Self-Attention is Not Only a Weight: Analyzing BERT with Vector Norms

Goro Kobayashi¹, Tatsuki Kuribayashi¹,², Sho Yokoi¹,³, Kentaro Inui¹,³

¹ Tohoku University, ² Langsmith Inc., ³ RIKEN

ACL Student Research Workshop 2020
July 6-8, 2020
Background: Success of Self-attention

Self-attention-based models have been successfully applied to a wide range of NLP tasks.

- **Transformer** [Vaswani+’17], **BERT** [Devlin+’19], **RoBERTa** [Liu+’19], etc.

Increasing research efforts on analysis of self-attention-based models  [Hewitt&Manning’19;Coenen+’19;Tenney+’19;etc.]

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>URL</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PING-AN Omni-Sinitic</td>
<td>ALBERT + DAAF + NAS</td>
<td></td>
<td>90.6</td>
</tr>
<tr>
<td>2</td>
<td>ERNIE Team - Baidu</td>
<td>ERNIE</td>
<td></td>
<td>90.4</td>
</tr>
<tr>
<td>3</td>
<td>Alibaba DAMO NLP</td>
<td>StructBERT</td>
<td></td>
<td>90.3</td>
</tr>
<tr>
<td>4</td>
<td>T5 Team - Google</td>
<td>T5</td>
<td></td>
<td>90.3</td>
</tr>
<tr>
<td>5</td>
<td>Microsoft D365 AI &amp; MSR AI &amp; GATECH</td>
<td>MT-DNN-SMART</td>
<td></td>
<td>89.9</td>
</tr>
</tbody>
</table>

(Leaderboard on June 14)
Previous studies: **Attention weight analysis**

One of the main analyses is to examine "**how self-attention mixes information**".

- Previous studies: Analysis of the magnitude of **attention weight** [Clark+’19; Kovaleva+’19; Reif+’19; Lin+’19; etc.]

#### Previous studies

![Diagram showing transformations and attention weights]

- Ignore the effects of input vectors and vector transformations
- Might lead to a misleading conclusion
Our contribution: Propose a novel analysis
Taking into account more effects

One of the main analyses is to examine "how self-attention mixes information".

- This study: Analysis of vector norms

This study

😊 Considers the effects of input vectors and vector transformations as well

→ not lead to a misleading conclusion
Self-attention is a weighted sum of vectors

By simply rewriting equations, self-attention can be regarded as a 3-step process.

1. Affine transformation (including transformation to Value vectors)
2. Weighting
3. Summation

Weighted sum of transformed vectors

\[ y_i = \sum_j \alpha_{i,j} f(x_j) + b^0 \]
Mixed amount $\neq$ Attention weight

**Attention weight analysis**
[Clark+'19;Kovaleva+'19;Reif+'19;etc.]

$$y_i = \sum_j \alpha_{i,j} f(x_j) + b^0$$

😢 Ignore the effect of transformed vector $f(x)$
Mixed amount ≠ Attention weight

Attention weight analysis
[Clark+’19;Kovaleva+’19;Reif+’19;etc.]

\[ y_i = \sum_j \alpha_{i,j} f(x_j) + b^0 \]

🤔 Ignore the effect of transformed vector \( f(x) \)

misunderstand that self-attention gathers a lot from \( x_5 \) to generate \( y_4 \) even if \( \alpha f(x_1) \) is predominant in \( y_4 \)
Propose a new analysis

- Focus on **the vector** to be actually summed

\[ y_i = \sum_j \alpha_{i,j} f(x_j) + b^0 \]

- Measure the mixed amount of each input by **norm** \( \| \alpha_{i,j} f(x_j) \| \)

😊 Consider the vector \( f(x) \) in addition to attention weight \( \alpha \)
Proposal: Norm analysis
Measure the norm of the vector actually summed

Propose a new analysis

• Focus on the vector to be actually summed

\[ y_i = \sum_j \alpha_{i,j} f(x_j) + b^0 \]

• Measure the mixed amount of each input

In addition to attention weight correctly understand that self-attention gathers the most from \( x_1 \) to generate \( y_4 \) (a little from \( x_5 \))
Experimental Setup

Investigate the behavior of self-attention with previous and proposed methods

- Models
  - pre-trained BERT-base (uncased)
    - 12 layers, 12 head (total of 144 self-attentions in the model)

- Data
  - 992 segments extracted from Wikipedia [Clark+’19]

Input segment: [CLS] paragraph1 [SEP] paragraph2 [SEP]

Token used for classification tasks
Separator tokens

https://github.com/clarkkev/attention-analysis
Previous result of attention weight analysis
[Clark+’19]

Average attention weight in each layer

- Attention weights are biased towards specific token categories
  - Early layers --> [CLS]
  - Middle layers --> [SEP]
  - Deep layers --> periods or commas
Different results between the methods

**Attention weight analysis [Clark+’19]**

**Proposed norm-based analysis**

Largely different results

- Self-attention gathers only a little from special tokens, periods, and commas, and most from the other tokens.
Detailed analysis ([SEP])

Why $\|\alpha f(x)\|$ is small despite its large weight $\alpha$?

• Attention weight $\alpha$ and norm of transformed vector $\|f(x)\|$ cancel each other out
  • Same tendency for [CLS], periods, and commas
Intuition: highly frequent words such as stop words have a little importance for pre-training tasks.

Strong positive correlation between frequency rank and $\|f(x)\|$ (Spearman’s $\rho = 0.75$)

Suggest that BERT discounts highly frequent words by adjusting $\|f(x)\|$.
Summary

- Proposed the norm-based analysis considering input vectors and vector transformations as well.
- Self-attentions in BERT gather only a little from specific tokens despite assigning high attention weights to them.
- Suggests that BERT discounts highly frequent words.

![Diagram showing attention weights and transformations](image)
Summary

• Proposed the norm-based analysis considering input vectors and vector transformations as well

• Self-attentions in BERT gather only a little from specific tokens and usually assign high attention weights to them

• Suggests that BERT discounts highly frequent words

Waiting for you in the following Q&A sessions!

• SRW session 6A (June 7)
• SRW session 12B (June 8)