Decoupled Context Processing for Context Augmented Language Modeling

Zonglin Li, Ruiqi Guo, Sanjiv Kumar

Google Research

Take Home Messages

- **Vanilla (decoupled) Encoder-Decoder Transformer** is a good choice for integrating context in retrieval augmentation language modeling

- Encoder-Decoder architecture further offers the opportunity to **cache** context encoding, which is **more efficient**

Background: Retrieval Augmented Language Modeling

- **Traditional LMs**

$$P(\mathbf{y}|\boxed{\mathbf{x},}\boxed{\theta'})$$

  1. **Local information** in the input
  2. **Internal knowledge** in the parameters (fixed)

- **Retrieval augmented models**

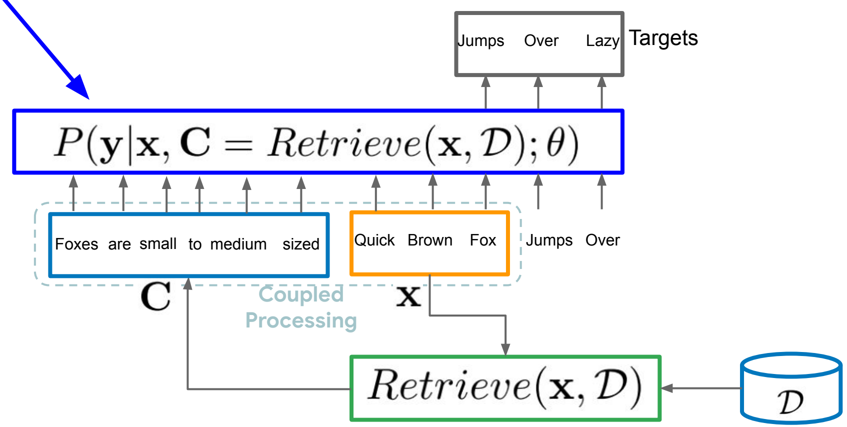
$$P(\mathbf{y}|\mathbf{x},\boxed{\mathbf{C}}=\boxed{Retrieve}(\mathbf{x},\boxed{\mathcal{D}});\theta)$$

  1. Adds retrieval as the third source
  2. Dynamically adapt to input with retriever
  3. External knowledge base can be updated

Motivation & Previous Works

Coupled processing of the context and input, after retrieval.

Method	Context Integration	Decoupled Context Encoding
kNN-LM [21]	Interpolation	Yes
Spalm [44]	Gating	Yes
Realm [15], RAG [26], FID [18]	Concat	No
Retro [4]	Chunked-Cross-Attention	No
Proposed	Encoder-Decoder Cross-Attention	Yes

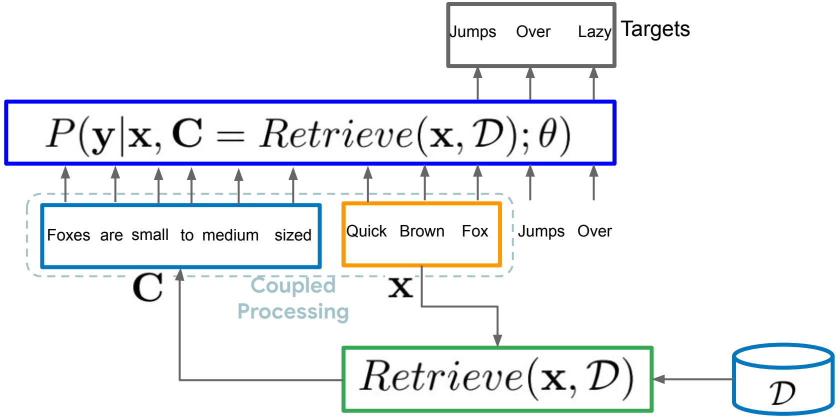


Motivation & Previous Works

Coupled processing of the context and input, after retrieval.

⇒ Context $\mathbf{C}$ conditioned on input $\mathbf{x}$ (Self attention in Realm, RAG, FiD and Chunked Cross Attention in Retro)

Method	Context Integration	Decoupled Context Encoding
kNN-LM [21]	Interpolation	Yes
Spalm [44]	Gating	Yes
Realm [15], RAG [26], FID [18]	Concat	No
Retro [4]	Chunked-Cross-Attention	No
Proposed	Encoder-Decoder Cross-Attention	Yes



Motivation & Previous Works

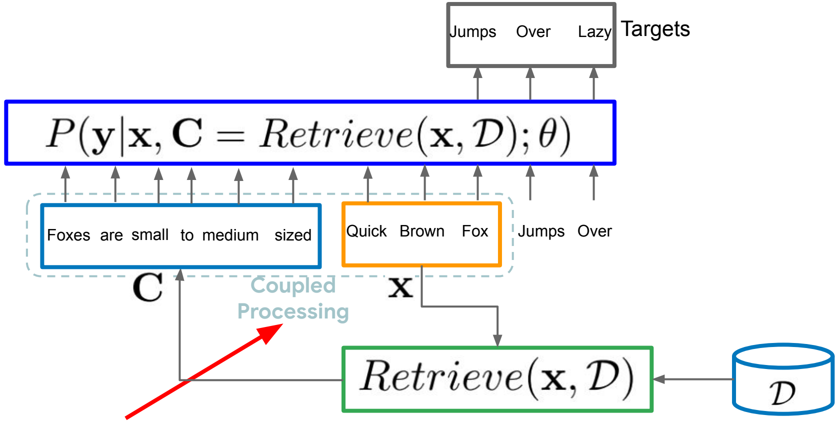
Coupled processing of the context and input, after retrieval.

⇒ Context $\mathbf{C}$ conditioned on input $\mathbf{x}$ (Self attention in Realm, RAG, FiD and Chunked Cross Attention in Retro)

⇒ Same context with different input with have different encodings

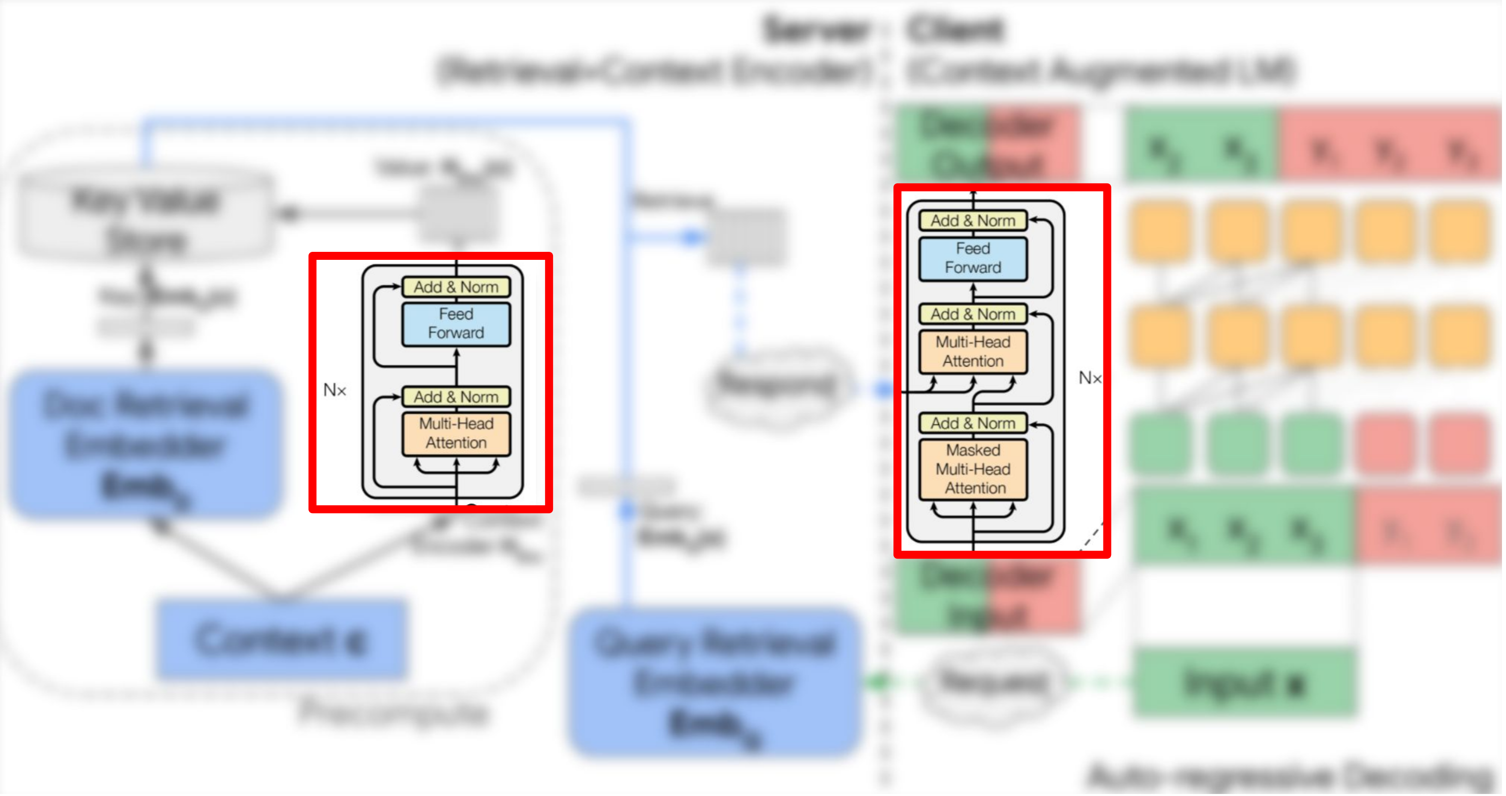
⇒ **Can't preprocess context and reuse it**

