EMNLP 2023, 5/5 Rating, 12 Citations

Adapting Language Models to Compress Contexts

Alexis Chevalier* **Alexander Wettig*** **Anirudh Ajith** **Danqi Chen**
Department of Computer Science & Princeton Language and Intelligence
Princeton University
{achevalier, anirudh.ajith}@princeton.edu
{awettig, danqic}@cs.princeton.edu

[Submitted on 24 May 2023 (v1), last revised 4 Nov 2023 (this version, v2)]

AutoCompressors ▶ TL;DR.

背景

- ➢ Transformer-based Language Models 上下文长度有限 (Finite Context Window)，长文本计算量大 (Computational Cost)。如何在不调整上下文长度的情况下，扩展上下文长度、降低计算量?

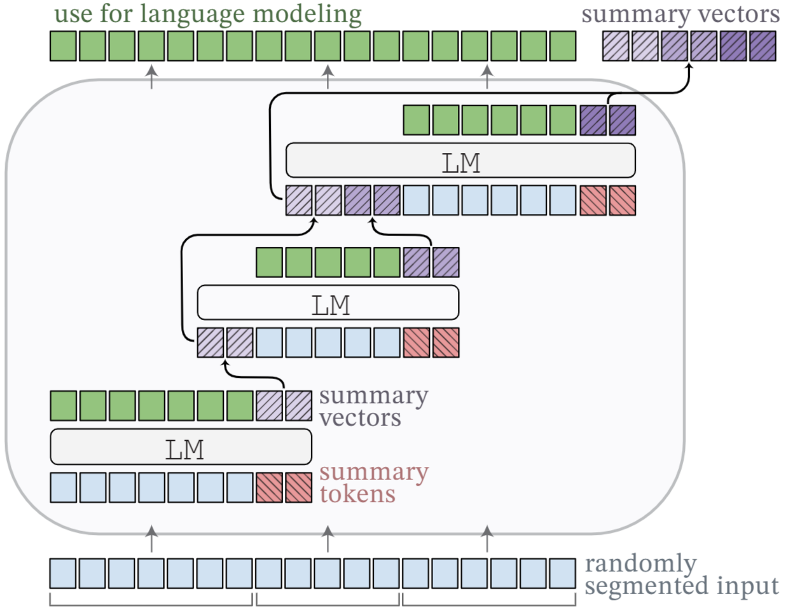
- ➢ Pre-trained Language Models 能 “压缩” 上下文信息成某一种形式，比如向量

> ***Conjecture —— Compressing Capability of PLMs***
>
> *Pretrained Language Models are capable of compressing long contexts into compact summary vectors, which are then accessible to the model as soft prompts.*

贡献

- ➢ **AutoCompressors**
  - ✓ PLM 可以无监督生成 Summary Vectors，压缩输入文本信
  - ✓ Summary Vectors 可以看成一种 Soft Prompt，是可学习的
  - ✓ Summary Vectors 能递归地作为输入生成新 Summary Vecto
- ➢ **Experiments**

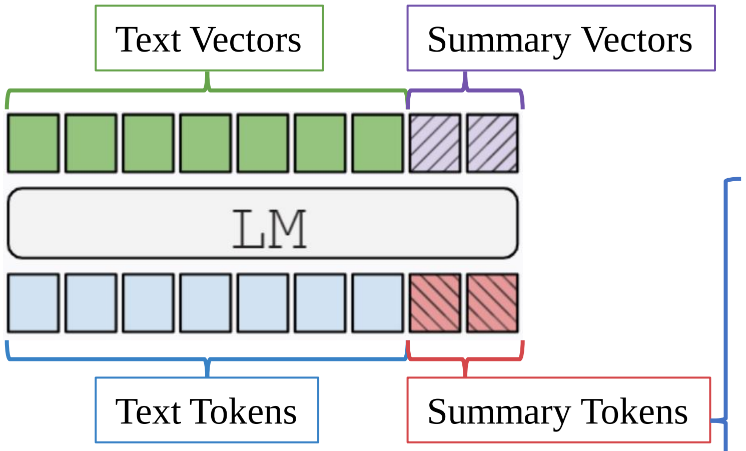
  Summary Vectors 能够编码有用的信息，扩展 In Context Learning 上下文长度，并减少推理计算量。



AutoCompressors ▶ Methodology

Rebirth of [CLS] ?!

