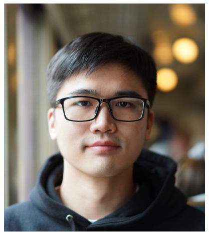
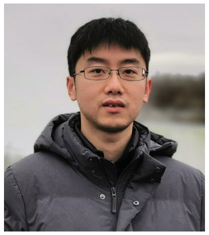
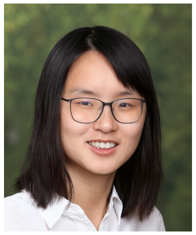
Training Language Models with Memory Augmentation





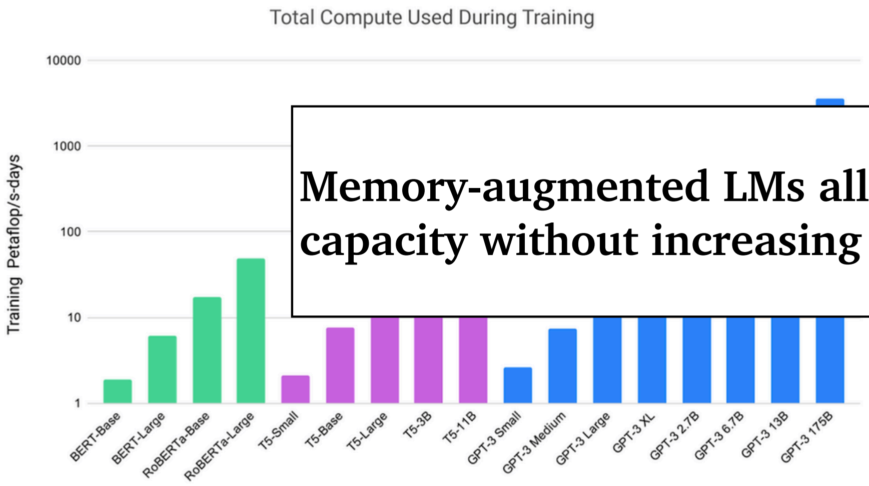


Zexuan Zhong Tao Lei Danqi Chen

Princeton University

Motivation

Large language models are very expensive to train/use!



(Brown et al., 2020)

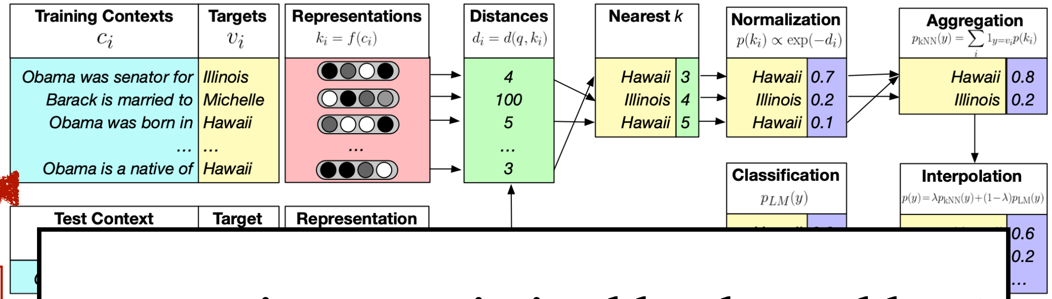
Gopher (280B)
2021/12

Memory-augmented LMs allow us to increase model capacity without increasing model size!

