Incorporating Residual and Normalization Layers into Analysis of Masked Language Models

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

[1] Tohoku University, [2] Langsmith Inc., [3] RIKEN

https://github.com/gorokoba560/norm-analysis-of-transformer

EMNLP 2021
November 7-11, 2021

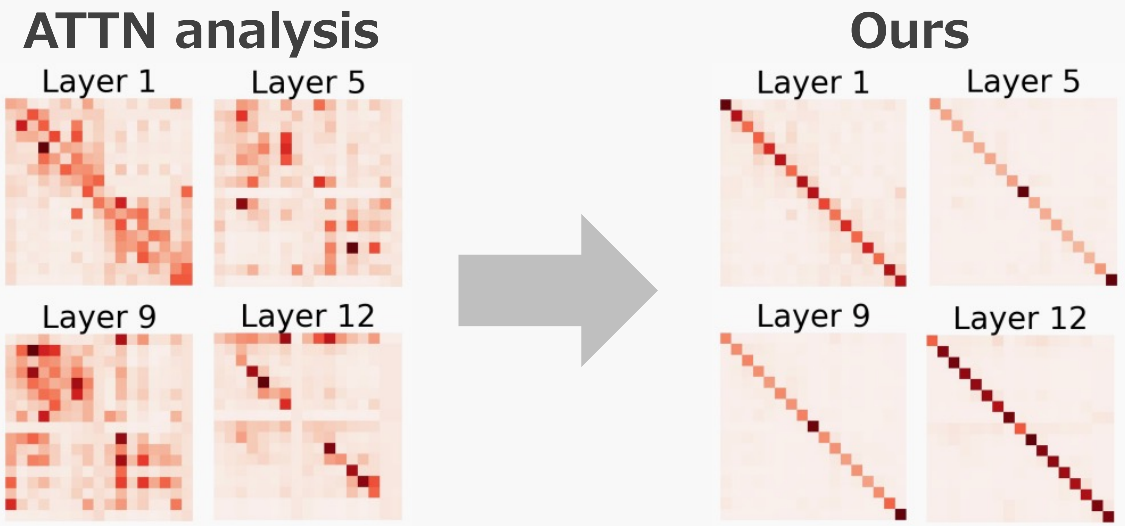
Overview

Propose to analyze Transformers considering:

- Multi-head attention (ATTN)
- Residual connection (RES) 👈 **new!**
- Layer normalization (LN) 👈 **new!**

Our analysis of Masked LMs reveals
weaker Mixing via Attention than previously assumed

Token-to-token interactions



Background: Success of Transformers

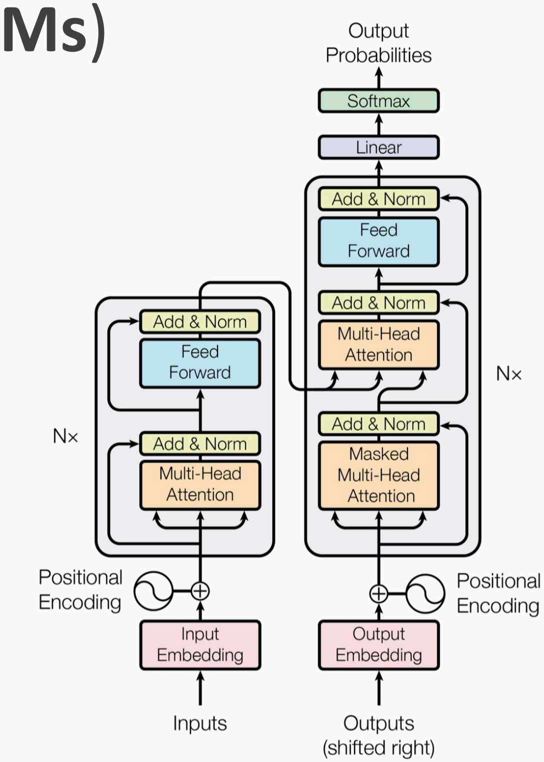
Transformer[Vaswani+'17] has been successfully applied to a wide range of NLP tasks.

- Especially **Masked language models** (**MLMs**)
  - **BERT**[Devlin+'19], **RoBERTa**[Liu+'19], etc.

GLUE (Leaderboard on October 12, 2021)

	Rank	Name	Model	URL	Score
	1	ERNIE Team - Baidu	ERNIE		91.1
	2	AliceMind & DIRL	StructBERT + CLEVER		91.0
	3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8
	4	liangzhu ge	DeBERTa + CLEVER		90.8
	5	HFL iFLYTEK	MacALBERT + DKM		90.7
+	6	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6

https://gluebenchmark.com/leaderboard

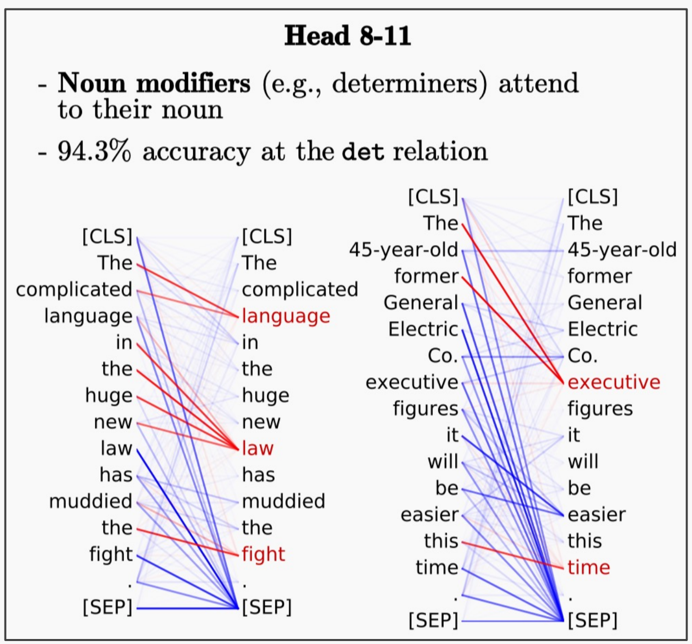


[Vaswani+'17]

