Dynamic Memory Networks for Question Answering over Text and Images

Richard Socher
Joint work with the MetaMind team
Caiming Xiong, Stephen Merity, James Bradbury, Ankit Kumar, Ozan Irsoy and others
Current Research

All NLP/AI tasks can be reduced to question answering.
QA Examples

I: Mary walked to the bathroom.  I: Jane has a baby in Dresden.
I: Sandra went to the garden.  Q: What are the named entities?
I: Daniel went back to the garden.  A: Jane - person, Dresden - location
I: Sandra took the milk there.  I: Jane has a baby in Dresden.
Q: Where is the milk?  Q: What are the POS tags?
A: garden  A: NNP VBZ DT NN IN NNP .
I: Everybody is happy.  I: I think this model is incredible
Q: What’s the sentiment?  Q: In French?
A: positive  A: Je pense que ce modèle est incroyable.
Goal

A joint model for general QA
First Major Obstacle

• For NLP no single model **architecture** with consistent state of the art results across tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>State of the art model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question answering (babI)</td>
<td>Strongly Supervised MemNN (Weston et al 2015)</td>
</tr>
<tr>
<td>Sentiment Analysis (SST)</td>
<td>Tree-LSTMs (Tai et al. 2015)</td>
</tr>
<tr>
<td>Part of speech tagging (PTB-WSJ)</td>
<td>Bi-directional LSTM-CRF (Huang et al. 2015)</td>
</tr>
</tbody>
</table>
Second Major Obstacle

- Fully joint multitask learning* is hard:
  - Usually restricted to lower layers
  - Usually helps only if tasks are related
  - Often hurts performance if tasks are not related

* meaning: same decoder/classifier and not only transfer learning
Tackling First Obstacle

Dynamic Memory Networks

An architecture for any QA task
High level idea for harder questions

• Imagine having to read an article, memorize it, then get asked various questions → Hard!
• You can't store everything in working memory
• **Optimal:** give you the input data, give you the question, allow as many glances as possible
Basic Lego Block: RNNs

- Gated Recurrent Unit (GRU), Cho et al. 2014
- A type of recurrent neural network (RNN), similar to the LSTM
- Consumes and/or generates sequences (chars, words, ...)
- The GRU updates an internal state $h$ according to the existing state $h$ and the current input $x$: $h_t = GRU(x_t, h_{t-1})$

![Diagram of GRU](image-url)
DMN Overview

Semantic Memory Module

(Glove vectors)

Episodic Memory Module

\[ e_i^t \]

\[ 0.0 \quad 0.3 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.9 \quad 0.0 \quad 0.0 \]

\[ m^t \]

\[ 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 1.0 \quad 0.0 \quad 0.0 \]

Answer module

Question Module

\( q \)

End of Sentence

Input Module

\[ w_i \]

\[ s_i \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \]

May got the milk there.

John moved to the bedroom.

Sandra went back to the kitchen.

Mary travelled to the hallway.

John got the football there.

John went to the hallway.

John put down the football.

Mary went to the garden.

Where is the football?

DMN Overview

Semantic Memory Module

(Glove vectors)

Episodic Memory Module

\[ e_i^t \]

\[ 0.0 \quad 0.3 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.9 \quad 0.0 \quad 0.0 \]

\[ m^t \]

\[ 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 1.0 \quad 0.0 \quad 0.0 \]

Answer module

Question Module

\( q \)

End of Sentence

Input Module

\[ w_i \]

\[ s_i \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \]

May got the milk there.

John moved to the bedroom.

Sandra went back to the kitchen.

Mary travelled to the hallway.

John got the football there.

John went to the hallway.

John put down the football.

Mary went to the garden.

Where is the football?

DMN Overview

Semantic Memory Module

(Glove vectors)

Episodic Memory Module

\[ e_i^t \]

\[ 0.0 \quad 0.3 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.9 \quad 0.0 \quad 0.0 \]

\[ m^t \]

\[ 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 1.0 \quad 0.0 \quad 0.0 \]

Answer module

Question Module

\( q \)

End of Sentence

Input Module

\[ w_i \]

\[ s_i \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \quad s_s \]

May got the milk there.

