Semester project presentation

Extending Dynamic Structure in Memory Network for Response Generation

- Supervisor: Mi Fei
- Director: Boi Faltings
- Myself: André Cibils

Plan:

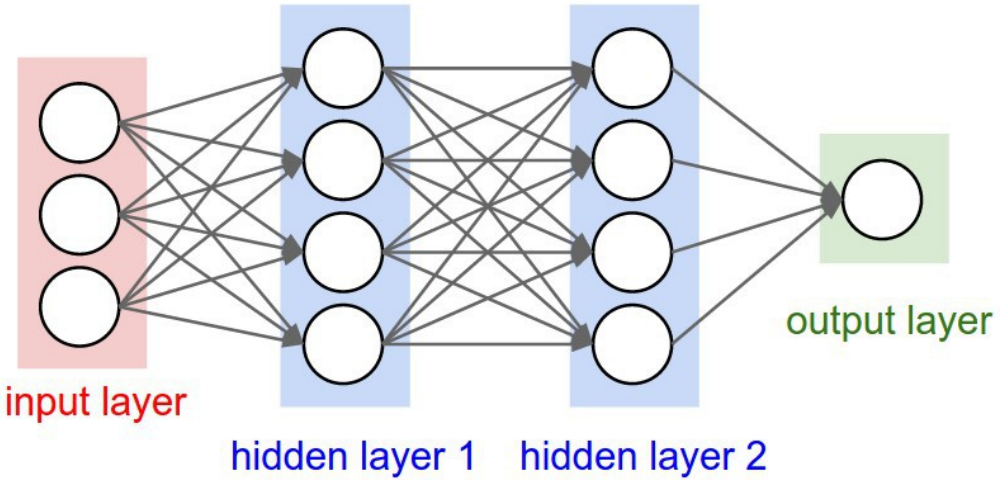
- 1. Introduction
- 2. RNN: LSTM, GRU and MemNN
- 3. DMN
- 4. DMN+
- 5. DMN&DMN results
- 6. Futur work & conclusion

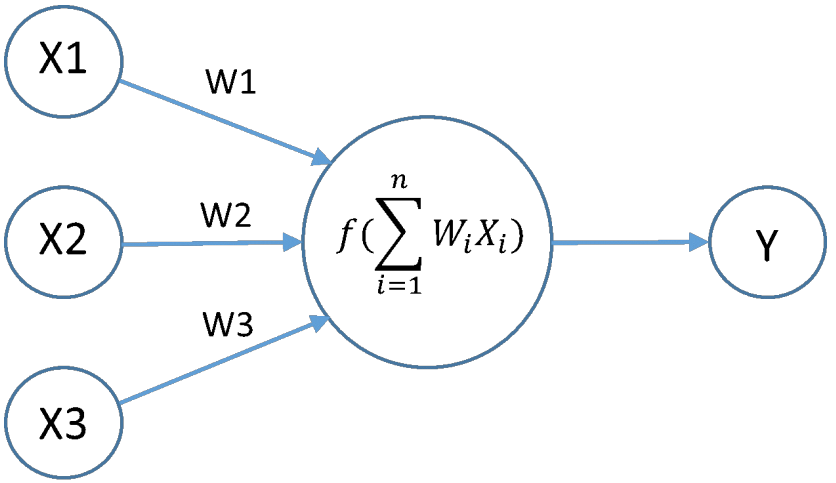
Quick Presentation

- Bachelor here at EPFL in informatics
- This is my 2nd semester as a master student in computer science
- Big interest in ML, and more precisely NN
- Goal of the project is to answer (rightfully!) with a sentence to a question about a text