OPT (175B)
2022/05

Prior works on memory-augmented LMs

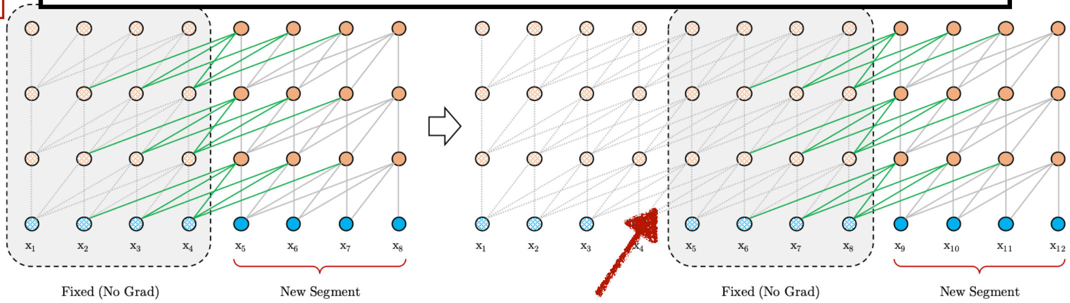
kNN-LM



Memory is constructed using a ***standard*** *LM only during* ***inference****!*

Memory is not optimized by the end loss!

Transformer-XL



Memory is constructed from ***previous*** *batches with a* ***stop gradient*** *operation!*

Khandelwal et al., 2021. Generalization through memorization: Nearest neighbor language models.
Dai et al., 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

Our approach: TRIME

Key idea: building a **training memory** from the same training batch on the fly

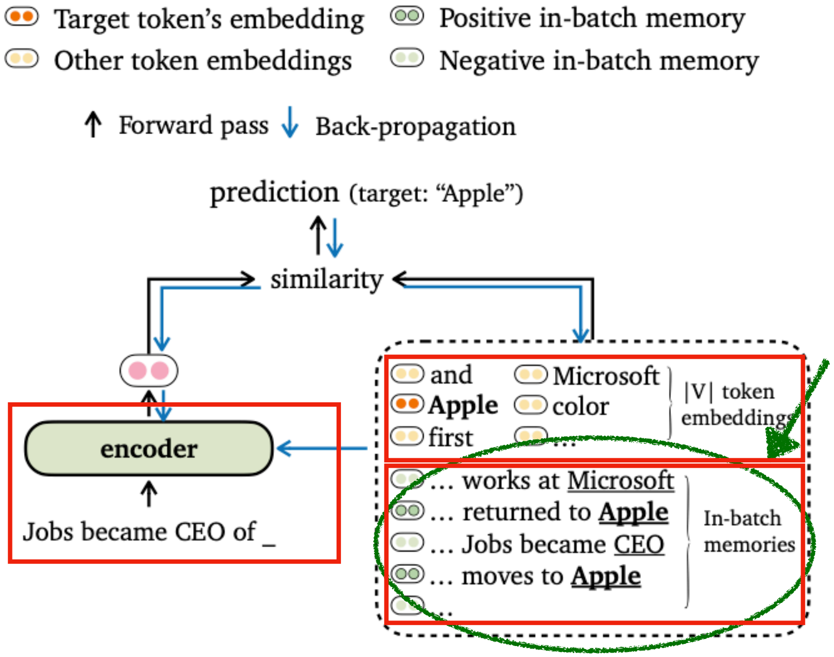
c_t: Jobs became CEO of ___

1) ●● **Apple** (output embedding)

