Visual Question Answering



Gera Versfeld

The VQA-v2 Challenge - June 25

Rank	Participant Team	yes/no	number	other	overall
1	anon_team	76.52	39.29	57.31	63.31
2	vqateam_mcb_benchmark	78.82	38.28	53.36	62.27
3	vqateam_deeper_LSTM_Q_norm_I	73.46	35.18	41.83	54.22
4	vqateam_language_only	67.01	31.55	27.37	44.26
5	vqateam_prior	61.20	0.36	1.17	25.98

The VQA-v2 Dataset

- Training: 4,437,570 answers, 443,757 questions, 82,783 images
- Validation: 2,143,540 answers, 214,354 questions, 40,504 images
- Test: 447,793 questions, 81,434 images

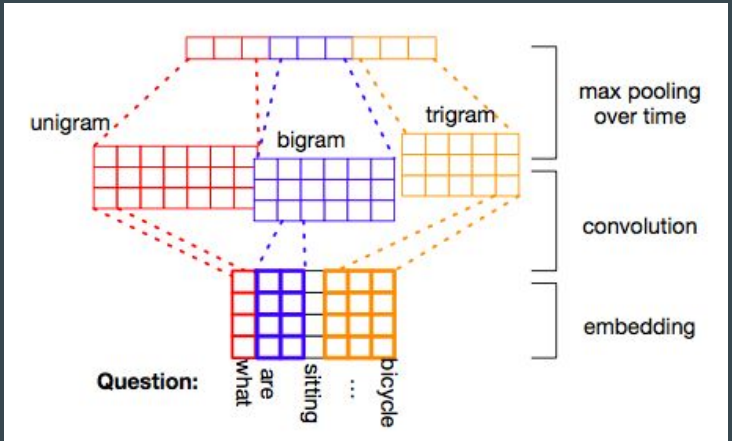
- MSCOCO images
- Balanced - answer for each question depends on image
- Top question types:
  - What - 41.2%
  - Is - 26.1%
  - How many - 10.3%
  - Are - 7.57%

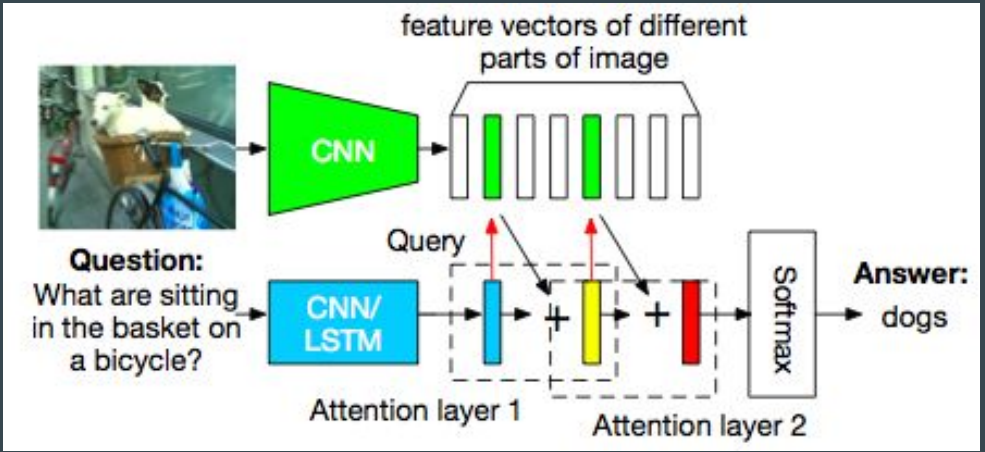
http://visualqa.org/visualize/

Stacked Attention Networks $(SAN)^2$

- Multiple attention layers allow SAN to gradually filter out noise
- Attention mechanism:

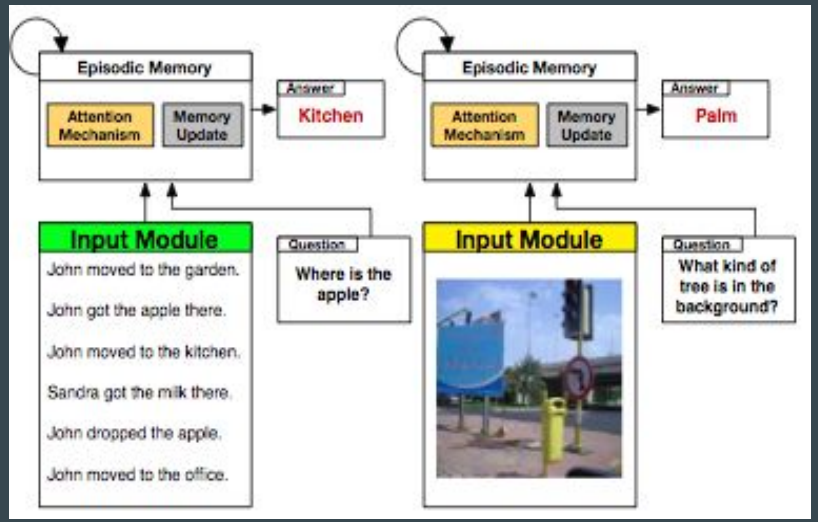
$$h_a = \tanh(W_{I,A} v_i \oplus (W_{Q,A} v_Q + b_A))$$
$$P_i = \text{softmax}(W_P h_a + b_P)$$



