CSE 561 Paper Review

Kriti Bhattarai

02/15/2024

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[†‡], Ethan Perez[⋆],

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel[†‡], Sebastian Riedel[†‡], Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; [⋆]New York University;
plewis@fb.com

GENERALIZATION THROUGH MEMORIZATION: NEAREST NEIGHBOR LANGUAGE MODELS

Urvashi Khandelwal[†*], Omer Levy[‡], Dan Jurafsky[†], Luke Zettlemoyer[‡] & Mike Lewis[‡]
[†]Stanford University
[‡]Facebook AI Research
{urvashik,jurafsky}@stanford.edu
{omerlevy,lsz,mikelewis}@fb.com

Overview

- Retrieval Augmented Generation (RAG)
- Model Setup
- Results
- Performance comparison

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[†‡], **Ethan Perez**[⋆],

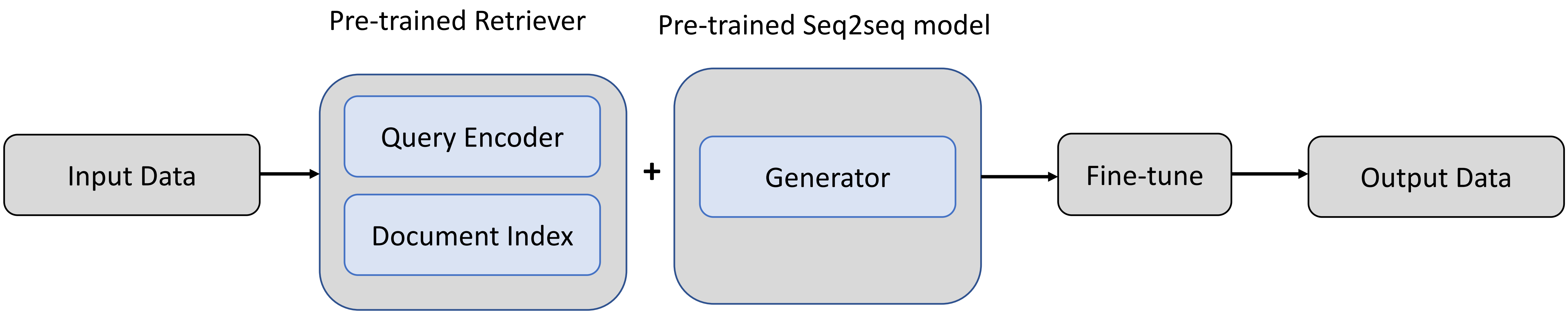
Aleksandra Piktus[†], **Fabio Petroni**[†], **Vladimir Karpukhin**[†], **Naman Goyal**[†], **Heinrich Küttler**[†],

Mike Lewis[†], **Wen-tau Yih**[†], **Tim Rocktäschel**[†‡], **Sebastian Riedel**[†‡], **Douwe Kiela**[†]

[†]Facebook AI Research; [‡]University College London; [⋆]New York University;
plewis@fb.com

