Image Captioning using Cross-Modal Distillation

-Harsh Shah(200050049)
-Shrey Bavishi(200050132)
under
Prof. Biplab Banerjee

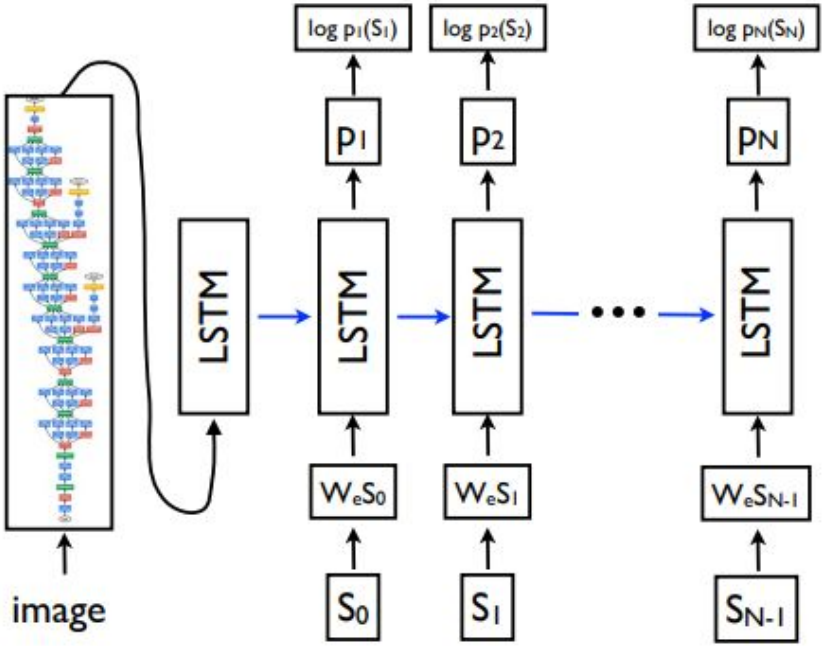
Problem Description

- Image captioning is a well studied problem much work has been done on it involving Grounded Image Captioning, Global and Local Distillation, etc.
- Encoder-Decoder models are often used for the image captioning, wherein a feature vector generated from image is passed into RNNs(often attention based LSTMs are used) to sequentially generate the captions
- The aim of this project is to give semantic meaning to the image feature by using distilling information from the text encoder to the image encoder(cross modal distillation) and then perform caption generation task

Overview

- Training text encoder-decoder model
- Training multi-label image classifier
- Cross-Modal distillation between image and text encoders
- Training model consisting of image encoder and text decoder
- Github repo: https://github.com/Harsh-Sensei/ImageCaptioning

Overview(cont.)



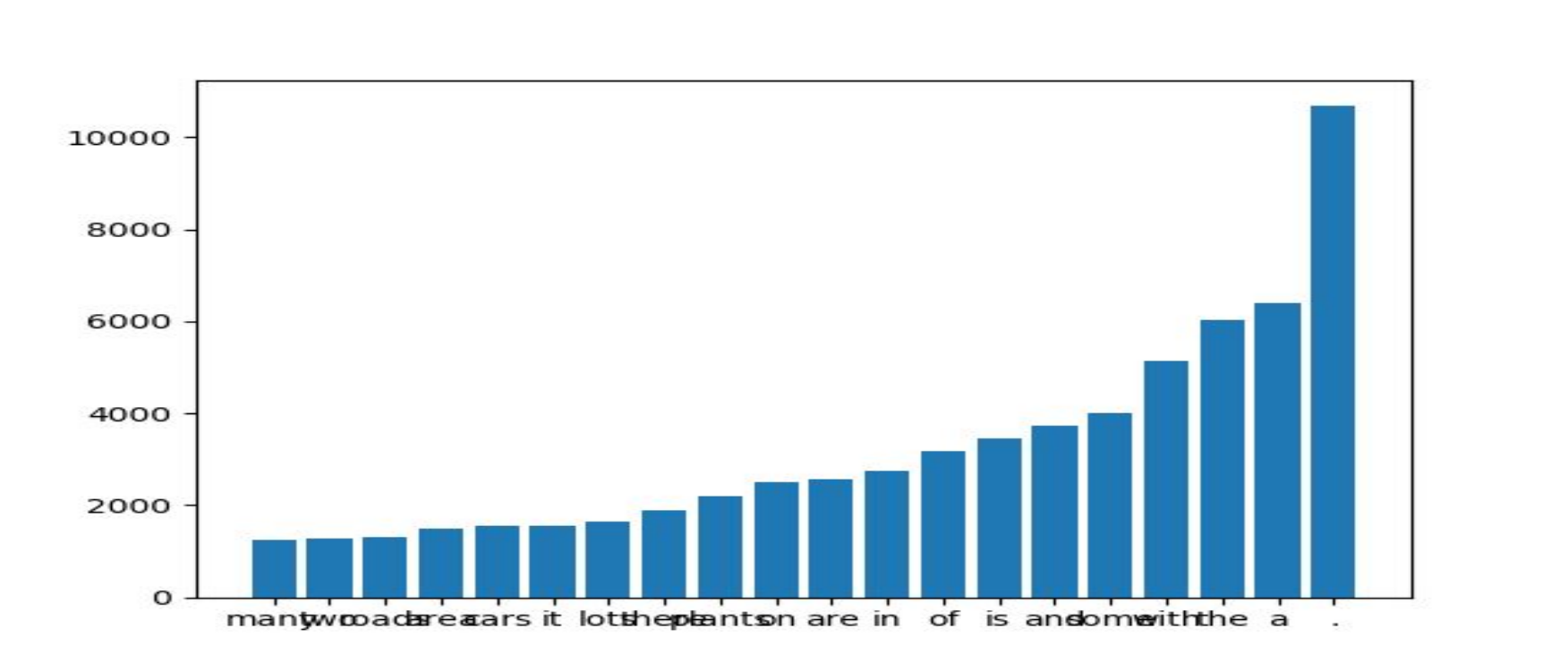
Source: “Show and Tell: A Neural Image Caption Generator” by *Vinyals, et al.*