Method	Context Integration	Decoupled Context Encoding
kNN-LM [21]	Interpolation	Yes
Spalm [44]	Gating	Yes
Realm [15], RAG [26], FID [18]	Concat	No
Retro [4]	Chunked-Cross-Attention	No
Proposed	Encoder-Decoder Cross-Attention	Yes



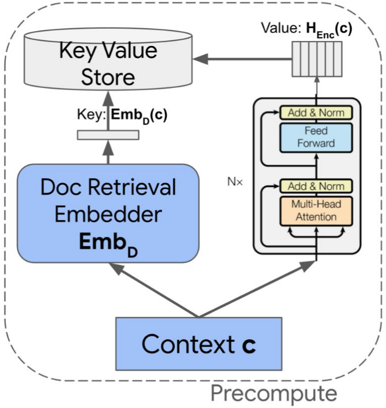
Can we avoid this?

Vanilla Encoder-Decoder Architecture



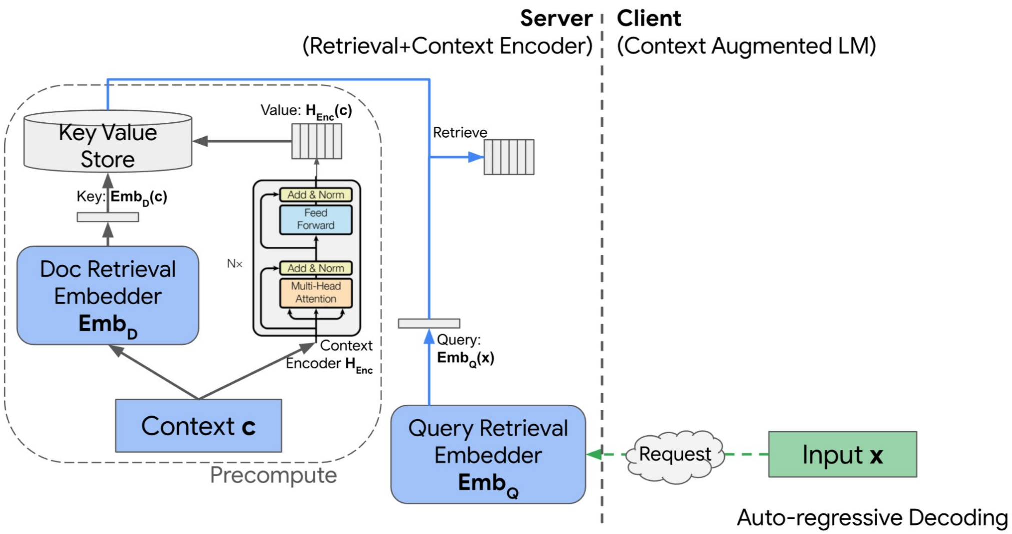
Step 0: Offline Processing

- Key: the retrieval embedding for this context
- Value: Encoder output embeddings for this context
- **Each context is encoded independently.**