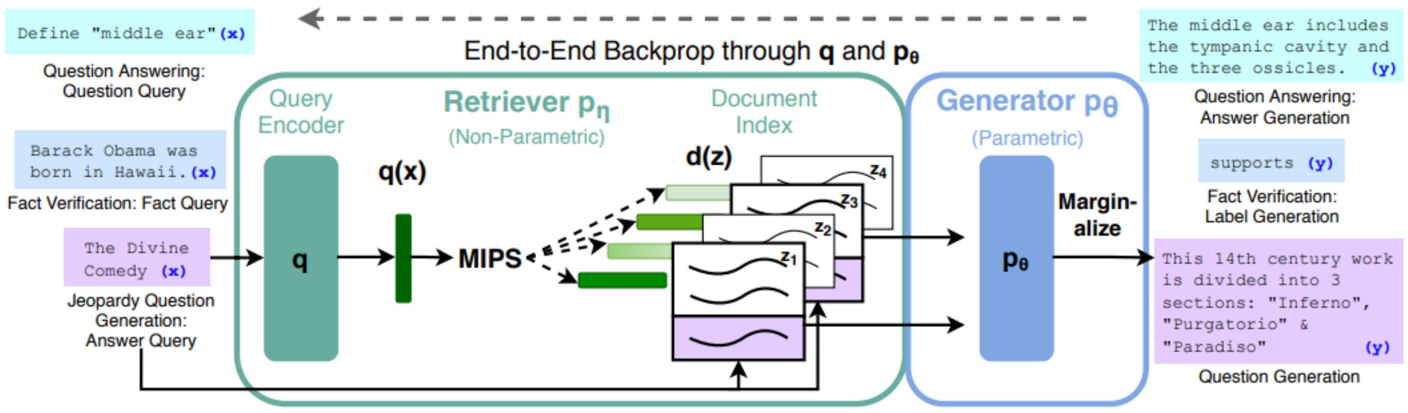
Retrieval Augmented Generation (RAG)

- Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources



Importance

- Updated world knowledge
- Providing insights into their predictions
- Hallucinations



Input data

Retriever Model Architecture

Generator Model Architecture

Output Data

For query *x*, Maximum Inner Product Search (MIPS) is used to find the top-K documents z_i.

For final prediction *y*, *z* is treated as a latent variable and marginalized over seq2seq predictions given different documents

RAG implementations showed better performance in all tested tasks .

Models

RAG

2 variations

RAG-Sequence

Retrieved passages are treated as sequences of text

Each retrieved passage is concatenated with the input data or prompt to form a longer sequence

This combined sequence is then fed into the generator model (e.g., BART) for generating the final output

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z\in\text{top-}k(p(\cdot|x))} p_\eta(z|x)p_\theta(y|x,z) = \sum_{z\in\text{top-}k(p(\cdot|x))} p_\eta(z|x)\prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

RAG-Token

Retrieved passages are represented as token-level embeddings

Instead of concatenating the passages with the input data, token-level representations of the passages are directly integrated into the input embeddings

The generator model (BART) then operates on these modified input embeddings, considering the additional information from the retrieved passages during generation.

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z\in\text{top-}k(p(\cdot|x))} p_\eta(z|x)p_\theta(y_i|x,z_i,y_{1:i-1})$$

Experimental Results

- Tasks:
  - Open-domain Question Answering
  - Abstractive Question Answering
  - Jeopardy Question Generation
  - Fact Verification

Experimental Results

Task: Open-domain Question Answering

	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

Table 1: Open-Domain QA Test Scores.

Model	Jeopardy B-1	Jeopardy QB-1	MSMARCO R-L	MSMARCO B-1	FVR3 Label	FVR2 Acc.
SotA	-	-	**49.8***	**49.9***	**76.8**	**92.2***
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	**17.3**	**22.2**	40.1	41.5	72.5	89.5
RAG-Seq.	14.7	21.4	40.8	44.2		

Table 2: Generation and classification Test Scores.

RAG implementations showed better performance showing improved performance on all except one open-domain question answering task.

Experimental Results

Task: Fact Verification

Document 1: his works are considered classics of American literature ... His wartime experiences formed the basis for his novel **”A Farewell to Arms”** (1929) ...

Document 2: ... artists of the 1920s ”Lost Generation” expatriate community. His debut novel, **”The Sun Also Rises”**, was published in 1926.