Speaking of NN

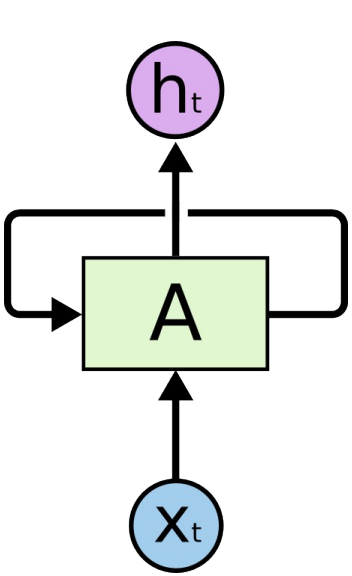
- You all know what NN are

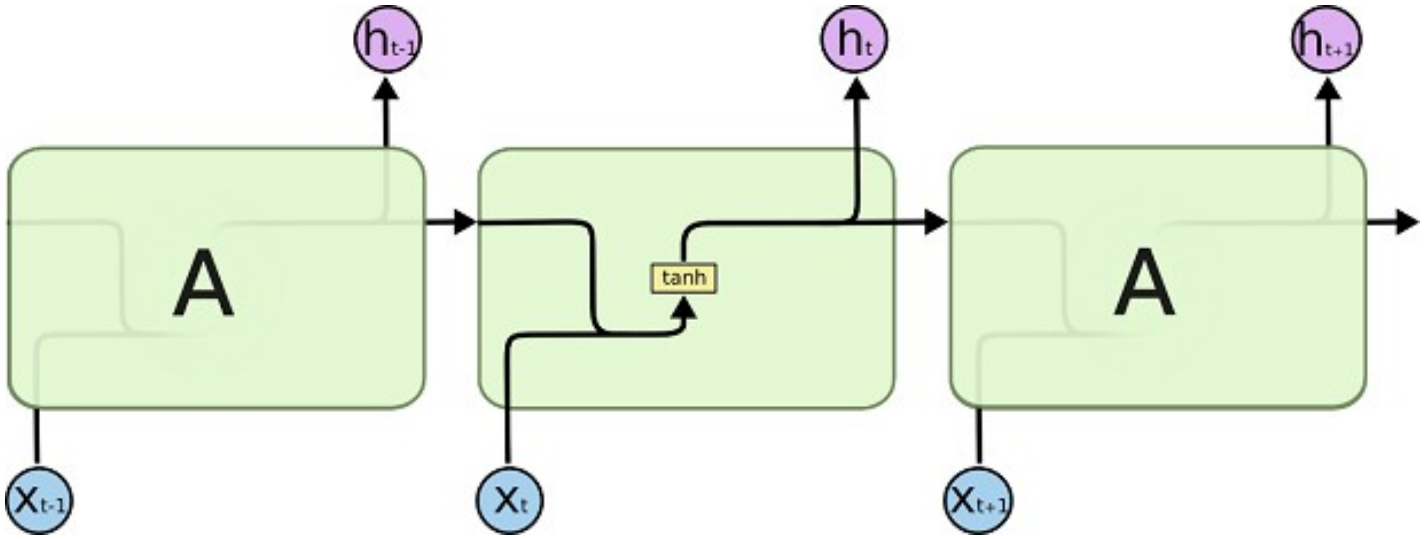




Recurrent Neural Networks (RNN)

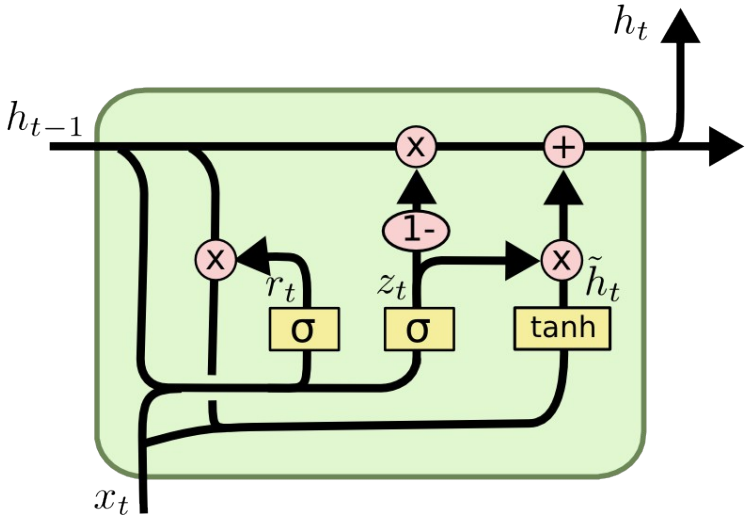
- A specific and powerful type of NN
- Main idea: the cells of the NN keep trace of theirs states and use it to do more complex computation





Long and Short Term Memory (LSTM) & Gated Recurrent Unit (GRU)

- LSTM & GRU are some specific RNN architecture
- Compute their states using different variables



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

- Zt is called the update gate
- Rt is called the reset gate
- ~ht is the potential state
- ht is the updated state

