Self-attention

Hung-yi Lee

李宏毅
Sophisticated Input

• Input is a vector

• Input is a set of vectors (may change length)
Vector Set as Input

**One-hot Encoding**

apple = [ 1 0 0 0 0 ...... ]

bag    = [ 0 1 0 0 0 ...... ]

cat    = [ 0 0 1 0 0 ...... ]

dog    = [ 0 0 0 1 0 ...... ]

elephant = [ 0 0 0 0 1 ...... ]

**Word Embedding**

To learn more: [https://youtu.be/X7PH3NuYW0Q](https://youtu.be/X7PH3NuYW0Q) (in Mandarin)
Vector Set as Input

10ms

25ms

1s → 100 frames

400 sample points (16KHz)
39-dim MFCC
80-dim filter bank output
Vector Set as Input

• Graph is also a set of vectors (consider each node as a vector)

Each profile is a vector
Vector Set as Input

• Graph is also a set of vectors (consider each node as a vector)

\[
H = [1 \ 0 \ 0 \ 0 \ 0 \ ...]
\]
\[
C = [0 \ 1 \ 0 \ 0 \ 0 \ ...]
\]
\[
O = [0 \ 0 \ 1 \ 0 \ 0 \ ...]
\]

One-hot vector
What is the output?

- Each vector has a label.

Example Applications

POS tagging

I saw a saw

HW2

not

buy

buy
**What is the output?**

- Each vector has a label.

- The whole sequence has a label.

**Example Applications**

- **Sentiment analysis**
  - this is good
  - positive

- **HW4**
  - speaker

- **hydrophilicity**
What is the output?

• Each vector has a label.

• The whole sequence has a label.

• Model decides the number of labels itself.
Sequence Labeling

Is it possible to consider the context?

FC can consider the neighbor

How to consider the whole sequence?

a window covers the whole sequence?
Self-attention

with context

Self-attention
Attention is all you need.

https://arxiv.org/abs/1706.03762
Self-attention

Can be either **input** or a **hidden layer**
Self-attention

Find the relevant vectors in a sequence

relevant?
Self-attention

**Dot-product**

\[ \alpha = q \cdot k \]

**Additive**
**Self-attention**

\[
\alpha_{1,2} = q^1 \cdot k^2 \\
\alpha_{1,3} = q^1 \cdot k^3 \\
\alpha_{1,4} = q^1 \cdot k^4
\]

Attention score

\[
q^1 = W^q a^1 \\
k^2 = W^k a^2 \\
k^3 = W^k a^3 \\
k^4 = W^k a^4
\]
Self-attention

\[ \alpha'_{1,i} = \frac{\exp(\alpha_{1,i})}{\sum_j \exp(\alpha_{1,j})} \]

\[ q^1 = W^q a^1 \]
\[ k^1 = W^k a^1 \]
\[ k^2 = W^k a^2 \]
\[ k^3 = W^k a^3 \]
\[ k^4 = W^k a^4 \]
Self-attention

Extract information based on attention scores

\[ b^1 = \sum_i \alpha'_{1,i} v^i \]

\[ \alpha'_{1,1} \]

\[ \alpha'_{1,2} \]

\[ \alpha'_{1,3} \]

\[ \alpha'_{1,4} \]

\[ q^1 \]

\[ k^1 \]

\[ v^1 \]

\[ a^1 \]

\[ k^2 \]

\[ v^2 \]

\[ a^2 \]

\[ k^3 \]

\[ v^3 \]

\[ a^3 \]

\[ k^4 \]

\[ v^4 \]

\[ a^4 \]

\[ v^1 = W^v a^1 \]

\[ v^2 = W^v a^2 \]

\[ v^3 = W^v a^3 \]

\[ v^4 = W^v a^4 \]
Self-attention

parallel

Can be either \textit{input} or a \textit{hidden layer}
Self-attention

\[ b^2 = \sum_i \alpha'_{2,i} v^i \]
Self-attention

\[ q^i = W^q a^i \]

\[ k^i = W^k a^i \]

\[ v^i = W^v a^i \]

\[
\begin{align*}
q^1 q^2 q^3 q^4 &= W^q a^1 a^2 a^3 a^4 \\
q^1 q^2 q^3 q^4 &= W^k a^1 a^2 a^3 a^4 \\
q^1 q^2 q^3 q^4 &= W^v a^1 a^2 a^3 a^4 
\end{align*}
\]
Self-attention

\[ \alpha_{1,1} = \begin{pmatrix} k^1 \end{pmatrix} q^1 \quad \alpha_{1,2} = \begin{pmatrix} k^2 \end{pmatrix} q^1 \]

\[ \alpha_{1,3} = \begin{pmatrix} k^3 \end{pmatrix} q^1 \quad \alpha_{1,4} = \begin{pmatrix} k^4 \end{pmatrix} q^1 \]
Self-attention

\[
\begin{align*}
\alpha_{1,1} &= \begin{bmatrix} k^1 \end{bmatrix} q^1 \\
\alpha_{1,2} &= \begin{bmatrix} k^2 \end{bmatrix} q^1 \\
\alpha_{1,3} &= \begin{bmatrix} k^3 \end{bmatrix} q^1 \\
\alpha_{1,4} &= \begin{bmatrix} k^4 \end{bmatrix} q^1
\end{align*}
\]

