Visual Recognition Mini- Project Report

Team ID: 4

Team Member 1: Kalahasti V V Depesh (IMT2019508)

Team Member 2: Maruthi Sriram (IMT2019)

Team Member 3: T Tarun Reddy (IMT2019)

Part 1:

Question:

Design a CNN-LSTM system that can perform image captioning [System 1] based on the following details:

a) You are free to decide the CNN and LSTM architecture that best suits your case. The objective is to achieve a good BLEU score on the test set of Flickr8K data.
b) Use the Flickr8K data for training and testing the model.

Solution:

We have used keras to design our CNN-LSTM system.

Our designed system performs image captioning on the Flickr8K dataset using the InceptionV3 pretrained CNN as a feature extractor. We used about 8000 images for training the image captioning system and about 2000 images for testing the system.

We have used BLEU score as an evaluation metric for the system’s performance on the Flickr8K dataset.

System Architecture:

Initially we read all the image files and respective captions from the file and store them in a dictionary. Each entry in this dictionary contains the image filename and all of its respective captions.

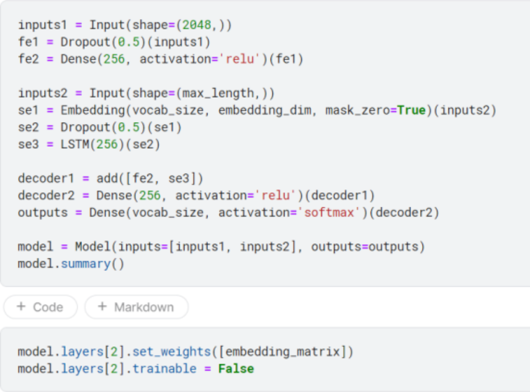
The next step is to perform some preprocessing on the captions: we convert the captions to lower case and remove any punctuation.

The next step is to create a vocabulary for all the words in the captions. This will be original vocabulary size considering all words regardless of their frequency.

Next, we split the images into train and test images and also get train image descriptors and test image descriptors.

Now we create a new vocabulary from all of the captions but we consider only those words whose frequency is more than 10. This will be useful while training as we do not have consider words that are rare.

Now we use the pretrained InceptionV3 CNN to get image features. Now we create a custom RNN LSTM model and use this model along with image features and captions vocabulary for training.



```
inputs1 = Input(shape=(2048,))
fe1 = Dropout(0.5)(inputs1)
fe2 = Dense(256, activation='relu')(fe1)

inputs2 = Input(shape=(max_length,))
se1 = Embedding(vocab_size, embedding_dim, mask_zero=True)(inputs2)
se2 = Dropout(0.5)(se1)
se3 = LSTM(256)(se2)

decoder1 = add([fe2, se3])
decoder2 = Dense(256, activation='relu')(decoder1)
outputs = Dense(vocab_size, activation='softmax')(decoder2)

model = Model(inputs=[inputs1, inputs2], outputs=outputs)
model.summary()
```

```
model.layers[2].set_weights([embedding_matrix])
model.layers[2].trainable = False
```

After fitting the training data to the model, we use Beam Search to get our final predictions.

Beam search is an algorithm used in many NLP and speech recognition models as a final decision-making layer to choose the best output given target variables.

BLEU Score:

After getting our final caption predictions, we calculate the BLEU score.

The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence. The approach works by counting matching n-grams in the candidate translation to n-grams in the reference text, where 1-gram or unigram would be each token and a bigram comparison would be each word pair. The comparison is made regardless of word order.

Objective Model results:

Our model gave a BLEU score of 49 percent on average.

Part-2:

Question:

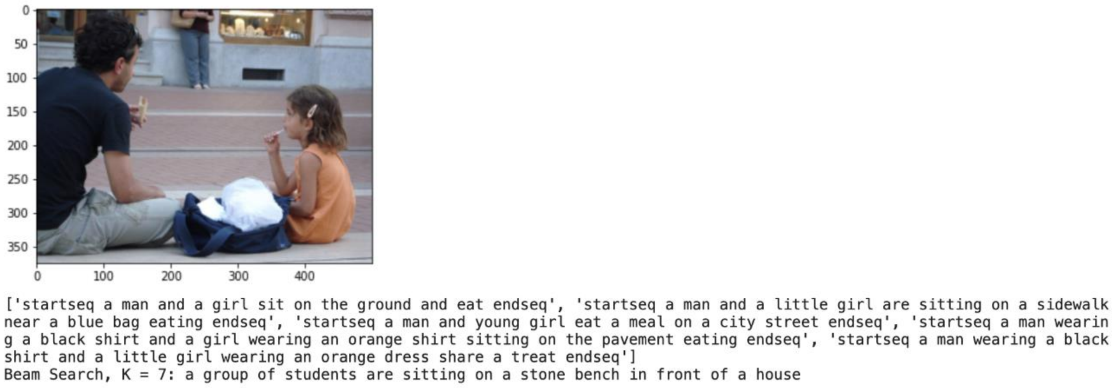
The LSTM based image captioning can 'blindly' learn the structure of the language and predict meaningful sentences even without learning much insight to the content of the image. This is termed as "language bias" of the system. Design a training experiment with Flickr8K data to assess the language bias of your CNN-LSTM system [System 1 Modified]. Provide objective and subjective analysis/comparison of the results of System 1 and System 1 Modified .

