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

Goro Kobayashi[1], Tatsuki Kuribayashi[1,2], Sho Yokoi[1,3], Kentaro Inui[1,3]

[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.]

GLUE (Leaderboard on June 14)

	Rank	Name	Model	URL	Score
+	1	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6
	2	ERNIE Team - Baidu	ERNIE	↗	90.4
+	3	Alibaba DAMO NLP	StructBERT	↗	90.3
	4	T5 Team - Google	T5	↗	90.3
	5	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART	↗	89.9

https://gluebenchmark.com/leaderboard

Previous studies: Attention weight analysis

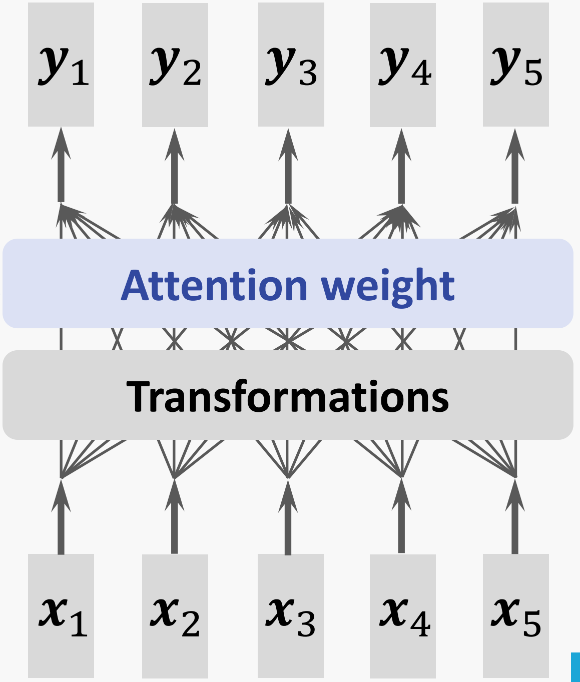
One of the main analyses is to examine "**how self-attention mixes information**".

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

Previous studies

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

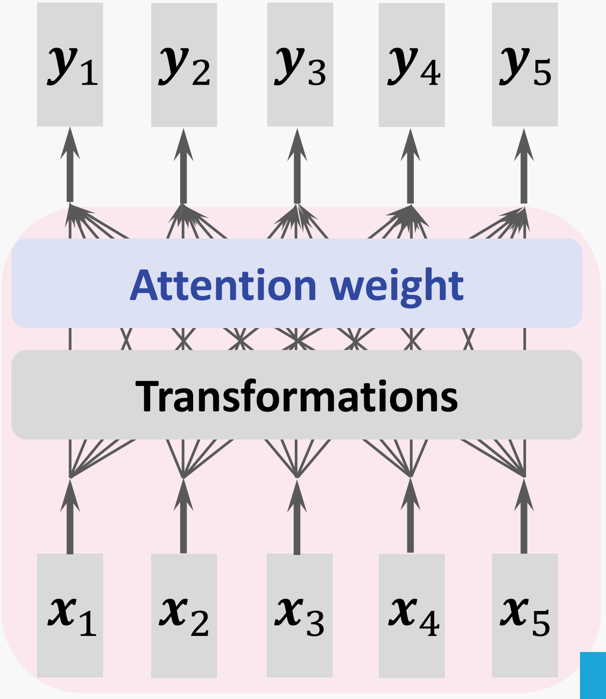
One of the main analyses is to examine "**how self-attention mixes information**”.

- 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**.

