Memorizing Transformers

Chris Winsor
DL4NLP COMP 5300
4/24/23

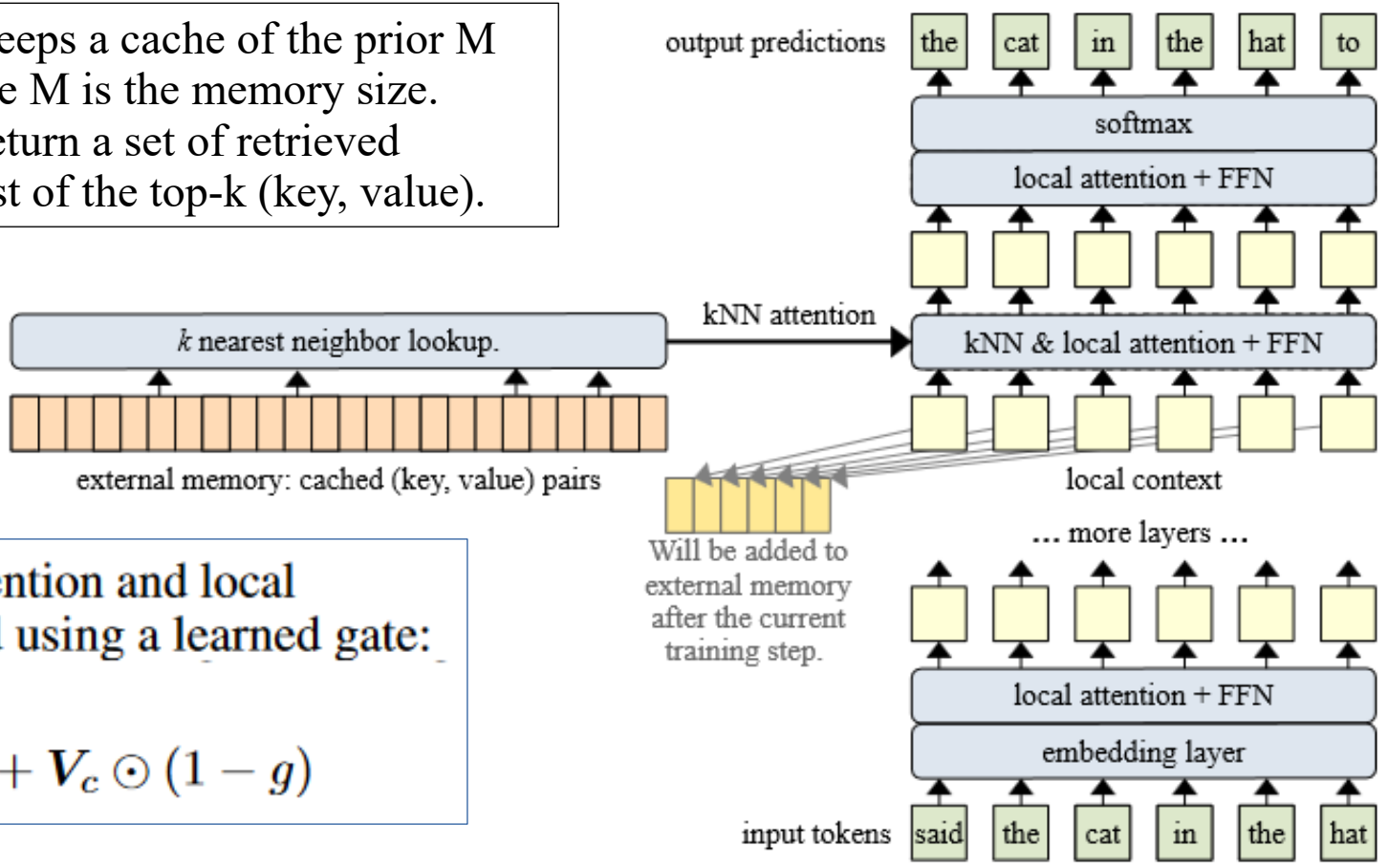
(1) Wu, Y., Rabe, M. N., Hutchins, D., & Szegedy, C. (2022). Memorizing Transformers (arXiv:2203.08913). arXiv. https://doi.org/10.48550/arXiv.2203.08913

Motivation

- **Problem**:
  - Long documents have references that are far away
    - Traditional transformer attention is limited to context length
  - A document may only have one occurrence of reference
    - Traditional transformer backprop requires thousands of examples
- **Examples**:
  - Novel, source code, theorem proofs… far away and infrequent
- **Idea**:
  - Add external memory to transformer for learning long distance relation
  - Runtime memorization - no training loop or backprop for these
- **Approach**:
  - FIFO memory + KNN lookup to retrieve prior key/value pairs

Main Contribution

- The external memory keeps a cache of the prior M (key, value) pairs, where M is the memory size.
- The kNN lookup will return a set of retrieved memories, which consist of the top-k (key, value).



The results of kNN-attention and local attention are combined using a learned gate:

$$g = \sigma(b_g)$$

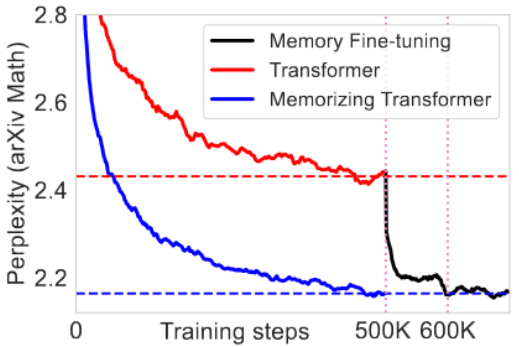
$$\boldsymbol{V}_a = \boldsymbol{V}_m \odot g + \boldsymbol{V}_c \odot (1 - g)$$

Results

- External memory resulted in significant perplexity reduction compared to vanilla transformer (all datasets)
- Best performance combined memory with XL cache (Dai et al. 2019)
- Non-memory transformer can be fine-tuned using memory

Context	Memory	XL cache	arXiv	PG19	C4(4K+)	GitHub	Isabelle
512	None	None	3.29	13.71	17.20	3.05	3.09
2048	None	None	2.69	12.37	14.81	2.22	2.39
512	None	512	2.67	12.34	15.38	2.26	2.46
2048	None	2048	2.42	11.88	14.03	2.10	2.16
512	1536	None	2.61	12.50	14.97	2.20	2.33
512	8192	None	2.49	12.29	14.42	2.09	2.19
512	8192	512	2.37	11.93	14.04	2.03	2.08
512	65K	512	2.31	11.62	14.04	1.87	2.06
2048	8192	2048	2.33	11.84	13.80	1.98	2.06
2048	65K	2048	**2.26**	**11.37**	**13.64**	**1.80**	**1.99**

Average token-level perplexities of each model when trained for 500k steps.



Finetuning a 1B vanilla Transformer model to use external memory of size 65K.