Visual Question Answering

Stefanie Anna Baby (stef96), Ashwini Pokle (ashwini1)
Department of Computer Science, Stanford University

CS 224n
Winter 2018

Motivation

- Appealing intersection of NLP and Computer Vision – a step towards general Artificial Intelligence
- Requires semantic understanding of images – tough!
- Interesting real world applications - helping visually impaired, improving image search

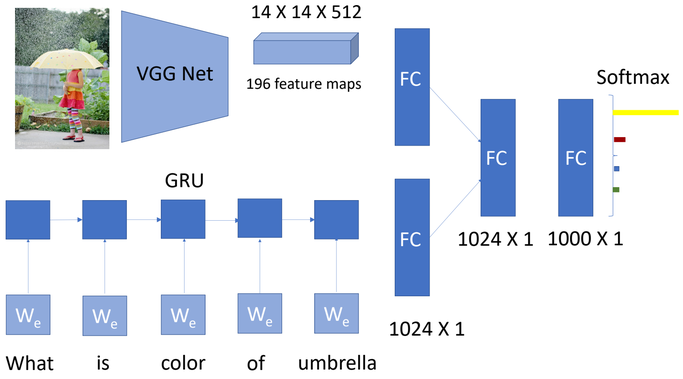
Approaches

- **Baseline** : CNN – GRU
- Stacked Attention Networks, Dynamic Memory Networks

Dataset

- VQA v2 dataset - Open Ended Questions
- 443,757 training questions, 82,783 training images
- 214,354 validation questions, 40,504 validation image (no test)
- Balanced dataset - minimizes influence of language priors
- Evaluation $\min\left(1, \frac{\#\text{ humans that provided that answer}}{3}\right)$

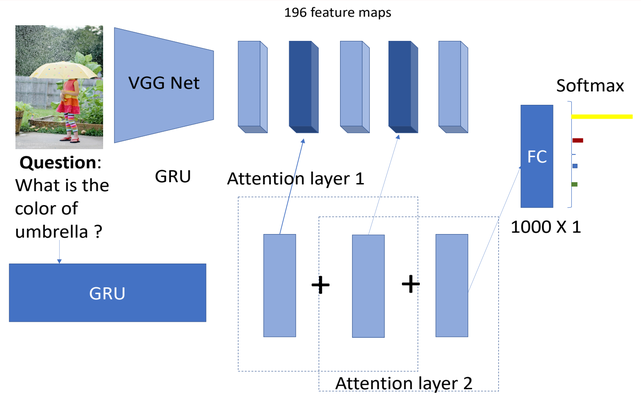
Baseline - CNN - GRU



$$v_I = \tanh(W_I f_I + b_I) \qquad v'_Q = tanh(W v_Q) + b$$
$$x_t = W_e q_t \qquad v_c = v_Q * v'_c$$
$$v_Q = GRU(x_t)$$

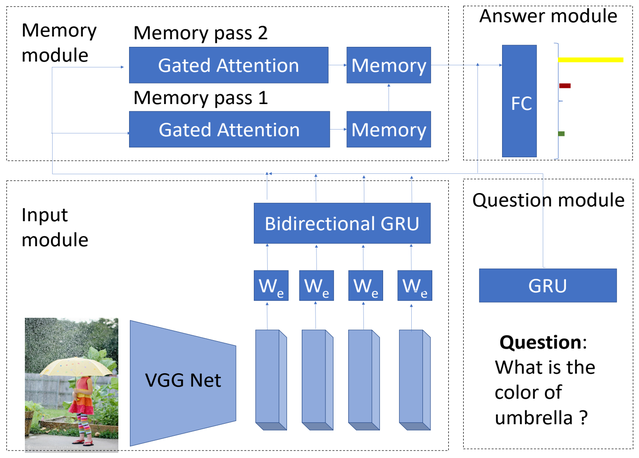
- Image features are 3D volumes extracted from the last max pooling layer of VGG Net - 19
- Embeddings are uniformly initialized between [-0.08, 0.08]
- Adam optimizer with learning rate 1e-4 and Cross entropy loss
- Softmax over 1000 answers as data is preprocessed to include only top 1000 most frequent answers
- Epochs : 10, Batch size : 100, Embedding size : 512

Stacked Attention Network



$$v_I = \tanh(W_I f_I + b_I) \qquad h_A = \tanh(W_{I,A} v_I \oplus (W_{Q,A} v_Q + b_A)),$$
$$x_t = W_e q_t \qquad p_I = \text{softmax}(W_P h_A + b_P)$$
$$v_Q = GRU(x_t) \qquad \tilde{v}_I = \sum_i p_i v_i \qquad u = \tilde{v}_I + v_Q$$

Dynamic Memory Network



- **Input module** $\overleftrightarrow{f_i} = GRU_{fwd}(\overrightarrow{f_{i-1}}, f_i) + GRU_{bwd}(\overleftarrow{f_{i+1}}, f_i)$
- **Question module** $q_t = GRU(q_t, q_{t-1})$
- **Episodic memory module** uses attention GRU to focus on relevant spatial regions; attention is computed through interaction between feature maps, question and previous memory state

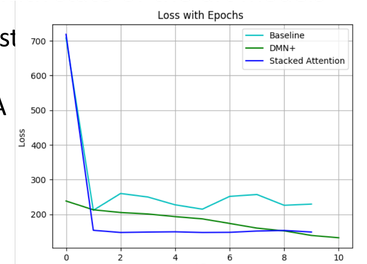
$$z_i^t = [\overleftrightarrow{f_i} \circ q; \overleftrightarrow{f_i} \circ m^{t-1}; |\overleftrightarrow{f_i} - q|; |\overleftrightarrow{f_i} - m^{t-1}|]$$
$$g_i^t = softmax(W^{(2)} tanh(W^{(1)} z_i^t + b^{(1)}) + b^{(2)})$$

- **Answer module** $y = softmax(W^{(a)} a), a = [q; m_T]$

Results

Architecture	Test Accuracy
Baseline – CNN-GRU (VQA v1)	50.21 %
Baseline – CNN-GRU (VQA v2)	40.06 %
Stacked Attention Network (VQA v2)	47.11 %
Dynamic Memory Network *(trained on 100K VQA v2)	21.65 %
Hierarchical Co-attention Network **(reported in [1])	51.88 %
Multimodal Compact Bilinear Pooling **(reported in [1])	56.08 %

Table 1: Performance of our implementation with state-of-the-art models



- Stacked Attention Network achieves highest test/validation set accuracy of 47.11%
- DMN was trained on a small subset of VQA dataset of size 100,000 and it achieves accuracy of 21.65% on validation set (first 80K samples)

Examples of Predictions made by SAN

How many cats are present in the image? A: 1
What type of cat is this ? A: Tabby

How many giraffe are pictured? A: 2
What is in front of giraffe? A: fence

Question type / format	Accuracy	Question Type / format	Accuracy
Yes / No	63.29 %	Count	33.82 %
What sport is	82.68 %	What is the name	7.65 %
Is this	63.20 %	Why	11.24 %
Has	64.37 %	What time	21.28 %
Was	65.33 %	How	20.19 %
What room is	82.03 %	Which	35.80 %

Table 2 : Question-type accuracy for Stacked Attention Networks

References

[1] Goyal, Yash, et al. “Making the V in VQA matter: Elevating the role of image understanding in. Visual Question Answering.” CVPR 2017
[2] Yang, Zichao, et al. "Stacked attention networks for image question answering." *CVPR* 2016
[3] Xiong, Caiming, Stephen Merity, and Richard Socher. "Dynamic memory networks for visual and textual question answering." ICML 2016