CS698O: Poster

Modified Stacked Attention Networks for VQA using Visual grounding of Phrases

Abhishek Verma(14026), Ayush Tulsyan(14167)

Introduction

The model we have used for visual question answering is called stacked attention networks (SANs) that learn to answer natural language questions from images. SANs use semantic representation of a question as query to search for the regions in an image that are related to the answer.We have implemented a multiple-layer SAN in which we query an image multiple times to infer the answer progressively. We also tried different models, namely LSTMs, LSTMs with attention, Seq2Seq LSTM for converting query phrase to a vector.

Stacked Attention Networks (SANs)

The overall architecture of SAN is shown in *Figure 1*(1).

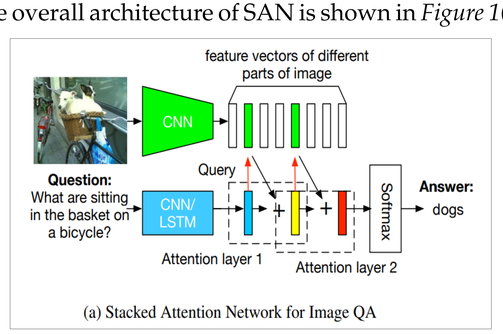

Figure 1: Stacked Attention Network

We will be describing all the components of our model in this section.

Image Model

This is the same model as mentioned in (1). The image model uses a CNN to get the representation of images. Specifically, the VGGNet (VGG19) is used to extract the image feature map from raw image. We then use a single layer perceptron to transform each feature vector to a new vector that has the same dimension as the question vector.

Question Model

We used two types of question models, one based of LSTM without attention(2) and the other based on LSTM with attention(3).

Stacked Attention Networks

Given the image feature matrix and the question feature vector, SAN predicts the answer via multi-step reasoning. In many cases, an answer is only related to a small region of an image. Therefore, using the one global image feature vector to predict the answer could lead to sub-optimal results due to the noises introduced from regions that are irrelevant to the potential answer. Instead, reasoning via multiple attention layers progressively, the SAN are able to gradually filter out noises and pinpoint the regions that are highly relevant to the answer. Mathematical details can be found in original paper(1).

Results

Following results were obtained when LSTMs without attention were used:

Methods	Accuracy	WUPS0.9	WUPS0.0
Guess	6.9	17.2	72.8
SAN(1,2)	36.4	45.3	77.1
SAN(2,2)	43.6	48.3	84.1

With attention mechanism incorporated in LSTM, the results improved significantly:

Methods	Accuracy	WUPS0.9	WUPS0.0
SAN(2,1)	47.4	45.4	85.2
SAN(2,2)	53.7	53.8	87.4

Usage of batch normalization in place of dropout led to improvement in results:

Methods	Accuracy	WUPS0.9	WUPS0.0
SAN(2,1)	49.1	47.3	86.8
SAN(2,2)	58.9	57.0	89.8

Here, *SAN(x,y)* is a notation for using x SAN layers and y attention layers in LSTM.

Observations

- Accuracy increased when attention mechanism was added in LSTM. The results significantly improved when Batch normalization was used with attention based LSTM. [*Figure 2*]

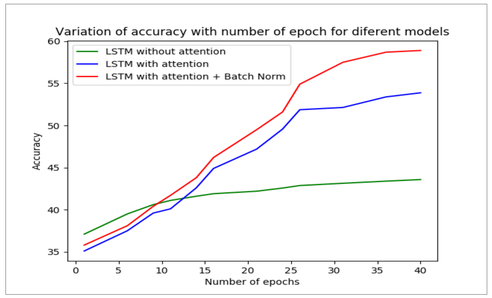

Figure 2: Accuracy for different LSTM models

- Training error decreased and accuracy increased with an increase in number of LSTM layers till a point after which it started increasing.
- Training error decreased when number of attention layers were increased from one to two. However it started increasing as the number of attention layers were increased further from two. The case with accuracy is vice versa. [*Figure 3*]

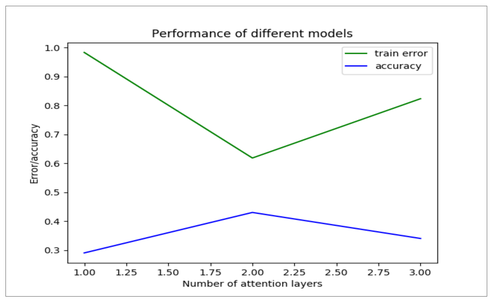

Figure 3: Accuracy for different LSTM models

- Variation of train error with number of epochs when diminishing learning rate mechanism was encapsulated. [*Figure 4*]

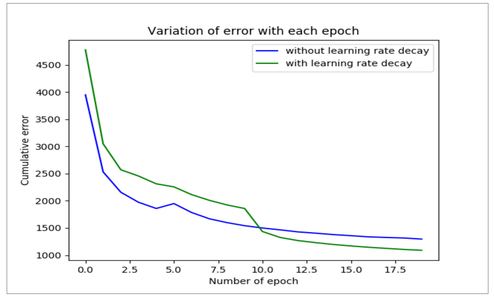

Figure 4: Error variation with diminishing learning rate

- Results of different optimization methods. [*Figure 5*]

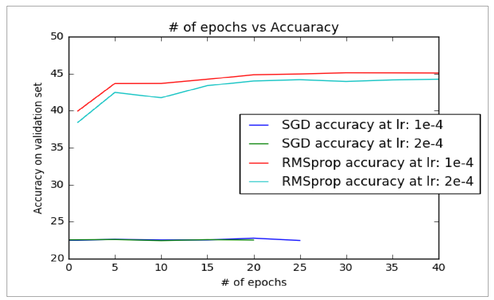

Figure 5: Effect of various optimizers on accuracy

References

[1] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng and Alex Smola. *Stacked Attention Networks for Image Question Answering*. arXiv:1511.02274

[2] Sepp Hochreiter and Jrgen Schmidhuber. *Long Short-Term Memory*. Journal of Neural Computation, 1997.

[3] Minh-Thang Luong, Hieu Pham and Christopher D. Manning. *Effective Approaches to Attention-based Neural Machine Translation*.

Advisors: Vinay P. Namboodiri , Badri Narayana Patro , Utkarsh Chauhan