Visual Question Answering

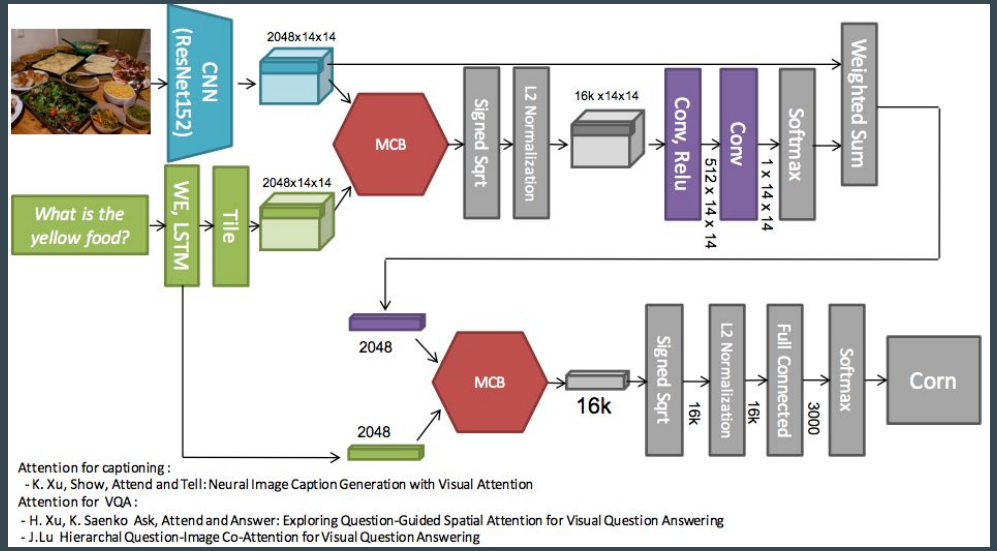


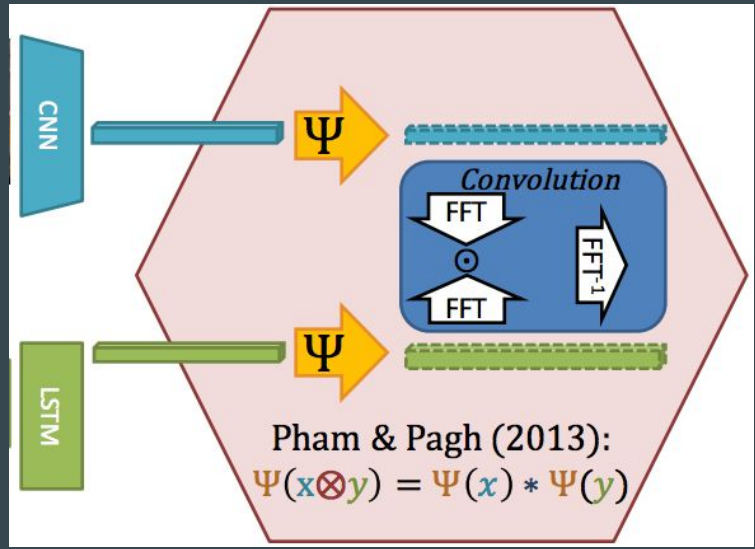
Gera Versfeld • 06.05.2017

Progress this week

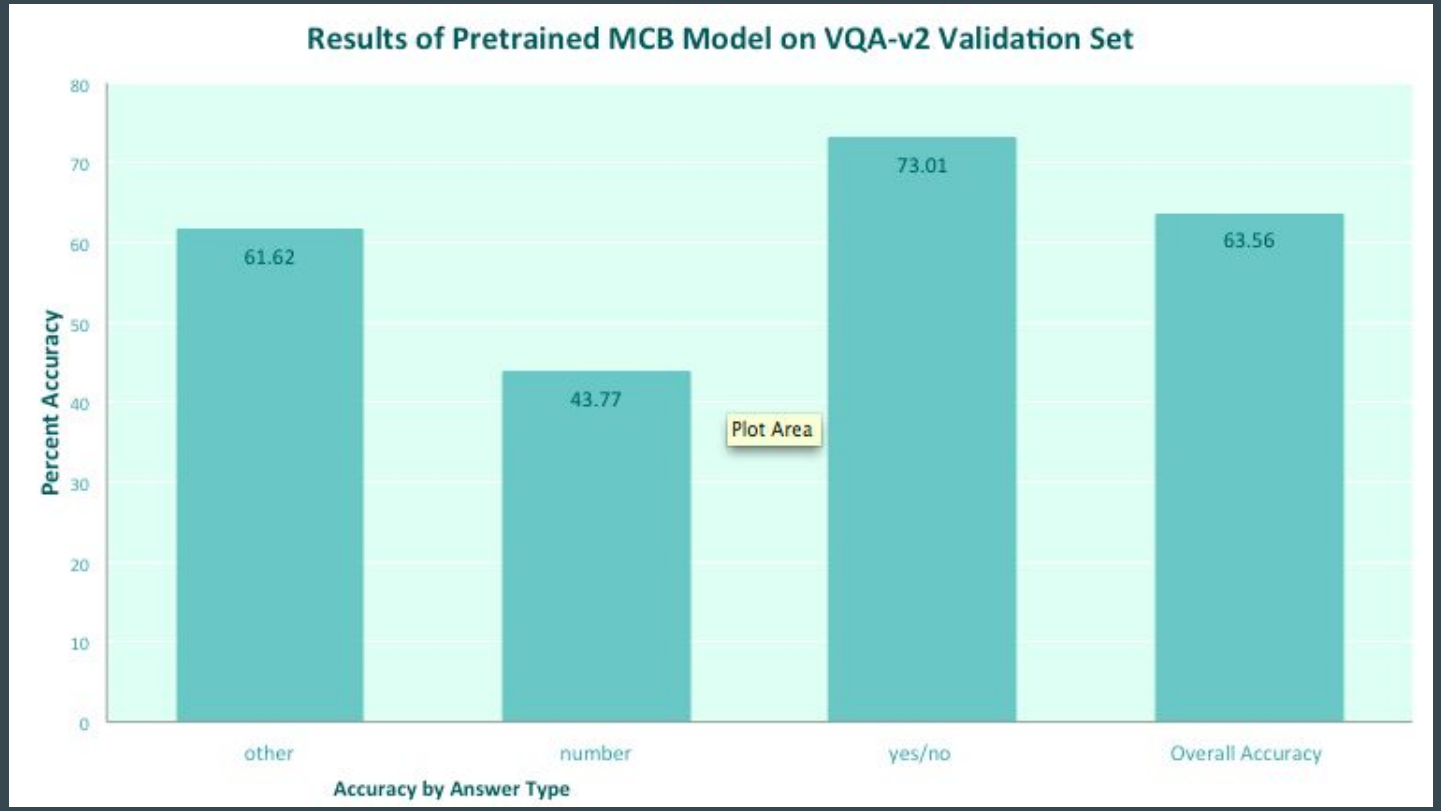
- Get cluster access and install necessary software
- Extract image features for new baseline model
- Evaluate validation data split on pre-trained baseline model
- Learn about different VQA architectures and attention mechanisms
- Begin developing ideas for potential new architectures

The Baseline Model: MCB - 2016 VQA-v1 winner[1]





