Deep NLP Applications and Dynamic Memory Networks

Richard Socher
richard@metamind.io
Why focus deep research on NLP?

- Image classification increasingly commoditized
  - Vision is more than classification but it’s central

- Demo: [https://www.metamind.io/vision/train](https://www.metamind.io/vision/train)
Overview

• Fun deep NLP applications:
  • Character RNNs on text and code
  • Image – Sentence mapping
  • Engagement
  • Question Answering

• Ask me Anything: Dynamic Memory Networks for NLP
Character RNNs on text and code

http://karpathy.github.io/2015/05/21/rnn-effectiveness/
Character RNNs on text and code

- Haven’t yet produced useful results on real datasets

- Shows that RNNs can memorize sequences and keep memory (mostly LSTMs)

- Most interesting results simply train on dataset and sample from it afterwards (first shown by Sutskever et al. 2011: Generating Text with Recurrent Neural Networks)

- Results from an LSTM (karpathy.github.io)
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.
Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict.
Proof. Omitted.

**Lemma 0.1.** Let $C$ be a set of the construction.

Let $C$ be a gerber covering. Let $F$ be a quasi-coherent sheaves of $O$-modules. We have to show that

$$ O_{O_X} = O_X(L) $$

Proof. This is an algebraic space with the composition of sheaves $F$ on $X_{etale}$ we have

$$ O_X(F) = \{ morph \times O_X(G,F) \} $$

where $G$ defines an isomorphism $F \to F$ of $O$-modules.

**Lemma 0.2.** This is an integer $Z$ is injective.

Proof. See Spaces, Lemma ??.

**Lemma 0.3.** Let $S$ be a scheme. Let $X$ be a scheme and $X$ is an affine open covering. Let $U \subset X$ be a canonical and locally of finite type. Let $X$ be a scheme. Let $X$ be a scheme which is equal to the formal complex.

The following to the construction of the lemma follows.

Let $X$ be a scheme. Let $X$ be a scheme covering. Let

$$ b : X \to Y' \to Y \to Y' \times_X Y \to X. $$

be a morphism of algebraic spaces over $S$ and $Y$.

Proof. Let $X$ be a nonzero scheme of $X$. Let $X$ be an algebraic space. Let $F$ be a quasi-coherent sheaf of $O_X$-modules. The following are equivalent

1. $F$ is an algebraic space over $S$.
2. If $X$ is an affine open covering.

Consider a common structure on $X$ and $X$ the functor $O_X(U)$ which is locally of finite type.

This since $F \in F$ and $x \in G$ the diagram

This is a limit. Then $G$ is a finite type and assume $S$ is a flat and $F$ and $G$ is a finite type $F'$. This is of finite type diagrams, and

- the composition of $G$ is a regular sequence,
- $O_{X'}$ is a sheaf of rings.

Proof. We have see that $X = Spec(R)$ and $F$ is a finite type representable by algebraic space. The property $F$ is a finite morphism of algebraic stacks. Then the cohomology of $X$ is an open neighbourhood of $U$.

Proof. This is clear that $G$ is a finite presentation, see Lemmas ??.

A reduced above we conclude that $U$ is an open covering of $C$. The functor $F$ is a "field"

$$ O_{X,0} \to \mathcal{F} \to O_{X,0}^{-1}(O_{X,0}) \to O_{X,0}^{-1}(O_{X,0}(O_{X,0}^{\mathbb{P}})) $$

is an isomorphism of covering of $O_{X,0}$. If $F$ is the unique element of $F$ such that $X$ is an isomorphism.

The property $F$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $O_X$-algebra with $F$ are opens of finite type over $S$.

If $F$ is a scheme theoretic image points.

If $F$ is a finite direct sum $O_{X,0}$ is a closed immersion, see Lemma ?? This is a sequence of $F$ is a similar morphism.
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */

static int indicate_policy(void)
{
    int error;
    if (fd == MARN_EPT) {
        /*
         * The kernel blank will coeld it to userspace.
         */
        if (ss->segment < mem_total)
            unblock_graph_and_set_blocked();
        else
            ret = 1;
        goto bail;
    }
    segaddr = in_SB(in.addr);
    selector = seg / 16;
    setup_works = true;
}
Question Answering: Quiz Bowl Competition

• **Iyyer et al. 2014**: A Neural Network for Factoid Question Answering over Paragraphs

• **QUESTION:**
  He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join. A more famous work by this author tells of the rise and fall of the composer Adrian Leverkühn. Another of his novels features the jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach. Name this German author of The Magic Mountain and Death in Venice.
• **QUESTION:**
  He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join. A more famous work by this author tells of the rise and fall of the composer Adrian Leverkühn. Another of his novels features the jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach. Name this German author of The Magic Mountain and Death in Venice.