3 Summation

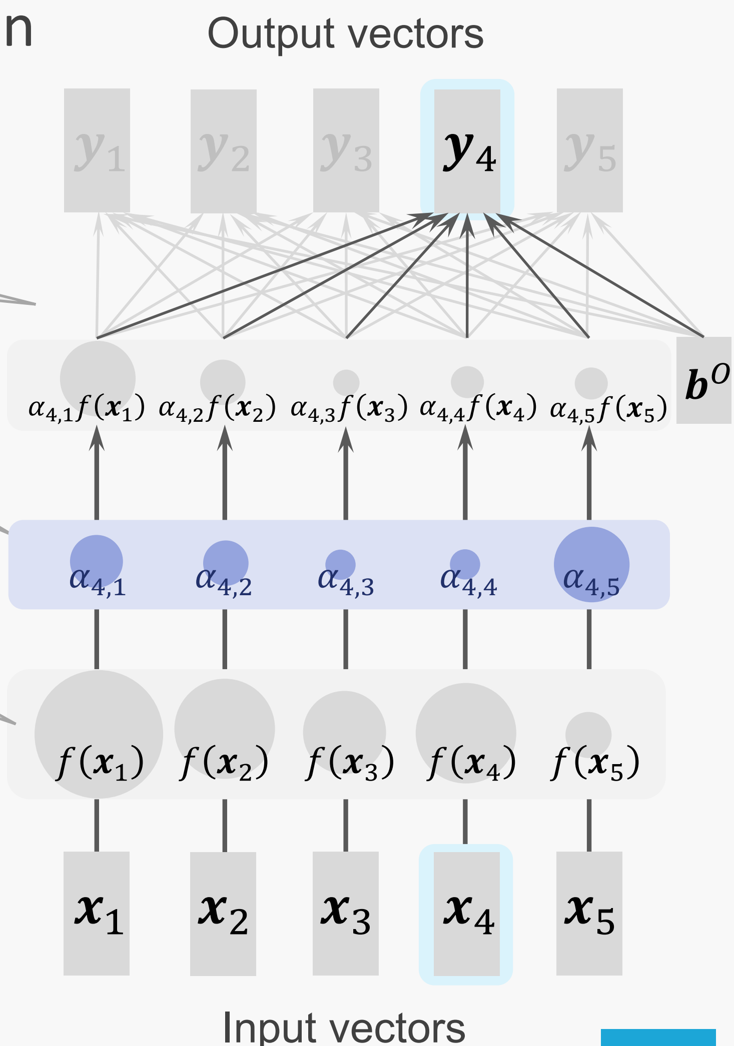
2 Weighting

1 Affine transformation (including transformation to Value vectors)

=

Weighted sum of transformed vectors

$$\boldsymbol{y}_i = \sum_j \alpha_{i,j} f(\boldsymbol{x}_j) + \boldsymbol{b}^O$$

