Generalization Through Memorization: Nearest Neighbor Language Models

Hyunjae Kim

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

Authors

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

Accepted at ICLR 2020

[−] **Paper Decision**

ICLR 2020 Conference Program Chairs

20 Dec 2019 (modified: 20 Dec 2019) ICLR 2020 Conference Paper1318 Decision Readers: Everyone

Decision: Accept (Poster)

Comment: This paper proposes an idea of using a pre-trained language model on a potentially smaller set of text, and interpolating it with a k-nearest neighbor model over a large datastore. The authors provide extensive evaluation and insightful results. Two reviewers vote for accepting the paper, and one reviewer is negative. After considering the points made by reviewers, the AC decided that the paper carries value for the community and should be accepted.

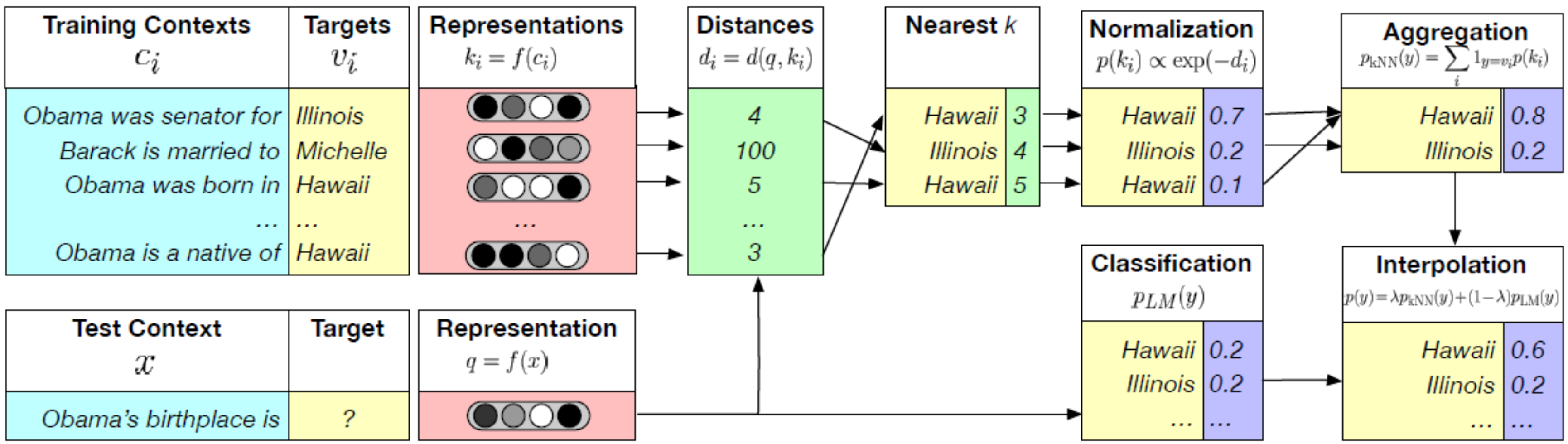
Approach : Datastore

- Language models (LMs) assign probabilities to sequences. Given a *context* sequence of tokens $c_t = (w_1, \ldots w_{t-1})$, autoregressive LMs estimate $p(w_t|c_t)$, the distribution over the *target* token w_t.
- Training example : $(c_i, w_i) \in \mathcal{D}$
- Datastore : $(\mathcal{K}, \mathcal{V}) = \{(f(c_i), w_i) | (c_i, w_i) \in \mathcal{D}\}$
- function *f* : intermediate state of the LM model

Approach : Inference

- kNN distribution : $p_{\mathrm{kNN}}(y|x) \propto \sum_{(k_i,v_i)\in\mathcal{N}} \mathbb{1}_{y=v_i} \exp(-d(k_i, f(x)))$
- Inference by interpolation : $p(y|x) = \lambda\, p_{\mathrm{kNN}}(y|x) + (1-\lambda)\, p_{\mathrm{LM}}(y|x)$
- distance function *d* : squared L2 distance

Illustration of kNN-LM



Motivation

- Neural language models (LMs) typically solve two subproblems
  (1) mapping sentence prefixes to fixed-sized representations
  (2) using these representations to predict the next word in the text

- The representation learning problem may be easier than the prediction problem.
  - *Dickens is the author of*
  - *Dickens wrote*

- The first problem → Improvement

Experimental Setup

- Data
  - WikiText-103, Books, Wiki-3B, Wiki-100M
- Model Architecture
  - Decoder-only Transformers
- Evaluation
  - Perplexity

Experiments - 1

- Using the Traninig Data as the Datastore
  - Performance on WikiText-103

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

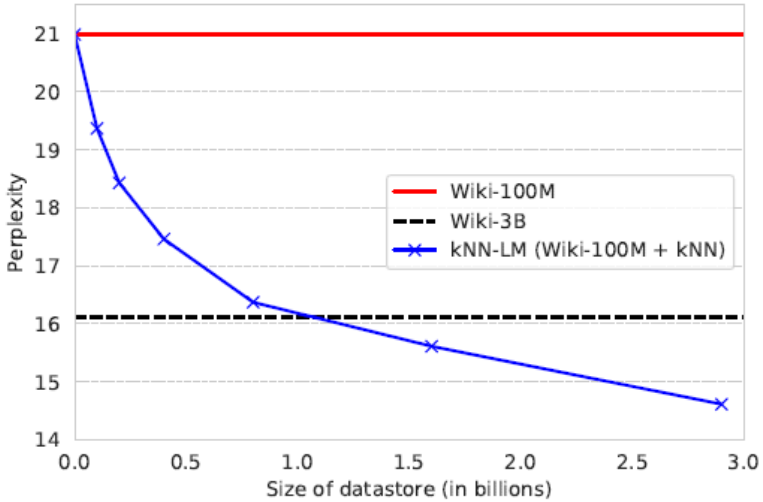
  - Performance on Books

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

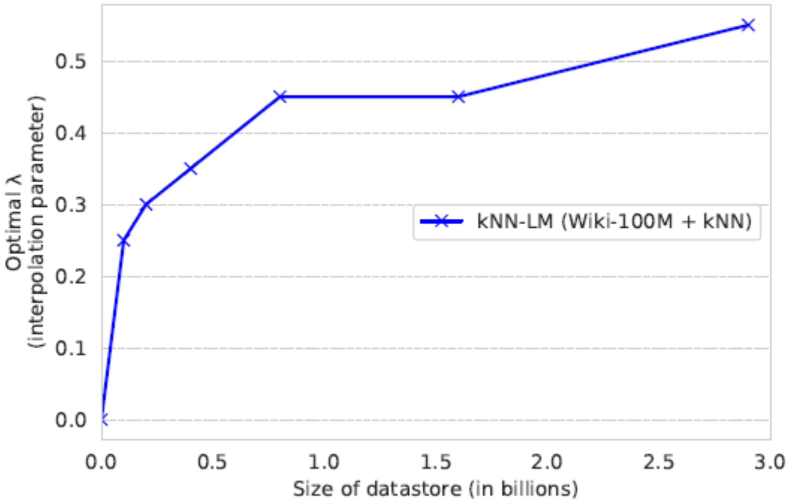
Experiments - 2

- More Data without Training

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.