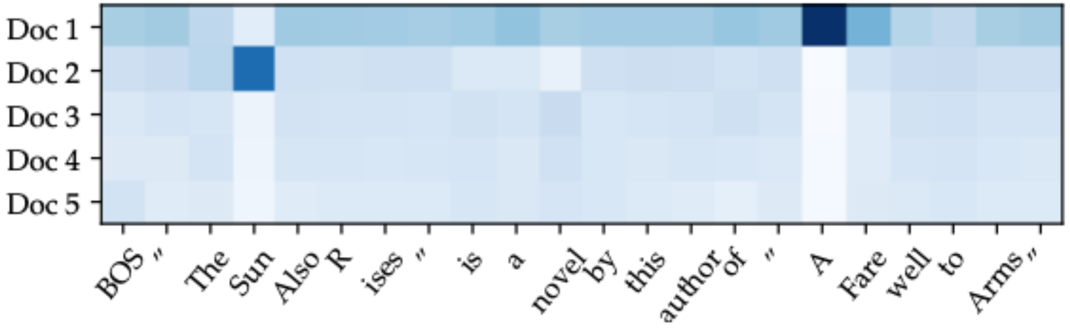


Figure 2: RAG-Token document posterior p(zi|x, yi, yi) for each generated token for input “Hemingway" for Jeopardy generation with 5 retrieved documents.

Experimental Results

Task: Fact Verification

Task	Input	Model	Generation
MS-MARCO	define middle ear	BART	?The middle ear is the part of the ear between the middle ear and the nose.
		RAG-T	The middle ear is the portion of the ear internal to the eardrum.
		RAG-S	The middle ear includes the tympanic cavity and the three ossicles.
	what currency needed in scotland	BART	The currency needed in Scotland is Pound sterling.
		RAG-T	Pound is the currency needed in Scotland.
		RAG-S	The currency needed in Scotland is the pound sterling.
Jeopardy Question Gener -ation	Washington	BART	?This state has the largest number of counties in the U.S.
		RAG-T	It's the only U.S. state named for a U.S. president
		RAG-S	It's the state where you'll find Mount Rainier National Park
	The Divine Comedy	BART	*This epic poem by Dante is divided into 3 parts: the Inferno, the Purgatorio & the Purgatorio
		RAG-T	Dante's "Inferno" is the first part of this epic poem
		RAG-S	This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso"

Table 3: Table 3: Examples from generation tasks. RAG models generate more specific and factually accurate responses. '?' indicates factually incorrect responses, * indicates partially correct responses.

Experimental Results

Task: Jeopardy Question Generation

	Factuality	Specificity
BART better	7.1%	16.8%
RAG better	**42.7%**	**37.4%**
Both good	11.7%	11.8%
Both poor	17.7%	6.9%
No majority	20.8%	20.1%

Table 4: Human assessments for the Jeopardy Question Generation Task

Task: Generation Diversity

	MSMARCO	Jeopardy QGen
Gold	89.6%	90.0%
BART	70.7%	32.4%
RAG-Token	77.8%	46.8%
RAG-Seq.	83.5%	53.8%

Table 5: Ratio of distinct to total tri-grams for generation tasks

Experimental Results

Task: Retrieval Ablations

Model	NQ	TQA	WQ	CT	Jeopardy-QGen		MSMarco		FVR-3	FVR-2
		Exact Match			B-1	QB-1	R-L	B-1	Label Accuracy	
RAG-Token-BM25	29.7	41.5	32.1	33.1	17.5	22.3	55.5	48.4	**75.1**	**91.6**
RAG-Sequence-BM25	31.8	44.1	36.6	33.8	11.1	19.5	56.5	46.9		
RAG-Token-Frozen	37.8	50.1	37.1	51.1	16.7	21.7	55.9	49.4	72.9	89.4
RAG-Sequence-Frozen	41.2	52.1	41.8	52.6	11.8	19.6	56.7	47.3		
RAG-Token	43.5	54.8	**46.5**	51.9	**17.9**	**22.6**	56.2	**49.4**	74.5	90.6
RAG-Sequence	**44.0**	**55.8**	44.9	**53.4**	15.3	21.5	**57.2**	47.5		