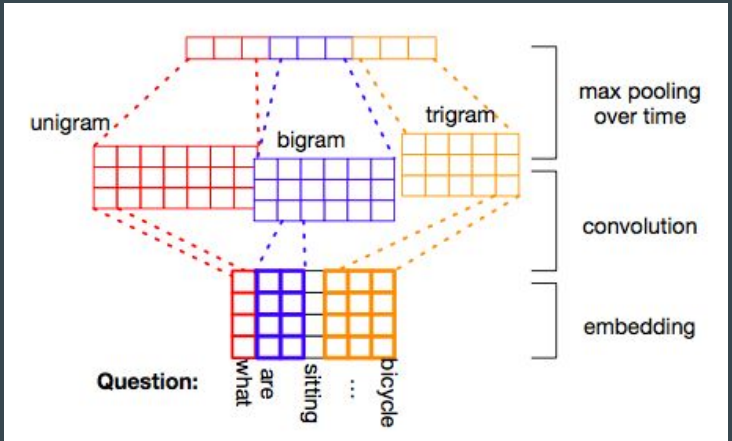
Results for MCB Baseline Model

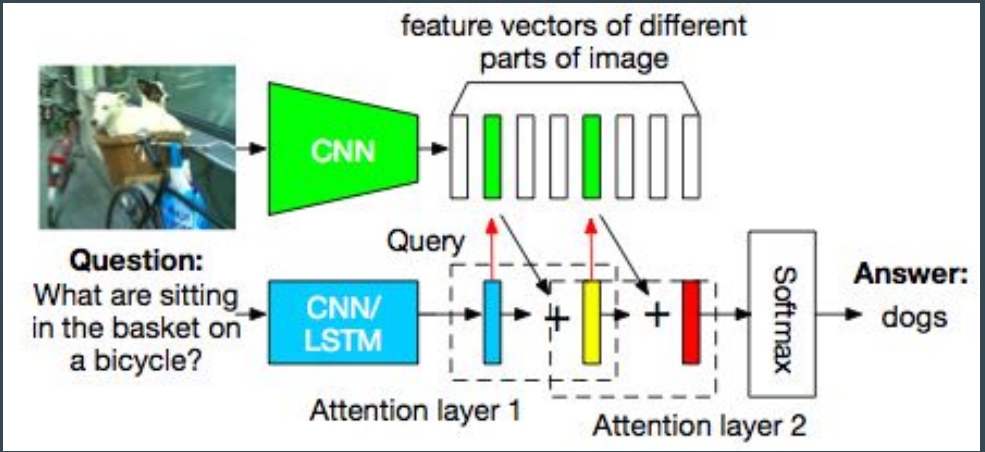


Stacked Attention Networks $(SAN)^2$

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

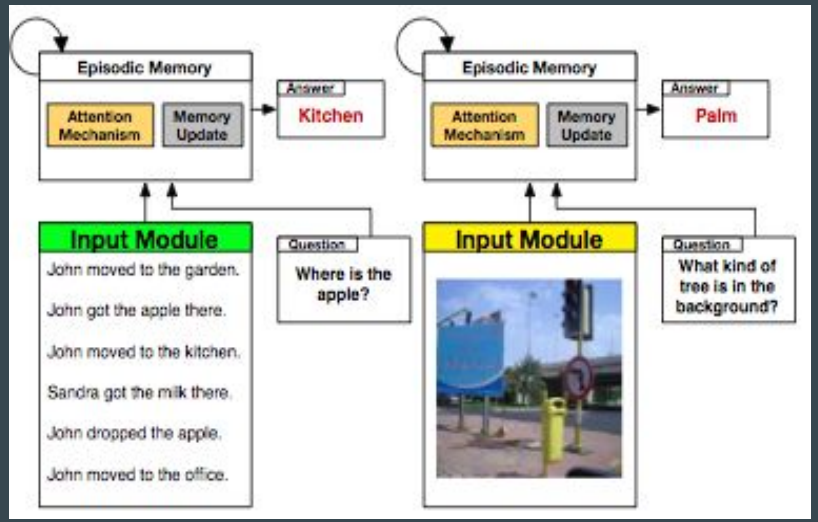
$$h_a = \tanh(W_{I,A} v_i \oplus (W_{Q,A} v_Q + b_A))$$
$$P_i = \text{softmax}(W_P h_a + b_P)$$





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



New Model Ideas:

- Categorical attention mechanism for question embedding (finetune pretrained POS classifier)
- New attention mechanism for MCB
- Stacked attention network with bilinear pooling instead of elementwise product in attention mechanism
- Ensemble of different models
- Question types with worst accuracies: ‘why’, ‘what number is’, ‘what does the’
- Treat answer generation as regression problem not classification?

Where we are at now:

- Have a good working baseline
- Ideas for direction of new model
- Comfortable using cluster and manipulating VQA dataset

Goals for next week

- Finalize new model plan
- Begin implementing new model
- Begin training 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.