Image Captioning

2021

Domain Background

The domain that I will be working on for this capstone project is image captioning, which is a multi modal task involving both image encoding and language generation. Traditionally CNN-RNN combination has been used for similar tasks. Using this project I want to implement a project based on the latest research which involves using transformers and has produced state of the art results.

I chose this particular project because it will give me exposure to working on transformers, NLP and Image recognition all together.

Problem Statement

Generating sentence based image descriptions or Captions for images as input.

Dataset and Inputs

Flickr30k dataset which is a standard benchmark database for image description based tasks will be used. The dataset can be found from kaggle[1]. The dataset is divided into 2 parts, first is the 30k image set and second is the csv file containing 3 columns, image_name which is the name of the image file, the comment id and the comment .

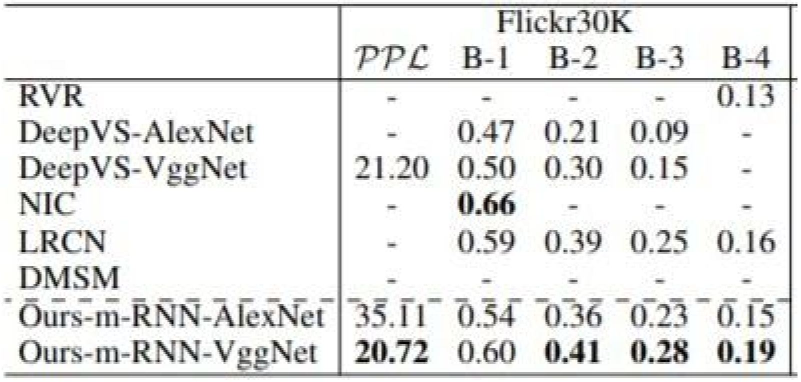
This dataset is an extension of Flickr 8K. For each image, it provides five sentences annotations. It consists of 158,915 crowd-sourced captions describing 31,783 images.

Solution Statement

Traditionally CNN's and RNN's have been used for image captioning tasks, In this solution a novel attention based transformer method will be used for the task. The work is inspired from [2]. The transformer based model incorporates two new features over all other image captioning models i) Image regions and their relationships are encoded in a multilevel fashion using memory based vectors which can encode relationships between two different parts of the image as well as the context or a priori knowledge ii) Generation of sentence utilises inputs from all layers which lets it take into account both low and high level relationships for better captions. I will explore this and other transformer based models for image caption generation.

Benchmark Model

M2 Transformer produces state of the art results on the COCO dataset, I will try replicating its results on the flickr 30k dataset. The other models that will be used for comparison of the performance will be CNN-RNN based models.



	Flickr30K				
	$\mathcal{PPL}$	B-1	B-2	B-3	B-4
RVR	-	-	-	-	0.13
DeepVS-AlexNet	-	0.47	0.21	0.09	-
DeepVS-VggNet	21.20	0.50	0.30	0.15	-
NIC	-	**0.66**	-	-	-
LRCN	-	0.59	0.39	0.25	0.16
DMSM	-	-	-	-	-
Ours-m-RNN-AlexNet	35.11	0.54	0.36	0.23	0.15
Ours-m-RNN-VggNet	**20.72**	0.60	**0.41**	**0.28**	**0.19**

[3]

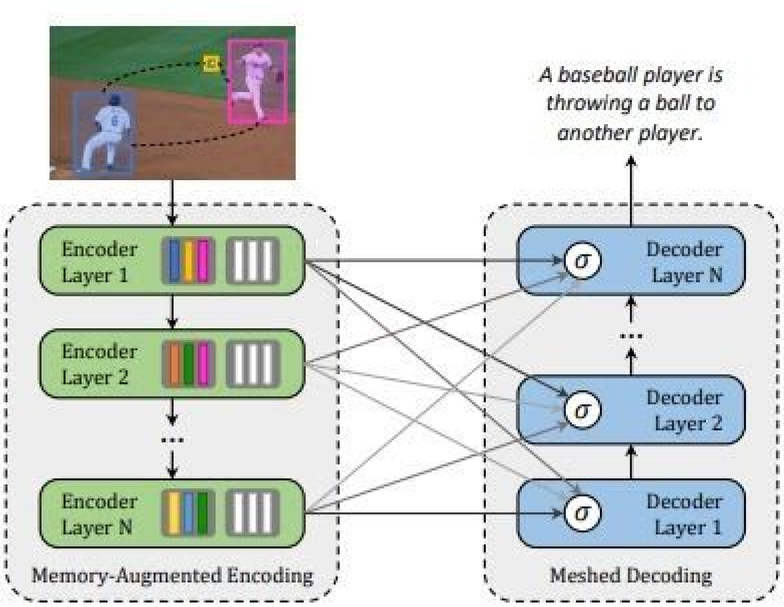
The above are the scores of some of the models on flickr 30k dataset, we will be using B1,B2,B3,B4 for comparison of our model performance, which represents the bleu scores.

Evaluation Metrics

For evaluating the model performance BLEU scores(Papineni et al. (2002)) will be used BLEU scores are standard for evaluation of machine translation based models, but can be used anywhere to compare sentence similarity.sentences might not contain all the possible descriptions in the image and BLEU might penalize some correctly generated sentences.

Project Design

Transformer based models contain an encoder part and a decoder part. The encoder part is responsible for processing regions of input image and the relationship between them, the decoder part is responsible for using the encoded output from all the encoder layers to generate captions.



The Memory part of this variation(M2 Transformer) tries to learn the context of the image, for example learning breakfast from eggs and plate to produce better captions.

The project flow will include:

i) Preprocessing the images

ii) Preprocessing the captions

iii) Split into train and test set

iv) Training the encoder-decoder model

I will explore normal encoder-decoder based models with attention and also the transformer models. The processing power can be a bottleneck for the large dataset, so training on a subset of the entire data might also be considered or simpler models,i.e. Without meshed connected layers, will be implemented, using which the output of encoder part will be cached which will ease the training effort.

References

[1] https://www.kaggle.com/adityajn105/flickr30k

[2] https://arxiv.org/abs/1912.08226v2

[3] https://arxiv.org/pdf/1412.6632v5.pdf