CS698A Final Project Report

Stacked Attention Networks for Image Questioning ANswering

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola

Prakhar K, Preetansh Goyal, R.N.Viswanadh

IIT KANPUR
India

8-11-2016

Outline

1. Image Question Answering: Introduction
    - Introduction: Image Question Answering
2. SOA Paper: Stacked Attention Networks
    - Introduction
    - Image Model
    - Question Model
        - CNN based question model
    - Stacked Attention Networks
3. Model 1
4. Model 2
5. Model 3
6. Model 4
7. Results
8. Contributions

Introduction: Image Question Answering

An Image QA system takes an input image and a natural language question pertaining to the image and produces an answer as the output.

Q: What type of animal is this?
Q: Is this animal alone?

Q: Is it snowing?
Q: Is this picture taken during the day?

Q: What kind of oranges are these?
Q: Is the fruit sliced?

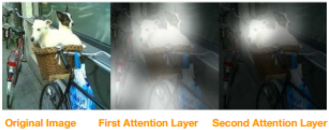
Q: What is leaning on the wall?
Q: How many boards are there?

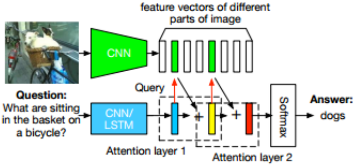
Figure: Sample images and questions in VQA dataset

In our presentation,we are interested in approaches with single word answer outputs.

Stacked Attention Networks

- proposed method that allows for multi-step reasoning for image QA




feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/
LSTM
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

- SAN consists of 3 major components
  1. image model
  2. question model
  3. stacked attention networks

Image Model

A CNN, VGGNet is used by the image model to extract the image feature map f_I from a raw image I, VGGNet is used:

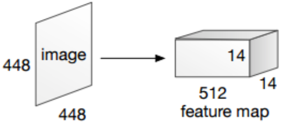


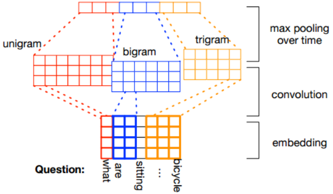
Figure: CNN based image model

$$f_I = CNN_{vgg}(I) \tag{1}$$

use a single layer perceptron to transform each feature vector to a new vector that has the dimension (dxm), d being the dimension of question vector, m being no. of regions.

$$v_I = tanh(W_I f_I + b_I) \in R^{dm} \tag{2}$$

CNN based question model



First embed words to vectors $x_t = W_e q_t$ and get the question vector by concatenating the word vectors:

$$x_{1:T} = [x_1, x_2, ..., x_T] \tag{3}$$

$$h = [\tilde{h}_1, \tilde{h}_2, \tilde{h}_3] \tag{4}$$

where $\tilde{h}_i$ is the output CNN model with i-gram convolution filter. Hence, $v_Q = h \in R^d$ is the CNN based question vector.

Stacked Attention Networks

For attention regions in the image, v_I and v_Q are fed into single layer neural network and softmax function

$$h_A = tanh(W_{I,A}v_I \oplus (W_{Q,A}v_Q + b_A)) \tag{5}$$

$$p_I = softmax(W_P h_A + b_P) \in R^m \tag{6}$$

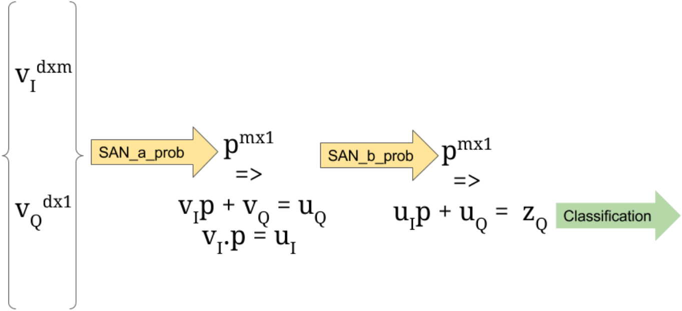
where $v_I \in R^{d \times m}$, $\{W_{I,A}, W_{Q,A}\} \in R^{k \times D}$, $W_P \in R^{1 \times k}$

$$s_I = \sum_i p_i v_i \in R^d, i = 1, 2..m \tag{7}$$

$$u = s_I + v_Q \tag{8}$$

$u \in R^d$ is the modified query vector. This process is repeated K times via K SAN layers to get u^K, which is then used for final classification.