• **ANSWER:** Thomas Mann
Recursive Neural Networks

- Follow dependency structure

Students ride bikes at night
Pushing Facts into Entity Vectors
### Table 1: Accuracy for history and literature at the first two sentence positions of each question and the full question.

<table>
<thead>
<tr>
<th>Model</th>
<th>Pos 1</th>
<th>Pos 2</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>bow</td>
<td>27.5</td>
<td>51.3</td>
<td>53.1</td>
</tr>
<tr>
<td>bow-dt</td>
<td>35.4</td>
<td>57.7</td>
<td>60.2</td>
</tr>
<tr>
<td>ir-qb</td>
<td>37.5</td>
<td>65.9</td>
<td>71.4</td>
</tr>
<tr>
<td>fixed-qanta</td>
<td>38.3</td>
<td>64.4</td>
<td>66.2</td>
</tr>
<tr>
<td>qanta</td>
<td>47.1</td>
<td>72.1</td>
<td>73.7</td>
</tr>
<tr>
<td>ir-wiki</td>
<td>53.7</td>
<td>76.6</td>
<td>77.5</td>
</tr>
<tr>
<td>qanta+ir-wiki</td>
<td>59.8</td>
<td>81.8</td>
<td>82.3</td>
</tr>
</tbody>
</table>

The top half of the table compares models trained on questions only, while the IR models in the bottom half have access to Wikipedia.

Qanta outperforms all baselines that are restricted to just the question data, and it substantially improves an IR model with access to Wikipedia despite being trained on much less data.

![History: Model vs. Human](image)

**Qanta Model Can Defeat Human Players**

Figure 4: Comparisons of qanta+ir-wiki to human quiz bowl players. Each bar represents an individual human, and the bar height corresponds to the difference between the model score and the human score. Bars are ordered by human skill. Red bars indicate that the human is winning, while blue bars indicate that the model is winning.

Figure 3: A question on the play “No Exit” with human buzz position marked as $^{3}\text{S.}$ Since the buzz occurs in the middle of the second sentence, our model is only allowed to see the first sentence.

#### 5.1 Experimental Results

Table 1 shows that when bag of words and information retrieval methods are restricted to question data, they perform significantly worse than qanta on early sentence positions. The performance of bow-dt indicates that while the dependency tree structure helps by itself, the compositional distributed representations learned by qanta are more useful. The significant improvement when we train answers as part of our vocabulary (see Section 3.2) indicates that our model uses answer occurrences within question text to learn a more informative vector space.

The disparity between ir-qb and ir-wiki indicates that the information retrieval models need lots of external data to work well at all sentence positions. ir-wiki performs better than other models because Wikipedia contains many more sentences that partially match specific words or phrases found in early clues than the question training set. In particular, it is impossible for all other models to answer clues in the test set that have no semantically similar
Table 1: Accuracy for history and literature at the first two sentence positions of each question and the full question. The top half of the table compares models trained on questions only, while the IR models in the bottom half have access to Wikipedia.

- **qanta** outperforms all baselines that are restricted to just the question data, and it substantially improves an IR model with access to Wikipedia despite being trained on much less data.

Figure 4: Comparisons of **qanta+ir-wiki** to human quiz bowl players. Each bar represents an individual human, and the bar height corresponds to the difference between the model score and the human score. Bars are ordered by human skill. Red bars indicate that the human is winning, while blue bars indicate that the model is winning.

**qanta+ir-wiki** outperforms most humans on history questions but fails to defeat the “average” human on literature questions.

A minor character in this play can be summoned by a bell that does not always work; that character also doesn’t have eyelids. Near the end, a woman who drowned her illegitimate child attempts to stab another woman in the Second Empire-style room in which the entire play takes place. For 10 points, Estelle and Ines are characters in which existentialist play in which Garcin claims “Hell is other people”, written by Jean-Paul Sartre?

