Attention-based Model
Hung-yi Lee
Attention

- Reasoning, memory
- Human’s memory
- Working memory

Information from the sensors (e.g. eyes, ears)

- Sensory Memory
- Attention
- Working Memory
- Encode
- Long-term Memory
- Retrieval
Attention on Sensory Information

Information from the sensors (e.g. eyes, ears)

- Sensory Memory
  - Attention
  - Encode
  - Retrieval
  - Long-term Memory
Machine Translation

- Sequence to sequence learning: Both input and output are both sequences *with different lengths*.
- E.g. 機器學習 → machine learning
Machine Translation

- Attention-based model

What is **match**?

- Cosine similarity of $z$ and $h$
- Small NN whose input is $z$ and $h$, output a scalar
- $\alpha = h^T W z$

How to learn the parameters?
Machine Translation

• Attention-based model

\[
c^0 = \sum \hat{\alpha}_0^i h^i = 0.5h^1 + 0.5h^2
\]

How to learn the parameters?
Machine Translation

• Attention-based model

\[ \alpha_1^1 \]

\[ z^0 \rightarrow z^1 \]

\[ c^0 \]
Machine Translation

- Attention-based model

\[ c^1 = \sum \hat{\alpha}_1^i h^i \]

\[ = 0.5h^3 + 0.5h^4 \]
Machine Translation

- Attention-based model

The same process repeat until generating

```
matched

h^1 \rightarrow h^2 \rightarrow h^3 \rightarrow h^4
```

```
\alpha_2^1
```

```
machine \rightarrow learning
```

```
z^0 \rightarrow z^1 \rightarrow z^2
```

```
c^0 \rightarrow c^1
```

```
======
```

```
機 器 學 習
```

```
……
```

```
……
```

```
……
```

```
=====
```
Speech Recognition

William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, “Listen, Attend and Spell”, arXiv’15
Image Caption Generation

• Input an image, but output a sequence of words

A vector for whole image

Input image

CNN

a woman is

.......
Image Caption Generation

A vector for each region

CNN

$z^0 \xrightarrow{\text{match}} 0.7$
Image Caption Generation

A vector for each region

CNN

weighted sum

0.7 0.1 0.1 0.1

z^0  z^1

Word 1
Image Caption Generation

A vector for each region

Word 1
Word 2

\( z^0 \) \( \rightarrow \) \( z^1 \) \( \rightarrow \) \( z^1 \)

weighted sum

0.0 0.8 0.2

0.0 0.0 0.0 0.0
Image Caption Generation

• Good captions

A woman is throwing a frisbee in a park. A dog is standing on a hardwood floor. A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear. A group of people sitting on a boat in the water. A giraffe standing in a forest with trees in the background.
Image Caption Generation

• Bad captions

A large white **bird** standing in a forest.

A woman holding a **clock** in her hand.

A man wearing a hat and a hat on a **skateboard**.

A person is standing on a beach with a **surfboard**.

A woman is sitting at a table with a large **pizza**.

A man is talking on his cell phone while another man watches.
Ref: A man and a woman ride a motorcycle

A **man** and a **woman** are **talking** on the **road**

Ref: A woman is frying food

Someone is **frying** a **fish** in a **pot**
Reading Comprehension

Sentence to vector can be jointly trained.

Document → Match → q

Extracted Information

\[ \sum_{n=1}^{N} \alpha_n x^n \]

Match → q → Answer

DNN

Please refer to the lecture on 2015/12/04
$$\sum_{n=1}^{N} \alpha_n h^n = x_n$$

Please refer to the lecture on 2015/12/04
Memory Network

• Performance of Hopping

<table>
<thead>
<tr>
<th>Story (16: basic induction)</th>
<th>Support</th>
<th>Hop 1</th>
<th>Hop 2</th>
<th>Hop 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brian is a frog.</td>
<td>yes</td>
<td>0.00</td>
<td>0.98</td>
<td>0.00</td>
</tr>
<tr>
<td>Lily is gray.</td>
<td></td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Brian is yellow.</td>
<td>yes</td>
<td>0.07</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Julius is green.</td>
<td></td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Greg is a frog.</td>
<td>yes</td>
<td>0.76</td>
<td>0.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

What color is Greg? Answer: yellow Prediction: yellow

Demo video: https://www.facebook.com/Engineering/videos/10153098860532200/
Special Attention: Spatial Transformers

Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu, Spatial Transformer Networks, arXiv’15

Jointly learned
Attention on Memory

Information from the sensors (e.g. eyes, ears)

- Sensory Memory
- Working Memory
- Long-term Memory

Encode ➔ Attention ➔ Retrieval

- Attention
Neural Turing Machine

• von Neumann architecture

Actually, Neural Turing Machine is an advanced RNN/LSTM.

https://www.quora.com/How-does-the-Von-Neumann-architecture-provide-flexibility-for-program-development
Neural Turing Machine

\[ r^0 = \sum \hat{\alpha}_0^i m_0^i \]

Retrieval process

Long term memory
Neural Turing Machine

\[ r^0 = \sum \hat{\alpha}^i_0 m^i_0 \]

(simplified)

\[ \alpha^i_1 = (1 - \lambda)\alpha^i_0 + \cos(m^i_0, k^1) \]
Neural Turing Machine

\[ m^i_1 = m^i_0 \times \begin{pmatrix} 1 & -\hat{\alpha}^i_1 & e^1 \end{pmatrix} + \hat{\alpha}^i_1 \alpha^1 \]

(element-wise)

Encode process
Wei Zhang, Yang Yu, Bowen Zhou, Structured Memory for Neural Turing Machines, arXiv’15
Stack RNN

0.7 0.2 0.1
Push, Pop, Nothing

Information to store

Push -1 -1
Pop -1
Nothing -1

Armand Joulin, Tomas Mikolov, Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets, 2015
Concluding Remarks

Information from the sensors (e.g. eyes, ears)

Sensory Memory

Attention

Working Memory

Encode

Retrieval

Long-term Memory

Translation
Speech Recognition
Caption Generation
Question Answering

Neural Turing Machine, Stack RNN
Reference


• Neural Turing Machines. Alex Graves, Greg Wayne, Ivo Danihelka. arXiv Pre-Print, 2014

• Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. Kumar et al. arXiv Pre-Print, 2015


• A Neural Attention Model for Abstractive Sentence Summarization. A. M. Rush, S. Chopra and J. Weston. EMNLP 2015.
Plan

• 1/8 (五) 23:59: Presentation team decided
• 1/13 (三) 23:59: Presentation slides deadline
• 1/15 (五)
  • 上課時間：Presentation
  • 返鄉投票
• 1/16 (六) : 投票
• 1/20 (三) 23:59: Report deadline
Teaching Machines to Read and Comprehend, Hermann et. al. (2015)