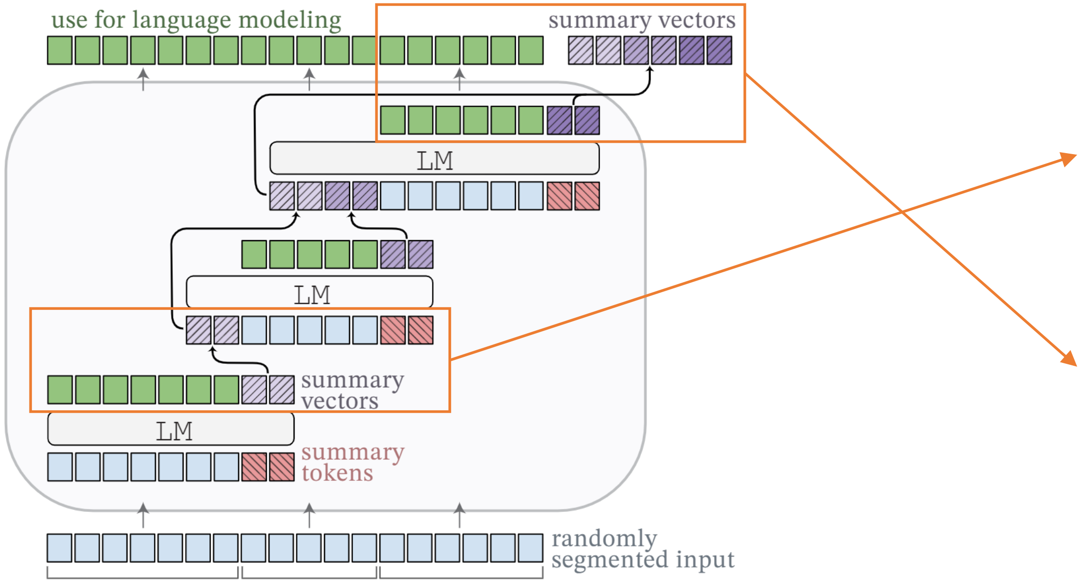
WHAT is Summary Tokens & Summary Vectors?



k 个特殊 token $\langle sum\rangle_i$ 以及它们对应的 Embedding $e_{\langle sum\rangle_i}$

不添加绝对位置编码，可添加相对位置编码（如 RoPE）

HOW to Use Summary Vectors?

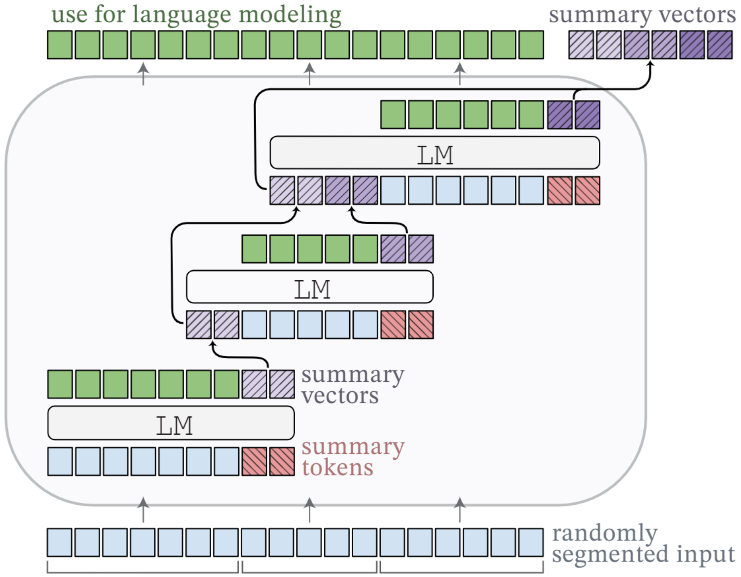


用法一：将 k 个 Summary Vectors 直接加入到其余文本的前面，用来补充缺失的前文信息（类似于 Soft Prompt）

用法二：全文分割为若干片段，通过左图的递归过程，将每个分段得到的 Summary Vectors 拼接，得到全文的 Summary Vectors，叫做 Summary Accumulation。

AutoCompressors ▶ Methodology

HOW to Train the Summary Vectors?