Figure 3: A question on the play “No Exit” with human buzz position marked as 3.

Since the buzz occurs in the middle of the second sentence, our model is only allowed to see the first sentence.

### 5.1 Experimental Results

Table 1 shows that when bag of words and information retrieval methods are restricted to question data, they perform significantly worse than **qanta** on early sentence positions. The performance of **bow-dt** indicates that while the dependency tree structure helps by itself, the compositional distributed representations learned by **qanta** are more useful. The significant improvement when we train answers as part of our vocabulary (see Section 3.2) indicates that our model uses answer occurrences within question text to learn a more informative vector space.

The disparity between **ir-qb** and **ir-wiki** indicates that the information retrieval models need lots of external data to work well at all sentence positions. **ir-wiki** performs better than other models because Wikipedia contains many more sentences that partially match specific words or phrases found in early clues than the question training set. In particular, it is impossible for all other models to answer clues in the test set that have no semantically similar
Visual Grounding

• Idea: Map sentences and images into a joint space

Socher et al. 2013:
Grounded Compositional Semantics for Finding and Describing Images with Sentences
Discussion: Compositional Structure

• Recursive Neural Networks so far used constituency trees which results in more syntactically influenced representations

• Instead: Use dependency trees which capture more semantic structure
Convolutional Neural Network for Images

- CNN trained on ImageNet (Le et al. 2013)
- RNN trained to give large inner products between sentence and image vectors:

\[
J(W_I, \theta) = \sum_{(i,j) \in \mathcal{P}} \sum_{c \in \mathcal{S}\setminus \mathcal{S}(i)} \max(0, \Delta - v_i^T y_j + v_i^T y_c)
\]
Results

A gray convertible sports car is parked in front of the trees. ✗
A close-up view of the headlights of a blue old-fashioned car. ✗
Black shiny sports car parked on concrete driveway. ✓
Five cows grazing on a patch of grass between two roadways. ✗

A jockey rides a brown and white horse in a dirt corral. ✓
A young woman is riding a Bay horse in a dirt riding-ring. ✗
A white bird pushes a miniature teal shopping cart. ✗
A person rides a brown horse. ✓

A motocross bike with rider flying through the air. ✓
White propeller plane parked in middle of grassy field. ✗
The white jet with its landing gear down flies in the blue sky. ✗
An elderly woman catches a ride on the back of the bicycle. ✗
### Results

<table>
<thead>
<tr>
<th>Describing Images</th>
<th>Mean Rank</th>
<th>Image Search</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>92.1</td>
<td>Random</td>
<td>52.1</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>21.1</td>
<td>Bag of Words</td>
<td>14.6</td>
</tr>
<tr>
<td>CT-RNN</td>
<td>23.9</td>
<td>CT-RNN</td>
<td>16.1</td>
</tr>
<tr>
<td>Recurrent Neural Network</td>
<td>27.1</td>
<td>Recurrent Neural Network</td>
<td>19.2</td>
</tr>
<tr>
<td>Kernelized Canonical Correlation Analysis</td>
<td>18.0</td>
<td>Kernelized Canonical Correlation Analysis</td>
<td>15.9</td>
</tr>
<tr>
<td>DT-RNN</td>
<td><strong>16.9</strong></td>
<td>DT-RNN</td>
<td><strong>12.5</strong></td>
</tr>
</tbody>
</table>

*Images:*
- People in an outrigger canoe sail on emerald green water.
- Two people sailing a small white sail boat behind a cliff, a boat sails away.
- Tourist move in on Big Ben on a typical overcast London day.
- A group of people sitting around a table on a porch.
- A group of four people walking past a giant mushroom.
- A man and women smiling for the camera in a kitchen.
- A group of men sitting around a table drinking while a man behind stands pointing.
Engagement
Demo
Several models came out simultaneously in 2015 that follow up
- Replace recursive neural network with LSTM and instead of only finding vectors they generate the description
- Mostly memorized training sequences (becomes similar again)

- Donahue et al. 2015: Long-term → Recurrent Convolutional Networks for Visual Recognition and Description
- Karpathy and Fei-Fei 2015: Deep Visual-Semantic Alignments for Generating Image Descriptions
Image – Sentence Generation (!)