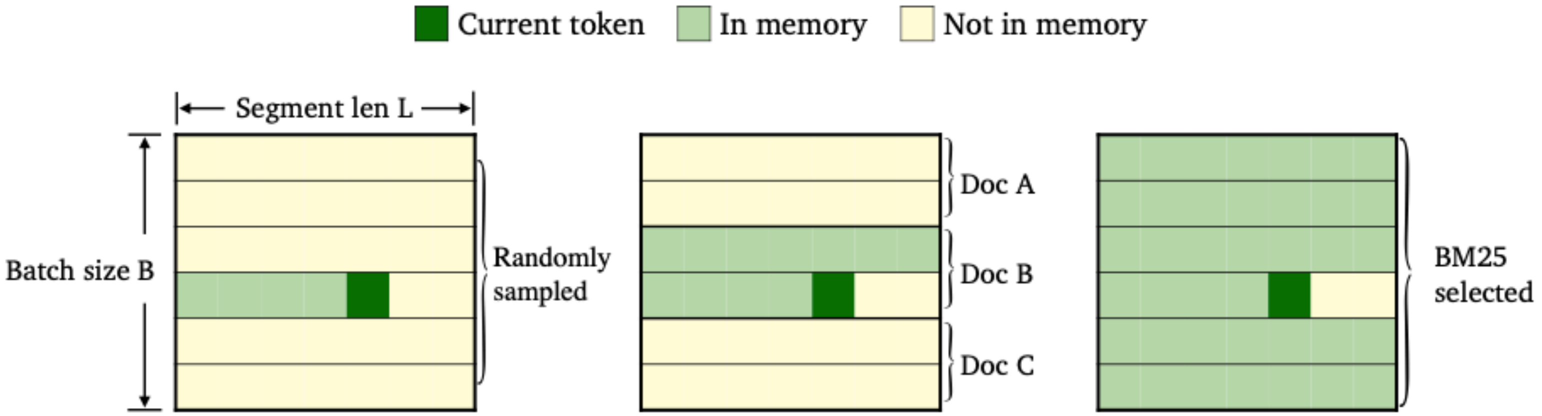
2) Other c in the **training memory** that share the same next word as x_t

   ●● … returned to **Apple**

   ●● … moves to **Apple**



Three TRIME language models



(a) Default batching

(b) Batching consecutive segments

(c) Batching lexically similar segments

TrimeLM
(local memory)

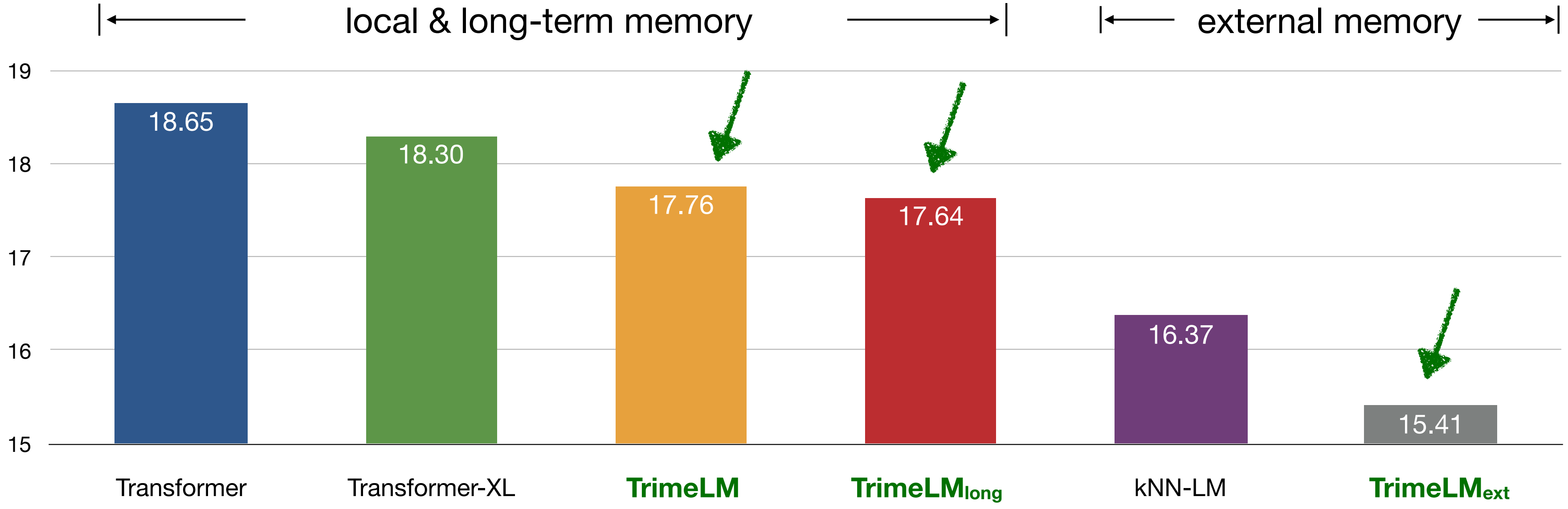
TrimeLM$_{\text{long}}$
(long-term memory)

