UCMerced Image Captioning

Dataset : UC Merced

This is a 21 class land use image dataset meant for research purposes.

There are 100 images for each of the following classes:

1. agricultural
2. airplane
3. baseballdiamond
4. beach
5. buildings
6. chaparral
7. denseresidential
8. forest
9. freeway
10. golfcourse
11. harbor
12. intersection
13. mediumresidential
14. mobilehomepark
15. overpass
16. parkinglot
17. river
18. runway
19. sparseresidential
20. storagetanks
21. tenniscourt

Each image measures 256x256 pixels.

The images were manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country. The pixel resolution of this public domain imagery is 1 foot.
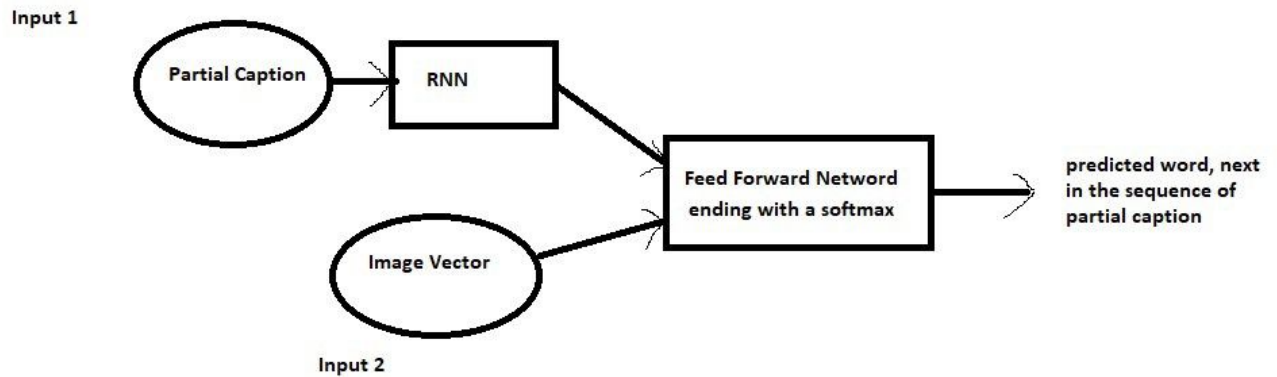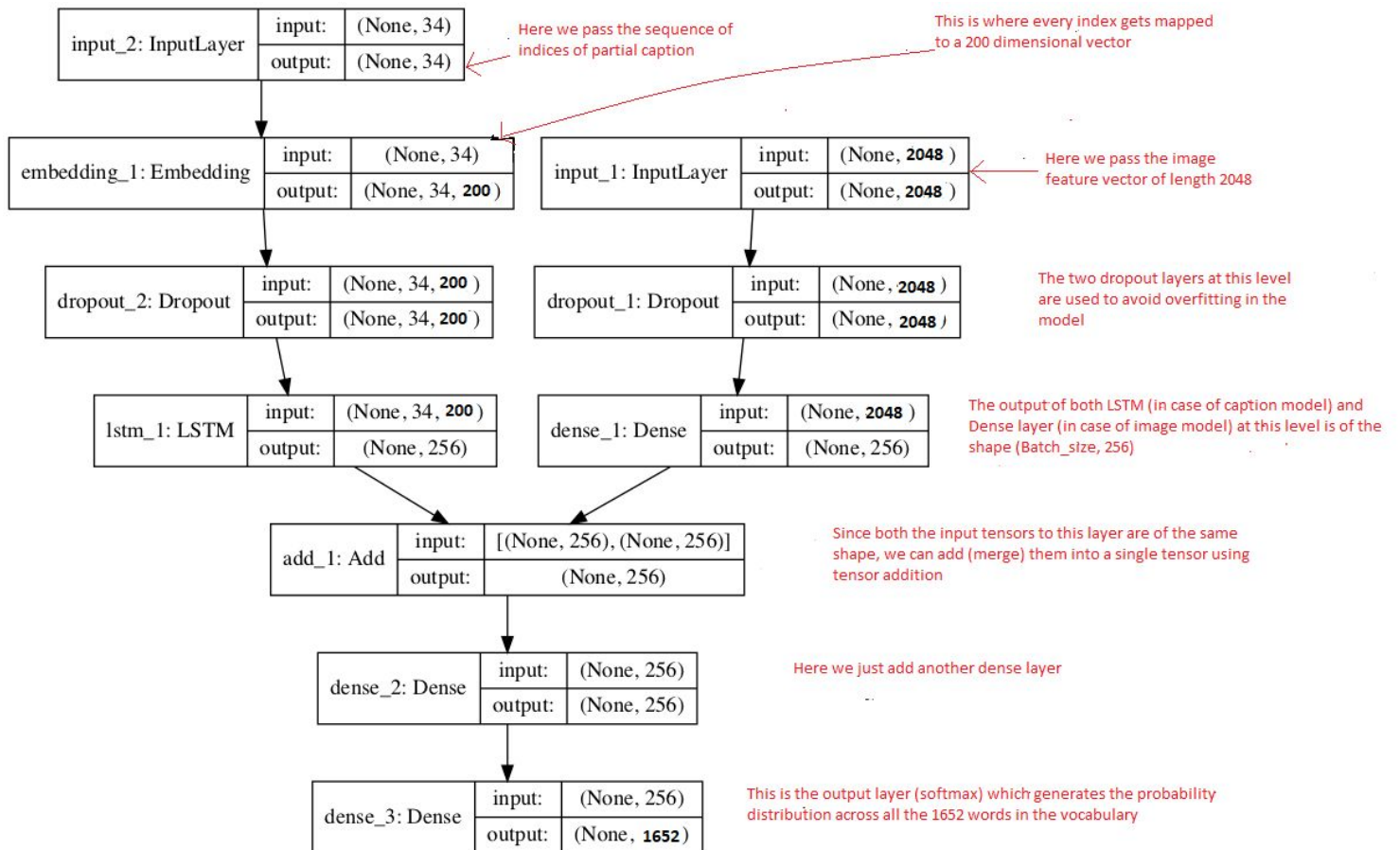
Sample Image and its respective caption