- "little girl is eating piece of cake."
- "baseball player is throwing ball in game."
- "woman is holding bunch of bananas."
- "a young boy is holding a baseball bat."
- "a cat is sitting on a couch with a remote control."
- "a woman holding a teddy bear in front of a mirror."
Dynamic Memory Networks

Richard Socher
richard@metamind.io
A new paradigm

All NLP tasks can be reduced to question answering
QA

- Question answering tackles complex questions over lots of text
  - Where was Obama's wife born?
- Machine translation
  - What is the translation into French?
- Sequence modeling tasks like named entity recognition (NER)
  - What are the named entity tags in this sentence?
- Classification problems like sentiment analysis
  - What is the sentiment?
- Even multi-sentence joint classification problems like coreference resolution
  - Who does "their" refer to?
Reduction to QA

Interesting but useless?
Yes, until a model makes it useful

Dynamic memory Network

• DMN, a neural network based model in which any QA task can be trained using input-question-answer triplets.
• Related to
  • Memory Networks, Weston et al. 2014
  • Neural Turing Machines, Graves et al. 2014
  • Teaching Machines to Read and Comprehend, Hermann et al. 2015
• as introduced by Phil yesterday but more general
Ask Me Anything: Dynamic Memory Networks for NLP
Joint Work with MetaMind intern team

- Ankit Kumar
- Ozan Irsoy
- Mohit Iyyer
- Peter Ondruska
- James Bradbury
- Ishaan Gulrajani
Example Input, Question, Answer

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: Everybody is happy.
Q: What’s the sentiment?
A: positive

I: Jane has a baby in Dresden.
Q: What are the named entities?
A: Jane - person, Dresden - location
I: Jane has a baby in Dresden.
Q: What are the POS tags?
A: NNP VBZ DT NN IN NNP .
I: I think this model is incredible
Q: In French?
A: Je pense que ce modèle est incroyable.
The DMN

Semantic Memory
- Word vectors
- Knowledge Basis

Episodic Memory

Input Text Sequence

Question

Answer

8/9/15
Richard Socher
The Modules: Input

- Responsible for computing representations of (audio, visual or) textual inputs such that they can be retrieved when needed later.
- Assume a temporal sequence indexable by a time stamp.
- For written language we have a sequence of words \((v_1, \ldots, v_{T_w})\)
- Both unsupervised and supervised learning
- Context-independent and context-dependent hidden states
- Word vectors from Glove model Pennington et al. (2014) → Stored in semantic memory module
- RNN computation for context states →
Reminder: Gated Recurrent Units in RNN

\[ h_t = GRU(x_t, h_{t-1}) \]

\[ z_t = \sigma \left( W^{(z)} x_t + U^{(z)} h_{t-1} + b^{(z)} \right) \]

\[ r_t = \sigma \left( W^{(r)} x_t + U^{(r)} h_{t-1} + b^{(r)} \right) \]

\[ \tilde{h}_t = \tanh \left( W x_t + r_t \circ U h_{t-1} + b^{(h)} \right) \]

\[ h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t, \]

For DMN input sequence: \( w_t = GRU(v_t, w_t). \)
The Modules: Question

- Simple GRU over question word vectors

\[ q_t = GRU(v_t, q_{t-1}) \]

Semantic Memory
- Word vectors
- Knowledge Basis

Episodic Memory

Answer

Input Text Sequence

Question
The Modules: Episodic Memory!

• Combines the previous three modules' outputs in order to reason over them and give the resulting knowledge to the answer module.

• Dynamically retrieves the necessary information over the sequence of words or sentences.

• If necessary to retrieve additional facts → iterate over inputs

• Needed for transitive inference (TI)
  • The hippocampus, the seat of episodic memory in humans, is active during this kind of inference and disruption of the hippocampus impairs TI
Gates over input sentences

- For each sentence in input:

\[
\begin{align*}
  z(s, m, q) &= [s \circ q, s \circ m, |s - q|, |s - m|, s, m, q, s^T W^{(b)} q, s^T W^{(b)} m] \\
  G(s, m, q) &= \sigma \left( W^{(2)} \tanh \left( W^{(1)} z(s, m, q) + b^{(1)} \right) + b^{(2)} \right)
\end{align*}
\]

- Summarize important facts in episode vector:

\[
e^1 = \sum_{t=1}^{T} \text{softmax}(g^1_t) s_t.
\]

- Done if only one pass over data was needed to answer question
Episodes

• What about: (from Facebook babI dataset)

