Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

Ankit Kumar Peter Ondruska Mohit Ivyer James Bradbury
Ishaan Gulrajani Victor Zhong Romain Paulus Richard Socher

MetaMind

ICML, 2017
Presenter: Tianlu Wang

Outline

1. Introduction
2. Dynamic Memory Network
   - Model Overview
   - Input Module
   - Question Module
   - Episodic Memory Module
   - Answer Module
3. Experiments
   - Compared to baselines
   - Qualitative Example

Introduction

- Tasks in natural language processing can be cast as a question answering problem:
  - Machine Translation $\Rightarrow$ What is the translation into French?
  - Name entity recognition $\Rightarrow$ What are the name entity tags in this sentence?

$$\begin{gathered}\text{Input sequences}+\text{Questions}\\ \downarrow \\ \text{Dynamic Memory Network(DMN)} \\ \downarrow \\ \text{Answers}\end{gathered}$$

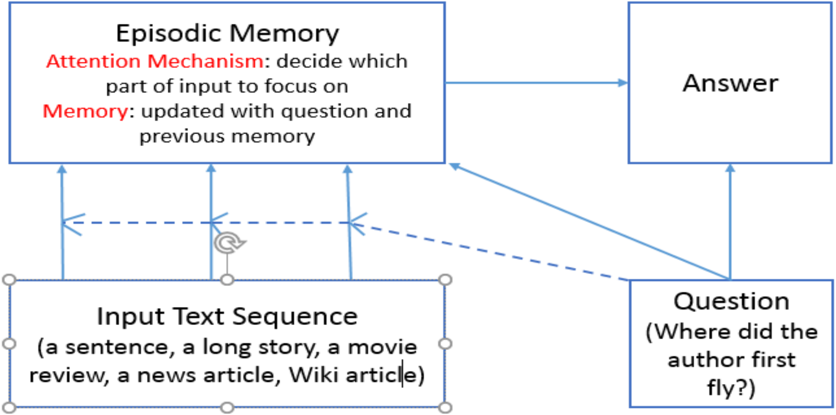
- State-of-the-art on multiple dataset:
  - Question answering(Facebook bAbI dataset)
  - Text classification for sentiment analysis(Stanford Sentiment Treebank)
  - Sequence modeling for part-of-speech tagging(WSJ-PTB)

Intuition from Neuroscience

- The episodic memory in humans stores specific experiences in their spatial and temporal context.
- Provide a vector representation to capture all relevant information from input sequences and questions.

Outline

Model Overview



Outline

1. Introduction
2. Dynamic Memory Network
   - Model Overview
   - Input Module
   - Question Module
   - Episodic Memory Module
   - Answer Module
3. Experiments
   - Compared to baselines
   - Qualitative Example

Input Module

- $h_t = GRU(x_t, h_{t-1})$, x_t is embedding of tth word
- output of this module is denoted as c, $|c| = T_c$
  - input is a single sentence:output all hidden states of RNN, $|c| = T_c$ is number of words
  - input is a list of sentences:concatenate, insert end-of-sentence tokens and output hidden states at end-of-sentence tokens, $|c| = T_c$ is number of sentences

Outline

Question Module

- $q_t = GRU(x_t^Q, q_{t-1})$, x_t^Q is embedding of tth word in the question
- output the final state of recurrent network, noted as q

Outline

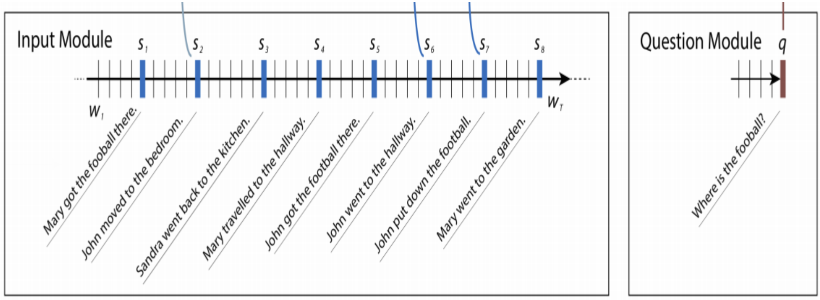
Need for Multiple Episodes

- In every iteration:**(note it is c, not c_t)**

$$c(\text{input sequence}) + q(\text{question}) + m^{i-1}(\text{previous memory}) \Rightarrow e^i(\text{episode memory})$$

$$m^i = GRU(e^i, m^{i-1})$$

- why we need multiple episodes?