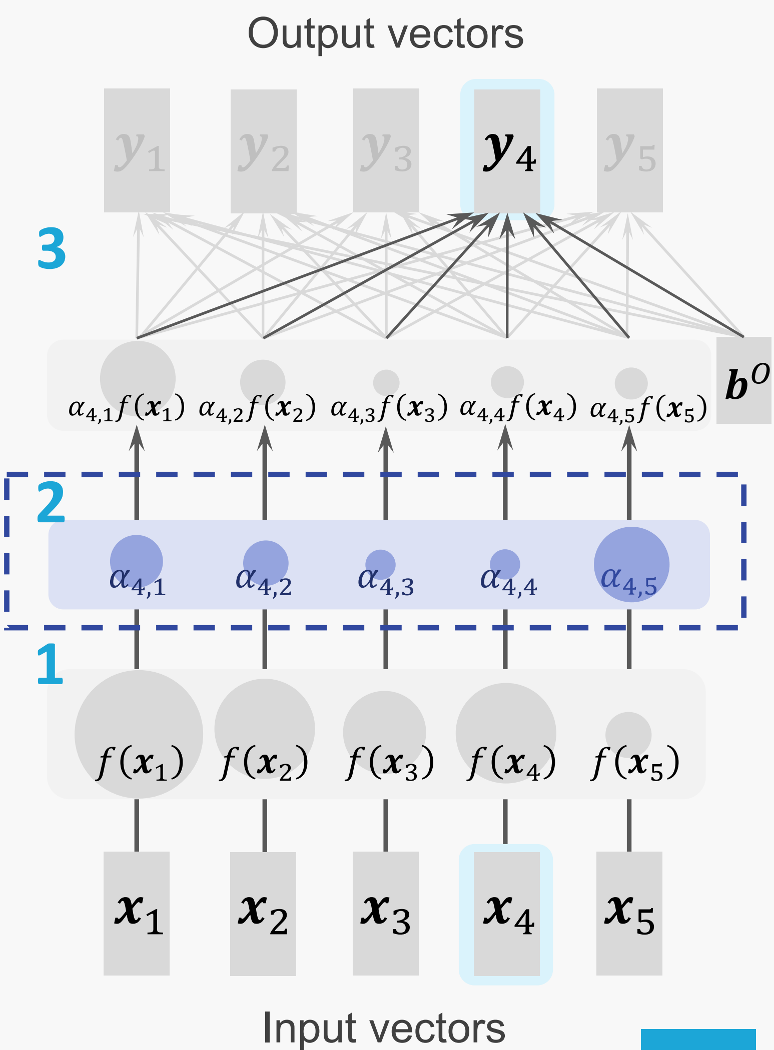


Mixed amount ≠ Attention weight

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

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

☹ Ignore the effect of transformed vector $f(x)$



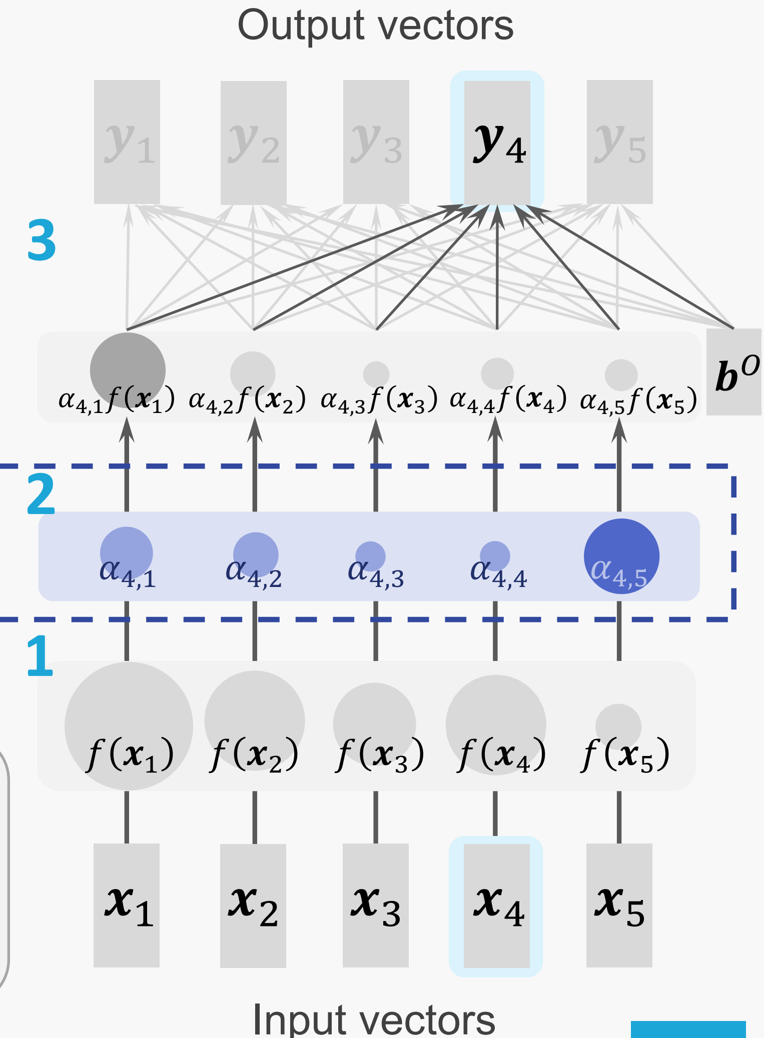
Mixed amount ≠ Attention weight

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

$$\boldsymbol{y}_i = \sum_j \alpha_{i,j} f(\boldsymbol{x}_j) + \boldsymbol{b}^O$$