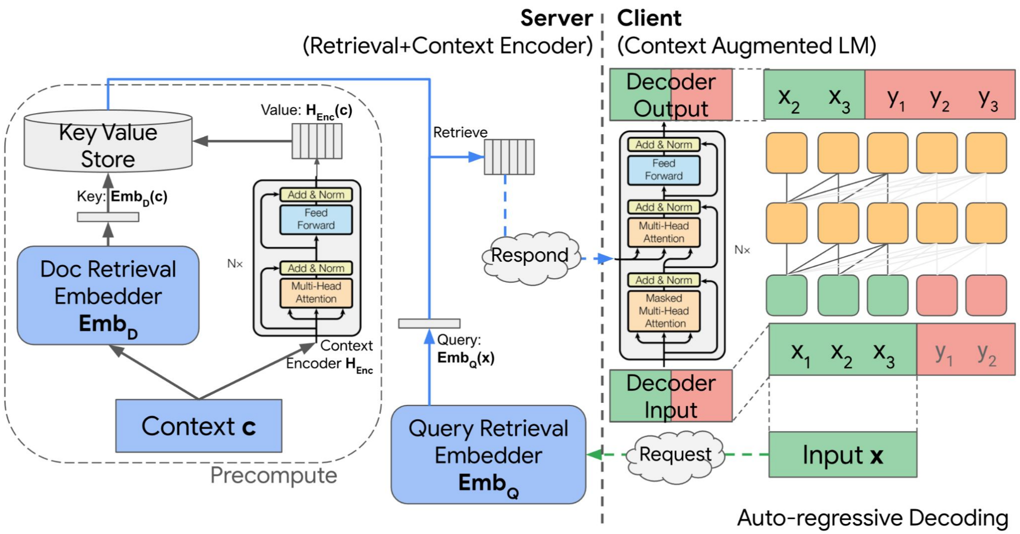
Step 1: Retrieval

- Input sequence x is converted to a query embedding to retrieve the context with largest dot-product similarity.
- Cached Encoder output is the retrieval result. **Encoder does not run during retrieval.**



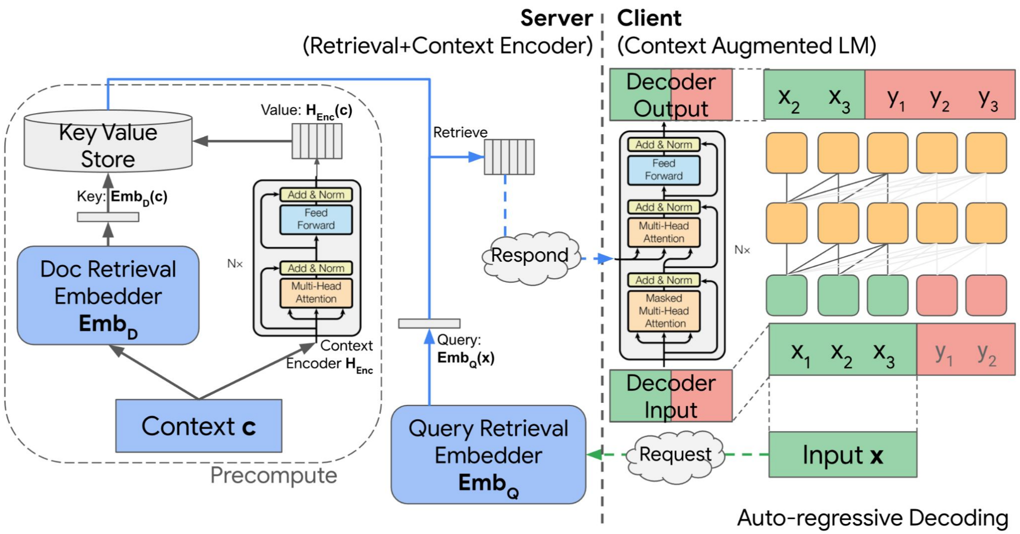
Step 2: Generation

- The retrieved Encoder output is directly passed to the decoder for inference.
- **Encoder does not run during generation either.**



