Seq2Seq - Attention

Renowned NLP architecture and mechanism

Content

1 RNN Revision

2 Seq2Seq architecture

3 Attention mechanism

RNN Revision

ANN

- Input: fixed-size

RNN Revision

- RNN: an architecture tailored for sequence data:

1) Doesn't depend on data length

2) Take advantage of past information

RNN

- Input: variable-size
- «Memory»

$$h_t = f_W(h_{t-1}, x_t)$$

$$\downarrow$$

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$$

$$y_t = W_{hy} h_t$$

1. Recurrent Neural Network

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

RNN Revision

RNN Revision

Seq2Seq architecture

Intuition

- Take machine translation task as an example: human would first read some parts of the text and then start to do the translation

The cat likes to eat pizza

el gato le gusta comer pizza

Seq2Seq architecture

-NLP researchers also employ that **idea** into designing a structure dubbed as Seq2Seq

- It consists of 2 **components**:

1) Encoder

2) Decoder

Teacher Forcing

-This is the architecture fully unfolded

-To train this architecture, we apply a technique called **Teacher Forcing**

Teacher Forcing

Without Teacher Forcing

Teacher Forcing

Encoder
Decoder
Single LSTM Cell
A
I
went
home
</s>
Ich
ging
nach
Hause
Softmax
</s>
Ich
ging
nach

target output words
Je
suis
étudiant
</s>
loss layer
|V|
projection layer
hidden layer 2
hidden layer 1
embedding layer
I
am
a
student
<s>
Je
suis
étudiant
encoding
decoding

[statically prepare the decoder input]

[dynamically insert the decoder input]

Attention mechanism

Alignment

-In sequence generation task, especially machine translation, some words in the **output** are **aligned** with specific phrases from the **input**

- Moreover, LSTM and GRU tend to **fail** to solve the dependency cases for **extremely long** texts

-> Researchers introduce a new technique to resolve named as **ATTENTION**

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.

Alignment

1
2
3
4
5
6
And
the
programme
has
been
implemented
Le
programme
a
été
mis
en
application
1
2
3
4
5
6
7

Attention mechanism

Attention mechanism

Attention mechanism

-Core idea: on each step of the decoder, we attend the current token to focus on a **particular part** of the sequence

- Attention helps solve the **bottleneck** problem of LSTM and GRU

Sequence-to-sequence: the bottleneck problem

Encoding of the source sentence. This needs to capture *all information* about the source sentence. Information bottleneck!

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Use the attention distribution to take a **weighted sum** of the encoder hidden states.

The attention output mostly contains information from the hidden states that received high attention.

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Sometimes we take the attention output from the previous step, and also feed it into the decoder (along with the usual decoder input). We do this in Assignment 4.

Decoder RNN

$$\alpha_{ts} = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \qquad \text{[Attention weights]} \qquad (1)$$

$$\boldsymbol{c}_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_s \qquad \text{[Context vector]} \qquad (2)$$

$$\boldsymbol{a}_t = f(\boldsymbol{c}_t, \boldsymbol{h}_t) = \tanh(\boldsymbol{W}_{\boldsymbol{c}}[\boldsymbol{c}_t; \boldsymbol{h}_t]) \qquad \text{[Attention vector]} \qquad (3)$$

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_{\boldsymbol{a}} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_{\boldsymbol{a}}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Sequence-to-sequence with attention

Decoder
Attention layer
decoder
hidden state
encoder
hidden state
Encoder

Decoder
Attention layer
decoder
hidden state
score
score
score
score
encoder
hidden state
Encoder

Decoder
Attention layer
multiplication
multiplication
multiplication
multiplication
softmax
softmax
softmax
softmax
score
score
score
score
decoder
hidden state
encoder
hidden state
Encoder

Decoder
Attention layer
addition
context vector
multiplication
multiplication
multiplication
multiplication
softmax
softmax
softmax
softmax
score
score
score
score
decoder
hidden state
encoder
hidden state
Encoder

Attention mechanism

- Attention could be divided based on various criteria:

1) Region

- Local attention
- Global attention

Attention mechanism

- Attention could be divided based on various criteria:

1) Region

- Local attention
- Global attention

2) Method

- Hard attention
- Soft attention

Summary

1) Seq2Seq architecture is great

2)Attention seems complicated, but fun to apply for tasks in real life like text summarization, machine translation, etc.

References

1)[Thang Luong et al. 2015] Effective Approaches to Attention-based Neural Machine Translation. EMNLP 2015.

2)[Sutskever et al. 2014] Sequence to sequence learning with

neural networks. In NIPS.

3) [d2l.ai] Dive into Deep Learning