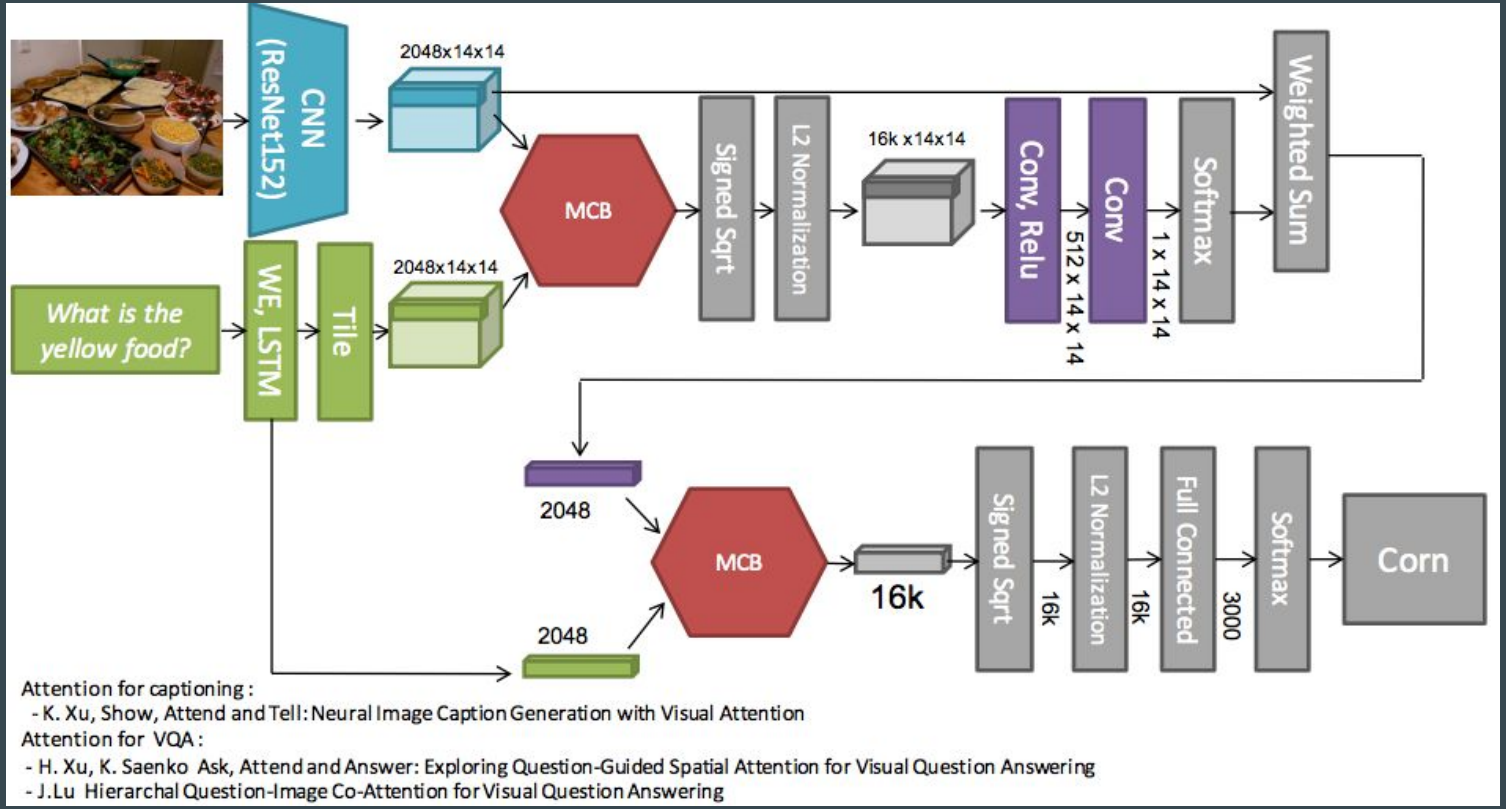


Dynamic Memory Networks for VQA[3]

- Input Module
  - Local feature extraction
  - Visual feature embedding
  - Input fusion layer
- Question Module
- Episodic Memory Module
  - Attention mechanism
  - Memory update mechanism
- Answer Module



The baseline model: Multimodal Compact Bilinear Pooling

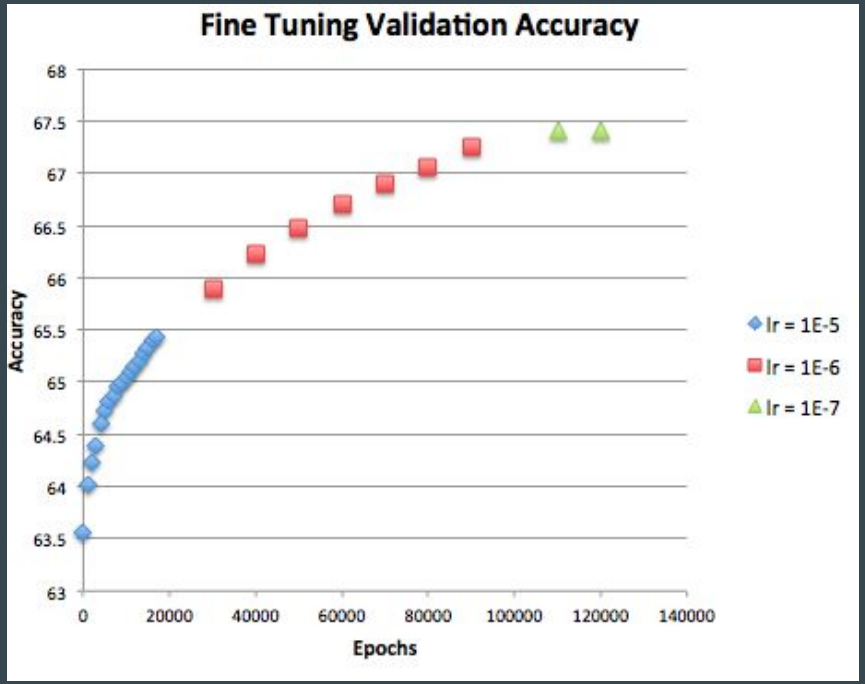


Attention for captioning :
- K. Xu, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