softmax

\[A' \quad \text{softmax} \quad A\]

\[
\begin{bmatrix}
\alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} \\
\alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} \\
\alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} \\
\alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\
\alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \alpha_{2,4} \\
\alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & \alpha_{3,4} \\
\alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & \alpha_{4,4}
\end{bmatrix}
\]

\[
\begin{bmatrix}
k^1 \\
k^2 \\
k^3 \\
k^4
\end{bmatrix}
\]

\[
\begin{bmatrix}
q^1 \\
q^2 \\
q^3 \\
q^4
\end{bmatrix}
\]

\[K^T \quad Q\]
Self-attention

\[
\begin{align*}
\alpha'_{1,1} &\quad b^1 &
\alpha'_{1,2} &\quad q^1 \\
\alpha'_{1,3} &\quad k^1 \\
\alpha'_{1,4} &\quad v^1
\end{align*}
\]

\[
\begin{align*}
\alpha'_{1,1} &\quad q^2 \\
\alpha'_{1,2} &\quad k^2 \\
\alpha'_{1,3} &\quad v^2 \\
\alpha'_{1,4} &\quad v^2
\end{align*}
\]

\[
\begin{align*}
\alpha'_{1,1} &\quad q^3 \\
\alpha'_{1,2} &\quad v^3 \\
\alpha'_{1,3} &\quad k^3 \\
\alpha'_{1,4} &\quad v^4
\end{align*}
\]

\[
\begin{align*}
\alpha'_{1,1} &\quad q^4 \\
\alpha'_{1,2} &\quad \times \\
\alpha'_{1,3} &\quad \times \\
\alpha'_{1,4} &\quad \times
\end{align*}
\]

\[
\begin{align*}
b^1 &\quad b^2 \\
b^3 &\quad v^1 \\
b^4 &\quad v^2 \\
O &\quad v^3 \\
V &\quad v^4 \\
A' &\quad \alpha'_{1,2} \\
&\quad \alpha'_{1,3} \\
&\quad \alpha'_{1,4}
\end{align*}
\]
Self-attention

\[
Q = W^q I \\
K = W^k I \\
V = W^v I
\]

Parameters to be learned

Attention Matrix

\[
A' = A = K^T Q \\
O = V = A'
\]
Multi-head Self-attention

Different types of relevance

\[ q^{i,1} = W^{q,1} q^i \]
\[ q^{i,2} = W^{q,2} q^i \]

(2 heads as example)
Multi-head Self-attention

\[ q^{i,1} = W^{q,1} q^i \]
\[ q^{i,2} = W^{q,2} q^i \]

(2 heads as example)

Different types of relevance
Multi-head Self-attention

Different types of relevance

\[ b^i = W^O \]

\[
\begin{align*}
q^{i,1} & \quad q^{i,2} & \quad k^{i,1} & \quad k^{i,2} & \quad \nu^{i,1} & \quad \nu^{i,2} \\
q^i & \quad k^i & \quad \nu^i & \quad q^{j,1} & \quad q^{j,2} & \quad k^{j,1} & \quad k^{j,2} & \quad \nu^{j,1} & \quad \nu^{j,2} \\
q^i & = W^q a^i & \quad (2 \text{ heads as example}) & \quad a^j
\end{align*}
\]
Positional Encoding

- No position information in self-attention.
- Each position has a unique positional vector $e^i$
- hand-crafted
- learned from data

Each column represents a positional vector $e^i$
Table 1. Comparing position representation methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Inductive</th>
<th>Data-Driven</th>
<th>Parameter Efficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinusoidal (Vaswani et al., 2017)</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Embedding (Devlin et al., 2018)</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Relative (Shaw et al., 2018)</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>This paper</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

(a) Sinusoidal

(b) Position embedding

(c) FLOATER

(d) RNN
Many applications ...

**Transformer**
https://arxiv.org/abs/1706.03762

**BERT**
https://arxiv.org/abs/1810.04805

Widely used in Natural Language Processing (NLP)!
Self-attention for Speech

Speech is a very long vector sequence.

If input sequence is length $L$

Attention Matrix

Truncated Self-attention

Attention in a range

Self-attention for Image

An **image** can also be considered as a **vector set**.

**Self-Attention GAN**

**DEtection Transformer (DETR)**

[Diagram of Self-Attention GAN and DETR]

https://arxiv.org/abs/1805.08318

Self-attention v.s. CNN

CNN: self-attention that can only attends in a receptive field

- CNN is simplified self-attention.

Self-attention: CNN with learnable receptive field

- Self-attention is the complex version of CNN.
Self-attention v.s. CNN

On the Relationship between Self-Attention and Convolutional Layers

https://arxiv.org/abs/1911.03584
Self-attention v.s. CNN

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

To learn more about RNN ......

https://youtu.be/xCGidAeyS4M (in Mandarin)

https://youtu.be/Jjy6ER0bHv8 (in English)
Self-attention for Graph

Consider \textbf{edge}: only attention to connected nodes

This is one type of \textbf{Graph Neural Network (GNN)}. 
Self-attention for Graph

• To learn more about GNN ...

https://youtu.be/eybCCtNKwzA
(in Mandarin)

https://youtu.be/M9ht8vsVEw8
(in Mandarin)
To Learn More …

Long Range Arena: A Benchmark for Efficient Transformers

Efficient Transformers: A Survey
Q&A