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
Episodes

• Iterate over multiple episodes

• Compute new gates (second episode) with previous memory vector:

\[
g_t^2 = G(s_t, m^1, q)
\]

• GRU over memories:

\[
m^1 = GRU(e^1, m^0)
\]
The Modules: Answer

- Simple GRU to produce an output at each of its time steps.
- Allow to predict EOS token and stop

$$a_t = \text{GRU}([y_{t-1}, q], a_{t-1}), \quad y_t = \text{softmax}(W^{(a)}a_t)$$
Putting it all together

- Training via cross-entropy errors and backpropagation

Semiotic Memory Module

(Semantic Memory Module)

(Glove vectors)

Episodic Memory Module

Answer module

Input Module

Question Module

<table>
<thead>
<tr>
<th>e_1'</th>
<th>e_2'</th>
<th>e_3'</th>
<th>e_4'</th>
<th>e_5'</th>
<th>e_6'</th>
<th>e_7'</th>
<th>e_8'</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>e_1</th>
<th>e_2</th>
<th>e_3</th>
<th>e_4</th>
<th>e_5</th>
<th>e_6</th>
<th>e_7</th>
<th>e_8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>s_1</th>
<th>s_2</th>
<th>s_3</th>
<th>s_4</th>
<th>s_5</th>
<th>s_6</th>
<th>s_7</th>
<th>s_8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary got the milk there.</td>
<td>John moved to the bedroom.</td>
<td>Sandra went back to the kitchen.</td>
<td>Mary travelled to the hallway.</td>
<td>John got the football there.</td>
<td>John put down the football.</td>
<td>Mary went to the garden.</td>
<td>Where is the football?</td>
</tr>
</tbody>
</table>

Training via cross-entropy errors and backpropagation
**Tasks with results above or near state of the art**

<table>
<thead>
<tr>
<th>Type</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA</td>
<td>babI - Facebook</td>
</tr>
<tr>
<td>Sequence</td>
<td>POS</td>
</tr>
<tr>
<td>Classification</td>
<td>Sentiment</td>
</tr>
<tr>
<td>Sequence</td>
<td>NER</td>
</tr>
<tr>
<td>MT</td>
<td>English-French</td>
</tr>
<tr>
<td>Coref</td>
<td>Guha et al. 2015</td>
</tr>
</tbody>
</table>
Details: QA on babI, POS and Sentiment

<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|

<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>97.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>MV-RNN</th>
<th>RNTN</th>
<th>DCNN</th>
<th>PVec</th>
<th>CNN-MC</th>
<th>DRNN</th>
<th>CT-LSTM</th>
<th>DMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary</td>
<td>82.9</td>
<td>85.4</td>
<td>86.8</td>
<td>87.8</td>
<td>88.1</td>
<td>86.6</td>
<td>88.0</td>
<td>88.3</td>
</tr>
<tr>
<td>Fine-grained</td>
<td>44.4</td>
<td>45.7</td>
<td>48.5</td>
<td>48.7</td>
<td>47.4</td>
<td>49.8</td>
<td>51.0</td>
<td>50.3</td>
</tr>
</tbody>
</table>
Dynamic Memory Network by MetaMind

Story
- wolves are afraid of mice.
- sheep are afraid of mice.
- winona is a sheep.
- mice are afraid of cats.
- cats are afraid of wolves.
- jessica is a mouse.
- emily is a cat.
- gertrude is a wolf.

Question
- what is winona afraid of?

Run DMN  Get new example

Answer: mouse

Episode 1
- 0.00: wolves are afraid of mice
- 0.00: sheep are afraid of mice
- 0.99: winona is a sheep
- 0.00: mice are afraid of cats
- 0.00: cats are afraid of wolves
- 0.00: jessica is a mouse
- 0.00: emily is a cat
- 0.01: gertrude is a wolf

Episode 2
- 0.00: wolves are afraid of mice
- 1.00: sheep are afraid of mice
- 0.00: winona is a sheep
- 0.00: mice are afraid of cats
- 0.00: cats are afraid of wolves
- 0.00: jessica is a mouse
- 0.00: emily is a cat
- 0.00: gertrude is a wolf

Model hidden state
Summary

- All (?) NLP tasks can be reduced to question answering
- The DMN can very accurately train with input-question-answer triplets
- Next steps: One very large multitask DMN