Model 1



v_I^{dxm}

v_Q^{dx1}

SAN_a_prob

p^{mx1} =>

$v_I p + v_Q = u_Q$

$v_I.p = u_I$

SAN_b_prob

p^{mx1} =>

$u_I p + u_Q = z_Q$

Classification

Figure: SAN_a obtains an m dimensional probability vector p^m to modify query vector v_Q to u_Q as given in original paper, and then $i^t h$ region of v_I is multiplied by $i^t h$ element of p^m to obtain modified image matrix u_I. SAN_b is normal SAN layer on u_I and u_Q to get z_Q

Model 2

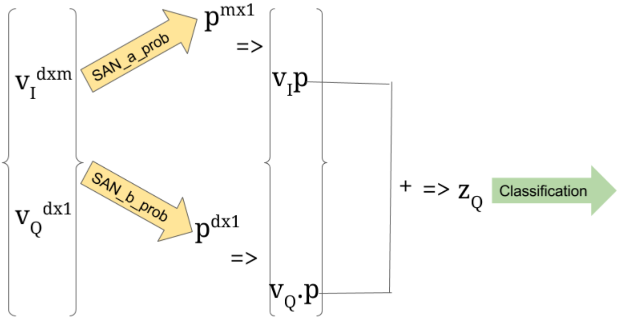


Figure: SAN_a obtains p^m to simply obtain modified query $v_I p$. SAN_b obtains probability vector p^d that modifies query to get $v_Q p$. Both these are added to get z_Q

Model 3

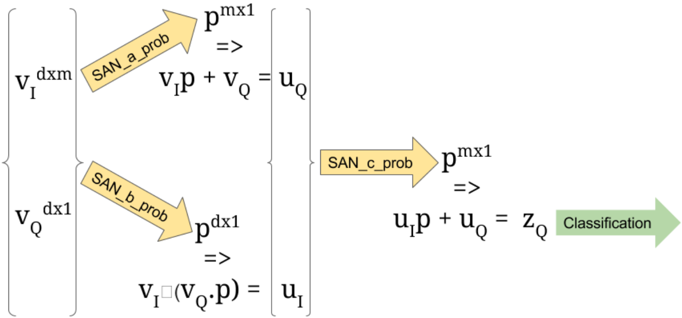


Figure: SAN_a obtains SAN_a obtains p^m to simply obtain modified query u_Q. SAN_b obtains p^d to modify image matrix v_I by adding $p.q$ to each row, to get u_I. Finally, u_I and u_Q are passed through normal SAN layer to get z_Q.

Model 4

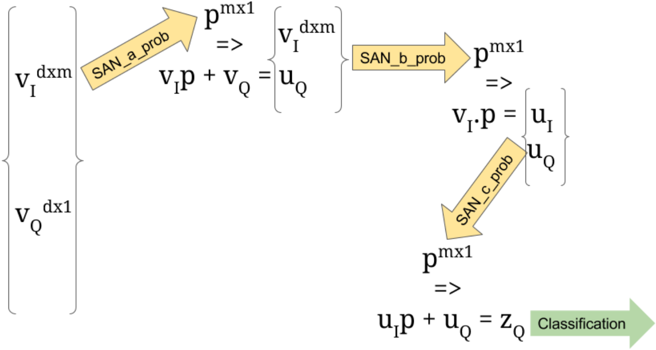


Figure: It is just an alternate version of Model 2 where SAN_a is used to get u_Q, then u_Q and v_I are passed in SAN_b to finally get u_I, and then we have normal SAN_c over u_I and u_Q to get z_Q

Results

Comparison of the results we obtained:

Method	Accuracy
Model Accuracy (SAN)	52.255% (50 epochs)
Model Accuracy (Model 1)	52.213%(50 epochs)
Model Accuracy (Model 2)	35.5% (20 epochs)
Model Accuracy (Model 3)	47.6% (4.12 epochs)
Model Accuracy (Model 4)	**52.424%** (50 epochs)

Summary and Contributions

1. We have designed and successfully implemented four different co-attention models over the existing SAN model.
2. Per person contribution:
    - Prakhar: Model Designing of model 3 and Implementation on GPU.
    - Preetansh: Model Designing of model 2,4 and Implementation on GPU.
    - Viswanadh: Model Designing of model 1, GPU configuration, Implementation on GPU.

References

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola, *Stacked Attention Networks for Image Question Answering*. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June, 2016.

End of Presentation

Thank you.