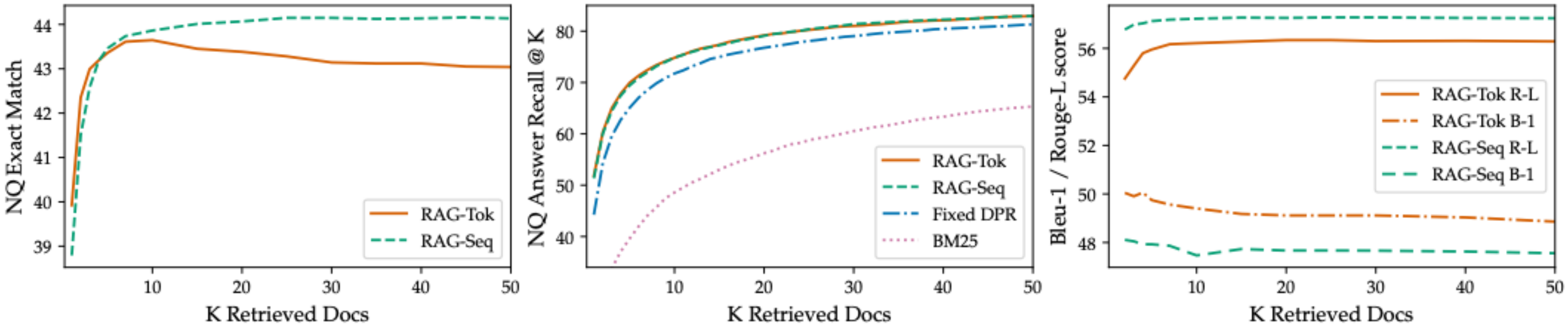
Table 6: Ablations on the dev set. As FEVER is a classification task, both RAG models are equivalent.

Experimental Results

Task: Index hot-swapping

- Built an index using the DrQA Wikipedia dump from December 2016 and compare outputs from RAG using this index to the newer index from our main results (December 2018).
- RAG answers 70% correctly using the 2016 index for 2016 world leaders and 68% using the 2018 index for 2018 world leaders.
- This shows that RAG's world knowledge can be updated by simply replacing its non-parametric memory.

Effect of Retrieving more documents



Retrieving more documents can lead to improved relevance of the retrieved passages. The model shows diminishing returns when it comes to the number of documents retrieved after retrieving a certain number of documents.

Conclusion

- Hybrid generation models with access to parametric and non-parametric memory.
- Obtains state of the art results on open-domain QA
- Improved generation compared to parametric BART, with RAG more factual and specific

Overview

- *k*NN-LM
- Model Setup
- Experimental Results
- Performance comparison

GENERALIZATION THROUGH MEMORIZATION: NEAREST NEIGHBOR LANGUAGE MODELS

Urvashi Khandelwal[†,*] **Omer Levy**[‡], **Dan Jurafsky**[†], **Luke Zettlemoyer**[‡] **& Mike Lewis**[‡]
[†]Stanford University
[‡]Facebook AI Research
{urvashik,jurafsky}@stanford.edu
{omerlevy,lsz,mikelewis}@fb.com

*k*NN-LM

- An approach that extends a pre-trained LM by linearly interpolating its next word distribution with a k-nearest neighbors (kNN) model
- The nearest neighbors are computed according to distance in the pre-trained embedding space and can be drawn from any text collection, including the original LM training data.
- This approach allows rare patterns to be memorized explicitly, rather than implicitly in model parameters.
- It also improves performance when the same training data is used for learning the prefix representations and the kNN model, strongly suggesting that the prediction problem is more challenging than previously appreciated.