Dataset used

- UCM captions consisting of satellite imagery mapped to 5 captions each
- Labels of images of UCM captions

- There is an airplane at the airport .
- An airplane is stopped at the airport and the ground is dark .
- A white airplane is stopped at the airport with a lot of luggage cars beside it .
- One airplane is stopped at the airport with many cars beside it .
- A white airplane is stopped at the airport and the ground is dark .

Vocabulary histogram



Text encoder-decoder model

- The text encoder-decoder model consists of 2-layered LSTM, wherein the feature vector of a given input captions is generated by the last output of the LSTM.
- 300-dimensional word embeddings were used to encode words(one-hot encoding produced similar results)
- Trained it on 80% of all the captions. Loss: Cross-Entropy Loss
- BLEU score of reconstructed sentence > 0.95 on train set

Multi-label Image Classification

- For the multi-label Image classification model, ResNet-50 model is used pretrained on ImageNet
- The classifier of the pre-trained model is changed to adapt to labels of UCM image labels(17 dimensional).
- The model is trained in two steps, first the model is trained with a low learning rate(for fine tuning) and then just the last custom classifier is trained.
- Training is done on 80% of the UCM image-labels dataset.
- Loss: Binary Cross Entropy Loss for each label as a class

Sample Output (Multi-label Image Classification)

Class	Prediction	GT
airplane	1.1790399e-05	0
bare-soil	4.9099140e-03	0
buildings	8.3452591e-04	0
cars	9.2407095e-01	1
chaparral	4.5988845e-08	0
court	1.9326635e-08	0
dock	5.5766459e-12	0
field	4.1546418e-06	0
grass	6.9034785e-01	1

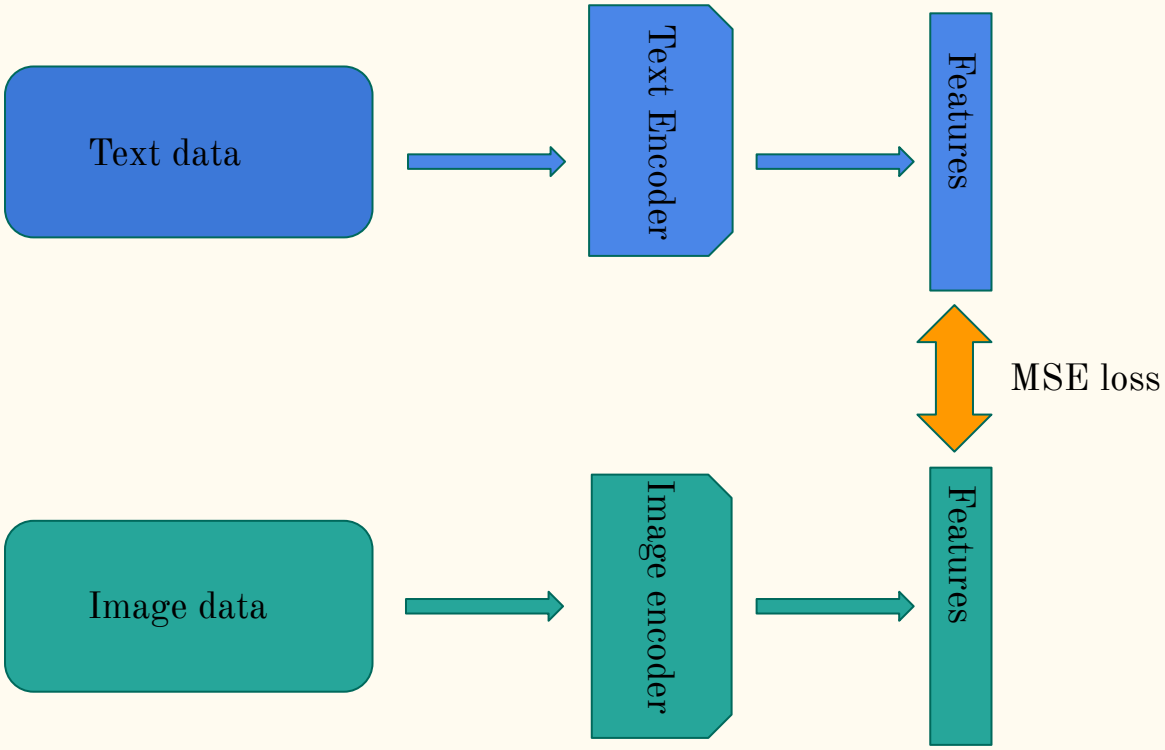
Sample Output(Contd.)