损失函数： Autoregressive Cross Entropy Loss

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{m_i}\log p(x_t^i \mid x_1^i, \ldots, x_{t-1}^i, \sigma_{<i})$$

增加难度： Randomized Segmenting Length

输入文本将被以随机长度切分成随机个数个 Segments

梯度控制： BPTT with stop-gradients

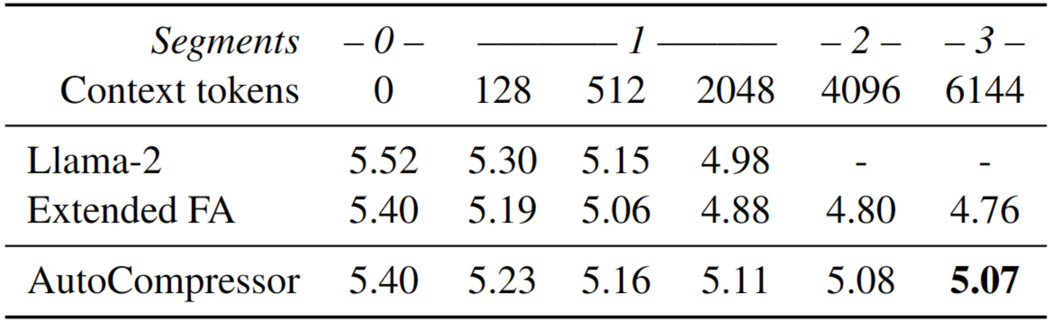
一组 Summary Vectors 在 2 个 Compression 操作后不再传播梯度

保证了这组 Summary Vectors 能压缩信息、还能用于预测 token

AutoCompressors ▶ Experiments

实验设置

- **Backbone：** Llama-2-7b，2048 max sequence length，RoPE position embedding
- **Metric：** Perplexity，↓ is better
- **Settings：** 输入一个 8192 token 的文档的前 6144 个 token，比较最后输出的 2048 token 的 Perplexity
- **Variety：**
  - **Extended FA：** RoPE 是一种相对位置编码，改变 θ 能使之扩展到更长的上下文
  - **AutoCompressor：** 将一定长度的上下文压缩到 $50 \times Segments$ 个 Summary Vectors 中，其他不变



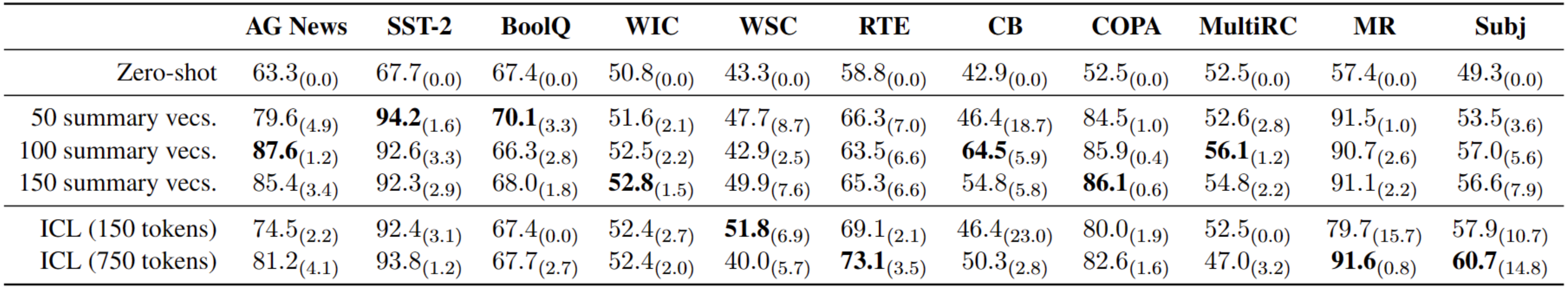
Segments	– *0* –		*1*		– *2* –	– *3* –
Context tokens	0	128	512	2048	4096	6144
Llama-2	5.52	5.30	5.15	4.98	-	-
Extended FA	5.40	5.19	5.06	4.88	4.80	4.76
AutoCompressor	5.40	5.23	5.16	5.11	5.08	**5.07**

对于使用 RoPE 的 Llama-2-7B 来说，将长度 6144 的上下文压缩成 150 个 Summary Vectors，能比较接近纯 RoPE 延拓的困惑度，并显著降低推理计算量（6144 计算量一定比 150 计算量大）

AutoCompressors ▶ Experiments

实验设置

- **Backbone：** Llama-2-7b，2048 max sequence length，RoPE position embedding
- **Metric：** 视数据集而定，越高越好
- **Settings：** 测试 Summary Vectors 对 In Context Learning 的帮助
    - **X Summary Vecs.:** 每 50 个 Summary Vectors 都来自于 750 tokens 的 Demonstrations;
    - **ICL (Y tokens):** 在上下文中加 Y tokens 的 Demonstrations.