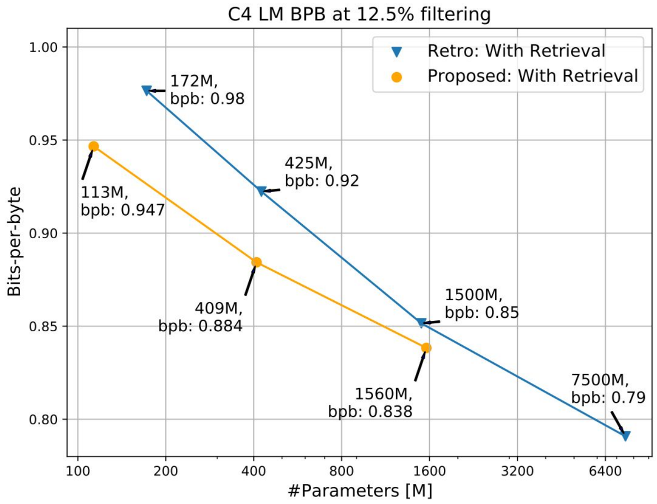
Vanilla Encoder-Decoder Architecture

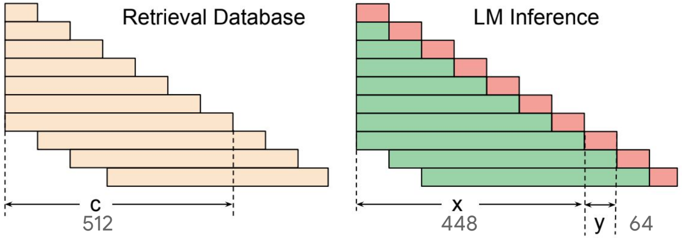
- Each context is encoded independently from the input and other context.
- **Context is encoded once and cached. Encoder is not needed for inference.**
- During training Encoder and Decoder are jointly trained.



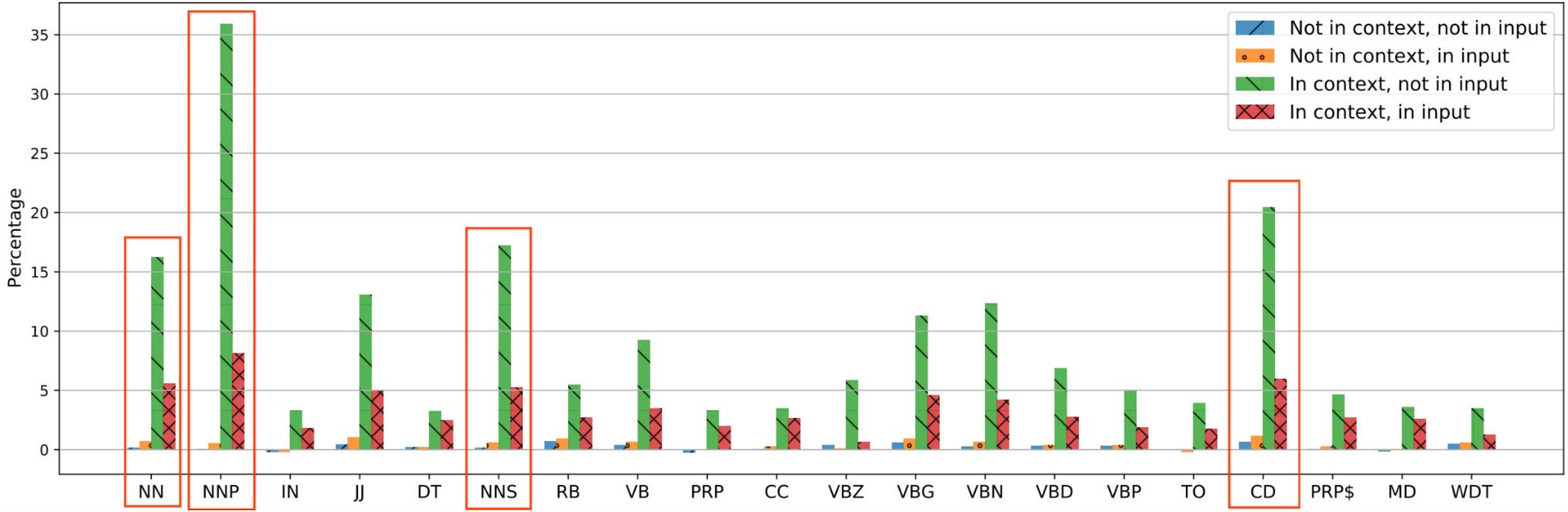
Experiments: Language Modeling

- English C4
- Measured by Bits-Per-Byte (BPB): the lower the better.
- Sliding window to create target **y**, input **x** and context **c**
- Filter the retrieved context that are too similar to the targets: remove the context and target pair from training and evaluation if they have more than 8 common consecutive tokens.





Breaking Down the Improvements



Percentage of improvement on log likelihood sliced by NLTK POS tags

Nouns and numbers benefit the most from having context. Please see our paper for the details.

Experiments: Natural Question

- Wikipedia chunks as context
- DPR retrieval, like in Retro and FiD
- 20 wiki chunks for each question. Independently encoded by the Encoder.
- Question goes to the Decoder.

Model	Model Size	Exact Match Accuracy
Realm [15]	110M	40.4
DPR [20]	110M	41.5
Ours (Large)	409M	44.35
RAG [26]	400M	44.5
Retro [4]	7.5B	45.5
Ours (XL)	1.56B	47.95
FiD [18]	770M	51.4
EMDR [37]	440M	52.5
FiD + Distill [17]	770M	54.4