John moved to the bedroom.

Sandra went back to the kitchen.

Mary travelled to the hallway.

John got the football there.

John went to the hallway.

John put down the football.

Mary went to the garden.

Where is the football?
The Modules: Input

Semantic Memory Module
(Glove vectors)

Episodic Memory Module

Answer module

Question Module

Input Module

Answer module

Semantic Memory Module

Episodic Memory Module

Question Module

Answer module

Input Module
Further Improvement: BiGRU
The Modules: Question

Standard GRU. Output: last hidden state $\rightarrow q$
The Modules: Episodic Memory

\[ h_i = g_i^t \circ \tilde{h}_i + (1 - g_i^t) \circ h_{i-1} \]
The Modules: Episodic Memory

- Gates are activated if relevant to the question

\[
\begin{align*}
    z_i^t &= [\vec{f}_i \circ q; \vec{f}_i \circ m^{t-1}; |\vec{f}_i - q|; |\vec{f}_i - m^{t-1}|] \\
    Z_i^t &= W^{(2)} \tanh \left( W^{(1)} z_i^t + b^{(1)} \right) + b^{(2)} \\
    g_i^t &= \frac{\exp(Z_i^t)}{\sum_{k=1}^{M_i} \exp(Z_k^t)}
\end{align*}
\]

- When the end of the input is reached, the relevant facts are summarized in another GRU or simple NNet

\[
m^t = \text{ReLU} \left( W^t[m^{t-1}; c^t; q] + b \right)
\]
The Modules: Episodic Memory

- If summary is insufficient to answer the question, repeat sequence over input
Inspiration from Neuroscience

• **Episodic memory** is the memory of autobiographical events (times, places, etc). A collection of past personal experiences that occurred at a particular time and place.

• The hippocampus, the seat of episodic memory in humans, is active during transitive inference

• In the DMN repeated passes over the input are needed for transitive inference
For datasets that mark which facts are important for a given question, such as Facebook’s bAbI, it described it here as its own computation to highlight the potential modularity of these subcomponents. Is equivalent to setting the memory to simply the attention mechanism’s final state, but we have de-attention mechanism’s state:

Finally, to summarize the episodic component allows its attention mechanism to attend more selectively to endowed with our gates. The episode is the final state of the GRU: 

In our work, we use a gating function as our attention mechanism. It takes as input, for each pass the input module, and a function that summarizes the episodes into a memory.

In its general form, the episodic memory module is characterized by an attention mechanism, a function which returns an episode given the output of the attention mechanism and the facts from sentence 2, which makes some intuitive sense, as sentence 2 is another place John had been. This behavior is indeed seen. Note that the second iteration has wrongly placed some weight in retrieve where John was. In this example, taken from a true test question on Facebook’s bAbI task, football. Only once the model sees that John is relevant can it reason the second iteration should the model ought attend to sentence 7 (4.1 for details on the dataset that this example comes from.)

For instance, in the example in Fig. 3, we are asked a type of transitive inference, since the first pass may uncover the need to retrieve additional facts. Each pass specific facts on each pass, as it can attend to other important facts at a later pass. It also allows for produces an...
Academic papers and related work

• For full details:
  • Ask Me Anything: Dynamic Memory Networks for Natural Language Processing (Kumar et al., 2015)
  • Dynamic Memory Networks for Visual and Textual Question Answering (Xiong et al., 2016)

• Sequence to Sequence (Sutskever et al. 2014)
• Neural Turing Machines (Graves et al. 2014)
• Teaching Machines to Read and Comprehend (Hermann et al. 2015)
• Learning to Transduce with Unbounded Memory (Grefenstette 2015)
• Structured Memory for Neural Turing Machines (Wei Zhang 2015)

• Memory Networks (Weston et al. 2015)
• End to end memory networks (Sukhbaatar et al. 2015)
Comparison to MemNets

Similarities:
• MemNets and DMNs have input, scoring, attention and response mechanisms

Differences:
• For input representations MemNets use bag of word, nonlinear or linear embeddings that explicitly encode position
• MemNets iteratively run functions for attention and response

• DMNs shows that neural sequence models can be used for input representation, attention and response mechanisms → naturally captures position and temporality
• Enables broader range of applications
Experiments: QA on babI (1k)