Solution:

To find different language bias in our CNN-LSTM system, we predicted the output of 200 test images, and compared the BLEU score of them. We took the threshold to tell whether the caption predicted is wrong as 0.25. We collected them and observed the following Biases in our predictions.

Bias 1:

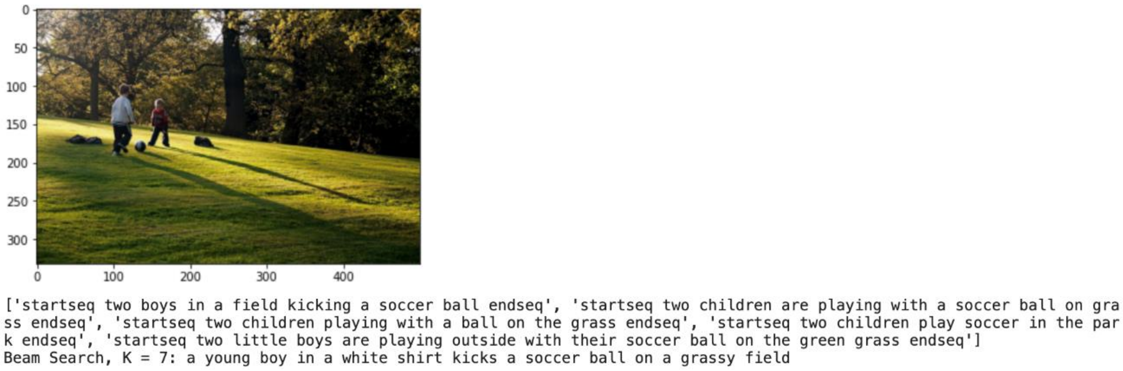
If there are a group of people in the picture, it would show that they are sitting on a bench. Example is given below



```
['startseq a man and a girl sit on the ground and eat endseq', 'startseq a man and a little girl are sitting on a sidewalk
near a blue bag eating endseq', 'startseq a man and young girl eat a meal on a city street endseq', 'startseq a man wearin
g a black shirt and a girl wearing an orange shirt sitting on the pavement eating endseq', 'startseq a man wearing a black
shirt and a little girl wearing an orange dress share a treat endseq']
Beam Search, K = 7: a group of students are sitting on a stone bench in front of a house
```

Bias 2:

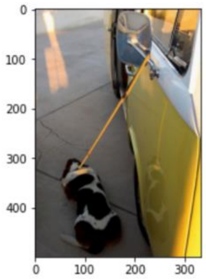
If Children are there it always thinks that one of them is playing soccer.



```
['startseq two boys in a field kicking a soccer ball endseq', 'startseq two children are playing with a soccer ball on gra
ss endseq', 'startseq two children playing with a ball on the grass endseq', 'startseq two children play soccer in the par
k endseq', 'startseq two little boys are playing outside with their soccer ball on the green grass endseq']
Beam Search, K = 7: a young boy in a white shirt kicks a soccer ball on a grassy field
```

Bias 3:

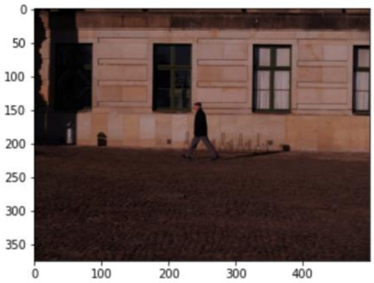
If a dog is there it always thinks it’s doing some sort of moving activity.



```
['startseq a basset hound is leashed to the rearview mirror of a yellow and white vehicle endseq', 'startseq a black and w
hite dog tied to a yellow and white van endseq', 'startseq a dog lying down tethered to the side mirror of a yellow vw bus
endseq', 'startseq a dog with black white and brown coloring is leashed up to a mirror endseq', 'startseq the black and wh
ite dog is tethered next to a yellow car endseq']
Beam Search, K = 7: a brown and white dog is jumping over a brick wall
```

Bias 4:

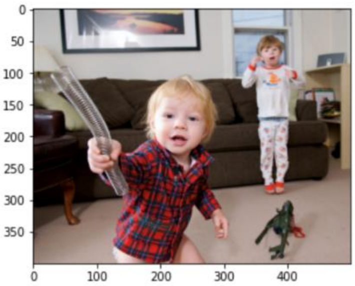
If a man is there it always thinks that there is another person with him and they are sitting on a bench or watching something



```
['startseq a man wearing a black hat walks along a road next to a building endseq', 'startseq an older man in a long sleev
ed black shirt is walking down a cobblestone street alone endseq', 'startseq a old man walks down the uncrowded road endse
q', 'startseq a person walking in front of a building endseq', 'startseq the man is walking near a brown building endseq']
Beam Search, K = 7: a man and a woman sitting on a bench in front of a tree
```

Bias 5:

Whenever there are little children, the model assumes that a little girl is playing with a leather toy.



```
['startseq an older child watching a toddler play with a long spring inside a house endseq', 'startseq a young child plays
with a slinky in the family room while an older sibling watches endseq', 'startseq two little children and one is playing
with a slinky endseq', 'startseq two young children are playing in the living room with some toys endseq', 'startseq two y
oung children playing in front of a couch endseq']
Beam Search, K = 7: a little girl in a blue dress is playing with a ribbon toy
```