Transformer Language Models Handle Word Frequency in Prediction Head

(Findings of ACL2023)

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

[1] Tohoku University [2] RIKEN [3] MBZUAI

Paper link

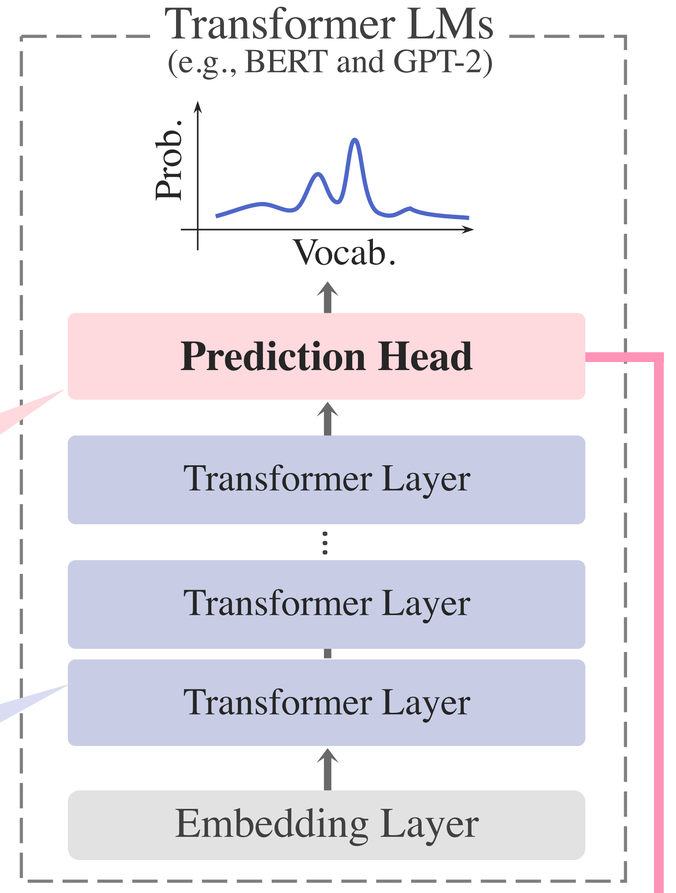
Overview

- Investigate **the inner workings of Prediction head** in Transformer LMs
- Show **bias vectors in prediction head handle word frequency** to adjust the prediction

Prediction head in Transformer LMs

- **Prediction head** is the last block of Transformer LMs
  ‖
  Can directly impact prediction
- **Has been overlooked in Transformer analyses**

- Transformer layer has been typically analyzed:
  - Attention mechanism [Clark+'19;Kobayashi+'20;etc.]
  - Feed-forward network [Geva+'21;Dai+'22;etc.]

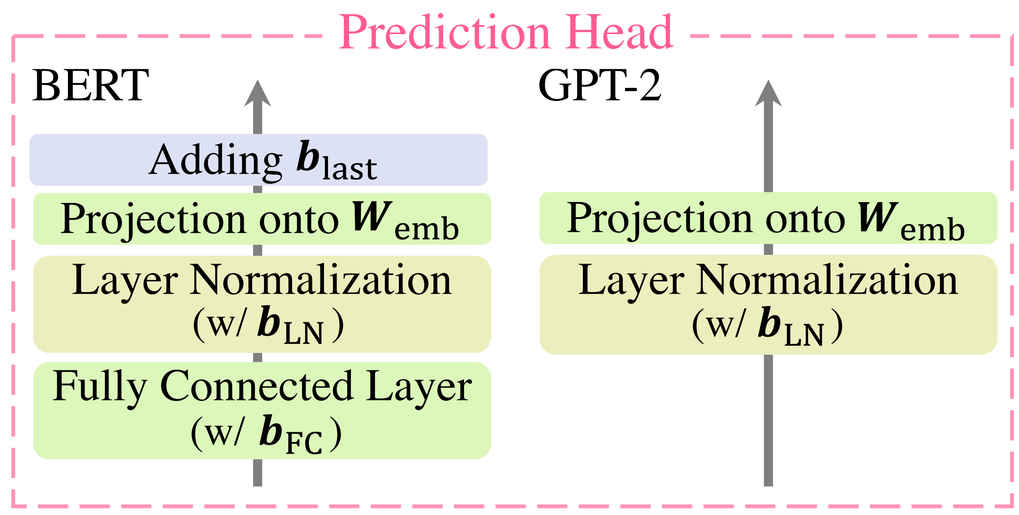


Our analysis on prediction head

- Analyze prediction head focusing on **bias vectors**
  - BERT has three bias vectors: $\boldsymbol{b}_{\mathrm{FC}}$, $\boldsymbol{b}_{\mathrm{LN}}$, $\boldsymbol{b}_{\mathrm{last}}$
  - GPT-2 has one bias vector: $\boldsymbol{b}_{\mathrm{LN}}$

  Bias parameters can be easily mapped to the output space (i.e., word prediction) and interpreted

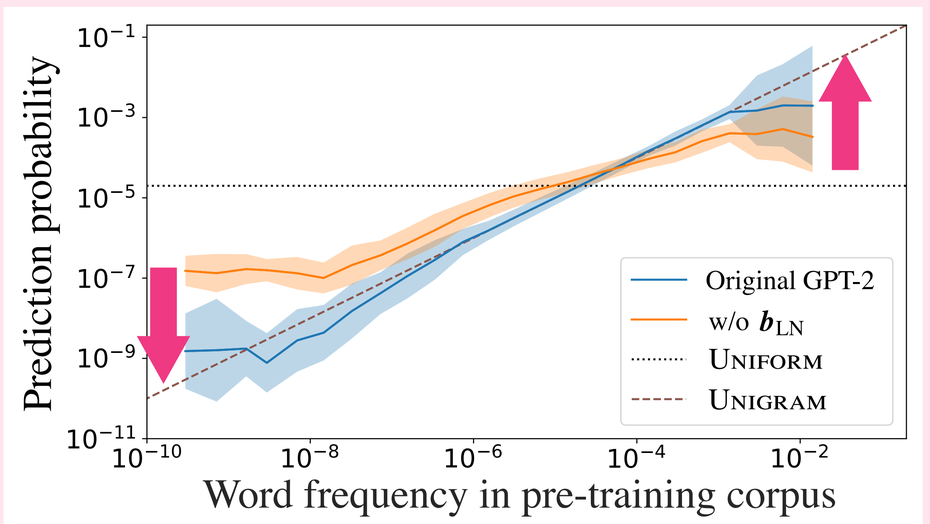
- Investigate the impact of these biases on the model's prediction and generation with respect to word frequency



Findings 1: Specific bias vector handles word frequency

- Compare the word prediction with/without a bias vector $\boldsymbol{b}_{\mathrm{LN}}$
- Bias vector $\boldsymbol{b}_{\mathrm{LN}}$ adjusts word prediction
  - To **promote high-frequency words** ↑
  - To **suppress low-frequency words** ↓
  - To be **closer to word frequency distribution** (UNIGRAM)

➡ **In the output embedding space, word frequency is encoded in the bias vector's direction**



Findings 2: Simply controlling the bias can improve generation

- Control the bias $\boldsymbol{b}_{\mathrm{LN}}$ with coefficient $\lambda \in [0,1]$

  $\lambda \times \boldsymbol{b}_{\mathrm{LN}}$

- Weakening the effect of the bias $\boldsymbol{b}_{\mathrm{LN}}$ can
  - **Improve diversity**
  - **Maintain quality** (for large models)

➡ **Simple way to make the model's generation more diverse**

- This can be seen as analogous to Logit adjustment methods

(More details in the paper)

Model	λ	Diversity ↑			Quality	
		D_1	D_2	D	MAUVE ↑	PPL ↓
large	1	0.04	0.30	0.47	0.90	**12.7**
	0.5	0.04	0.36	0.50	**0.91**	12.9
	0	**0.04**	**0.42**	**0.54**	0.86	13.6
xl	1	0.04	0.30	0.47	0.90	**11.4**
	0.7	0.04	0.34	0.49	**0.92**	11.5
	0	**0.04**	**0.41**	**0.53**	0.86	12.1