A Read-Write Memory Network for Movie Story Understanding

Barış Çağlar Ayyıldız
1941798
Outline

- Introduction - Related work
- Read-Write Memory Network (RWMN)
- Experiments – Results
- Conclusion
Introduction

• MovieQA challenge
  - Understand movies over two hours long and answer questions about the movie content.
  - Six tasks according to source of information about the content.
    • **Video + subtitles**
    • Subtitles only
    • DVS only
    • Scripts only
    • Plot synopses only
    • Open-ended
Introduction (cont’d.)

- Trinity contacts him confirming that Morpheus can…
- Neo meets Morpheus,…
- The Matrix is revealed to be a shared simulation of the world as it was in 1999…
- … secretly betrayed Morpheus to Agent Smith in exchange for a comfortable…
- Morpheus and Trinity exit the Matrix, but Smith ambushes and kills Neo before he can…
- He ends the call and flies into the sky.

Quiz

<table>
<thead>
<tr>
<th>What is the Matrix?</th>
<th>Who kills Neo in the Matrix?</th>
<th>Why does Cypher betray Morpheus?</th>
<th>How does the movie end?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: A shared simulation of the world</td>
<td>A: Smith kills Neo</td>
<td>A: In exchange for a comfortable life</td>
<td>A: With Neo flying into the sky</td>
</tr>
<tr>
<td>A: A group of robots</td>
<td>A: Trinity kills Neo</td>
<td>A: In exchange for money</td>
<td>A: With the Machines chasing after Neo</td>
</tr>
<tr>
<td>A: A human body</td>
<td>A: Morpheus kills Neo after he realizes that Neo is not the one</td>
<td>A: Because he is threatened by Agent Smith</td>
<td>A: We see Mr. Smith torture Morpheus</td>
</tr>
<tr>
<td>A: A set of numbers stored as a table</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Introduction (cont’d.)

• Understanding a movie is highly challenging.

• It is not enough to understand the content of individual frames like actions of the characters or places.

• Need to infer more abstract and high-level knowledge such as reasons of a characters’ behaviors, and relationships between them.

• What kind of methods do we need?
Introduction (cont’d.)

• QA problems generally require memorizing a large amount of information, and correctly access the most relevant information to a given question.

• Processing, representing, and storing long sequential information
  – RNNs (LSTM, GRU)
    • Store information into a fixed length hidden memory(state).
    • Fail to utilize far-distant information.
  – Neural Memory Networks
    • Cache sequential inputs in memory slots, and explicitly utilize even far early information.
Introduction (cont’d.)

• Memory Networks
  - Input feature map (I)
  - Generalization (G)
  - Output feature map (O)
  - Response (R)

• Supporting fact + query

Fred moved to the bedroom.
Joe went to the kitchen.
Joe took the milk.
Dan journeyed to the bedroom.

Where is the milk now?
Introduction (cont’d.)

- End-to-End Memory Networks
  - Soft Attention
Introduction (cont’d.)

- **Neural Turing Machines**
  - Rewritable memory by controller.
  - Different memory image at every time step.
  - Location based vs. Content based memory access.
Introduction (cont’d.)

- Paper proposes a novel memory network model to perform QA tasks for large-scale, multimodal movie story understanding: **Read-Write Memory Networks (RWMN)**
- What is different in terms of **memory**?
• Existing NMN models treat each memory slot as independent block. However, adjacent memory blocks often have strong correlations.

• When we understand a story or a movie, we generally recognize it as a sequence of closely interconnected abstract events.

• **Memory networks need to read and write sequential memory cells as chunks.**
Introduction (cont’d.)

• Existing NMN models treat each memory slot as independent block. However, adjacent memory blocks often have strong correlations.

• When we understand a story or a movie, we generally recognize it as a sequence of closely interconnected abstract events.

• **Memory networks need to read and write sequential memory cells as chunks.**
Read-Write Memory Network (RWMN)

- A novel memory network which has a flexible and high capacity read and write networks.
- Trained to store the movie content with proper representation in the memory, extract relevant information from memory cells in response to a given query, and select correct answer from five choices.
Read-Write Memory Network (RWMN) (cont’d.)