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

Mean Accuracy (%)                     | 93.3  | 93.6 |

This still requires that relevant facts are marked during training to train the gates.
Despite the glowing reviews, this movie wasn't an especially surprising or interesting experience.
Experiments: Sentiment Analysis

- Stanford Sentiment Treebank
- Test accuracies:
  - MV-RNN and RNTN: Socher et al. (2013)
  - DCNN: Kalchbrenner et al. (2014)
  - PVec: Le & Mikolov. (2014)
  - CNN-MC: Kim (2014)
  - CT-LSTM: Tai et al. (2015)

<table>
<thead>
<tr>
<th>Task</th>
<th>Binary</th>
<th>Fine-grained</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV-RNN</td>
<td>82.9</td>
<td>44.4</td>
</tr>
<tr>
<td>RNTN</td>
<td>85.4</td>
<td>45.7</td>
</tr>
<tr>
<td>DCNN</td>
<td>86.8</td>
<td>48.5</td>
</tr>
<tr>
<td>PVec</td>
<td>87.8</td>
<td>48.7</td>
</tr>
<tr>
<td>CNN-MC</td>
<td>88.1</td>
<td>47.4</td>
</tr>
<tr>
<td>DRNN</td>
<td>86.6</td>
<td>49.8</td>
</tr>
<tr>
<td>CT-LSTM</td>
<td>88.0</td>
<td>51.0</td>
</tr>
<tr>
<td>DMN</td>
<td>88.6</td>
<td>52.1</td>
</tr>
</tbody>
</table>
## Analysis of Number of Episodes

- How many attention + memory passes are needed in the episodic memory?

<table>
<thead>
<tr>
<th>Max passes</th>
<th>task 3 three-facts</th>
<th>task 7 count</th>
<th>task 8 lists/sets</th>
<th>sentiment (fine grain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 pass</td>
<td>0</td>
<td>48.8</td>
<td>33.6</td>
<td>50.0</td>
</tr>
<tr>
<td>1 pass</td>
<td>0</td>
<td>48.8</td>
<td>54.0</td>
<td>51.5</td>
</tr>
<tr>
<td>2 pass</td>
<td>16.7</td>
<td>49.1</td>
<td>55.6</td>
<td>52.1</td>
</tr>
<tr>
<td>3 pass</td>
<td>64.7</td>
<td>83.4</td>
<td>83.4</td>
<td>50.1</td>
</tr>
<tr>
<td>5 pass</td>
<td><strong>95.2</strong></td>
<td><strong>96.9</strong></td>
<td><strong>96.5</strong></td>
<td>N/A</td>
</tr>
</tbody>
</table>
Analysis of Attention for Sentiment

- Sharper attention when 2 passes are allowed.
- Examples that are wrong with just one pass

1-iter DMN (pred: negative, ans: positive)

2-iter DMN (pred: positive, ans: positive)
Analysis of Attention for Sentiment

1-iter DMN (pred: very positive, ans: negative)

2-iter DMN (pred: negative, ans: negative)
Analysis of Attention for Sentiment

- Examples where full sentence context from first pass changes attention to words more relevant for final prediction.

1-iter DMN (pred: negative, ans: positive)

2-iter DMN (pred: positive, ans: positive)
Analysis of Attention for Sentiment

- Examples where full sentence context from first pass changes attention to words more relevant for final prediction

1-iter DMN (pred: positive, ans: negative)

2-iter DMN (pred: negative, ans: negative)
Despite the glowing reviews, this movie wasn't an especially surprising or interesting experience.
Experiments: POS Tagging

- PTB WSJ, standard splits
- Episodic memory does not require multiple passes, single pass enough

<table>
<thead>
<tr>
<th>Model</th>
<th>SVMTool</th>
<th>Sogaard</th>
<th>Suzuki et al.</th>
<th>Spoustova et al.</th>
<th>SCNN</th>
<th>DMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc (%)</td>
<td>97.15</td>
<td>97.27</td>
<td>97.40</td>
<td>97.44</td>
<td>97.50</td>
<td><strong>97.56</strong></td>
</tr>
</tbody>
</table>
Despite the glowing reviews, this movie wasn't an especially surprising or interesting experience.

