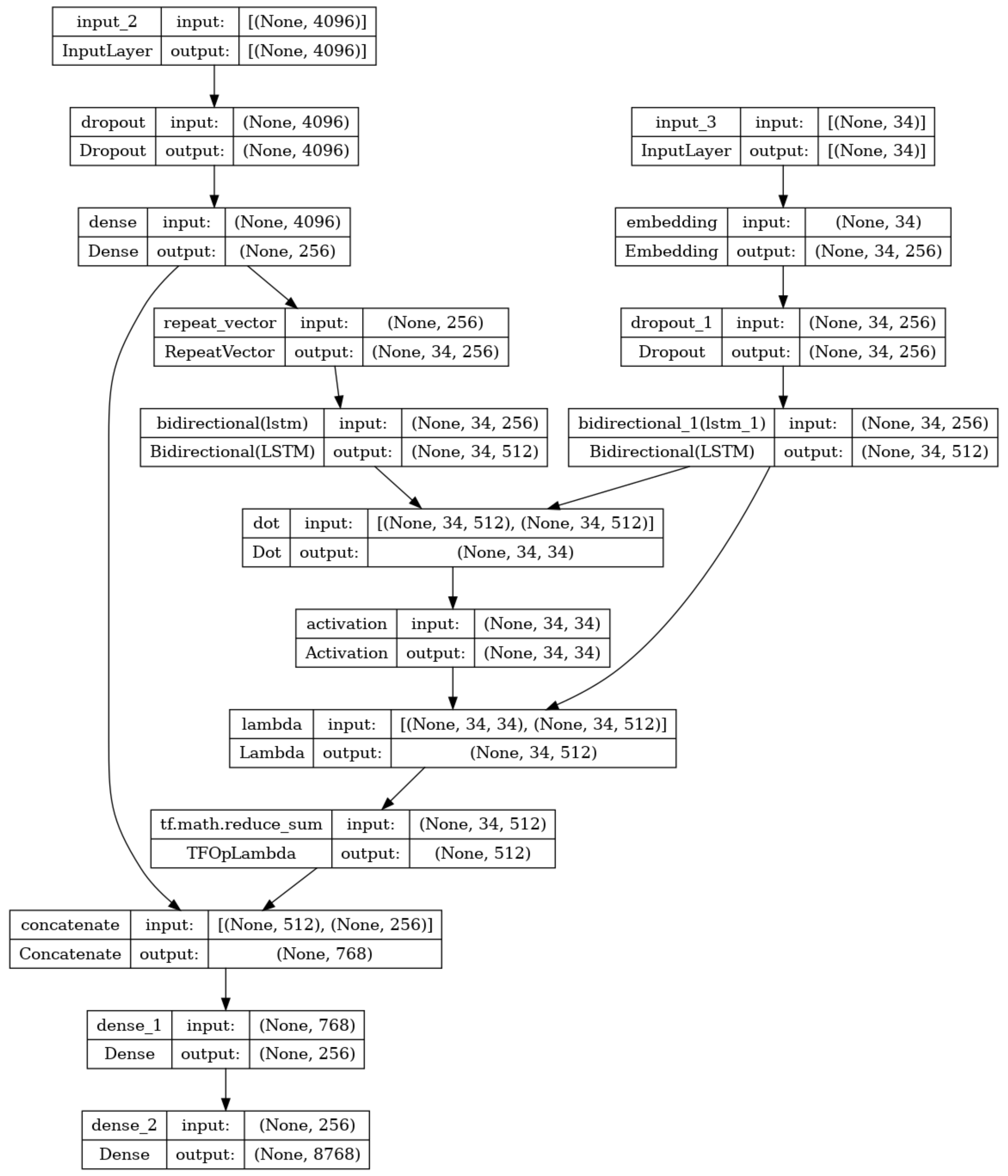
```
---------------------Actual--------------------
startseq woman crouches near three dogs in field endseq
startseq three dogs are playing on grassy hill with blue sky endseq
startseq three dogs are standing in the grass and person is sitting next to them endseq
startseq three dogs on grassy hill endseq
startseq three dogs stand in grassy field while person kneels nearby endseq
--------------------Predicted-------------------
startseq three dogs stand near grassy field endseq
```



```
---------------------Actual--------------------
startseq man in hat is displaying pictures next to skier in blue hat endseq
startseq man skis past another man displaying paintings in the snow endseq
startseq person wearing skis looking at framed pictures set up in the snow endseq
startseq skier looks at framed pictures in the snow next to trees endseq
startseq man on skis looking at artwork for sale in the snow endseq
--------------------Predicted-------------------
startseq person in hat to skier next to skier next to skier to skier to trees endseq
```



```
---------------------Actual--------------------
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl endseq
startseq little girl is sitting in front of large painted rainbow endseq
startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it endseq
startseq there is girl with pigtails sitting in front of rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
--------------------Predicted-------------------
startseq girl with pigtails sitting in front of painted rainbow endseq
```