Attention Mechanism

- gating function: $g_t^i = G(c_t, m^{i-1}, q)$, output a scalar score
- $G(c, m, q) = \sigma(W^{(2)} tanh(W^{(1)} z(c, m, q) + b^{(1)}) + b^{(2)})$, 2-layer nn
- $z(c, m, q) = [c, m, q, c \circ q, c \circ m, |c - q|, |c - m|, c^T W^{(b)} q, c^T W^{(b)} m]$
- output is a scalar score g_t^i for every c_t in c

Memory Update Mechanism

- episode vector is the final state of GRU

$$h_t^i = g_t^i GRU(c_t, h_{t-1}^i) + (1 - g_t^i) h_{t-1}^i \tag{1}$$

$$e^i = h_{T_c}^i \tag{2}$$

$$m^i = GRU(e^i, m^{i-1}) \tag{3}$$

Outline

1. Introduction
2. Dynamic Memory Network
   - Model Overview
   - Input Module
   - Question Module
   - Episodic Memory Module
   - Answer Module
3. Experiments
   - Compared to baselines
   - Qualitative Example

Answer Module

- A *GRU* whose initial state is initialized to the last memory $a_0 = m^{T_M}$

$$y_t = softmax(W^{(a)} a_t) \tag{4}$$

$$a_t = GRU([y_{t-1}, q], a_{t-1}) \tag{5}$$

Outline

Compared to baselines

Task	MemNN	DMN
1: Single Supporting Fact	100	100
2: Two Supporting Facts	100	98.2
3: Three Supporting Facts	100	95.2
4: Two Argument Relations	100	100
5: Three Argument Relations	98	99.3
6: Yes/No Questions	100	100
7: Counting	85	96.9
8: Lists/Sets	91	96.5
9: Simple Negation	100	100
10: Indefinite Knowledge	98	97.5
11: Basic Coreference	100	99.9
12: Conjunction	100	100
13: Compound Coreference	100	99.8
14: Time Reasoning	99	100
15: Basic Deduction	100	100
16: Basic Induction	100	99.4
17: Positional Reasoning	65	59.6
18: Size Reasoning	95	95.3
19: Path Finding	36	34.5
20: Agent's Motivations	100	100
Mean Accuracy (%)	93.3	**93.6**

Table 1. Test accuracies on the bAbI dataset. MemNN numbers taken from Weston et al. (Weston et al., 2015a). The DMN passes (accuracy $> 95\%$) 18 tasks, whereas the MemNN passes 16.

Compared to baselines

Model	Acc (%)
SVMTool	97.15
Sogaard	97.27
Suzuki et al.	97.40
Spoustova et al.	97.44
SCNN	97.50
DMN	**97.56**

Table 3. Test accuracies on WSJ-PTB

Compared to baselines

Max passes	task 3 three-facts	task 7 count	task 8 lists/sets	sentiment (fine grain)
0 pass	0	48.8	33.6	50.0
1 pass	0	48.8	54.0	51.5
2 pass	16.7	49.1	55.6	**52.1**
3 pass	64.7	83.4	83.4	50.1
5 pass	**95.2**	**96.9**	**96.5**	N/A

Table 4. Effectiveness of episodic memory module across tasks. Each row shows the final accuracy in term of percentages with a different maximum limit for the number of passes the episodic memory module can take. Note that for the 0-pass DMN, the network essential reduces to the output of the attention module.

Outline

1 Introduction

2 Dynamic Memory Network
- Model Overview
- Input Module
- Question Module
- Episodic Memory Module
- Answer Module

3 Experiments
- Compared to baselines
- Qualitative Example

Qualitative Examples