• Using read/write networks of multi-layered CNNs, it abstracts a given series of frames stepwise to capture higher-level sequential information and stores it into memory slots. It eventually helps answer complex questions of movie QAs.
RWMN Input-Output format

- Based on MovieQA specifications
- Input format
  - A sequence of video segment and subtitle pairs.
    \[ S_{\text{movie}} = \{(v_1, s_1), \ldots, (v_n, s_n)\} \] (n ~ 1558 on average)
    \[ v_i = \{v_{i1}, \ldots, v_{im}\} \] (frames sampled at 6 fps)
  - A question q for the movie.
  - 5 answer candidates
    \[ a = \{a_1, \ldots, a_5\} \]
Movie Embedding

- Converting subshot $v_i$
  - For each frame $v_{ij} \in v_i$ obtain feature vector $v_{ij}$ by applying ResNet-152 pretrained on ImageNet.
  - Then mean-pool over all frames $v_i = \sum_j v_{ij} \in \mathbb{R}^{7 \times 7 \times 2,048}$

- Converting subtitle $s_i$
  - Divide sentence into words.
  - Apply pretrained Word2Vec. (ndim = 300)
  - Mean-pool with position encoding $s_i = \sum_j \text{PE}(s_{ij}) \in \mathbb{R}^{300}$

$$\text{PE}(s_{ij}) = s_{ij} \cdot ((1 - j/J) - (i/I) \cdot (1 - 2j/J))$$
Movie Embedding (cont’d.)

• In order to get a multimodal embedding of $s_i$ and $v_i$, Compact Linear Pooling (CLP) is used.

$$E[i] = CBP(v_i, s_i) \in \mathbb{R}^{4,096}$$

• Performing this for all $n$ pairs, we get 2D movie embedding matrix $E \in \mathbb{R}^{n \times 4,096}$ which is input to write network.
The Write Network (cont’d.)

Algorithm 1 Multimodal Compact Bilinear

1: input: $v_1 \in \mathbb{R}^{n_1}, v_2 \in \mathbb{R}^{n_2}$
2: output: $\Phi(v_1, v_2) \in \mathbb{R}^d$
3: procedure MCB($v_1, v_2, n_1, n_2, d$)
4:     for $k \leftarrow 1 \ldots 2$ do
5:         if $h_k, s_k$ not initialized then
6:             for $i \leftarrow 1 \ldots n_k$ do
7:                 sample $h_k[i]$ from $\{1, \ldots, d\}$
8:                 sample $s_k[i]$ from $\{-1, 1\}$
9:     $v'_k = \Psi(v_k, h_k, s_k, n_k)$
10:    $\Phi = \text{FFT}^{-1}(\text{FFT}(v'_1) \odot \text{FFT}(v'_2))$
11: return $\Phi$

procedure $\Psi(v, h, s, n)$
13: $y = [0, \ldots, 0]$
14: for $i \leftarrow 1 \ldots n$ do
15: $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$
16: return $y$
Takes movie embedding matrix $E$ and generates memory tensor $M$

Paper proposes a memory model in which each memory cell associates neighboring movie embeddings instead of storing each of $n$ embeddings separately using a CNN as the write network.

Write network pipeline:
- Apply a Fully Connected layer to $E \in \mathbb{R}^{n \times 4,096}$ in order to project each $E[i]$ into $d$ dimensional vector $E \in \mathbb{R}^{n \times d}$
- Apply Convolutional layer with three filters with kernel size $(40,d)$ and stride $(30,1)$

$$M = \text{ReLU}(\text{conv}((EW_c + b_c), w^w_{\text{conv}}, b_w)), \quad M \in \mathbb{R}^{m \times d \times 3}$$
where $m = \left\lceil \frac{(n - 1)}{s^w_v} + 1 \right\rceil$
The Read Network

- Takes question $q$ and generates answer from a compatibility between $q$ and $M$.
  - **Question $q$ embedding**
    - Divide question into words.
    - Apply pretrained *Word2Vec*. (ndim = 300)
    - Project $q \in \mathbb{R}^{300 \times 1}$ into $d$ dimensional vector $u$.

\[
  u = W_q q + b_q \quad \text{where} \quad W_q \in \mathbb{R}^{d \times 300} \quad \text{and} \quad b_q \in \mathbb{R}^d
\]
The Read Network (cont’d.)

• Next the read network takes the memory $M$ and the query embedding $u$ as input, and generates the confidence score vector $o \in \mathbb{R}^d$

  - **Query-dependent memory embedding**
    
    • Transform memory $M$ to be query dependent as different types of questions require different types of information retrieved from memory slots.