5 Captions related to image in the left.

1. There are two airplanes closed to each other in the airport.
2. Two airplanes are stopped at the airport.
3. There are one big airplane and one small airplane stopped at the airport.
4. Two airplanes stopped at the airport closed to each other.
5. Two white airplanes are stopped at the airport.
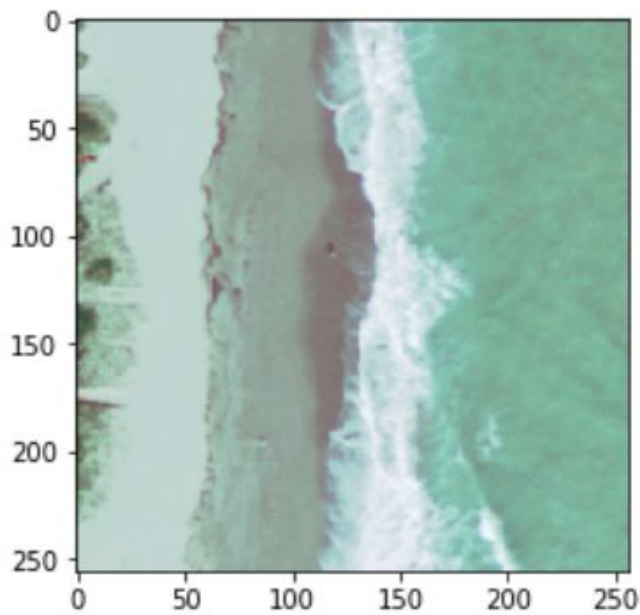
Image Captioning Model Architecture :

Input 1

Partial Caption → RNN

Image Vector

Feed Forward Network ending with a softmax → predicted word, next in the sequence of partial caption

Input 2

Flowchart of the architecture

| input_2: InputLayer | input: | (None, 34) |
| | output: | (None, 34) |

Here we pass the sequence of indices of partial caption

This is where every index gets mapped to a 200 dimensional vector

| embedding_1: Embedding | input: | (None, 34) |
| | output: | (None, 34, **200**) |

| input_1: InputLayer | input: | (None, **2048**) |
| | output: | (None, **2048**) |

Here we pass the image feature vector of length 2048

| dropout_2: Dropout | input: | (None, 34, **200**) |
| | output: | (None, 34, **200**) |

| dropout_1: Dropout | input: | (None, **2048**) |
| | output: | (None, **2048**) |

The two dropout layers at this level are used to avoid overfitting in the model

| lstm_1: LSTM | input: | (None, 34, **200**) |
| | output: | (None, 256) |

| dense_1: Dense | input: | (None, **2048**) |
| | output: | (None, 256) |

The output of both LSTM (in case of caption model) and Dense layer (in case of image model) at this level is of the shape (Batch_size, 256)

| add_1: Add | input: | [(None, 256), (None, 256)] |
| | output: | (None, 256) |

Since both the input tensors to this layer are of the same shape, we can add (merge) them into a single tensor using tensor addition

| dense_2: Dense | input: | (None, 256) |
| | output: | (None, 256) |

Here we just add another dense layer

| dense_3: Dense | input: | (None, 256) |
| | output: | (None, **1652**) |

This is the output layer (softmax) which generates the probability distribution across all the 1652 words in the vocabulary
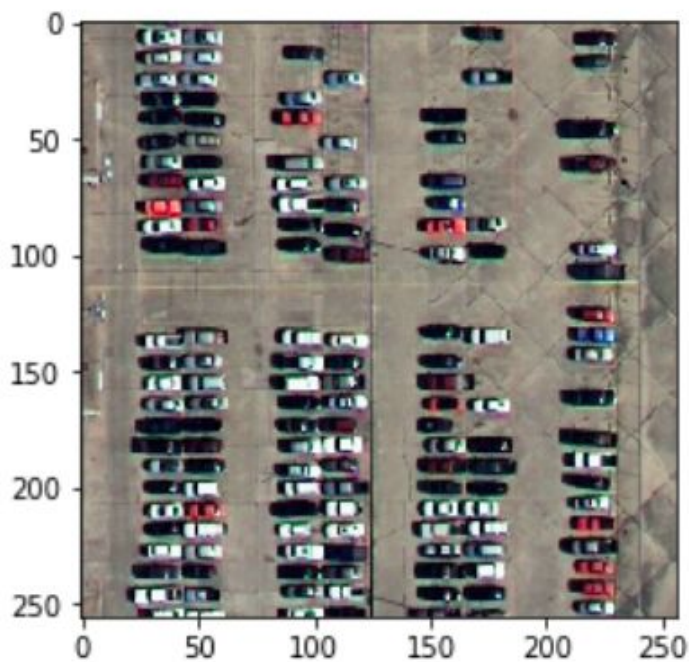
1. It is Simple Encoder - Decoder Model built using Keras Framework trained over 300 epochs.
2. Use **"UCMerced-ImgCaptioning-Training.pynb"** notebook to train the model and test samples.
3. Use **"UCMerced-ImgCaptioning-Demo/Testing.pynb"** notebook for demo purpose or just test the samples.
4. All the necessary files including the training/testing images, captions and trained weights are provided in the folder

Sample Results :

Beach

Caption: the waves slapping white sand beach over and over again

Parking Lot

Caption: many cars densely arranged in the parking lot