Memory Network (MemNN)

- Really recent! ~2015 [1]
- Base idea is to have a memory *m* (an array of objects) and some components interacting with it, for example:
  - Input feature map
  - Generalization
  - Output feature map
  - Response

Dynamic Memory Network (DMN & DMN+)

- An enhanced version of MemNN
- Even more recent! ~March 2016 [2]
- This is what I'm doing & working with
- Four parts:
  - Input Module
  - Question Module
  - Episodic Memory Module
  - Answer Module

Reminder: goal of the project

- Given some facts, answer a question with a sentence.
- State of the art:

I: Jane went to the hallway.
I: Mary walked to the bathroom.
I: Sandra went to the garden.
I: Daniel went back to the garden.
I: Sandra took the milk there.
Q: Where is the milk?
A: garden

- I'll be using the bAbi set
- Later, I'll try training my network with my own generated dataset or other datasets found online

Input Module

- In NLP, the input is a sequence of words (w_1, ... w_{Ti})
- At each timestep *t,* update the hidden state

  $h_t = RNN(L[w_t], h_{t-1})$

  L is the embedding matrix, w_t is the word index of the t^{th} word of the input sequence
- Uses a GRU as RNN
- Output a sequence of T_C facts representations

Question Module

- Similar to input module, input is a sequence of words (= the question)
- Similar to input module, this module encodes the question using $q_t = GRU(L[w_t], q_{t-1})$
- Output a representation of the question q

Episodic Memory Module

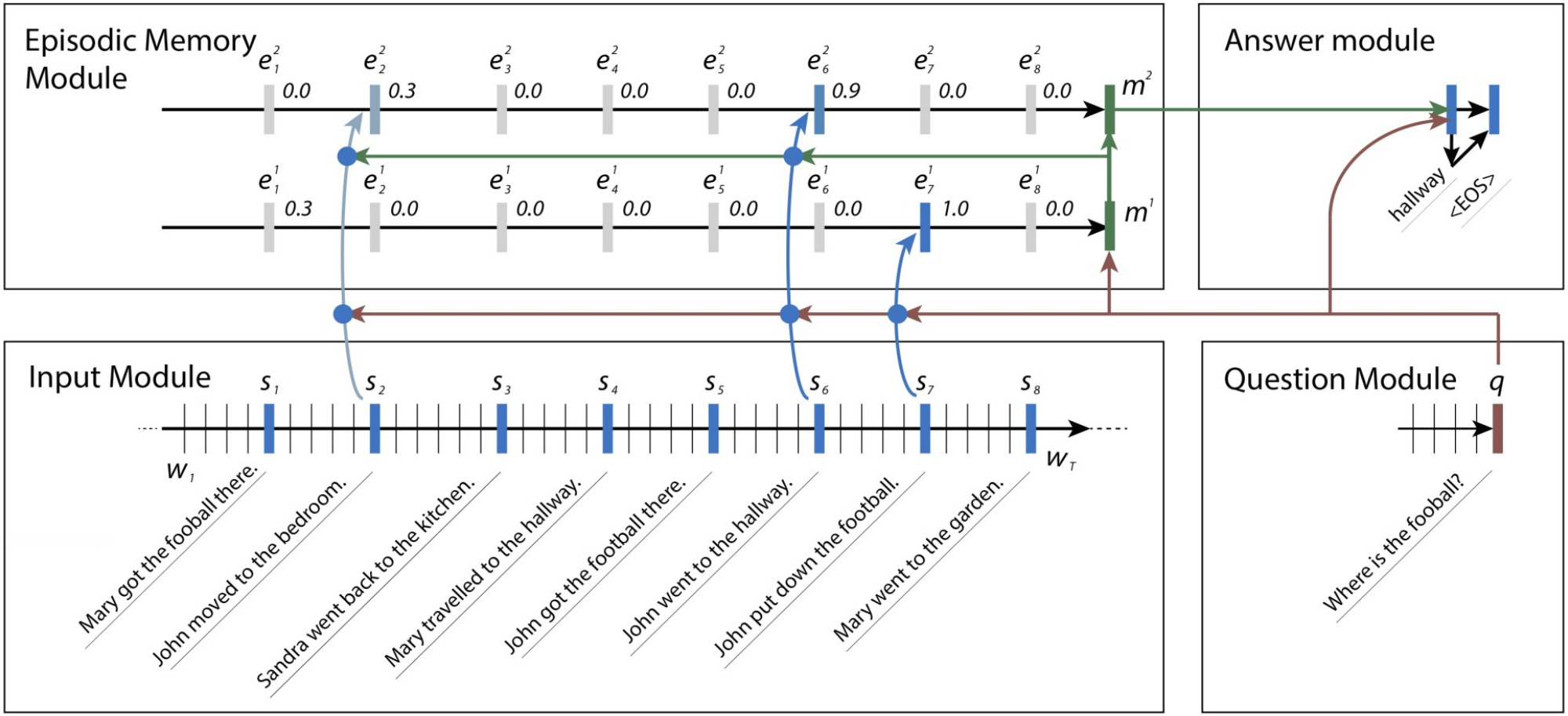
- Attention mechanism: a "simple" two-layer feed forward NN which gives a score to each fact c_t given the memory state m_{t-1} and the question q
- Memory update mechanism: a modified GRU
- The EMM iterates over representation outputted by the input module

