 THE UNIVERSITY OF TENNESSEE KNOXVILLE

AICIP RESEARCH

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

AICIP RESEARCH

- [D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR 2015*.](#)
- [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


 THE UNIVERSITY OF TENNESSEE KNOXVILLE

2

Outline

AICIP RESEARCH

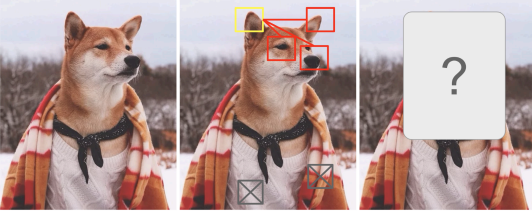
- What is Attention?
- Why Attention?
- How does Attention work?
- Self-Attention
- Where is Attention used?

 THE UNIVERSITY OF TENNESSEE KNOXVILLE

What is Attention?

AICIP
RESEARCH

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



Shiba Inu

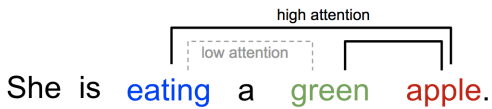
<https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>



What is Attention?

AICIP
RESEARCH

... correlate words in one sentence.



- 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?

AICIP
RESEARCH

- 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

<https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>



Outline

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

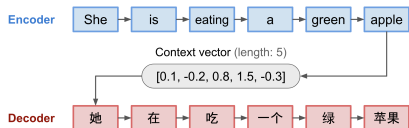
Why Attention? – The Seq2Seq Model

- 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

I. Sutskever, O. Vinyals, Q.V. Le, "Sequence to sequence learning with neural networks," NIPS 2014.

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



AICIP RESEARCH

Why Attention?

$f = (\text{La, croissance, économique, s'est, ralentie, ces, dernières, années, ...})$

$e = (\text{Economic, growth, has, slowed, down, in, recent, years, ...})$

THE UNIVERSITY OF TENNESSEE
Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015) 10

AICIP RESEARCH

Why Attention?

(Target) Y_{t-1}, Y_t

Decoder: RNN with input from previous state + dynamic context vector.

Context vec

Global alignment weights

Attention layer: parameterized by a simple feed-forward network

Additive Attention

Encoder: bidirectional RNN [Bahdanau et al. 2015]

$X_1, X_2, X_3, \dots, X_n$ (Source)

$h_1, h_2, h_3, \dots, h_r$

$\alpha_{t,1}, \alpha_{t,2}, \alpha_{t,3}, \dots, \alpha_{t,r}$

S_{t-1}, S_t

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR 2015*.

THE UNIVERSITY OF TENNESSEE

AICIP RESEARCH

Outline

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

THE UNIVERSITY OF TENNESSEE

How does Attention work?

AICIP RESEARCH

The diagram illustrates the attention mechanism for the source sentence "I am a student" and the target word "Je". The source words are processed by an encoder (green boxes) to produce hidden states. The target word "Je" is processed by a decoder (blue box) to produce a context vector. This context vector is then used to calculate attention weights for each source word. The weights are shown as 0.5 for "I", 0.3 for "am", 0.1 for "a", and 0.1 for "student". The context vector is a weighted sum of the source hidden states.

<https://medium.com/syncoedreview/a-brief-overview-of-attention-mechanism-13c578ba9129>

THE UNIVERSITY OF TENNESSEE

How does Attention work?

AICIP RESEARCH

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

The diagram illustrates the attention mechanism for the source sentence "I am a student" and the target word "Je". The source words are processed by an encoder (green boxes) to produce hidden states. The target word "Je" is processed by a decoder (blue box) to produce a context vector. This context vector is then used to calculate attention weights for each source word. The weights are shown as 0.5 for "I", 0.3 for "am", 0.1 for "a", and 0.1 for "student". The context vector is a weighted sum of the source hidden states.

$$\alpha_{ts} = \frac{\exp(\text{score}(h_t, \tilde{h}_s))}{\sum_{s'=1}^S \exp(\text{score}(h_t, \tilde{h}_{s'}))} \quad \text{[Attention weights]} \quad (1)$$

$$c_t = \sum_s \alpha_{ts} \tilde{h}_s \quad \text{[Context vector]} \quad (2)$$

$$a_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t]) \quad \text{[Attention vector]} \quad (3)$$

<https://medium.com/syncoedreview/a-brief-overview-of-attention-mechanism-13c578ba9129>

THE UNIVERSITY OF TENNESSEE

How does Attention work?

AICIP RESEARCH

The alignment matrix shows the relationship between the French sentence "L'accord sur la zone économique européenne a été signé en août 1992" and its English translation "The agreement on the European Economic Area was signed in August 1992". The matrix is a heatmap where the diagonal elements are white, indicating a strong alignment between corresponding words in the two sentences.

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"

THE UNIVERSITY OF TENNESSEE

**AICIP
RESEARCH**

T THE UNIVERSITY OF
TENNESSEE

**AICIP
RESEARCH**

Outline

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

T THE UNIVERSITY OF
TENNESSEE

**AICIP
RESEARCH**

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
animal
didn't
cross
the
street
because
it
was
too
wide
.

The
animal
didn't
cross
the
street
because
it
was
too
wide
.

The
animal
didn't
cross
the
street
because
it
was
too
wide
.

T THE UNIVERSITY OF
TENNESSEE

Self-Attention

AICIP
RESEARCH

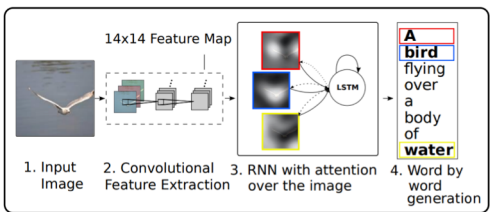
The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .

Machine reading:
 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



Self-Attention

AICIP
RESEARCH



[Xu et al. 2015]

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



Self-Attention

AICIP
RESEARCH

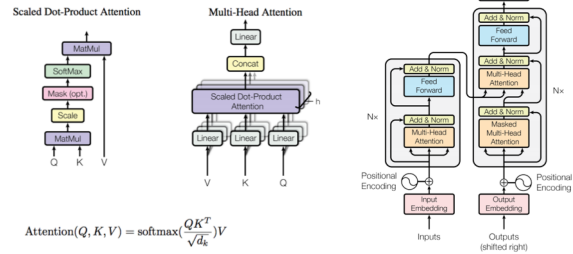


[K. Xu, et al. Show, attend and tell: Neural image caption generation with visual attention. ICML 2015.](#)



Self Attention – The Transformer (No RNN or alignment needed)

AICIP RESEARCH



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Vaswani, et al. Attention Is All You Need. NIPS, 2017

Where is Attention used?

AICIP RESEARCH

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](#)

1. Vaswani, et al. Attention Is All You Need. NIPS, 2017
2. N. Mishra, et al. A simple neural attentive meta-learner. ICLR, 2018
3. H. Zhang, I. Goodfellow, et al. Self-Attention Generative Adversarial Networks. arXiv, 2018