    • To transform $M$ to be query dependent, CBP between each memory cell of $M$ and $u$ is applied.

    $$M_q[i, :, j] = CBP(M[i, :, j], u) \text{ for } i = 1, \cdots, m, \text{ and } j = 1, 2, 3.$$
In order to connect and relate a series of scenes as a whole, CNN architecture again used to associate query dependent sequential memory cells.

- **Convolutional memory read**
  - Convolutional layer is applied to \( \mathbf{M}_q \in \mathbb{R}^{m \times d \times 3} \) with three filters with kernel size \((3,d)\) and stride \((1,1)\).
  - The reconstructed memory is \( \mathbf{M}_r \in \mathbb{R}^{c \times d \times 3} \) where \( c = \left\lfloor \frac{(m - 1)}{s^r} + 1 \right\rfloor \)

\[
\mathbf{M}_r = \text{ReLU}(\text{conv}(\mathbf{M}_q, \mathbf{w}_{\text{conv}}, \mathbf{b}_r))
\]
Answer Selection

• Compute attention matrix \( p \in \mathbb{R}^{c \times 3} \) by applying the softmax to the dot product between the query embedding \( u \) and each memory cell of \( M_r \)

\[
p[i, j] = \text{softmax}(M_r[i, :, j] \cdot u)\
\]

• Finally, the output vector \( o \in \mathbb{R}^d \) is calculated through a weighted sum between each memory cell of \( M_r \) and attention vector \( p \).

\[
o[i] = \sum_{j=1}^{c} \sum_{k=1}^{3} M_r[j, i, k] p[j, k].
\]
Answer Selection (cont’d.)

- Finally, embedding of five answer sentences are obtained and projected into $d$ dimension as before with same parameters $W_q$ and $b_q$. We get answer embedding matrix $g \in \mathbb{R}^{5 \times d}$.
- Confidence vector $z \in \mathbb{R}^5$ is calculated by finding similarity between $g$ and the weighted sum of $o$ and $u$
  
  $$z = \text{softmax}((\alpha o + (1 - \alpha)u)^T g)$$

- Predict the answer $y$ with highest confidence score.
  
  $$y = \arg\max_{i \in [1, 5]}(z_i)$$
Overall Network Architecture

(a) Movie Embedding
(b) Write Network
(c) Read Network
(d) Question Answering
Training

• Minimize the **Softmax cross-entropy** between the prediction $\mathbf{z}$ and the ground-truth one-hot vector $\mathbf{z}_{gt}$. 
Experiments

- There are 408 movies and 14,944 multiple choice QA pairs, each of which consists of five answer choices with only one correct answer in the dataset.
- Five types of story sources associated with movie.

<table>
<thead>
<tr>
<th>Story sources</th>
<th># movie</th>
<th># QA pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Videos and subtitles</td>
<td>140</td>
<td>6,462</td>
</tr>
<tr>
<td>Subtitles</td>
<td>408</td>
<td>14,944</td>
</tr>
<tr>
<td>DVS</td>
<td>60</td>
<td>2,446</td>
</tr>
<tr>
<td>Scripts</td>
<td>199</td>
<td>7,810</td>
</tr>
<tr>
<td>Plot synopses</td>
<td>408</td>
<td>14,944</td>
</tr>
</tbody>
</table>
Experiments (cont’d.)

- MovieQA Challenge tasks
  - Video + subtitle → The only VQA
  - Subtitles only
  - DVS only
  - Scripts only
  - Plot synopses only → little movie understanding
  - Open-ended

- For text-only QA tasks, movie embedding $E$ is constructed only with textual $\{s_1, ..., s_n\}$ information without CBP.
Experiments (cont’d.)

- Baselines
  - All the methods proposed in the original MovieQA paper or in the official MovieQA leaderboard.

- Seven variants of RWMN to observe the effect of key components
  - RWMN-noRW
  - RWMN-noR
  - RWMN-noQ
  - RWMN-noVid
  - RWMN
  - RWMN-bag
  - RWMN-ensemble
## Quantitative Results

- **Results of VQA task**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Video+Subtitle val</th>
<th>Video+Subtitle test</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVQAP</td>
<td>–</td>
<td>23.61</td>
</tr>
<tr>
<td>Simple MLP</td>
<td>–</td>
<td>24.09</td>
</tr>
<tr>
<td>LSTM + CNN</td>
<td>–</td>
<td>23.45</td>
</tr>
<tr>
<td>LSTM + Discriminative CNN</td>
<td>–</td>
<td>24.32</td>
</tr>
<tr>
<td>VCFSM</td>
<td>–</td>
<td>24.09</td>
</tr>
<tr>
<td>DEMN</td>
<td>–</td>
<td>29.97</td>
</tr>
<tr>
<td>MEMN2N</td>
<td>34.20</td>
<td>–</td>
</tr>
<tr>
<td>RWMN-noRW</td>
<td>34.20</td>
<td>–</td>
</tr>
<tr>
<td>RWMN-noR</td>
<td>36.50</td>
<td>–</td>
</tr>
<tr>
<td>RWMN-noQ</td>
<td>38.17</td>
<td>–</td>
</tr>
<tr>
<td>RWMN-noVid</td>
<td>37.20</td>
<td>–</td>
</tr>
<tr>
<td>RWMN</td>
<td><strong>38.67</strong></td>
<td><strong>36.25</strong></td>
</tr>
<tr>
<td>RWMN-bag</td>
<td>38.37</td>
<td>35.69</td>
</tr>
<tr>
<td>RWMN-ensemble</td>
<td>38.30</td>
<td>–</td>
</tr>
</tbody>
</table>
Quantitative Results (cont’d.)

- Results of text-only tasks.
  - Script (2877 sentences on average.)
  - Subtitle (1558 sentences on average.)
  - DVS (636 sentences on average.)
  - Plot Synopses (35 sentences on average.)

<table>
<thead>
<tr>
<th>Method</th>
<th>Subtitle</th>
<th>Script</th>
<th>DVS</th>
<th>Plot Synopses</th>
<th>Open-end</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>val</td>
<td>test</td>
<td>val</td>
<td>test</td>
<td></td>
</tr>
<tr>
<td>MEMN2N</td>
<td>38.0</td>
<td>36.9</td>
<td>42.3</td>
<td>37.0</td>
<td>40.6</td>
</tr>
<tr>
<td>SSCB-W2V</td>
<td>24.8</td>
<td>23.7</td>
<td>25.0</td>
<td>24.4</td>
<td>45.1</td>
</tr>
<tr>
<td>SSCB-TF-IDF</td>
<td>27.6</td>
<td>26.5</td>
<td>26.1</td>
<td>23.9</td>
<td>48.5</td>
</tr>
<tr>
<td>SSCB Fusion</td>
<td>27.7</td>
<td>–</td>
<td>28.7</td>
<td>–</td>
<td>56.7</td>
</tr>
<tr>
<td>CNN Word Matching</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>72.1</td>
</tr>
<tr>
<td>Convnet Fusion (TF-IDF + Word2Vec)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>77.6</td>
</tr>
<tr>
<td>Longest Answer</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>RWMN</td>
<td><strong>40.4</strong></td>
<td><strong>38.5</strong></td>
<td><strong>44.0</strong></td>
<td><strong>39.4</strong></td>
<td><strong>34.2</strong></td>
</tr>
</tbody>
</table>