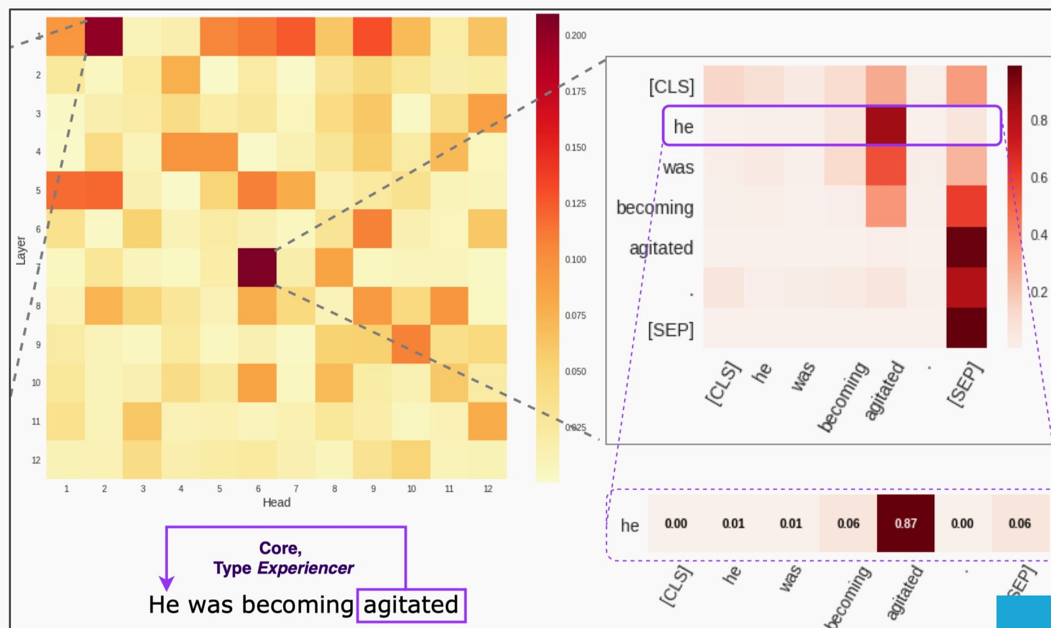
Big goal: Understand successful Transformers

Reveal mechanisms/characteristics of Transformers

- analyzed and probed by many studies
  [Hewitt&Manning'19;Reif+'19;Tenney+'19; etc.]
  - Typically **focused on "Mixing" at Attention** (e.g., Attention weight) [Clark+'19;Kovaleva+'19;Reif+'19;etc.]



[Clark+'19]

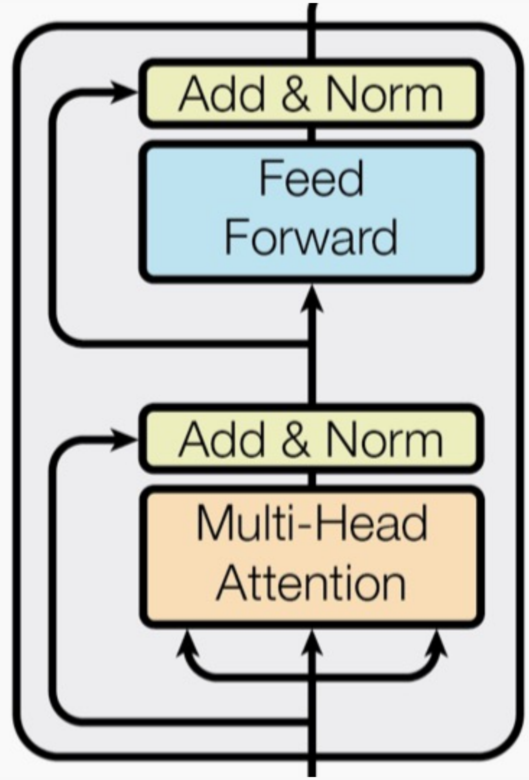


[Kovaleva+'19]

Transformer architecture

Transformer layer consists of:

- Multi-head attention (ATTN)
- Residual connection (RES)
- Layer normalization (LN)
- Feed-forward network (FF)

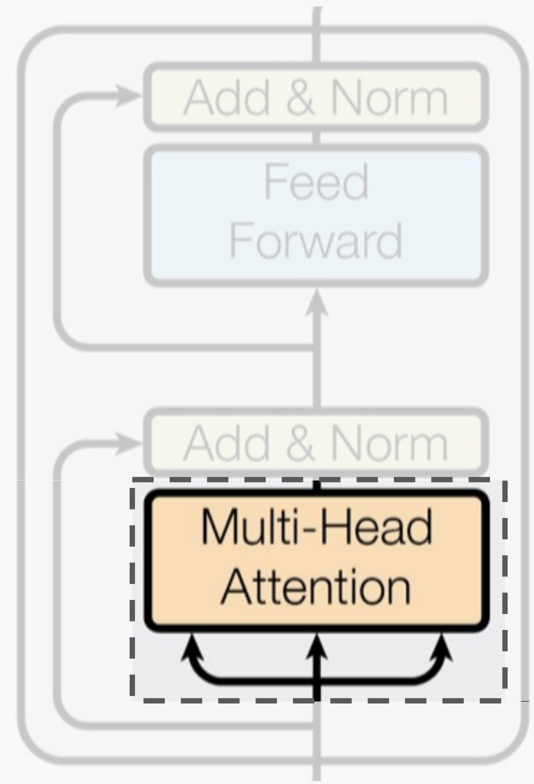


[Vaswani+'17]

Scope of existing Transformer analysis: Only attention

Transformer layer consists of:

