Visual Recognition

Mini Project

Team ID - 19

IMT2019041 - Kasturi Siva Hitesh
IMT2019039 - K.S.S.V.Naveen
IMT2019045 - Kopparapu Sai Krishna

Link to code - https://www.kaggle.com/code/ksaikrishna/image-caption-generator

Problem - 1:

Process Flow :

Processing -

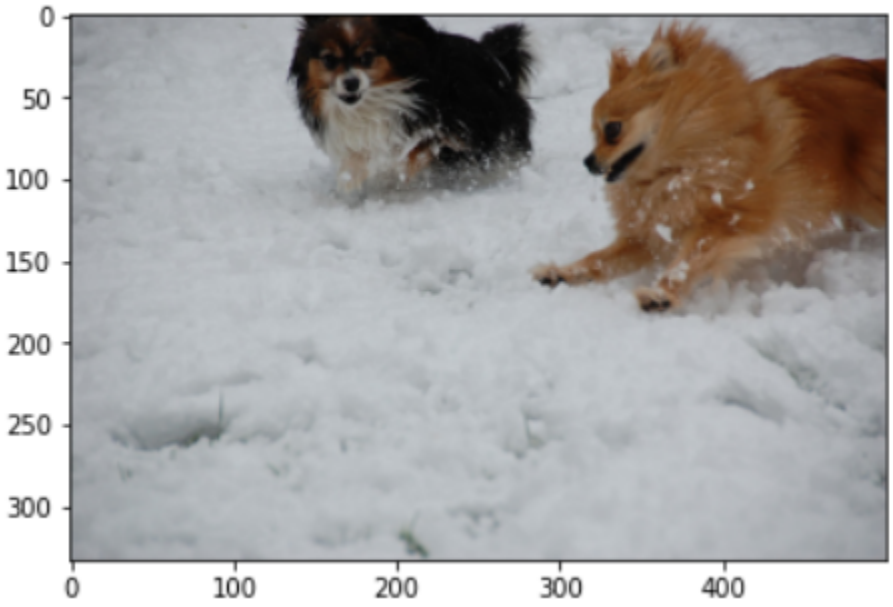
1. Extracting and deleting unnecessary spaces and punctuations in the descriptions of the images.
2. The descriptions are stored into a dictionary with image names as keys and values as the list of 5 descriptions for the image as values.
3. A vocabulary unique words is created from the list of 40000 (8000 * 5) image cations.
4. No. of distinct words found = 8828.
5. Two tokens ‘startseq’ and ‘endseq’ are added to each caption.
6. To make the model robust, the words that occur at least 10 times in the entire 40000 captions are chosen. Filtering, final no. of words = 1660.
7. To predict a caption limited to a length, maximum length of the caption is chosen as the maximum length of the caption predicted. Maximum length = 38
8. The 38 length caption is mapped to a 200 dimension vector using Glove. Where similar words are clustered together and different words are separated.

Model -

1. Inception v-3 network is used that is pre-trained on Image-Net classification. This network accepts images of shape (299, 299).
2. Image vectors of shape (2048,1) are extracted from the training and testing images.
3. Model is created as a merge model with the image vector and partial caption combined. Steps to form the model -
    a. Process sequence from the text.
    b. Extract feature vectors from the text.
    c. Decode the output using softmax by concatenating the above two layers.
4. Partial caption of maximum length 34 is fed to the embedding layer to map the word to 200d Glove embedding. A dropout of 0.5 is chosen to avoid overfitting.
5. Image vector extracted by the inception v-3 network is stored in input_2 and fed into the fully connected layer.
6. Image and Language models are concatenated by adding and fed into another fully connected layer which gives the softmax probabilities.
7. Adam optimiser and Categorical_Crossentropy loss function are used.
8. Greedy search picks the word with the highest probability and predicts the next word.
9. In Beam Search, top k predictions are chosen and fed into the model and sort them using the probabilities returned by the model.
10. Hyper Parameters -
    a. Epochs = 5, batch_size = 3, no. of steps = len(train_descriptions)//batch_size, k = 3, 10.

Outputs of image captioning :

Image-1

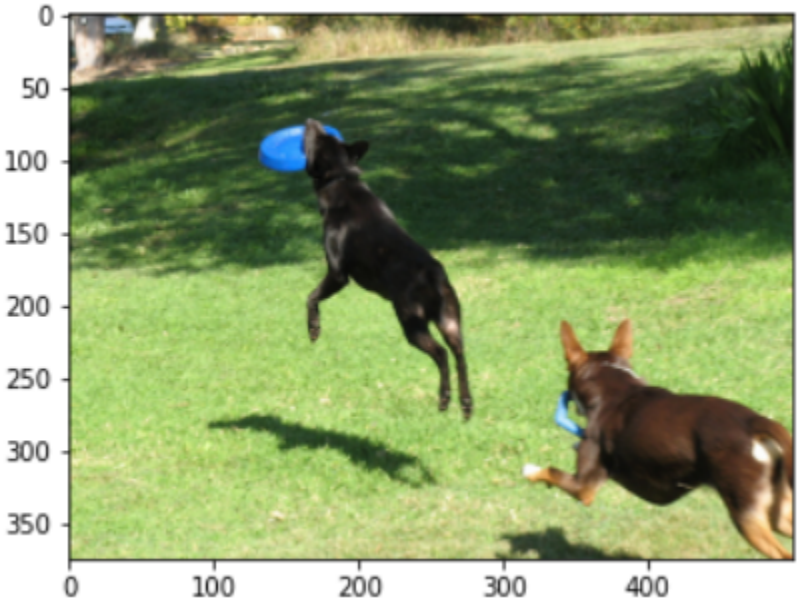


Actual captions of image: ['a black dog and a brown dog in snow ', 'the small dogs play in the snow ', 'two longhaired puppy dogs have a romp in the snow ', 'two pomeranian dogs playing in the snow ', 'two small dogs playing in the snow together ']

```
***** Greedy Search *****
Predicted sentence: a dog is running through the snow
Bleu score:  0.5773502691896257
***** Beam Search K = 3 *****
Predicted sentence: a brown and white dog is running through the snow
Bleu score:  0.5773502691896257
***** Beam Search K = 10 *****
Predicted sentence: a brown and white dog is running through the snow
Bleu score:  0.5773502691896257
```

- For this image, it is clearly evident that the captions generated are correct(i.e they are relevant to the input image) and the bleu score can be seen above. One of the dogs' color has matched and playing can be related to running.