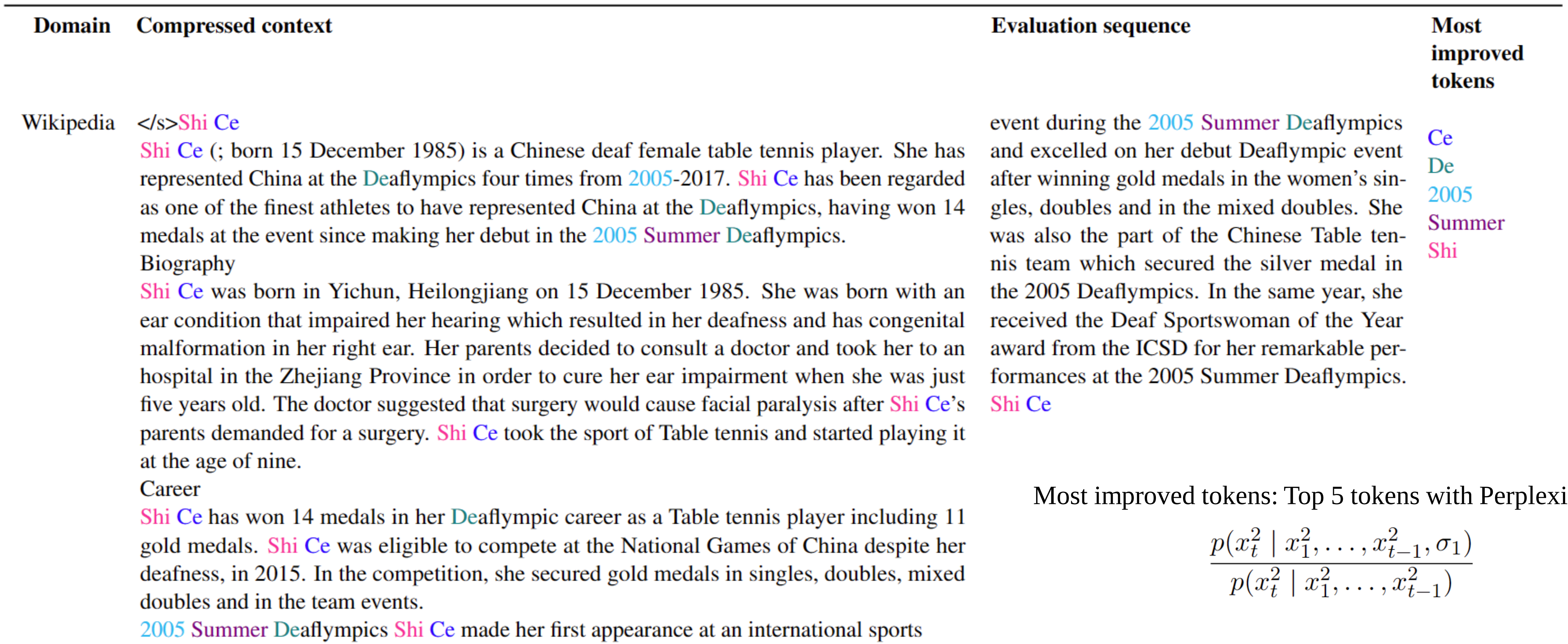


	AG News	SST-2	BoolQ	WIC	WSC	RTE	CB	COPA	MultiRC	MR	Subj
Zero-shot	63.3$_{(0.0)}$	67.7$_{(0.0)}$	67.4$_{(0.0)}$	50.8$_{(0.0)}$	43.3$_{(0.0)}$	58.8$_{(0.0)}$	42.9$_{(0.0)}$	52.5$_{(0.0)}$	52.5$_{(0.0)}$	57.4$_{(0.0)}$	49.3$_{(0.0)}$
50 summary vecs.	79.6$_{(4.9)}$	**94.2**$_{(1.6)}$	**70.1**$_{(3.3)}$	51.6$_{(2.1)}$	47.7$_{(8.7)}$	66.3$_{(7.0)}$	46.4$_{(18.7)}$	84.5$_{(1.0)}$	52.6$_{(2.8)}$	91.5$_{(1.0)}$	53.5$_{(3.6)}$
100 summary vecs.	**87.6**$_{(1.2)}$	92.6$_{(3.3)}$	66.3$_{(2.8)}$	52.5$_{(2.2)}$	42.9$_{(2.5)}$	63.5$_{(6.6)}$	**64.5**$_{(5.9)}$	85.9$_{(0.4)}$	**56.1**$_{(1.2)}$	90.7$_{(2.6)}$	57.0$_{(5.6)}$
150 summary vecs.	85.4$_{(3.4)}$	92.3$_{(2.9)}$	68.0$_{(1.8)}$	**52.8**$_{(1.5)}$	49.9$_{(7.6)}$	65.3$_{(6.6)}$	54.8$_{(5.8)}$	**86.1**$_{(0.6)}$	54.8$_{(2.2)}$	91.1$_{(2.2)}$	56.6$_{(7.9)}$
ICL (150 tokens)	74.5$_{(2.2)}$	92.4$_{(3.1)}$	67.4$_{(0.0)}$	52.4$_{(2.7)}$	**51.8**$_{(6.9)}$	69.1$_{(2.1)}$	46.4$_{(23.0)}$	80.0$_{(1.9)}$	52.5$_{(0.0)}$	79.7$_{(15.7)}$	57.9$_{(10.7)}$
ICL (750 tokens)	81.2$_{(4.1)}$	93.8$_{(1.2)}$	67.7$_{(2.7)}$	52.4$_{(2.0)}$	40.0$_{(5.7)}$	**73.1**$_{(3.5)}$	50.3$_{(2.8)}$	82.6$_{(1.6)}$	47.0$_{(3.2)}$	**91.6**$_{(0.8)}$	**60.7**$_{(14.8)}$

现象一： 并非越多 Summary Vectors 越好

现象二： Summary Vectors 在一些数据集上并不优于 ICL，即使 Summary Vectors 来自更多的 Demo

AutoCompressors ▶ Case Study



Domain	Compressed context	Evaluation sequence	Most improved tokens
Wikipedia	</s>Shi Ce Shi Ce (; born 15 December 1985) is a Chinese deaf female table tennis player. She has represented China at the Deaflympics four times from 2005-2017. Shi Ce has been regarded as one of the finest athletes to have represented China at the Deaflympics, having won 14 