- Multi-head attention (ATTN)
- Residual connection (RES)
- Layer normalization (LN)
- Feed-forward network (FF)



[Vaswani+'17]

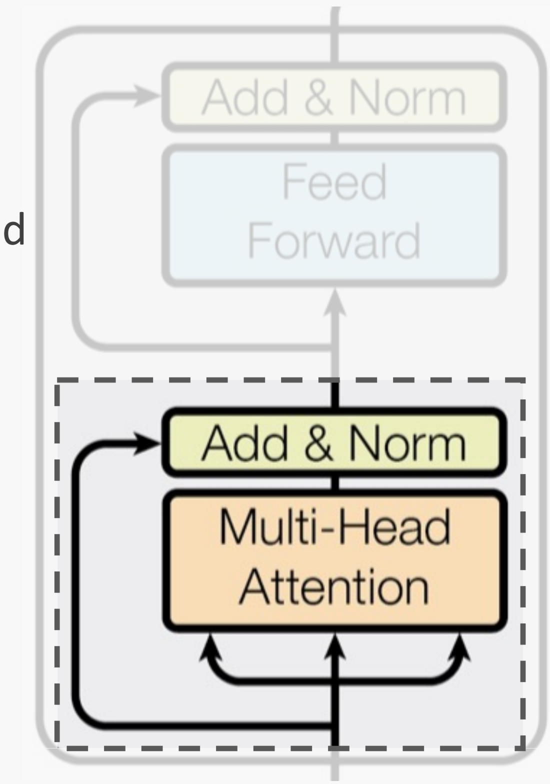
Problem:
Ignored components can overwrite attention's process

Our scope of Transformer analysis

Transformer layer consists of:

- Multi-head attention (ATTN)
- Residual connection (RES)
- Layer normalization (LN)
- Feed-forward network (FF)