Image-2



Actual captions of image: ['a dog with a frisbee in front of a brown dog ', 'a large black dog is catching a frisbee while a large brown dog follows shortly after ', 'two dark colored dogs romp in the grass with a blue frisbee ', 'two dogs are catching blue frisbees in grass ', 'two dogs are playing one is catching a frisbee ']

```
***** Greedy Search *****
Predicted sentence: a black and white dog is jumping up at a larger dog
Bleu score:  0.5773502691896257
***** Beam Search K = 3 *****
Predicted sentence: a brown and white dog is jumping over a red ball
Bleu score:  0.5773502691896257
***** Beam Search K = 10 *****
Predicted sentence: a brown and white dog is playing with a soccer ball in the grass
Bleu score:  0.5773502691896257
```

- For this image, it is clearly evident that the captions generated are correct(i.e they are relevant to the input image) and the bleu score can be seen above.

Image-3

Actual captions of image: ['a little boy in a red jacket plays on a jungle gym ', 'a little boy playing on a playground ', 'a small black child plays on an outdoor jungle gym ', 'a young boy in a black sweatshirt and red vest plays on a red jungle gym ', 'boy playing on some metal bars']

```
***** Greedy Search *****
Predicted sentence: a young boy jumps on a trampoline
Bleu score:  0.6147881529512643
***** Beam Search K = 3 *****
Predicted sentence: a young boy in a red shirt is playing on a swing
Bleu score:  0.6147881529512643
***** Beam Search K = 10 *****
Predicted sentence: a little boy plays on a swing
Bleu score:  0.6147881529512643
```

- For this image, the caption prediction is somewhat similar to actual captions. The main thing in the above image is a boy playing something. The captions we generated are stating that but are not telling correctly what he is playing.

Problem - 2:

The below are some instances where the caption generation went wrong (i.e didn't give correct predictions according to the given picture), but the captions generated are meaningful. This is called language bias.

Image-4

Actual captions of image: ['a woman holding onto a camera smiling at the camera ', 'a woman with a blue hat and blue and red jacket setting up a camera on a tripod ', 'a woman with a camera on a tripod is smiling for another camera ', 'people at a photo shoot ', 'the woman in blue is operating a camera in front of two other women ']

```
***** Greedy Search *****
Predicted sentence: a man in a red shirt and a woman in a white shirt and a woman in a white shirt are walking down a street
Bleu score:  0.668740304976422
***** Beam Search K = 3 *****
Predicted sentence: a group of people stand on a street
Bleu score:  0.668740304976422
***** Beam Search K = 10 *****
Predicted sentence: a man and a woman pose for a picture on a street
Bleu score:  0.668740304976422
```

- For this image, it is clearly evident that the captions generated are not correct because;
  - The colors of the shirts of both men and women are wrong.
  - There is no street in the image and they are not walking.

- But, the captions predicted are meaningful according to the language. This is called language bias.

Image-5

Actual captions of image: ['a long haired drummer plays music outdoors ', 'a man with long hair and tattoos plays a drum outdoors ', 'a musician plays a drum while his hair covers his face and tattoos dot his arms ', 'a tatooed person shakes their hair and beats a drum ', 'the drummer beats the drums at an outdoor concert ']

```
***** Greedy Search *****
Predicted sentence: a man in a black shirt and jeans is standing on a bench
Bleu score:  0.5623413251903491
***** Beam Search K = 3 *****
Predicted sentence: a boy in a red shirt is playing with a red ball in a park
Bleu score:  0.6147881529512643
***** Beam Search K = 10 *****
Predicted sentence: a group of people are sitting on a bench in front of a store
Bleu score:  0.6147881529512643
```

- For this image, it is clearly evident that the captions generated are not correct because;
    - There is no beach in the above image.
    - There is no stone in the above image.
    - There is no red ball in the above image.
- But, the captions predicted are meaningful according to the language. This is called language bias.

Image-6

Actual captions of image: ['a crowd is watching a dog climb up a staircase ', 'a dog is running through an obstacle course in front of a group of people ', 'a dog performs a trick on a ladder in front of a crowd observing his talent ', 'large brown dog walks up a blue staircase ', 'onlookers watch a german shephard climb steps on an obstacle course ']

```
***** Greedy Search *****
Predicted sentence: a man is sitting on a bench next to a man in a white shirt
Bleu score:  0.5946035575013605
***** Beam Search K = 3 *****
Predicted sentence: a man in a blue shirt is sitting on a bed
Bleu score:  0.5946035575013605
***** Beam Search K = 10 *****
Predicted sentence: a group of people sit on top of a truck
Bleu score:  0.5946035575013605
```

- For this image, it is clearly evident that the captions generated are not correct because;
    - There is no truck in the above image.
    - The main thing in the above image is a dog doing something, but no predicted caption contains it.
    - There is no beach in the above image.
- But, the captions predicted are meaningful according to the language. This is called language bias.