What is the sentiment?

Run DMN

Get new example
Modularization Allows for Different Inputs

Episodic Memory → Answer: Kitchen

Input Module
- John moved to the garden.
- John got the apple there.
- John moved to the kitchen.
- Sandra picked up the milk there.
- John dropped the apple.
- John moved to the office.

Question: Where is the apple?

Episodic Memory → Answer: Palm

Input Module
- Question: What kind of tree is in the background?

Figure 1. Question Answering over text and images using a Dynamic Memory Network (DMN).
The VGG-19 model (from the image, we use a convolutional neural network)

Local region feature extraction:

in Sec. feature embedding, and the input fusion layer introduced

The input module for VQA is composed of three parts, il-
gion equivalent to a sentence in the input module for text.
an image into small local regions and considers each re-

3. DMN Input Module for VQA

Episode Memory Updates

hidden state of the attention based GRU.

tion of facts into account, which the soft attention model
based GRU can now take positional and ordering informa-
to update decisions. Additionally, the attention
gate can be more detailed, we speculate it allows bet-
the output of the attention gates

We propose replacing the update gates

the question or previous episode memory.
The episodic memory for

is computed by

for untied experiments where

. For untied experiments where

is computed using only the current input and the hidden
isl is used for this purpose. The episodic memory for

is computed by

as the input to the up-

As the bi-directional GRU is one dimensional and
allows for information propagation from neighboring image

To solve this, we add an input fusion layer similar to that

Input fusion layer:

Visual feature extraction

Input Module for Images

Input Module

Input fusion layer

Feature embedding

Visual feature extraction

CNN
### Accuracy: Visual Question Answering

VQA test-dev and test-standard:
- Antol et al. (2015)
- ACK Wu et al. (2015);
- iBOWIMG - Zhou et al. (2015);
- DPPnet - Noh et al. (2015); D-NMN - Andreas et al. (2016);
- SAN - Yang et al. (2015)

<table>
<thead>
<tr>
<th>Method</th>
<th>test-dev</th>
<th>test-std</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Y/N</td>
</tr>
<tr>
<td>VQA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>28.1</td>
<td>64.0</td>
</tr>
<tr>
<td>Question</td>
<td>48.1</td>
<td>75.7</td>
</tr>
<tr>
<td>Q+I</td>
<td>52.6</td>
<td>75.6</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td>53.7</td>
<td>78.9</td>
</tr>
<tr>
<td>ACK</td>
<td>55.7</td>
<td>79.2</td>
</tr>
<tr>
<td>iBOWIMG</td>
<td>55.7</td>
<td>76.5</td>
</tr>
<tr>
<td>DPPnet</td>
<td>57.2</td>
<td>80.7</td>
</tr>
<tr>
<td>D-NMN</td>
<td>57.9</td>
<td>80.5</td>
</tr>
<tr>
<td>SAN</td>
<td>58.7</td>
<td>79.3</td>
</tr>
<tr>
<td>DMN+</td>
<td><strong>60.3</strong></td>
<td>80.5</td>
</tr>
</tbody>
</table>
What is the main color on the bus? Answer: blue
What type of trees are in the background? Answer: pine
How many pink flags are there? Answer: 2
Is this in the wild? Answer: no
Attention Visualization

Which man is dressed more flamboyantly? Answer: right

Who is on both photos? Answer: girl

What time of day was this picture taken? Answer: night

What is the boy holding? Answer: surfboard
Attention Visualization

What is this sculpture made out of? Answer: metal

What color are the bananas? Answer: green

What is the pattern on the cat's fur on its tail? Answer: stripes

Did the player hit the ball? Answer: yes
Live Demo
What is the girl holding?  tennis racket
What is the girl doing?  playing tennis
Is the girl wearing a hat?  yes
What is the girl wearing?  shorts

What is the color of the ground?  brown
What color is the ball?  yellow
What color is her skirt?  white
What did the girl just hit?  tennis ball
Summary

• Most NLP tasks can be reduced to QA
• DMN accurately solves variety of QA tasks
• Next goals: One joint multitask DMN