ECE599/692 - Deep Learning

Lecture 17 – Attention!

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
http://www.eecs.utk.edu/faculty/qi
Email: hqi@utk.edu

References

- K. Xu, et al. Show, attend and tell: Neural image caption generation with visual attention. ICML 2015.
- Vaswani, et al. Attention Is All You Need. NIPS, 2017

Outline

- What is Attention?
- Why Attention?
- How does Attention work?
- Self-Attention
- Where is Attention used?
What is Attention?

Visual attention to different regions of an image or ...


What is Attention?

... correlate words in one sentence.

She is eating a green apple.

- Widely used in NLP (machine translation).
- Allows machine translator to look over all the information the original sentence holds, locally or globally.

What is Attention?

- In a nutshell, attention in the deep learning can be broadly interpreted as a vector of importance weights: in order to predict or infer one element, we estimate using the attention vector how strongly it is correlated with (or “attends to”) other elements and take the sum of their values weighted by the attention vector as the approximation of the target

Outline

• What is Attention?
• Why Attention?
• How does Attention work?
• Self-Attention
• Where is Attention used?

Why Attention? – The Seq2Seq Model

• RNN: can map sequences to sequences whenever the alignment between the inputs and the outputs is known
• What if the input and the output sequences have different lengths with complicated and non-monotonic relationships?
• General sequence learning:
  – Map the input sequence to a fixed-sized vector using one RNN
  – Map the vector to the target sequence with another RNN


Why Attention?

• Drawbacks of Seq2Seq model
  – Fixed length context vector
  – Difficulty in modeling long dependency
  – Gradient vanishing/exploding, hard to train when sentences are long

Encoder

Decoder

Context vector (length 5)

[0.1, -0.2, 0.8, 1.5, -0.3]
Why Attention?

Outline

- What is Attention?
- Why Attention?
- How does Attention work?
- Self-Attention
- Where is Attention used?
How does Attention work?

- Context vector takes all cells' outputs as input to compute the probability distribution of source language words for EACH target word.

  Capture global information

\[
\alpha_t = \frac{\exp(\text{score}(h_t, h_s))}{\sum_{s=1}^{S} \exp(\text{score}(h_t, h_s))}
\]

[Attention weights] (1)

\[
\alpha_t = \sum_{s=1}^{S} \alpha_t h_s
\]

[Context vector] (2)

\[
\alpha_t = f(c_t, h_s) = \text{tanh}(W_s[c_t; h_s])
\]

[Attention vector] (3)

Alignment matrix of “L'accord sur l'Espace économique européen a été signé en août 1992” (French) and its English translation “The agreement on the European Economic Area was signed in August 1992”
Outline

• What is Attention?
• Why Attention?
• How does Attention work?
• Self-Attention
  – E.g., Transformer
• Where is Attention used?

Self-Attention

Self-attention, also known as intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. It has been shown to be very useful in machine reading, abstractive summarization, or image description generation.
The self-attention mechanism learns the correlation between the current word and the previous part of the sentence. The current word is in red and the size of the blue shade indicates the activation level.

Machine reading:

A woman is throwing a frisbee in a park.

K. Xu, et al. Show, attend and tell: Neural image caption generation with visual attention. ICML 2015.

[Xu et al. 2015]
Self Attention – The Transformer (No RNN or alignment needed)

Where is Attention used?

Attention is not mysterious or complex. It is just an interface formulated by parameters and delicate math. You could plug it anywhere you find it suitable, and potentially, the result may be enhanced.

- Machine translation: Attention Is All You Need
- Meta-learning: A Simple Neural Attentive Meta-Learner
- Image → Text and Text → Image: Many works
- Image generation: Self-Attention Generative Adversarial Networks
- Visual attention: Extra material

References: