Deep Learning CS6910, Assignment 4

November 6, 2020

1 Image Captioning

- A dictionary of image file name and corresponding list of captions is created. The entire dataset is split into 'train', 'val' and 'test'. A set of 800 images each for validation and testing and the remaining images are used for training. For each split, image name and the corresponding captions are saved in a JSON file. GloVe embeddings are used as word representation. All captions are saved in a CSV file (for pre-processing). Vocabulary is built from 100-dimensional pre-trained GloVe vectors. Finally, a dataset is created with images, corresponding tokens for captions and caption length saved as H5 data file.

- **Task 1**: Image captioning using a CNN with NetVLAD as encoder and a single hidden layer LSTM based decoder.
  Based on the previous model, only the RNN layer is replaced by the LSTM.

Module	**Layer**	**Dimension**
Encoder	`CNN (ResNet)`	`Input:(224, 224, 3)` `Output: (2048, 7, 7)`
	`NetVLAD`	`Input: (49, 2048)` `Output: (4, 2048)`

Decoder	`Embedding Layer`
	`LSTM layer`
	
	`Output Layer`
		`(100,None)`
		`Input:(100+2048*4)`
		`Hidden:1000`
		`(1000, 9973)`

Table 1: Image captioning model summary

1.1 Results

- Example caption

Figure 1: Example image

Predicted caption: a tan dog is standing in front of some plants
Reference caption: a brown and white animal surrounded by trees