*Bold values indicate the best performance in each task.*
## Ablation Results

- Results of VQA task

<table>
<thead>
<tr>
<th># Layers</th>
<th>Write network $(f_{vi}^w, s_{vi}^w, f_{ci}^w)$</th>
<th>Read network $(f_{vi}^r, s_{vi}^r, f_{ri}^r)$</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
<td>–</td>
<td>–</td>
<td>34.2</td>
</tr>
<tr>
<td>1 0</td>
<td>(40,7,1)</td>
<td>–</td>
<td>33.9</td>
</tr>
<tr>
<td>1 0</td>
<td>(40,30,3)</td>
<td>–</td>
<td>36.5</td>
</tr>
<tr>
<td>1 1</td>
<td>(40,30,3)</td>
<td>(3,1,1)</td>
<td><strong>38.6</strong></td>
</tr>
<tr>
<td>1 1</td>
<td>(40,60,3)</td>
<td>(3,1,1)</td>
<td>33.6</td>
</tr>
<tr>
<td>2 1</td>
<td>(40,10,3), (10,5,3)</td>
<td>(3,1,1)</td>
<td>37.2</td>
</tr>
<tr>
<td>2 1</td>
<td>(5,3,1), (5,3,1)</td>
<td>(3,1,1)</td>
<td>37.3</td>
</tr>
<tr>
<td>2 2</td>
<td>(4,2,1), (4,2,1)</td>
<td>(3,1,1), (3,1,1)</td>
<td>36.9</td>
</tr>
<tr>
<td>2 2</td>
<td>(4,2,1), (4,2,1)</td>
<td>(4,2,1), (4,2,1)</td>
<td>37.3</td>
</tr>
<tr>
<td>3 1</td>
<td>(10,3,3), (40,3,3), (100,3,3)</td>
<td>(3,1,1)</td>
<td>35.1</td>
</tr>
<tr>
<td>3 1</td>
<td>(40,3,3), (10,3,3), (10,3,3)</td>
<td>(3,1,1)</td>
<td>37.9</td>
</tr>
<tr>
<td>3 1</td>
<td>(40,3,3), (40,3,3), (40,3,3)</td>
<td>(3,1,1)</td>
<td>35.7</td>
</tr>
<tr>
<td>3 1</td>
<td>(100,3,3), (40,3,3), (10,3,3)</td>
<td>(3,1,1)</td>
<td>35.8</td>
</tr>
</tbody>
</table>
Ablation Results (cont’d.)

- MEMN2N and our RWMN model according to question types in the video+subtitle task.
Qualitative Results

Q. Why does Amy’s disappearance receive heavy press coverage?
0. Because her parents are popular
1. Because Amy was the inspiration for the popular “Amazing Amy” children books
2. Because Amy is a popular actress
3. Because it happened on the day of her wedding anniversary
4. Because her husband is popular

Q. Where does the Joker set a trap for Vicki?
0. At the Gotham Museum of Art
1. At her house
2. At Gotham Police Station
3. At the Gotham Museum of History
4. At Bruce’s mansion

Q. What does Gandalf learn from Pippin’s visions?
1. Sauron will attack Minas Tirith
2. Sauron will hide in Minas Tirith
3. Sauron will attack Erebor
4. Sauron will attack The Shire
5. Sauron will flee from Minas Tirith

Q. How does Travis think Miley knows Hannah Montana?
0. He thinks that Miley and Hannah are friends from school
1. He thinks that Hannah saved Miley’s life in a surfing accident
2. He thinks that Miley and Hannah are cousins
3. He thinks that Miley saved Hannah’s life in a car accident
4. He thinks that Miley saved Hannah’s life in a car accident
Qualitative Results (cont’d.)

Q. Why did Lillian run away from her wedding?
A1. Because she spilled something on her dress right before the ceremony and was too embarrassed of everyone seeing
A2. Because of Annie's extravagant planning and out of fear of leaving her life in Milwaukee
A3. Because it didn't feel right without Annie there
A4. No reason in particular
A5. Because of Helen's extravagant planning and out of fear of leaving her life in Milwaukee

Q. How does Forrest get admitted to public school despite his low IQ?
[0] His mother agrees to pay more money
[1] His mother agrees to a one night stand with the school principal
[2] He gets a football scholarship because he runs very fast
[3] His mother begs the principal and he takes mercy on her
[4] Forrest is very good in football so the school accepts him on this account
Q. Why did Lillian run away from her wedding?
A1. Because she spilled something on her dress right before the ceremony and was too embarrassed of everyone seeing
A2. Because of Annie’s extravagant planning and out of fear of leaving her life in Milwaukee
A3. Because it didn’t feel right without Annie there
A4. No reason in particular
A5. Because of Helen’s extravagant planning and out of fear of leaving her life in Milwaukee

Q. How does Forrest get admitted to public school despite his low IQ?
[0] His mother agrees to pay more money
[1] His mother agrees to a one night stand with the school principal
[2] He gets a football scholarship because he runs very fast
[3] His mother begs the principal and he takes mercy on her
[4] Forrest is very good in football so the school accepts him on this account
Conclusion

- A novel memory network, RWMN, which has CNN-based read/write network that enable the model to have highly-capable and flexible read/write operations is proposed.
- Performance on visual question answering tasks for large-scale, multimodal movie story understanding is validated.
Conclusion

- A novel memory network, RWMN, which has CNN-based read/write network that enable the model to have highly-capable and flexible read/write operations is proposed.

- Performance on visual question answering tasks for large-scale, multimodal movie story understanding is validated.
Thank you!

Any questions?