mobile-home	9.9877137e-01	1
pavement	9.9946982e-01	1
sand	2.4260514e-06	0
sea	4.6555716e-08	0
ship	1.1258932e-10	0
tanks	3.3558578e-10	0
trees	8.4277868e-01	1
water	6.5137774e-06	0

Cross Modal Knowledge Distillation

- For the purpose of distillation, the image encoder and text encoder(pre-trained) were fine-tuned again, to minimize the MSE loss between the encodings
- The procedure is carried out to close the semantic gap between image encoding and text encoding
- Experiments were performed in this step with different learning rates for image encoder and text encoder
    - Text encoder non-trainable-gave high and unstable losses while training
    - Text encoder trainable and with same learning rate as image encoder - Low loss but inferior downstream performance of captioning
    - Text encoder trainable and with 100 times lower learning rate than that of image encoder-Low loss and better performance than earlier case

Cross-Modal Knowledge Distillation(cont.)



Cross Modal Knowledge Distillation(cont.)

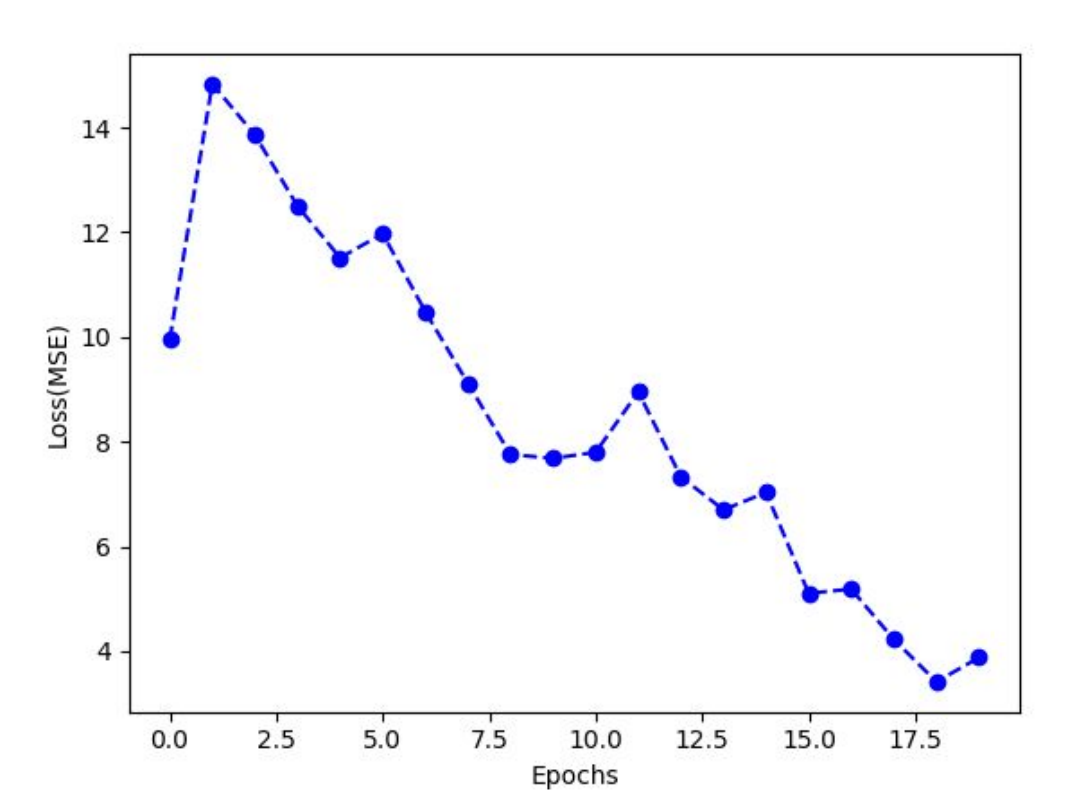
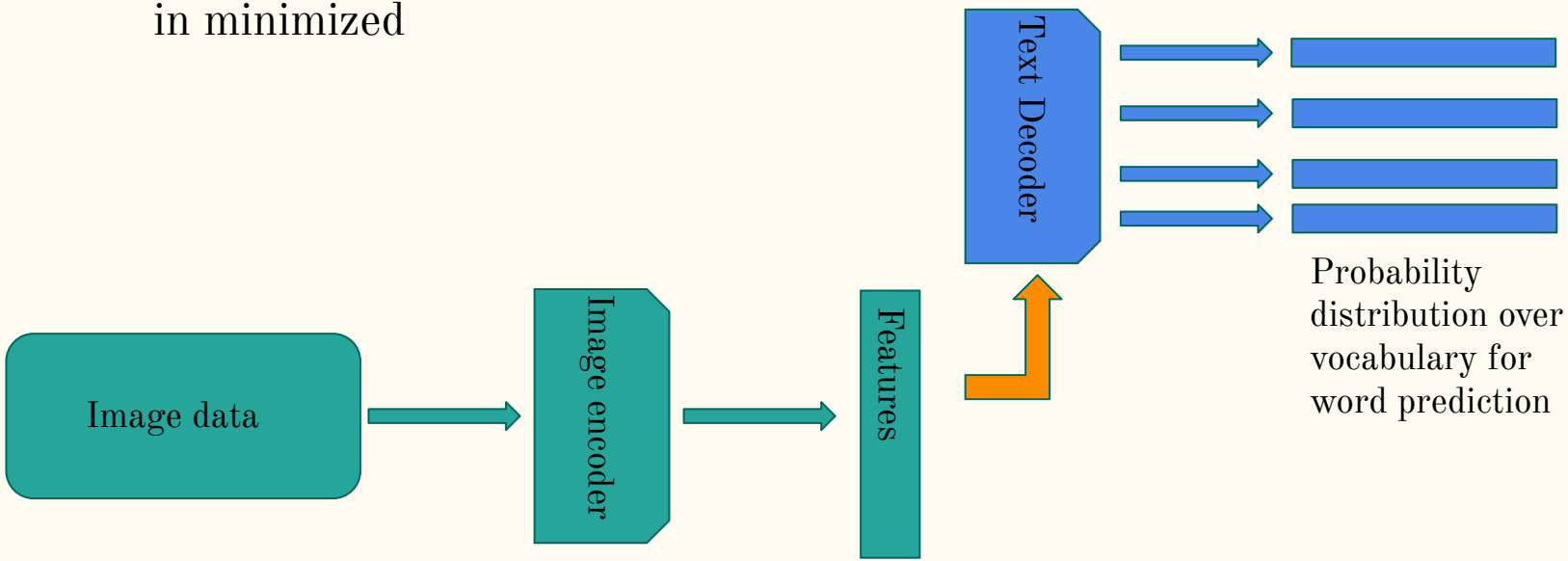


Image encoder to text decoder

- Finally, the image encoder and text decoder are combined to generate the captions
- Both the pre-trained models are fine-tuned in this step and cross entropy loss in minimized



Results