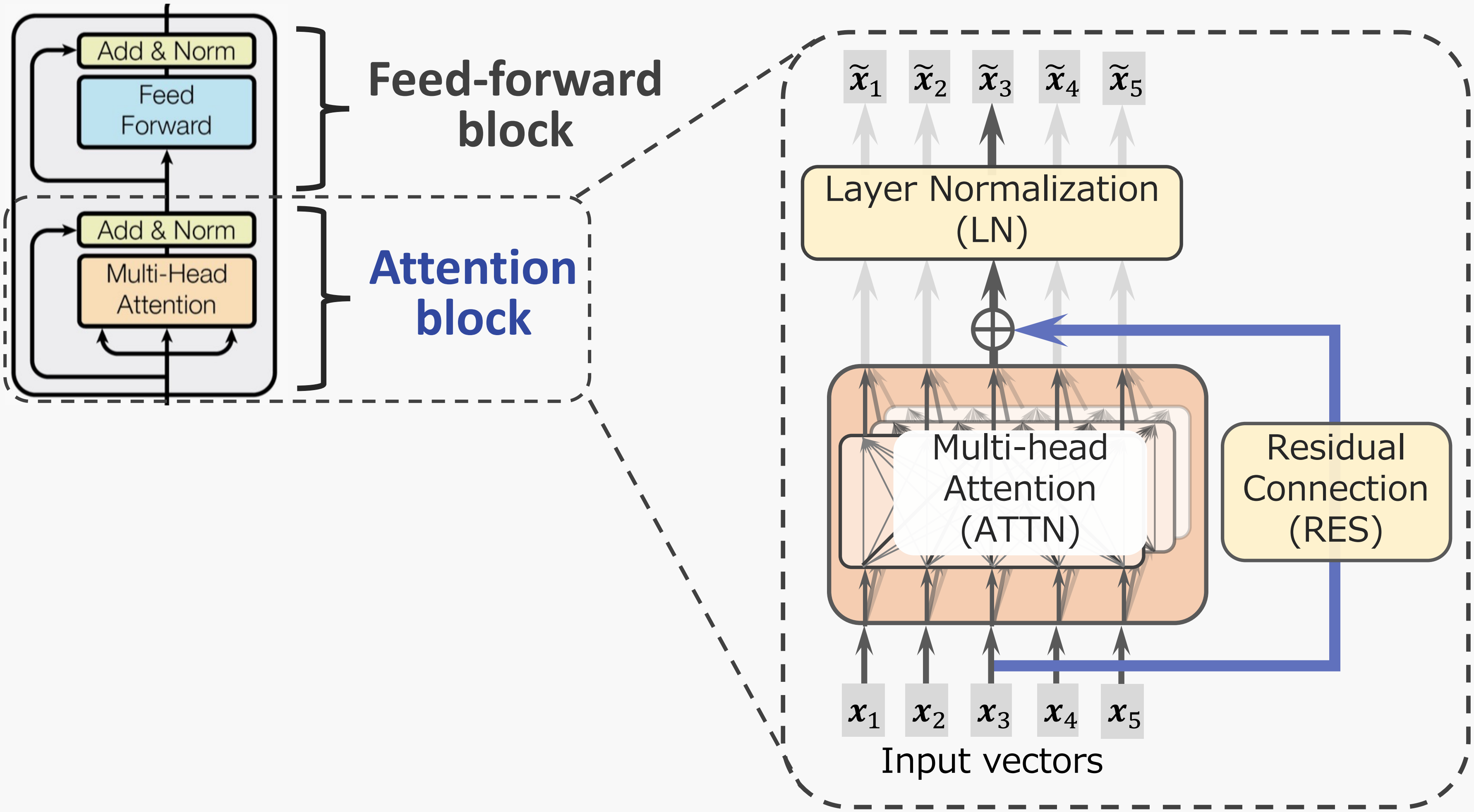
expand



expand

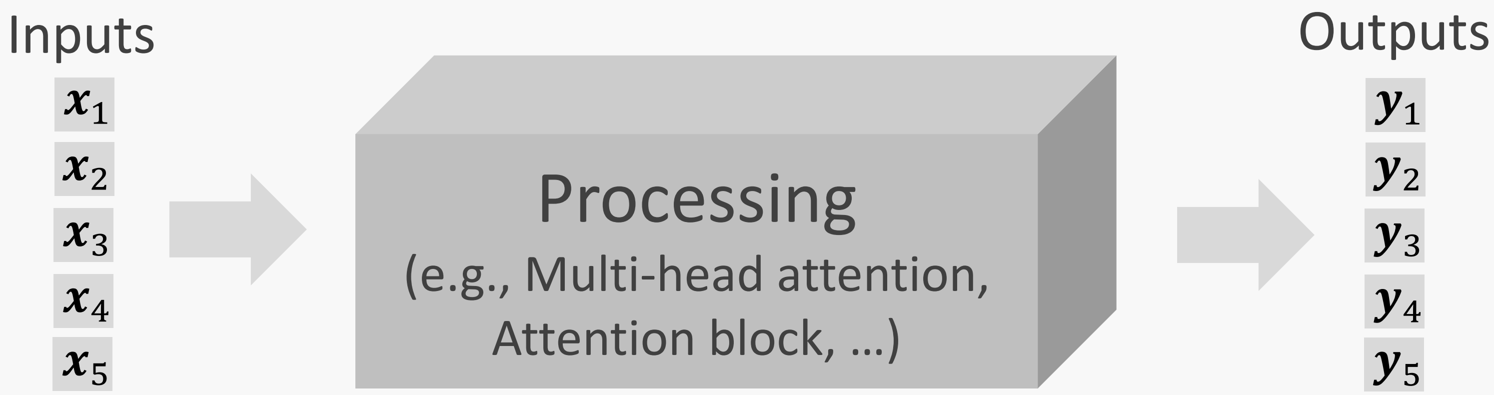
[Vaswani+'17]

Our scope of Transformer analysis: Attention block



Including Feed-forward block is future work

Strategy: Norm-based analysis [Kobayashi+'20]



Compute the contribution of each input $\boldsymbol{x}_j$ to the output $\boldsymbol{y}_i$:

1. Decompose $\boldsymbol{y}_i$ into **the sum of transformed input vectors**

$$\boldsymbol{y}_i = \sum_j F(\boldsymbol{x}_j)$$

Sum of transformed vectors

2. Measure the **norm** $\|F(\boldsymbol{x}_j)\|$

Decomposition of processing at attention block

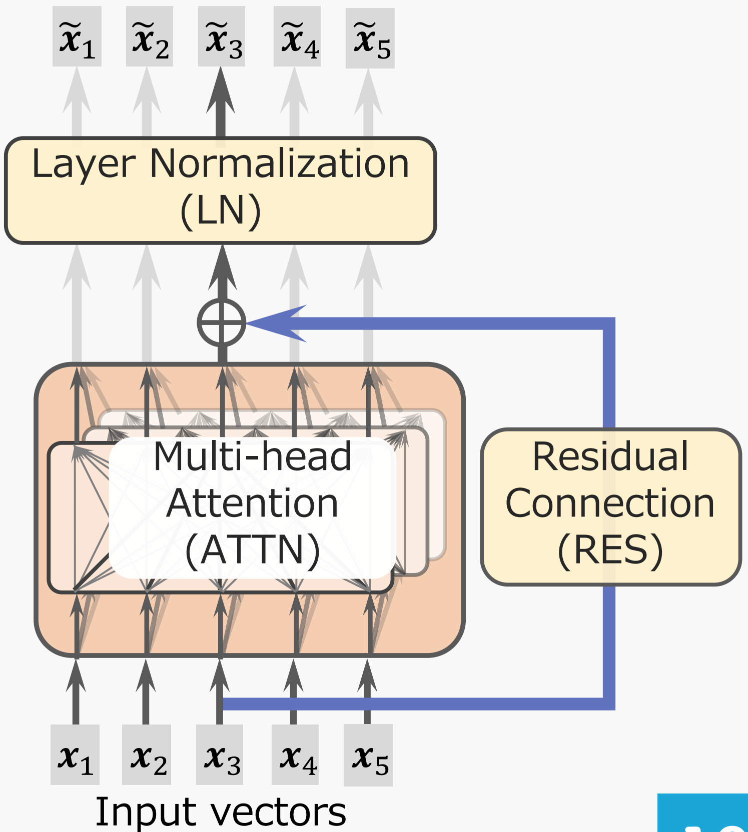
Express the processing at the attention block as "**the sum of transformed input vectors**"

Input vectors: $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \in \mathbb{R}^{n \times d}$

$$\widetilde{\boldsymbol{x}}_i = \mathrm{LN}\left(\mathrm{RES}(\mathrm{ATTN}(\boldsymbol{X}))\right)$$