TrimeLM$_{\text{ext}}$
(External memory)

Compared to Vanilla LM · Transformer-XL · kNN-LM

Experiments: WikiText-103

(247M model, L=3072)

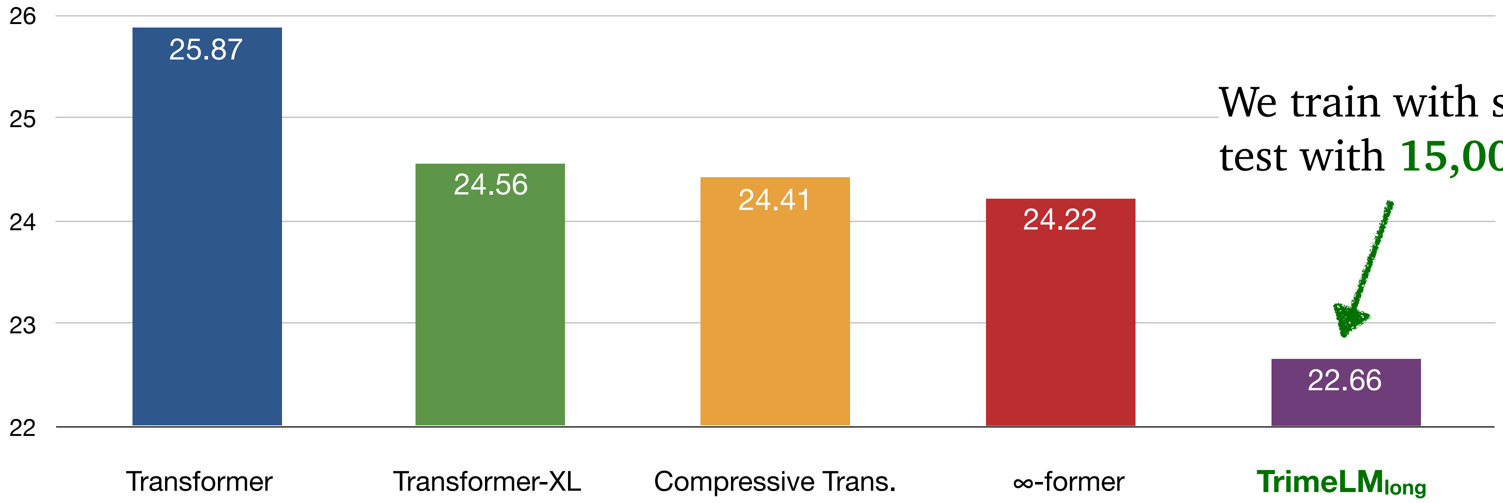


Metric: perplexity

Khandelwal et al., 2021. Generalization through memorization: Nearest neighbor language models.
Dai et al., 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

Experiments: WikiText-103

(150M model, **L=150**)



We train with segment length **150;** test with **15,000** tokens!

Metric: perplexity

Dai et al., 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context
Rae et al., 2019. Compressive Transformers for long-range sequence modeling.
Martins et al., 2022. ∞-former: Infinite memory Transformer.

Summary: TRIME

TRIME: a **training objective** which leverages in-batch memories + three ways of memory construction

- adds very little computational overhead
- does not modify model architectures
- is compatible with other model architectures and techniques
- outperforms existing methods!

Check out more results on **domain adaptation**, **machine translation**, and **character-level LM** in the paper!

Paper: https://arxiv.org/pdf/2205.12674.pdf
Code & models: https://github.com/princeton-nlp/TRIME
Email: zzhong@cs.princeton.edu