- The text encoder-decoder model gave a BLEU score of ~96.
- Multi-label Image classifier produced outputs with a F1 score of 0.94

  Recall: 0.92, Precision: 0.96

- Cross-Modal distillation: Decreasing trend of loss
- Final task of image captioning: Model predicting high frequency words very often

Results(cont.)

Notice that the model was able to detect roads in the image and hence has learnt some semantic features in the image

Ground truth:

```
['two', 'curved', 'freeways', 'with',
'some', 'plants', 'beside', 'them',
'while', 'some', 'cars', 'on', 'them',
'.']
```

Predicted:

```
['there', 'is', 'a', 'a', 'roads',
'at', 'the', 'the', 'the', '<EOS>']
```

Future Work

- We plan to incorporate attention in decoding stage so that the LSTM could get information regarding which part of the image to focus on while generating the wod at that stage. This is also known as grounded image captioning.
- In order to improve distillation various ways are proposed, one of them being usage of teacher-assistant models instead of directly distilling knowledge to the student model.

References

- “More Grounded Image Captioning by Distilling Image-Text Matching Mode” by *Zhou, et al.*
- “Show and Tell: A Neural Image Caption Generator” by *Vinyals, et al.*
- “Densely Guided Knowledge Distillation using Multiple Teacher Assistants” by *Son, et al.*

Thank You