$$= \sum_j F(\boldsymbol{x}_j)$$

Sum of transformed vectors

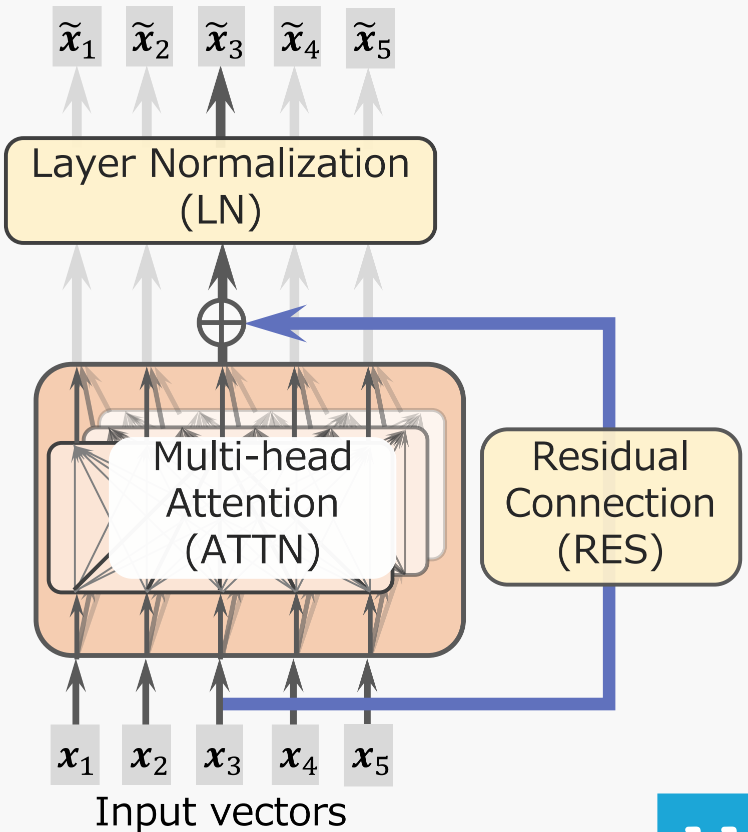


Decomposition of processing at attention block

Express the processing at the attention block as “**the sum of transformed input vectors**”

Role of each component:

- **ATTN** → **Mixing** the surrounding inputs
- **RES** → **Preserving** the original input
- **LN** → **Normalizing** and **Scaling** each vector



Decomposition of processing at attention block

Express the processing at the attention block as "**the sum of transformed input vectors**"

No non-linear calculations

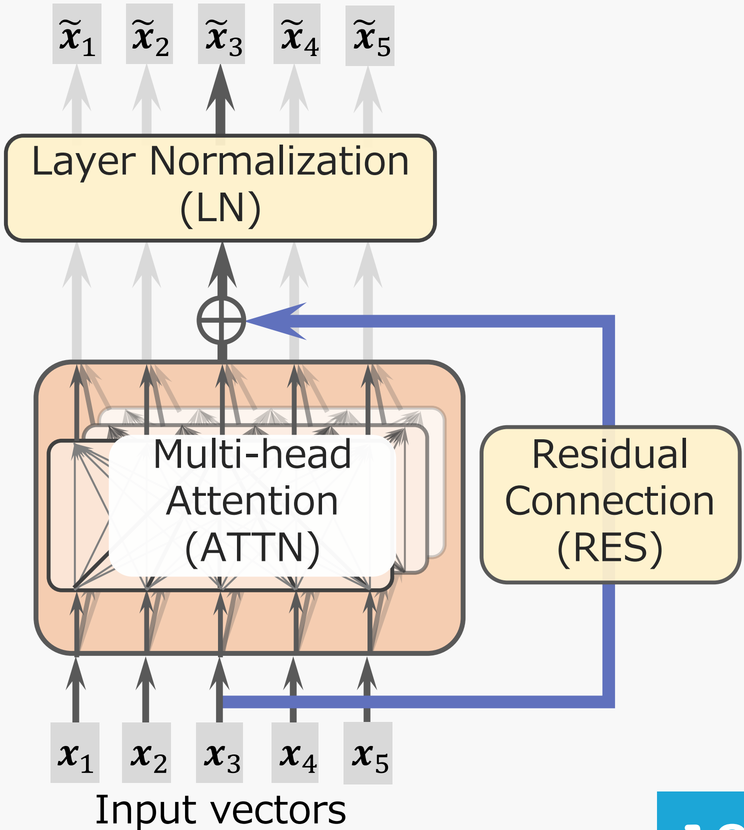
Able to decompose:

$$\tilde{\boldsymbol{x}}_i = \mathrm{LN}\Big(\mathrm{Res}\big(\mathrm{ATTN}(\boldsymbol{X})\big)\Big)$$

⋮ without approximation

$$= \sum_j F(\boldsymbol{x}_j) + \boldsymbol{\beta}$$

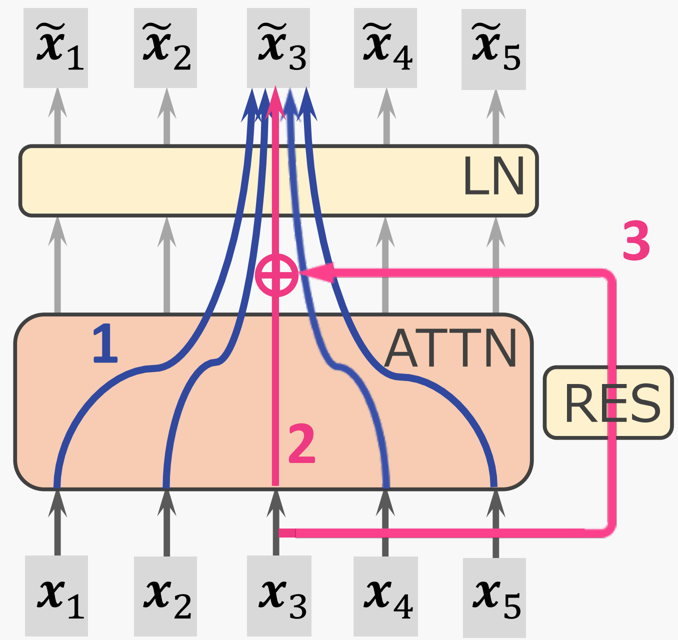
Sum of transformed vectors



Our interest:
Relationship between Mixing and Preserving

Effects in Attention block:

1. **Mixing contexts via ATTN**
2. **Preserving the original via ATTN**
3. **Preserving the original via RES**



Contextualized representations have been successful

Interested in **strength of the context mixing**

=

Power relationship between mixing and preserving

Mixing ratio: Relationship between Mixing and Preserving

Able to decompose the process of Attention block into two effects and bias:

$$\widetilde{\boldsymbol{x}}_i = \underbrace{\widetilde{\boldsymbol{x}}_{i \leftarrow \text{context}}}_{\text{Mixing}} + \underbrace{\widetilde{\boldsymbol{x}}_{i \leftarrow i}}_{\text{Preserving}} + \underbrace{\boldsymbol{\beta}}_{\text{bias}}$$