Model Architecture

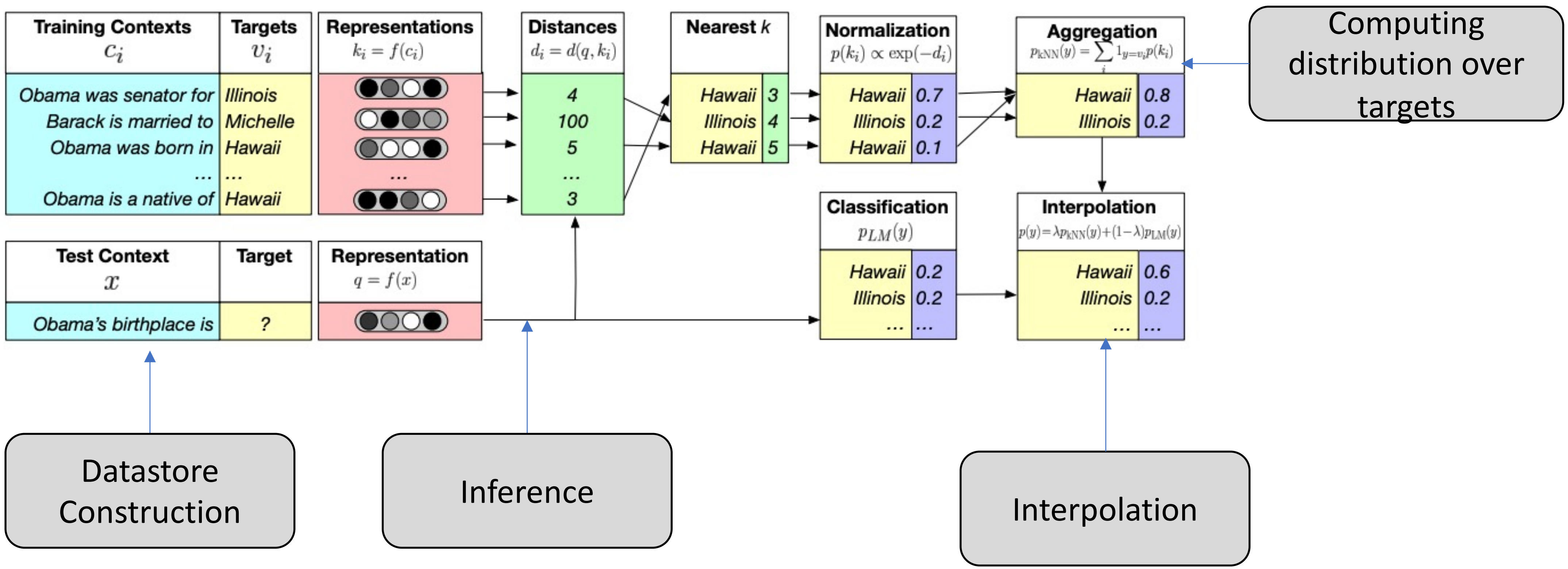


Figure 1. Illustration of kNN-LM

Experimental Results

Model	Perplexity (↓) Dev	Test	# Trainable Params
Baevski & Auli (2019)	17.96	18.65	247M
+Transformer-XL (Dai et al., 2019)	-	18.30	257M
+Phrase Induction (Luo et al., 2019)	-	17.40	257M
Base LM (Baevski & Auli, 2019)	17.96	18.65	247M
+kNN-LM	**16.06**	**16.12**	247M
+Continuous Cache (Grave et al., 2017c)	17.67	18.27	247M
+kNN-LM + Continuous Cache	**15.81**	**15.79**	247M

Table 1. Performance on WIKITEXT-103

Model	Perplexity (↓) Dev	Test	# Trainable Params
Base LM (Baevski & Auli, 2019)	14.75	11.89	247M
+kNN-LM	**14.20**	**10.89**	247M