☹ Ignore the effect of transformed vector $f(\boldsymbol{x})$

misunderstand that self-attention gathers a lot from $\boldsymbol{x}_5$ to generate $\boldsymbol{y}_4$ even if $\alpha f(\boldsymbol{x}_1)$ is predominant in $\boldsymbol{y}_4$



Proposal: Norm analysis
Measure the norm of the vector actually summed

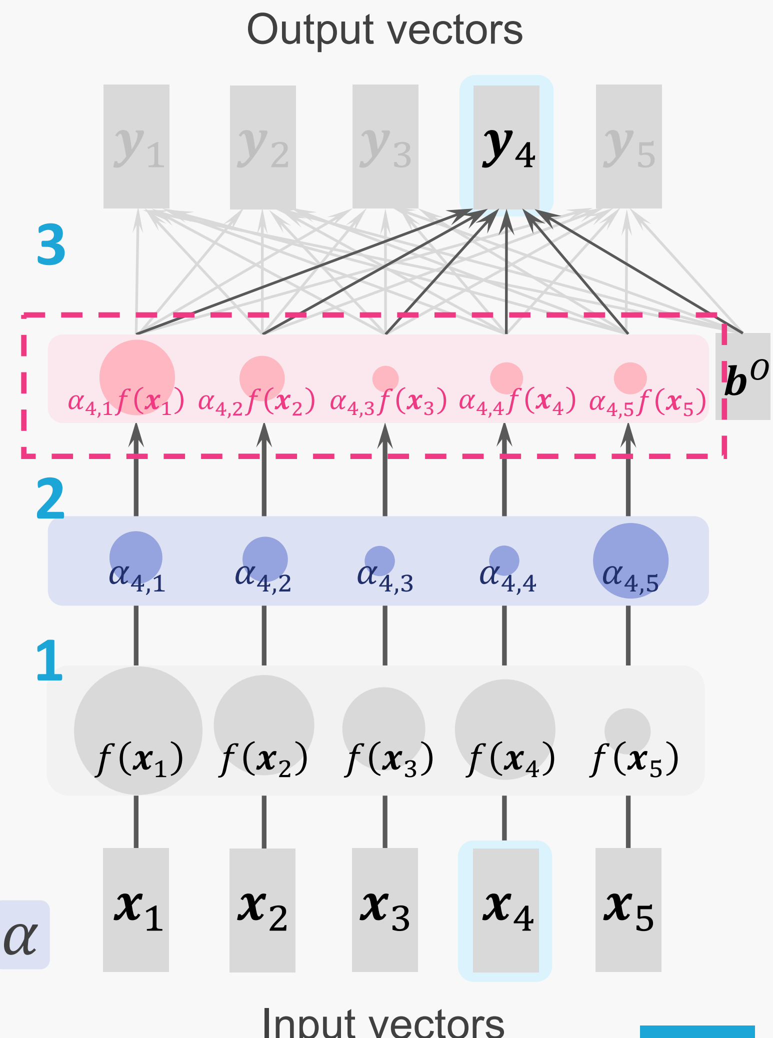
Propose a new analysis

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

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

- 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 α



Proposal: Norm analysis
Measure the norm of the vector actually summed

Propose a new analysis