Measure each magnitude by its vector norm

- **Magnitude of Mixing**: $\left\| \underline{\widetilde{\boldsymbol{x}}_{i \leftarrow \text{context}}} \right\|$
- **Magnitude of Preserving**: $\left\| \underline{\widetilde{\boldsymbol{x}}_{i \leftarrow i}} \right\|$

Mixing ratio:
Relationship between Mixing and Preserving

Able to decompose
into two effects an

$\tilde{x}_i =$

Mixing ratio:

$$r = \frac{\left\|\tilde{\boldsymbol{x}}_{i \leftarrow \text{context}}\right\|}{\left\|\tilde{\boldsymbol{x}}_{i \leftarrow \text{context}}\right\| + \left\|\tilde{\boldsymbol{x}}_{i \leftarrow i}\right\|}$$

- If $r = 0.5$, mixing and preserving are 1:1

Measure each magnitude by its vector norm

- **Magnitude of Mixing**: $\left\|\underline{\tilde{\boldsymbol{x}}_{i \leftarrow \text{context}}}\right\|$
- **Magnitude of Preserving**: $\left\|\underline{\tilde{\boldsymbol{x}}_{i \leftarrow i}}\right\|$

Experiments

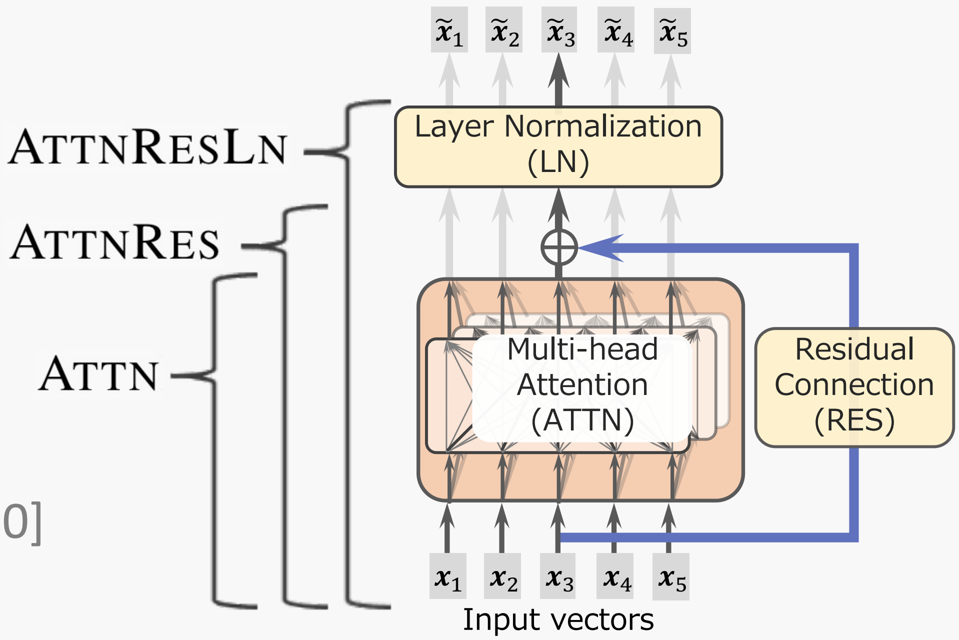
Experiment setup

Measure mixing ratio at each attention block of MLMs

- Models
  - Pre-trained BERT [Devlin+'19; Turc+'19]
    - BERT-tiny, BERT-small, BERT-medium, **BERT-base**, BERT-large
    - 25 BERT-base models trained with different seeds [Sellam+'21]
  - Pre-trained RoBERTa [Liu+'19]
    - RoBERTa-base, RoBERTa-large
- Data
  - **Excerpts from Wikipedia** [Clark+'19]
  - SST-2 [Socher+'03]
  - MNLI [Williams+'18]
  - CoNLL-2003 NER dataset [Sang&Meulder'03]

Compare mixing ratio computed with different analysis methods

- ATTN-W
- ATTN-N [Kobayashi+'20]
- ATTNRES-W [Abnar+'20]
- ATTNRES-N (Ours)
- ATTNRESLN-N (Ours)



Strategy:

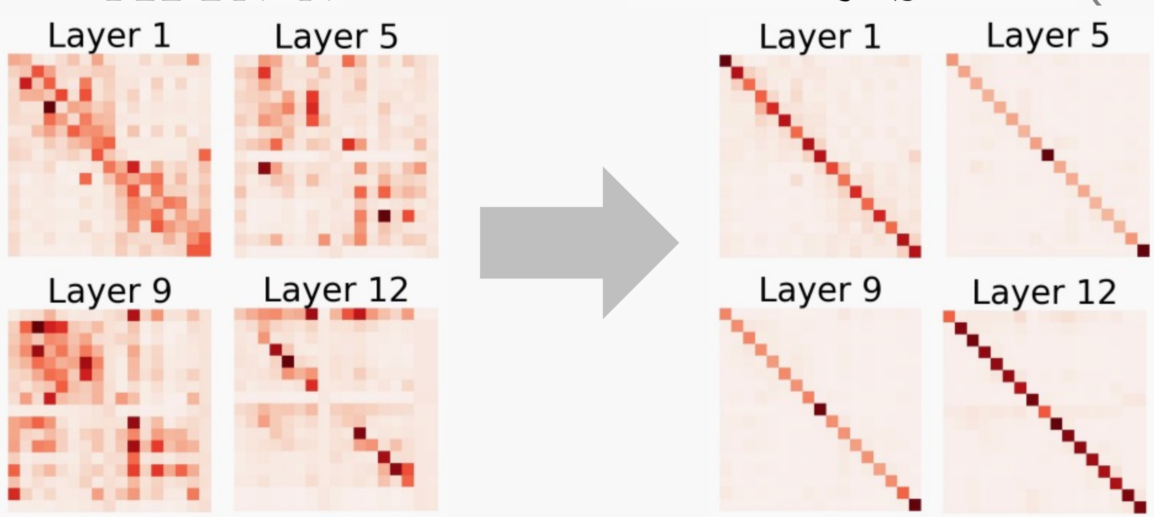
- -W : based on attention Weights
- -N : use Norm-based method