Answer Module

- Generates an answer given a vector
- Uses a modified GRU

- This is the important part to produce sentences rather than words!

Summary of DMN

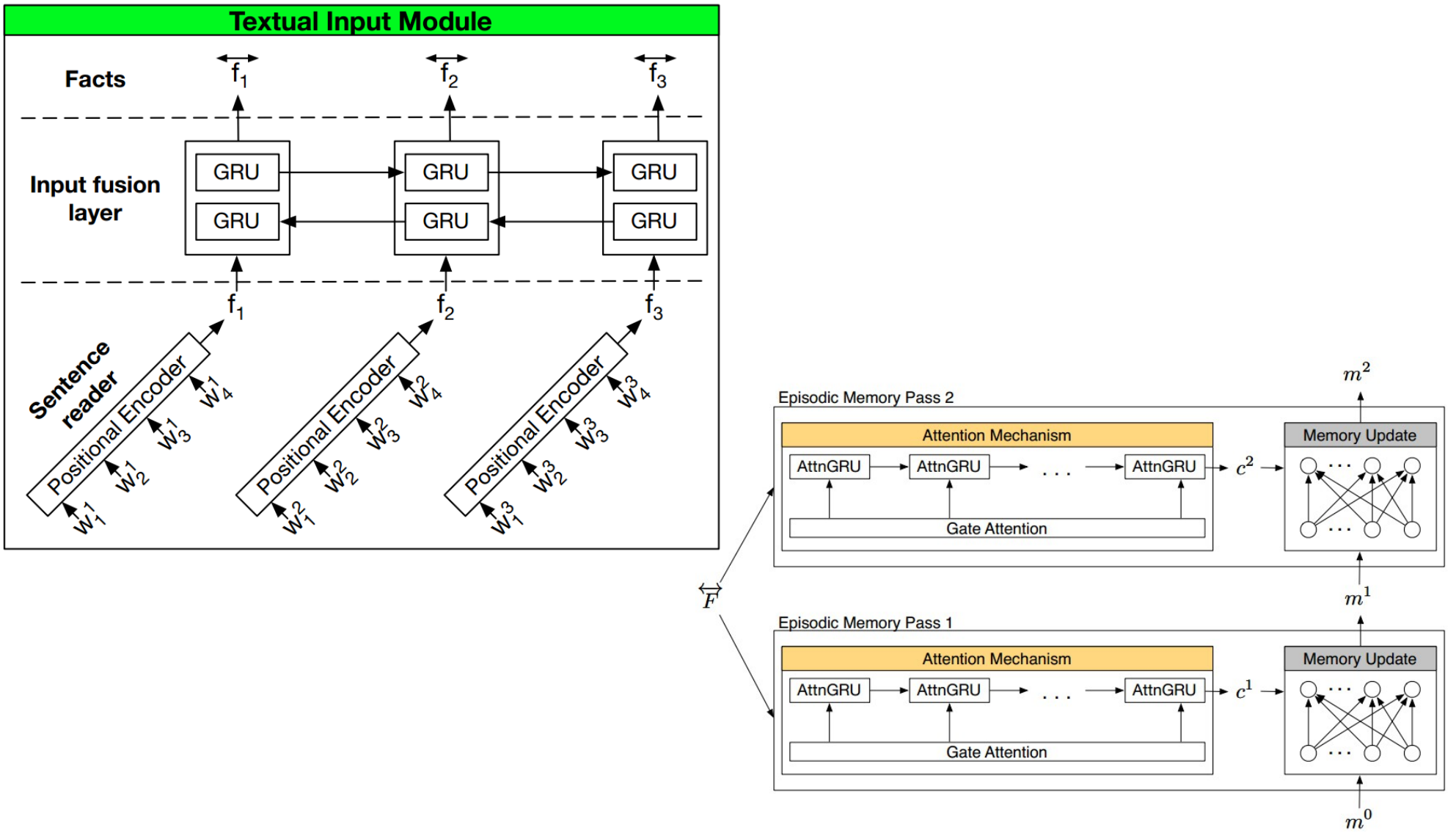


DMN+ improvements

- Modification to the input module and the episodic memory module [3].
  - IM is now made of 2 components, one for encoding words into a sentence embedding and the second is responsible for the interaction between sentences
  - EMM has now an attention based GRU (slightly modified GRU) and a ReLU memory update mechanism

    $m_t = ReLU(W(t)[m_{t-1};c_t;q] + b)$

DMN+ improvements



Results of DMN & DMN+

Model	ODMN	DMN2	DMN3	DMN+
Input module	GRU	Fusion	Fusion	Fusion
Attention	$\sum g_i f_i$	$\sum g_i f_i$	AttnGRU	AttnGRU
Mem update	GRU	GRU	GRU	ReLU
Mem Weights	Tied	Tied	Tied	Untied
bAbI English 10k dataset				
QA2	36.0	0.1	0.7	0.3
QA3	42.2	19.0	9.2	1.1
QA5	0.1	0.5	0.8	0.5
QA6	35.7	0.0	0.6	0.0
QA7	8.0	2.5	1.6	2.4
QA8	1.6	0.1	0.2	0.0
QA9	3.3	0.0	0.0	0.0
QA10	0.6	0.0	0.2	0.0
QA14	3.6	0.7	0.0	0.2
QA16	55.1	45.7	47.9	45.3
QA17	39.6	5.9	5.0	4.2
QA18	9.3	3.8	0.1	2.1
QA20	1.9	0.0	0.0	0.0
Mean error	11.8	3.9	3.3	2.8

- Obtained using 3 passes
- For some tasks (QA3, QA17, Q18) accuracy is not stable accross multiple runs

Results of DMN & DMN+