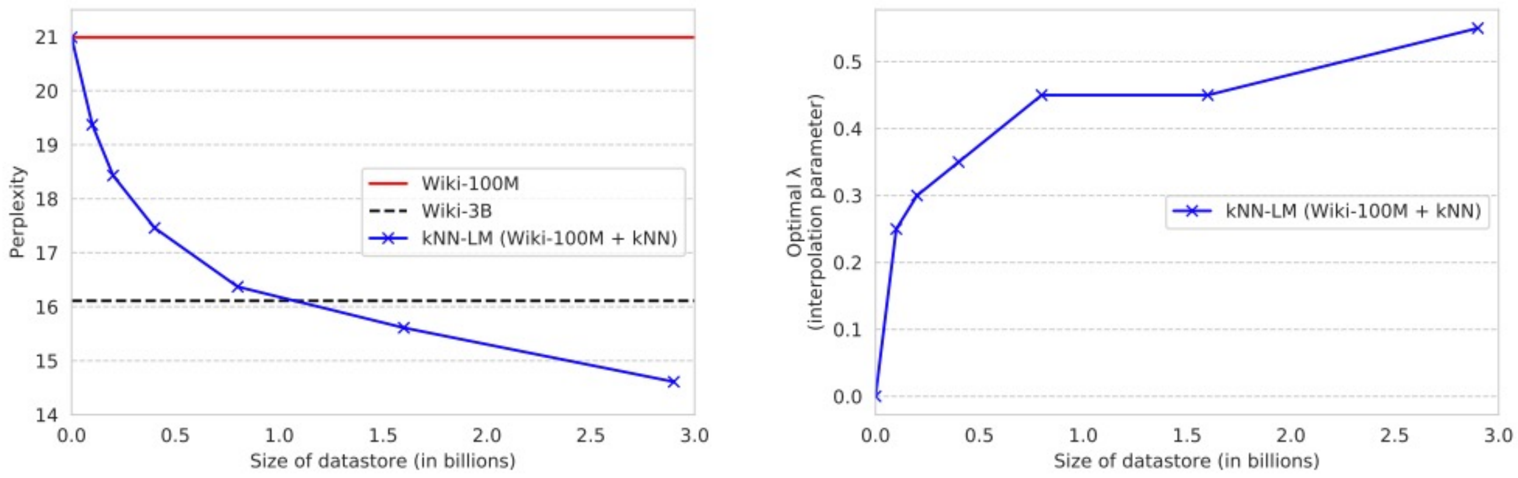
Table 2. Performance on BOOKS

The kNN-LM model shows improvement compare to the baselines with lower perplexity scores

Experimental Results

Task : Training Data as the datastore

Training Data	Datastore	Perplexity (↓) Dev	Test
WIKI-3B	-	16.11	15.17
WIKI-100M	-	20.99	19.59
WIKI-100M	WIKI-3B	14.61	13.73



(a) Effect of datastore size on perplexities. (b) Tuned values of λ for different datastore sizes.

Table 3. Experimental results on WIKI-3B

Figure 2. Varying size on the datastore

As the size of the datastore increases, a higher weight on the retrieved training examples (controlled by λ) becomes more beneficial in improving model performance

Experimental Results

Task : Additional Data without training

Training Data	Datastore	Perplexity (↓) Dev	Test
WIKI-3B	-	37.13	34.84
BOOKS	-	14.75	11.89
WIKI-3B	BOOKS	24.85	20.47

Table 4. Performance on in-domain BOOKS data

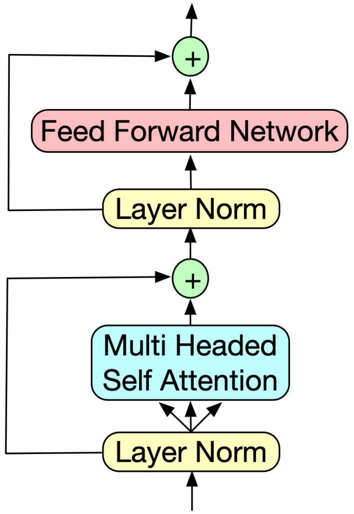


Figure 3. Transformer layer of the LM

Key Type	Dev ppl. (↓)
No datastore	17.96
Model output	17.07
Model output layer normalized	17.01
FFN input after layer norm	**16.06**
FFN input before layer norm	17.06
MHSA input after layer norm	16.76
MHSA input before layer norm	17.14

Table 5. WIKITEXT-103 validation results using different states from the final layer of the LM as the representation function for keys and queries