medals at the event since making her debut in the 2005 Summer Deaflympics. Biography Shi Ce was born in Yichun, Heilongjiang on 15 December 1985. She was born with an ear condition that impaired her hearing which resulted in her deafness and has congenital malformation in her right ear. Her parents decided to consult a doctor and took her to an hospital in the Zhejiang Province in order to cure her ear impairment when she was just five years old. The doctor suggested that surgery would cause facial paralysis after Shi Ce's parents demanded for a surgery. Shi Ce took the sport of Table tennis and started playing it at the age of nine. Career Shi Ce has won 14 medals in her Deaflympic career as a Table tennis player including 11 gold medals. Shi Ce was eligible to compete at the National Games of China despite her deafness, in 2015. In the competition, she secured gold medals in singles, doubles, mixed doubles and in the team events. 2005 Summer Deaflympics Shi Ce made her first appearance at an international sports	event during the 2005 Summer Deaflympics and excelled on her debut Deaflympic event after winning gold medals in the women's singles, doubles and in the mixed doubles. She was also the part of the Chinese Table tennis team which secured the silver medal in the 2005 Deaflympics. In the same year, she received the Deaf Sportswoman of the Year award from the ICSD for her remarkable performances at the 2005 Summer Deaflympics. Shi Ce	Ce De 2005 Summer Shi

Most improved tokens: Top 5 tokens with Perplexity Gain

$$\frac{p(x_t^2 \mid x_1^2, \ldots, x_{t-1}^2, \sigma_1)}{p(x_t^2 \mid x_1^2, \ldots, x_{t-1}^2)}$$

可以看出人名、时间、地名等 token 从 Summary Vectors 的收益大；

而且生成序列的确能继续承接原上下文的语义，说明 PLMs 的确可以做到压缩。