Mean of mixing ratio: Lower mixing ratio than previously assumed

- More expanded method shows the lower mixing ratio
  - 19% → Mixing ≪ Preserving
  - RES largely decreases the ratio
  - LN decreases slightly the ratio

Methods	Mean
— BERT-base —	
ATTN-W	97.1
ATTN-N	85.2
ATTNRES-W	48.6
ATTNRES-N	22.3
ATTNRESLN-N	**18.8**

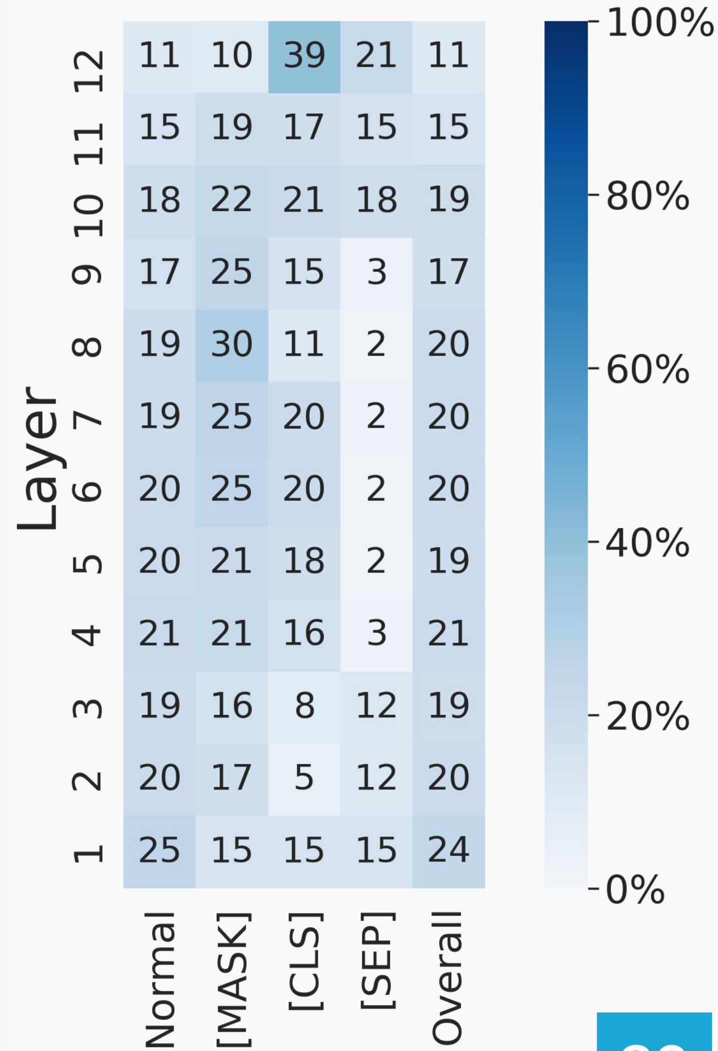
ATTN-N ATTNRESLN-N (Ours)



Detailed analysis 1:
Differences by layers and tokens

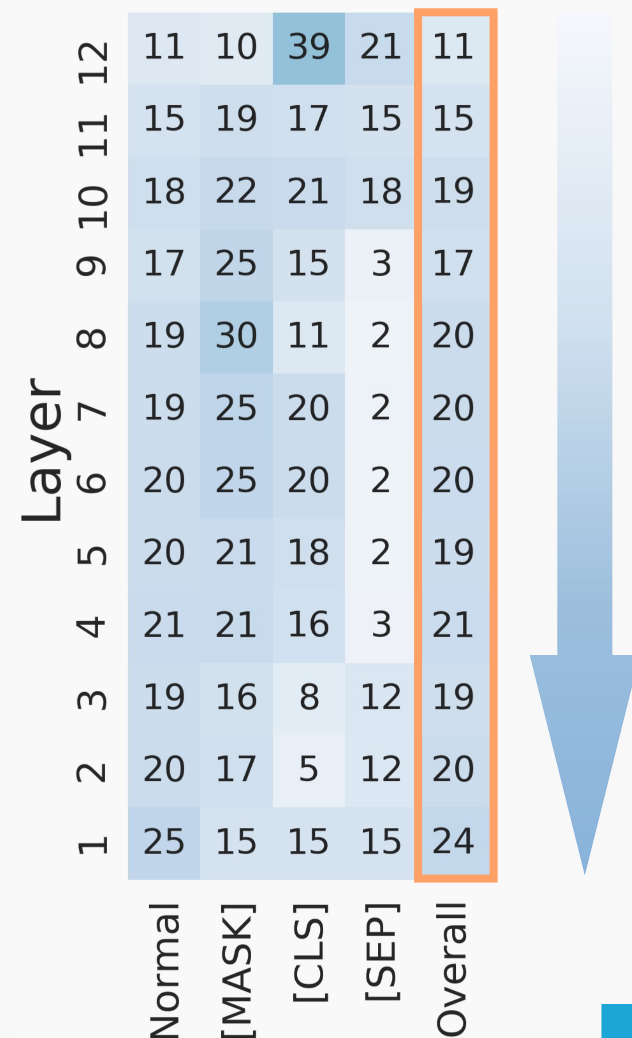
- **Mixing ratio at each layer computed with our method**

- **Token categories**
  - Normal: non-special tokens
  - [MASK]
  - [CLS]
  - [SEP]