Experimental Results

Task: Tuning Nearest neighbor search

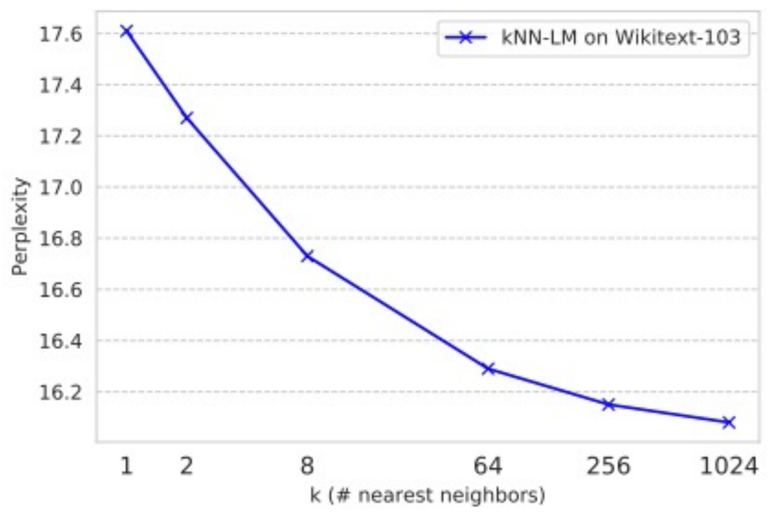


Figure 4. Effect of the number of nearest neighbors returned per word on WIKITEXT-103 (validation set).

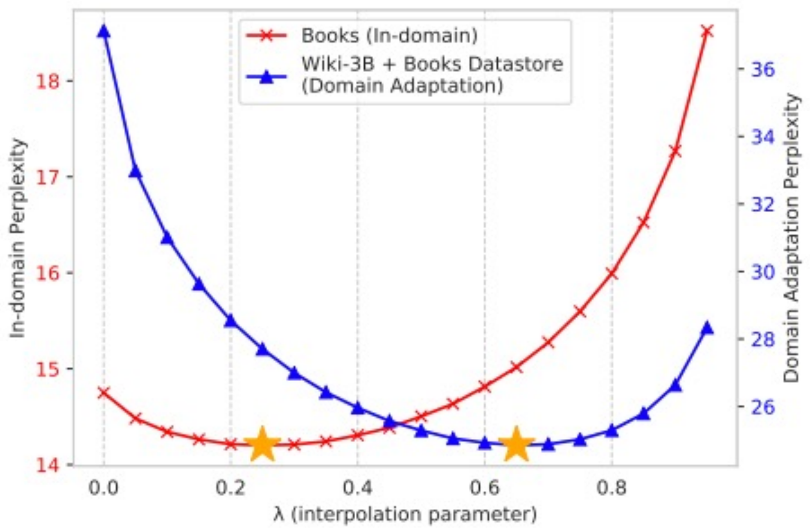


Figure 4. Effect of interpolation parameter λ on in-domain (left y-axis) and out-of-domain (right y-axis) validation set performances.

Experimental Results

Test Context ($p_{\text{kNN}} = 0.998$, $p_{\text{LM}} = 0.124$)	**Test Target**	
it was organised by New Zealand international player Joseph Warbrick, promoted by civil servant Thomas Eyton, and managed by James Scott, a publican. The Natives were the first New Zealand team to perform a haka, and also the first to wear all black. They played 107 rugby matches during the tour, as well as a small number of Victorian Rules football and association football matches in Australia. Having made a significant impact on the...	development	
Training Set Context	**Training Set Target**	**Context Probability**
As the captain and instigator of the 1888-89 Natives – the first New Zealand team to tour the British Isles – Warbrick had a lasting impact on the...	development	0.998
promoted to a new first grade competition which started in 1900. Glebe immediately made a big impact on the...	district	0.00012
centuries, few were as large as other players managed. However, others contend that his impact on the...	game	0.000034
Nearly every game in the main series has either an anime or manga adaptation, or both. The series has had a significant impact on the...	development	0.00000092

Figure 6: Example where the kNN model has much higher confidence in the correct target than the LM. Although there are other training set examples with similar local n-gram matches, the nearest neighbour search is highly confident of specific and very relevant context.

