

Adapting Language Models to Compress Contexts



* equal contribution

Alexis Chevalier* Alexander Wettig* Anirudh Ajith Danqi Chen Princeton Language and Intelligence, Princeton University Summary use for language modeling summary vectors LМ summary vectors LM summary tokens **Evaluation** plain-text demonstrations Zero-shot ICL (150 tokens) ICL (750 tokens) 50 summary vecs. 100 summary vecs. 150 summary vecs. 50/100 1 50/150 50 55 60 **Extended Full-Attention** RMT AutoCompressor 1. Compress all the documents in corpus to summary vectors 2048 512 4096 6144 20 420 420 421 Context Length (#Tokens) 11/ 100 AutoCompressor

randomly segmented input Fine-tuning an AutoCompressor from a pre-trained LM In-Context Learning Can summary vectors encode task semantics? • Encoding demonstrations as summary vectors can be seen as zero-shot soft prompt tuning • On 5/7 SuperGLUE tasks, conditioning on 150 summary vectors outperforms 750 tokens worth of Llama2-7B Extended Full-Attention (initialize new 65 70 OPT-2.7B Avg. Accuracy (11 tasks) **Retrieval-Augmented LM** 2. Retrieve passages and use their summary vectors for efficient inference Generate... AutoCompressor pre-computed summary Fetch vectors top-k Off-the-shelf Retriever Prompt Using summary vectors outperforms retrieving Pareto front same-size token passages at same inference speed AutoComp-2.7B 20 summary vecs AutoComp-2.7B 50 summary vecs OPT-2.7E Retrieve 50 token passages # Retrieved Passages OPT-1.3B 250 Retrieve 50 summary vectors 100 100<mark>250</mark> 150 OPT-350M 2 OPT-125M 100 OPT-2.7B 250 5 250 50 100 100 -50 10 50 50 50 0.0 0.5 3.0 1.0 1.5 2.0 2.5 3.5 10 12 6 8 Retrieval-Augmented Perplexity Gain (%)

Goals:

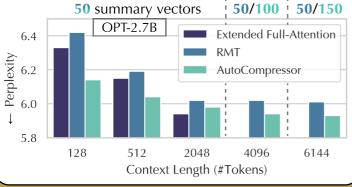
- Speed up attention over long contexts for both training and inference
- Reduce memory cost of storing past key-value cache

Introducing *AutoCompressors*:

- Compress text into compact summary vectors which the model can read as soft prompts
- Summary vectors can be *accumulated* to compress long documents
- Simple training objective: Fine-tune from pre-trained model with next-token prediction
- Summary vectors of a document can be *cached* and re-used as a concise context to relevant prompts

Language Modeling

- We present AutoCompressors based on OPT-1.3B, OPT-2.7B and Llama-2-7B, fine-tuned on sequences of up to 30k tokens (OPT) and 6k tokens (Llama) All models available at huggingface.co/princeton-nlp!
- Recursively compress up to 2048 tokens into 50 tokens
- Measure gains in perplexity from adding raw context tokens vs. summary vectors
- Baselines: •
 - **RMT** [Bulatov et al., 2022]
 - (no summary accumulation, fixed segment length)
 - position embeddings / RoPE interpolation)



Passage Re-ranking

- Re-rank passages based on *p*(*query* | *passage*)
- Preprocess: 512 token passage \rightarrow 50 summary vectors
- Inference with large models based on summary vectors is superior to small models based on full passages