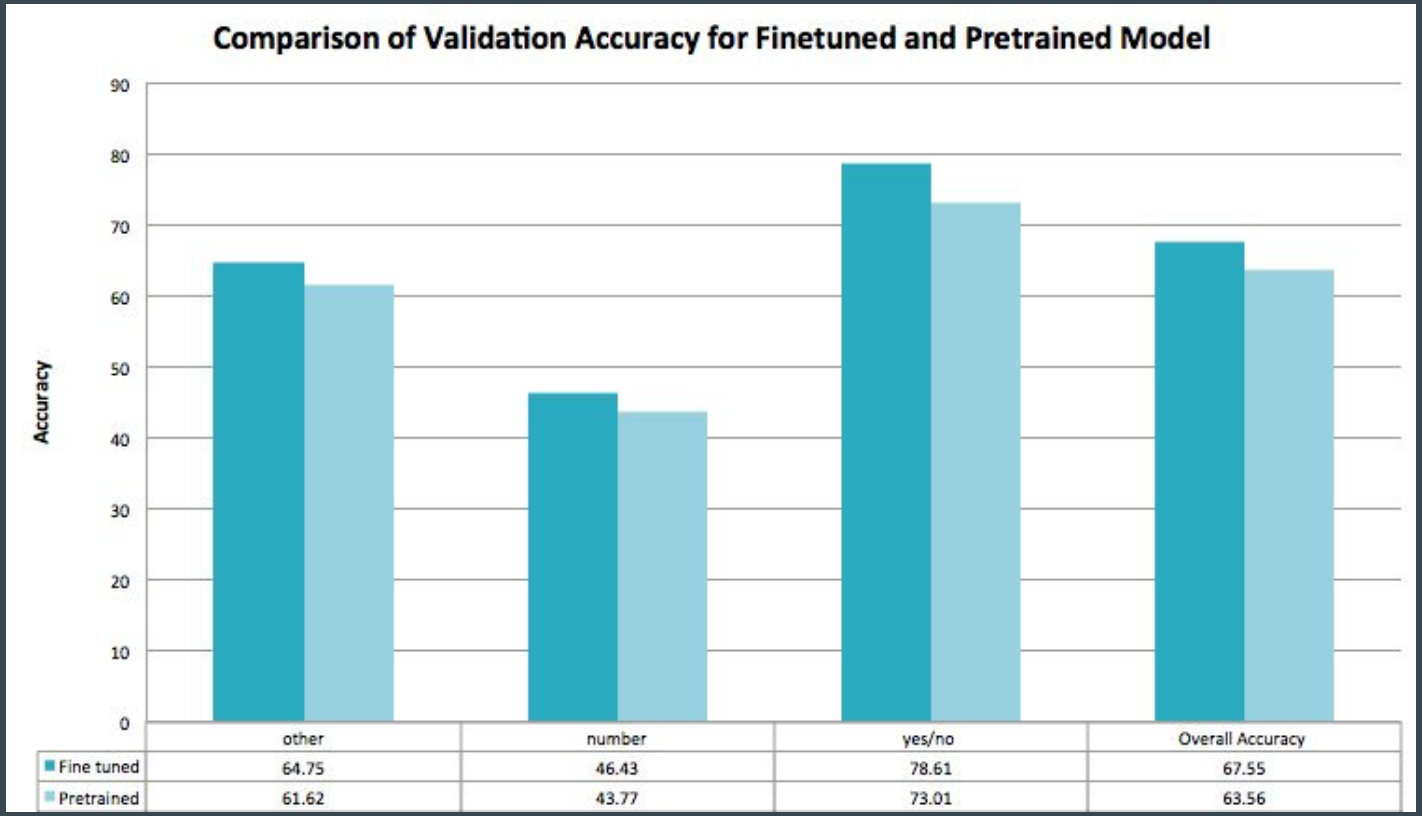
Attention for VQA :
- H. Xu, K. Saenko Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering
- J.Lu Hierarchal Question-Image Co-Attention for Visual Question Answering

Fine-tuning:


Training loss
Loss
1.9
1.85
1.8
1.75
1.7
1.65
1.6
1.55
0
20000
40000
60000
80000
100000
120000
140000
Epochs
lr = 1 E-5
lr = 1E-6
lr = 1E-7


Fine Tuning Validation Accuracy
Accuracy
68
67.5
67
66.5
66
65.5
65
64.5
64
63.5
63
0
20000
40000
60000
80000
100000
120000
140000
Epochs
lr = 1E-5
lr = 1E-6
lr = 1E-7

Fine-tuned Model on Validation Set



	other	number	yes/no	Overall Accuracy
Fine tuned	64.75	46.43	78.61	67.55
Pretrained	61.62	43.77	73.01	63.56

Next steps:

- Implement attention mechanism for question based on part of speech
- Implement regression based answer generator
- Begin training
- Continue fine-tuning baseline model

References

1. Akira Fukui, Doing Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach. Multimodal Compact Bilinear Pooling. 2016.
2. Zichao Yang, Xiaodong He, Jianfeng Gong, Li Deng, Alex Smola. Stacked Attention Networks for Image Question Answering. 2016.
3. Caiming Xiong, Stephen Merity, Richard Socher. Dynamic Memory Networks for Visual Textual Question Answering.