Figure 6. Example where the kNN model has much higher confidence in the correct target than the LM.

Experimental Results

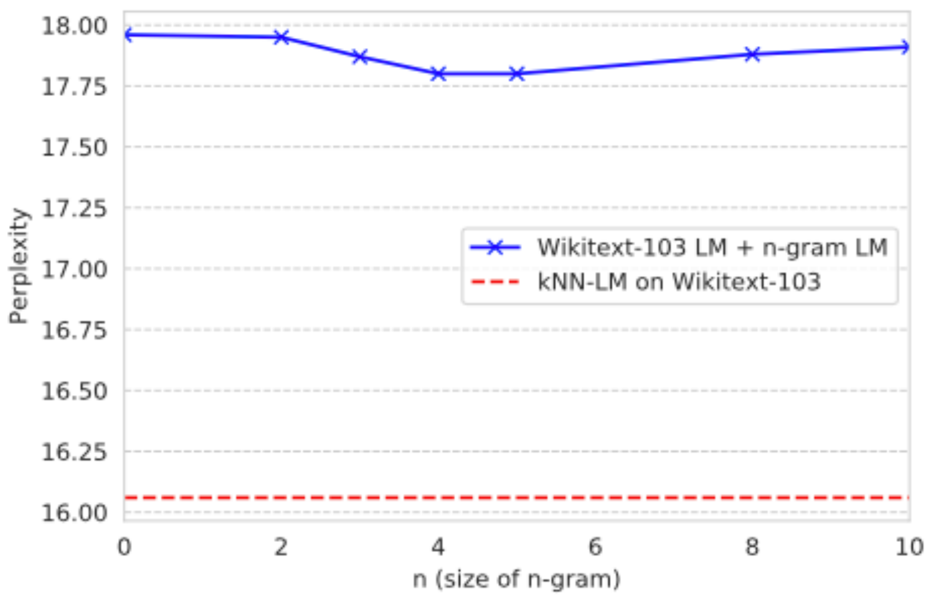


Figure 7. Interpolating the Transformer LM with n-gram LMs on WIKITEXT-103 (validation set) .

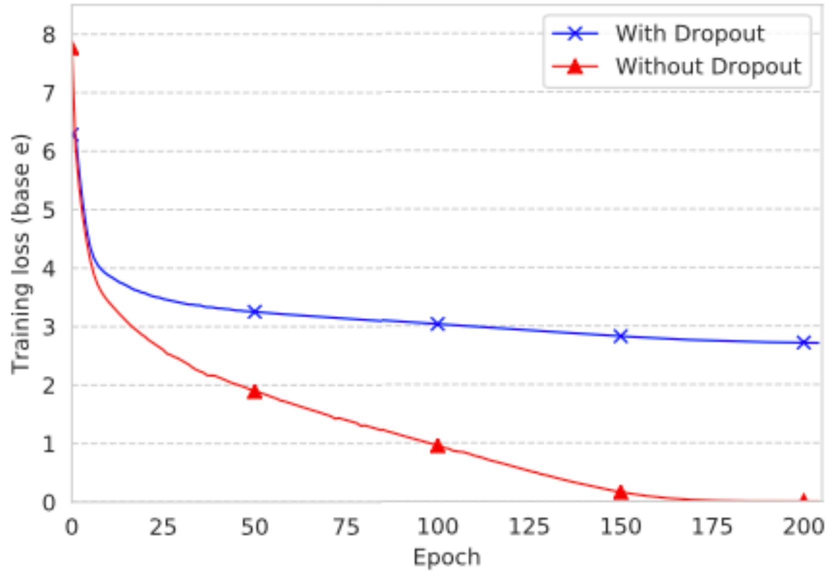


Figure 8. Training curves for the Transformer LM with and without dropout.

Conclusion

- kNN-LM outperform standard language models by directly quering training examples at test time

- Learning similarity functions between contexts may be an easier problem than predicting the next word from some given context.

Questions?