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

Summary

- Proposed to analyze Transformers considering:
  - Multi-head attention (ATTN)
  - Residual connection (RES) 👈 **new!**
  - Layer Normalization (LN) 👈 **new!**

Analysis of Masked LMs revealed:

- Mixing via ATTN is weaker than previously assumed
- Strength of mixing is related to word frequency

Analysis scope

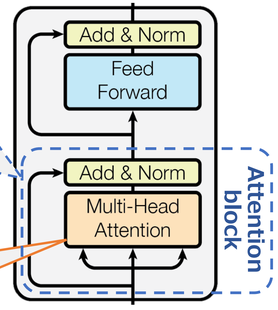
Succes of Transformers [vaswani+'17]

- Especially Masked LMs (e.g., BERT [devlin+'17])

➡ **Their mechanisms/properties should be clarified**

This study:
analyzes the whole **Attention block**
- Multi-head attention (**ATTN**)
- Residual connection (**RES**)
- Layer normalization (**LN**)

Existing studies:
typically analyzed the **ATTN** alone

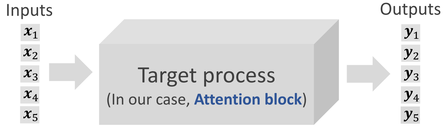


Analysis method

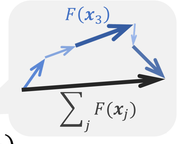
Norm-based strategy [Kobayashi+'20]

Compute the contribution of each iput to the output:

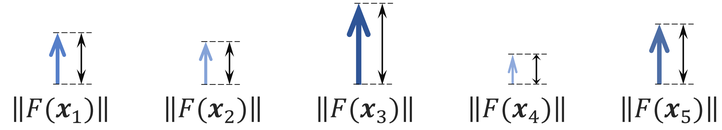
1) Decompose the target process into the sum of transformed input vectors



$$y_i = \sum_j F(x_j)$$



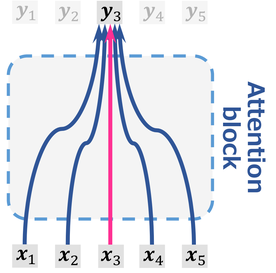
2) Compute each contribution by the norm



Mixing ratio

Information flow ($y_i \leftarrow X$) can be decomposed into:

$$y_i = \underbrace{\tilde{x}_{i \leftarrow \text{context}}}_{\text{Mixing}} + \underbrace{\tilde{x}_{i \leftarrow i}}_{\text{Preserving}} + \underbrace{\beta}_{\text{bias}}$$



Mixing ratio:

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

Experiment setup

Compute mixing ratio at each attention block
- Model: pre-trained BERT-base [devlin+'19]
- Data: Excerpts from Wikipedia [Clark+'19]

Results with more MLMs & more data are in the paper!

Experiment 1: Mean of mixing ratio

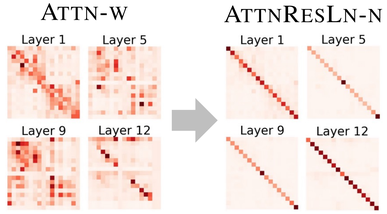
Expanded method shows lower mixing ratio

18.8% computed with ours

➡ **Mixing ≪ Preserving**

Methods	Mean
— BERT-base —	
ATTN-W	97.1
ATTN-N	85.2
ATTNRESLN-N (Ours)	**18.8**

Mixing via ATTN is weaker than previously assumed



Experiment 2: Mixing ratio and frequency

Explored the relationship with the frequency rank

Negative correlation
(Spearman's $\rho = -0.54$)

More contexts are gathered at the higher frequent word