Model	ODMN	DMN2	DMN3	DMN+
Input module	GRU	Fusion	Fusion	Fusion
Attention	$\sum g_i f_i$	$\sum g_i f_i$	AttnGRU	AttnGRU
Mem update	GRU	GRU	GRU	ReLU
Mem Weights	Tied	Tied	Tied	Untied
bAbI English 10k dataset				
QA2	36.0	0.1	0.7	0.3
QA3	42.2	19.0	9.2	1.1
QA5	0.1	0.5	0.8	0.5
QA6	35.7	0.0	0.6	0.0
QA7	8.0	2.5	1.6	2.4
QA8	1.6	0.1	0.2	0.0
QA9	3.3	0.0	0.0	0.0
QA10	0.6	0.0	0.2	0.0
QA14	3.6	0.7	0.0	0.2
QA16	55.1	45.7	47.9	45.3
QA17	39.6	5.9	5.0	4.2
QA18	9.3	3.8	0.1	2.1
QA20	1.9	0.0	0.0	0.0
Mean error	11.8	3.9	3.3	2.8

- QA16: Basic induction. An simpler model achieve an error rate of 0.4 [3]
- Overall, really good results!

Potential Modifications

- Final Goal: Use of Extended DMN(+) as a chatbot!
- Modify the answer module to produce a sentence
  - Similar architecture to recurrent encoder-decoder? Used with success in the query suggestion task [4]
- Mark the sentence so the model know if a sentence is from the user or itself (and be able to use this information!)

References

- [1] Memory Networks, 29 Nov 2015
- [2] Ask Me Anything: Dynamic Memory Network for Question Answering, 5 Mar 2016
- [3] Dynamic Memory Network for Visual and Textual Question Answering, 4 Mar 2016
- [4] A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion, 8 Jul 2015

Questions?