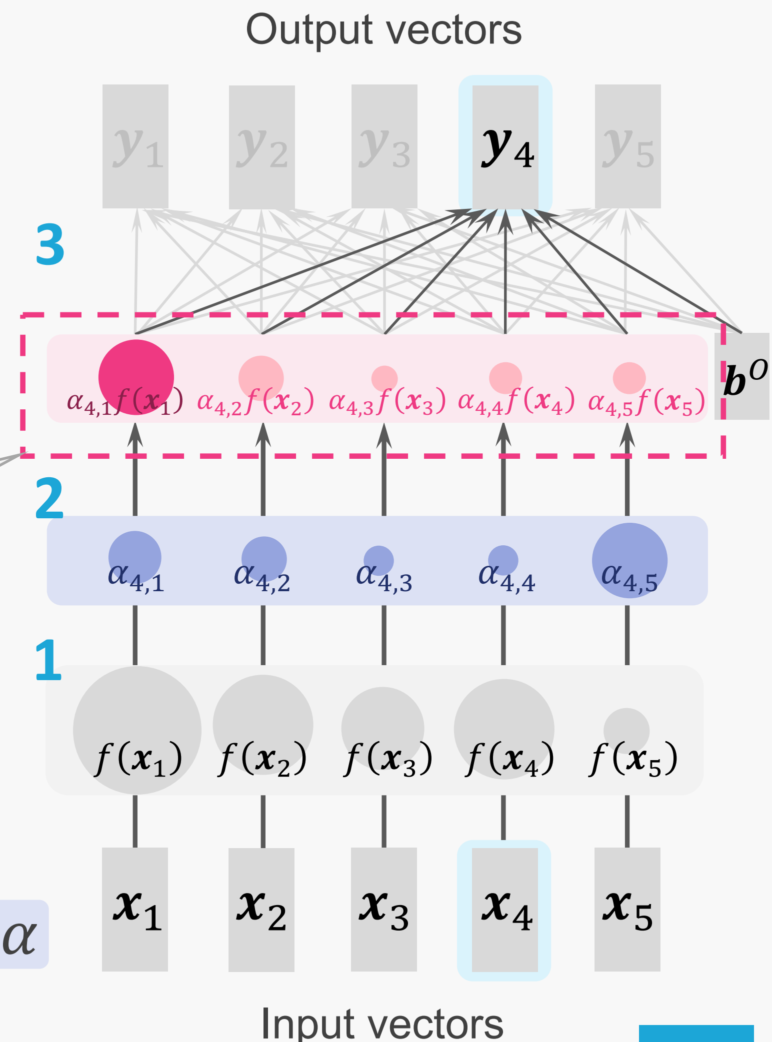
- Focus on **the vector to be actually summed**

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

- Measure the mixed ...

correctly understand that self-attention gathers the most from x_1 to generate y_4 (a little from x_5)

in addition to attention weight α



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]

https://github.com/clarkkev/attention-analysis

token used for classification tasks

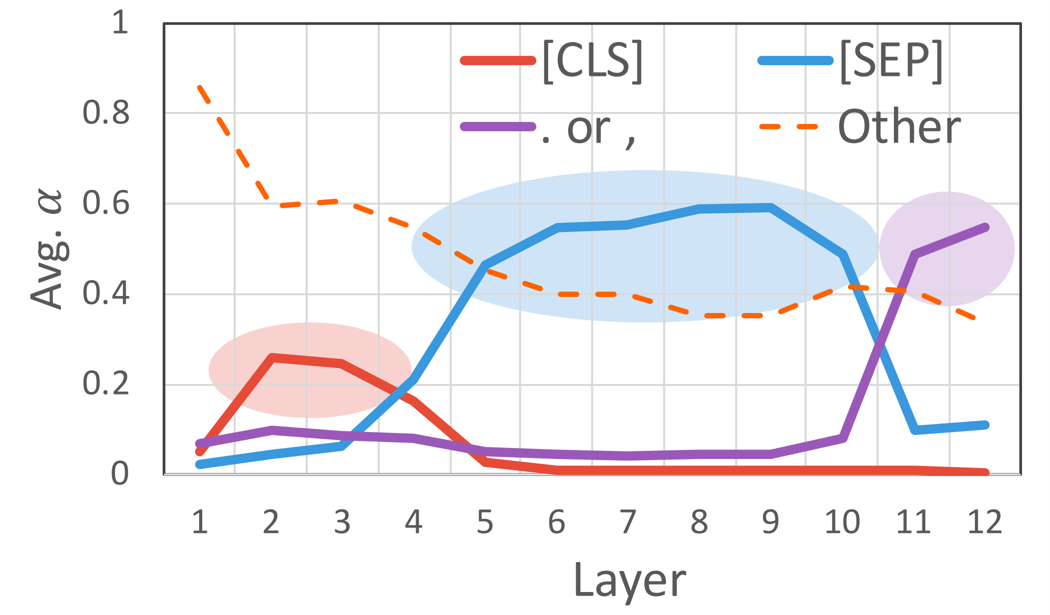
separator tokens

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

Previous result of attention weight analysis

[Clark+'19]