Detailed analysis 1: Differences by layers and tokens

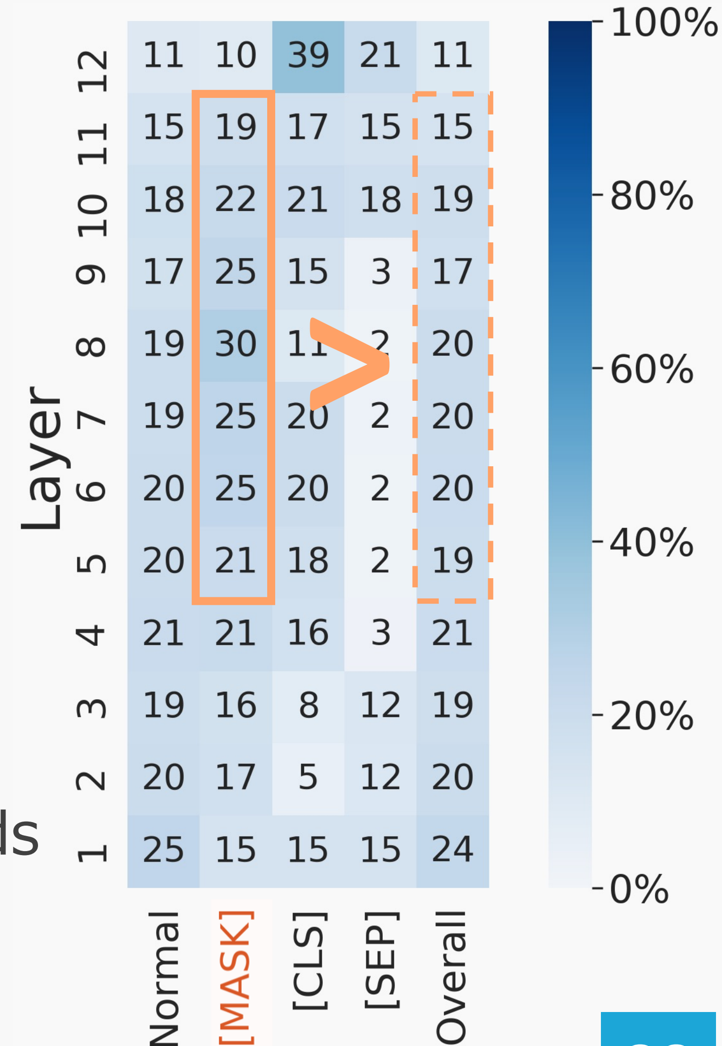
- **Mixing ratio is relatively higher in the earlier layers**



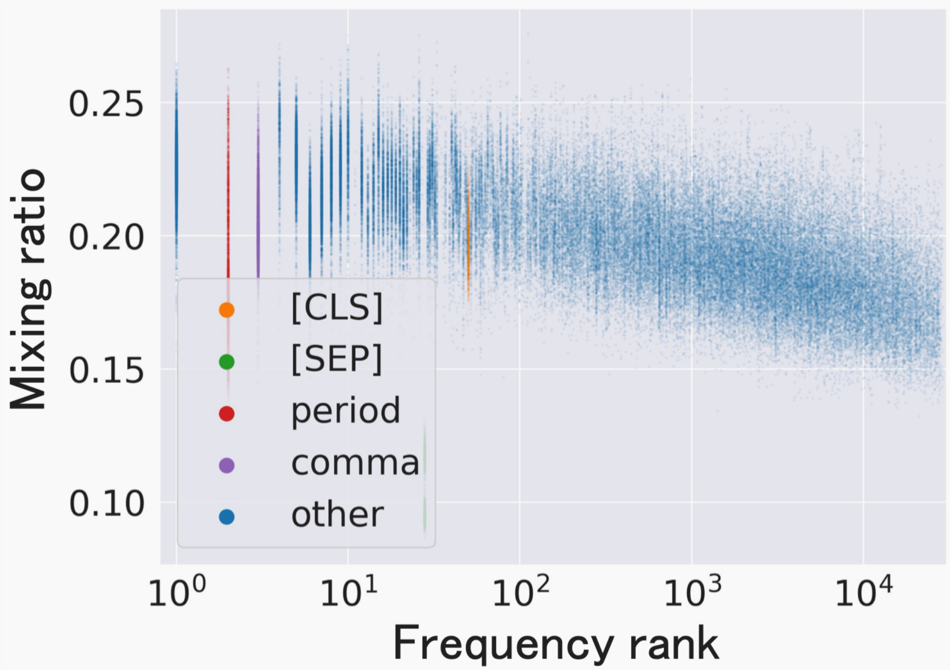
Detailed analysis 1: Differences by layers and tokens

- Mixing ratio is relatively higher in the earlier layers
- **Mixing ratio for [MASK] is relatively high in the middle and deep layers**

These layers refer to more contextual information for predicting masked words



Detailed analysis 2: Relationship with the word frequency



- **Strong negative correlation** (Spearman's $\rho = -0.54$)
- **Higher frequent word tends to gather more contextual information than lower frequent word**

→ Suggests that BERT discounts the information of high-frequency words

Summary

- Propose to analyze Transformers considering RES and LN in addition to ATTN
- Our analysis of MLMs reveals:
    - Mixing ratio is lower than previously assumed
    - Mixing is relatively strong to update MASK tokens
    - Contribution of contextual information is related to word frequency

https://github.com/gorokoba560/norm-analysis-of-transformer

Summary

- Propose to analyze Transformers considering RES and LN in addition to ATTN
- Our analysis of MLMs reveals:
  - Mixing ratio is lower than previously assumed
  - Mixing is relatively strong to update MASK tokens
  - Contribution of contextual information is related to word frequency

Questions & comments are welcome!!
I'm not a native speaker of English.
Please speak simply and slowly 🙏

https://github.com/gorokoba560/norm-analysis-of-transformer