input_2	InputLayer	input:	[(None, 4096)]
		output:	[(None, 4096)]
dropout	Dropout	input:	(None, 4096)
		output:	(None, 4096)
dense	Dense	input:	(None, 4096)
		output:	(None, 256)
repeat_vector	RepeatVector	input:	(None, 256)
		output:	(None, 34, 256)
bidirectional(lstm)	Bidirectional(LSTM)	input:	(None, 34, 256)
		output:	(None, 34, 512)
input_3	InputLayer	input:	[(None, 34)]
		output:	[(None, 34)]
embedding	Embedding	input:	(None, 34)
		output:	(None, 34, 256)
dropout_1	Dropout	input:	(None, 34, 256)
		output:	(None, 34, 256)
bidirectional_1(lstm_1)	Bidirectional(LSTM)	input:	(None, 34, 256)
		output:	(None, 34, 512)
dot	Dot	input:	[(None, 34, 512), (None, 34, 512)]
		output:	(None, 34, 34)
activation	Activation	input:	(None, 34, 34)
		output:	(None, 34, 34)
lambda	Lambda	input:	[(None, 34, 34), (None, 34, 512)]
		output:	(None, 34, 512)
tf.math.reduce_sum	TFOpLambda	input:	(None, 34, 512)
		output:	(None, 512)
concatenate	Concatenate	input:	[(None, 512), (None, 256)]
		output:	(None, 768)
dense_1	Dense	input:	(None, 768)
		output:	(None, 256)
dense_2	Dense	input:	(None, 256)
		output:	(None, 8768)

RESULTS

The project's success is evaluated using BLEU (Bilingual Evaluation Understudy) scores, a metric commonly employed in machine translation tasks. BLEU scores assess the quality of generated captions by comparing them to the ground truth captions.

BLEU-1 (Unigrams): 0.429907

Indicates that, on average, about 43% of the unigrams (individual words) in the model's predicted captions match those in the reference captions.

BLEU-2 (Bigrams): 0.212081

Shows that roughly 21% of bigrams in the predicted captions overlap with those in the reference captions.

FUTURE WORK

- Improve Model: Make the computer's image descriptions even better by adjusting how it learns.
- More Data, More Smarts: Teach the computer with extra pictures from datasets like Flickr30k, which has over twice the amount of data.
- Compare the performance with other models.
- Smart Choices for Captions: Get the computer to produce many captions for a picture and pick the best one using beam search.
- Upgrade App Features: Make the app (seen on laptop below) cooler with image previews and confidence scores for the computer's captions.
- Speak Different Languages: Teach the computer to talk about pictures in various languages using datasets with lots of languages.

REFERENCES

[1] Kawshik B. et.al (2023). Deep learning based automated image caption generator. *International Journal for Innovative Engineering and Management Research*, 12(2), 573-5 https://doi.org/10.48047/IJIEMR/V12/ISSUE 02/88

[2] Jha D. et.al (2022) Image caption generator using Deep learning. *International Journal for Research in Applied Science and Engineering technology*, 10(X), 621 626 https://doi.org/10.22214/ijraset.2022.47058

[3] Sudeep V.P et.al (2022). Image Captioning Encoder–Decoder Models Using CNN-RNN Architectures: A Comparative Study. https://doi.org/10.1007/s00034-022-02050-2

[4] Chang, Y.-H., Chen, Y.-J., Huang, R.-H., & Yu, Y.-T. (2022). Enhanced Image Captioning with Color Recognition Using Deep Learning Methods. *Applied Sciences*, 12(1), 209. https://doi.org/10.3390/app12010209

[5] Sri Neha, V., Nikhila, B., Deepika, K., Subetha, T. (2022). A Comparative Analysis on Image Caption Generator Using Deep Learning Architecture—ResNet and VGG16. In: Smys, S., Tavares, J.M.R.S., Balas, V.E. (eds) Computational Vision and Bio-Inspired Computing. Advances in Intelligent Systems and Computing, vol 1420. Springer, Singapore. https://doi-org.libezproxy2.syr.edu/10.1007/978-981-16-9573-5_15

[6] Ganesan, J., Azar, A. T., Alsenan, S., Kamal, N. A., Qureshi, B., & Hassanien, A. E. (2022). Deep Learning Reader for Visually Impaired. *Electronics*, *11*(20), 3335. https://doi.org/10.3390/electronics11203335

cbataa@syr.edu, rkuszaje@syr.edu, hkushta@syr.edu