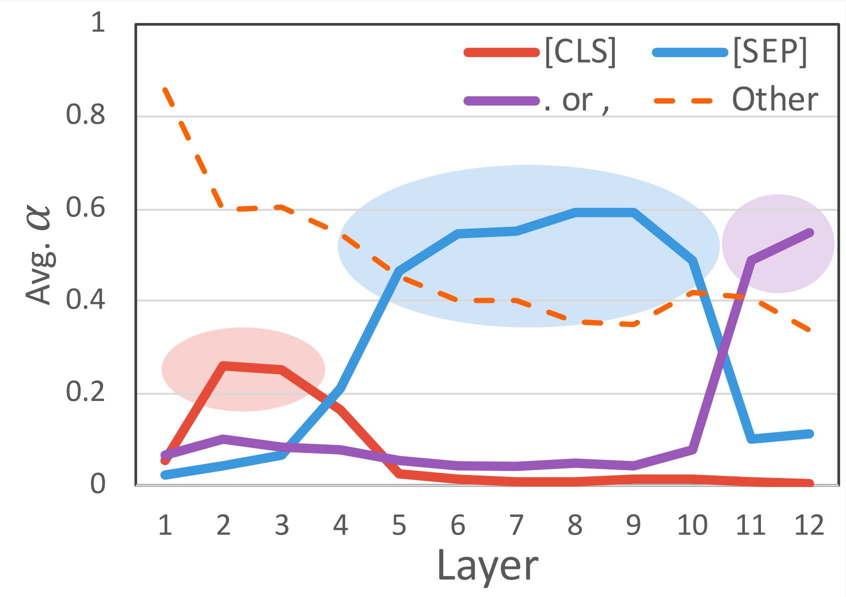
Average attention weight in each layer



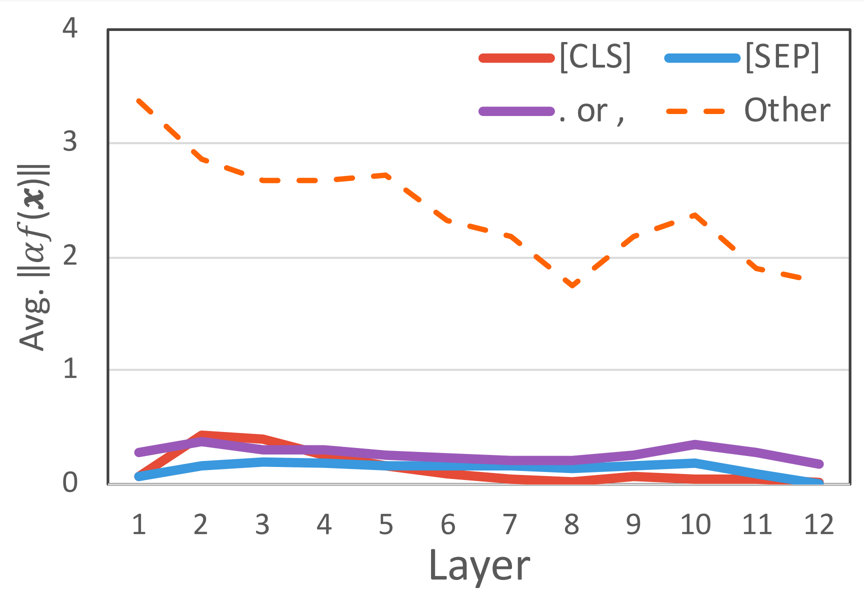
- Attention weights are biased towards specific token categories
  - Early layers --> [CLS]
  - Middle layers --> [SEP]
  - Deep layers --> periods or commas

Different results between the methods

Attention weight analysis [Clark+'19]



Proposed norm-based analysis



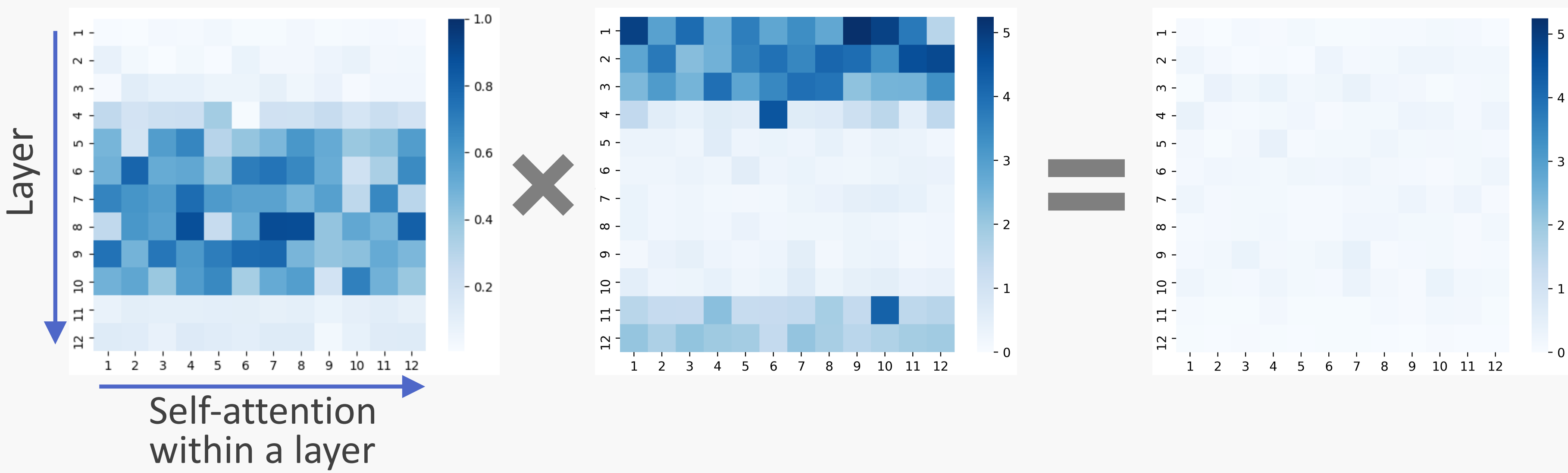
Largely different results

- Self-attention gathers only a little from special tokens, periods, and commas, and most from the other tokens.

Detailed analysis ([SEP])

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

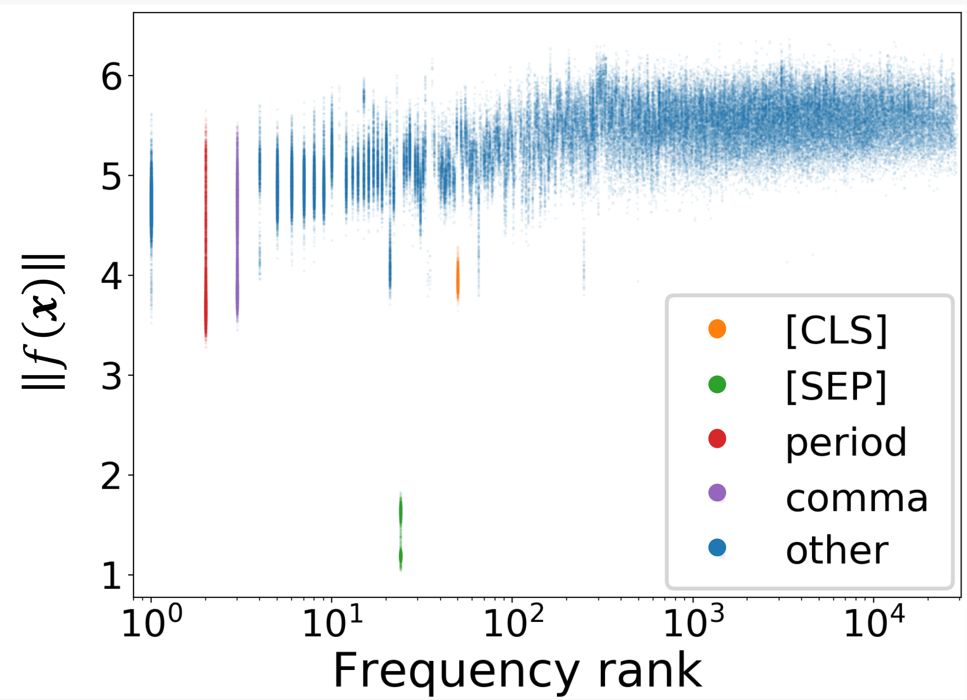
α Attention weight (previous analysis) | $\|f(\boldsymbol{x})\|$ Norm of transformed vector (previously ignored element) | $\|\alpha f(\boldsymbol{x})\|$ Norm of weighted vector (proposed analysis)



- Attention weight α and norm of transformed vector $\|f(\boldsymbol{x})\|$ cancel each other out
  - Same tendency for [CLS], periods, and commas

Relation with frequency

Intuition: highly frequent words such as stop words have a little importance for pre-training tasks

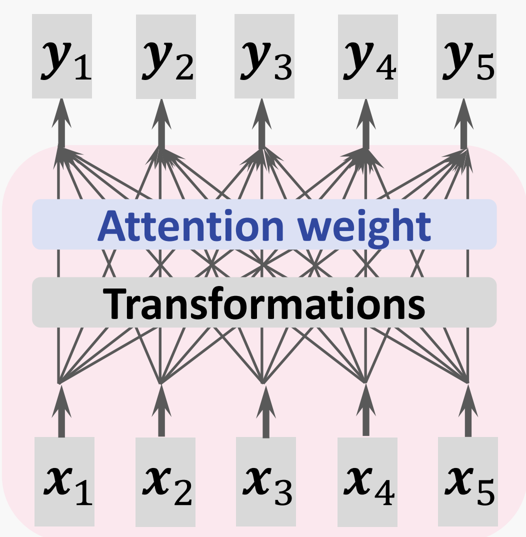


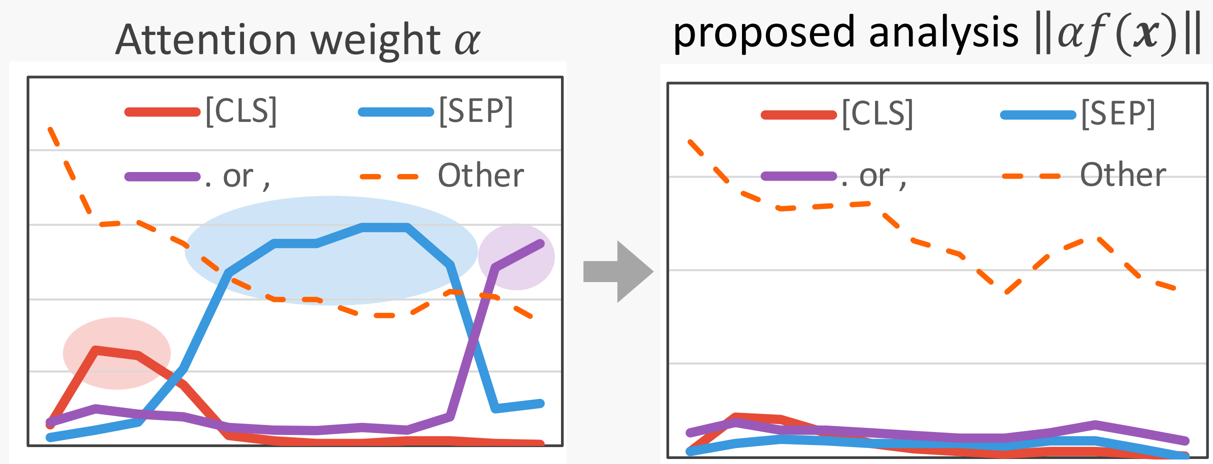
Strong positive correlation between frequency rank and $\|f(\boldsymbol{x})\|$ (Spearman's $\rho = 0.75$)

Suggest that BERT discounts highly frequent words by adjusting $\|f(\boldsymbol{x})\|$

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





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

Waiting for you in the following Q&A sessions!

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