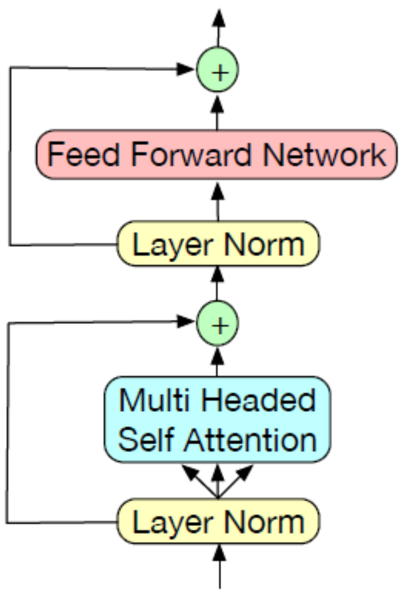
Experiments - 3

- Domain Adaptation

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

Experiments - 4

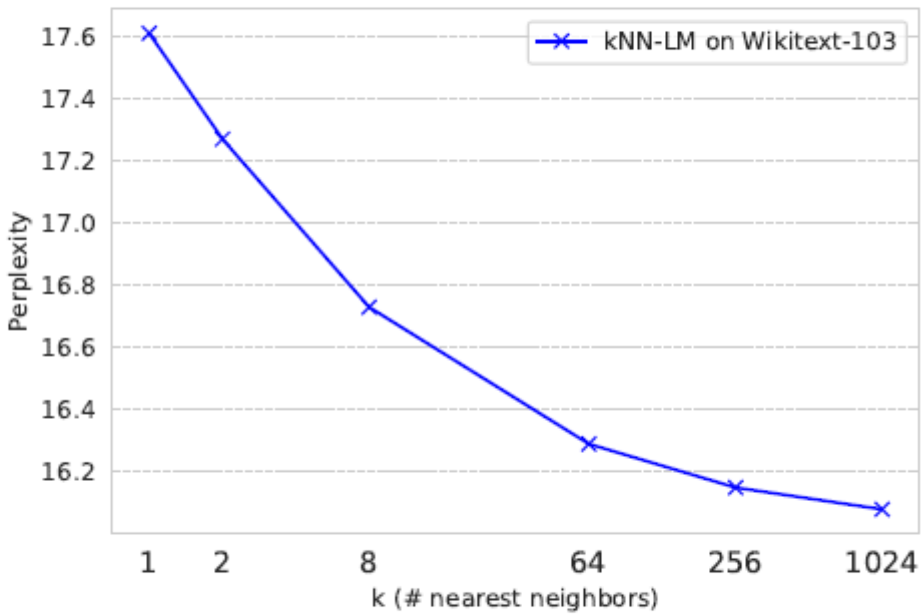
- Choice of Key Function



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

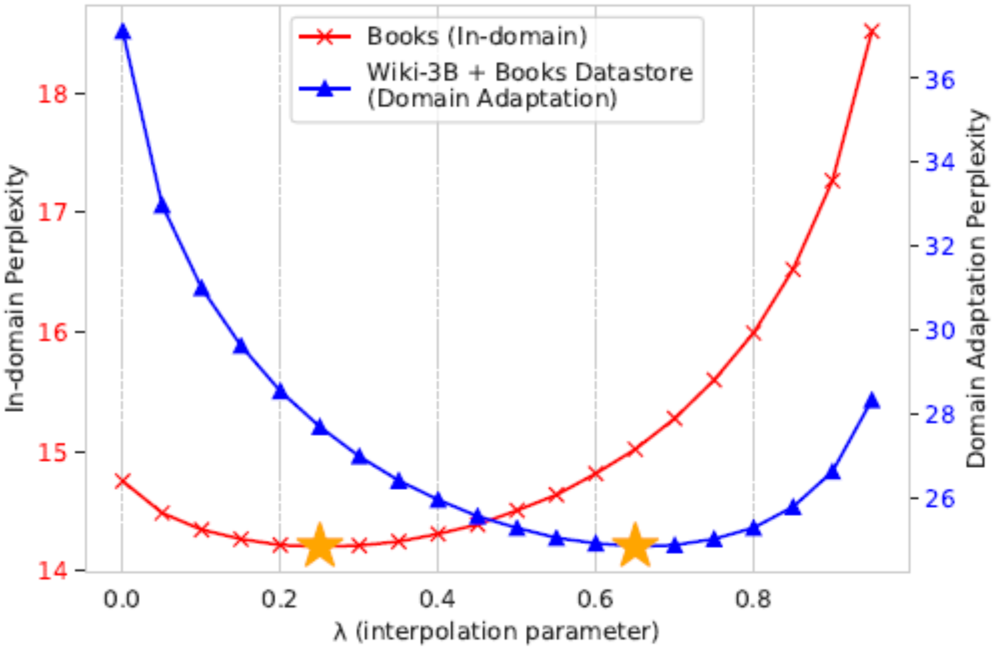
Experiments - 5

- <u>Number of Neighbors per Query</u>



Experiments - 6

- <u>Interpolation Parameter</u>



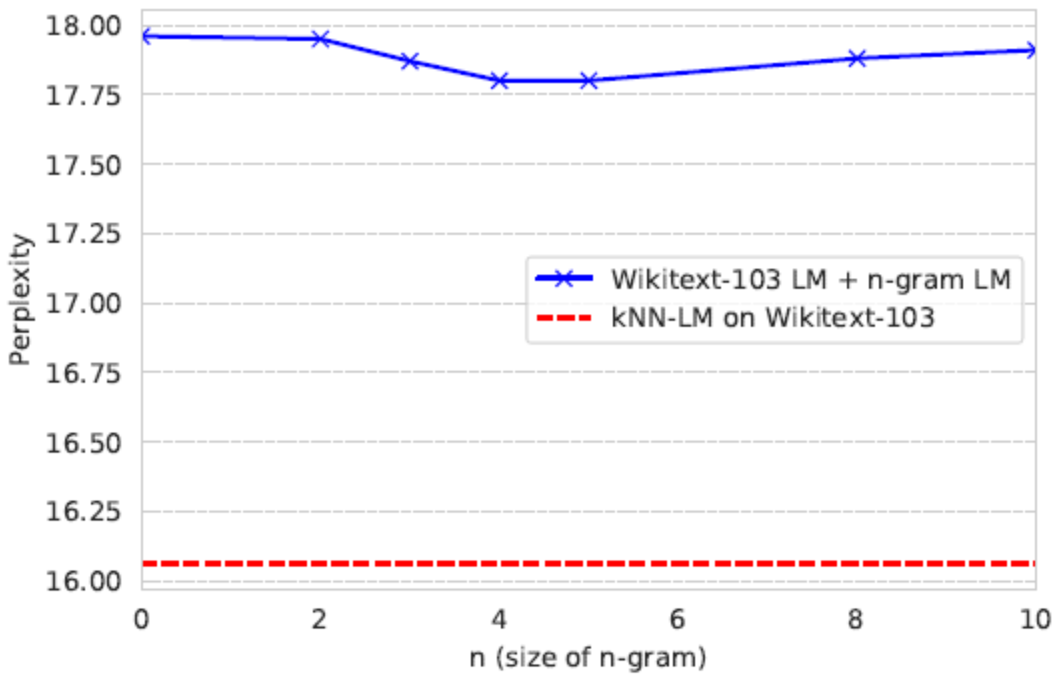
Qualitative Analysis - 1

Test Context ($p_{\mathrm{kNN}} = 0.998$, $p_{\mathrm{LM}} = 0.124$)	**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

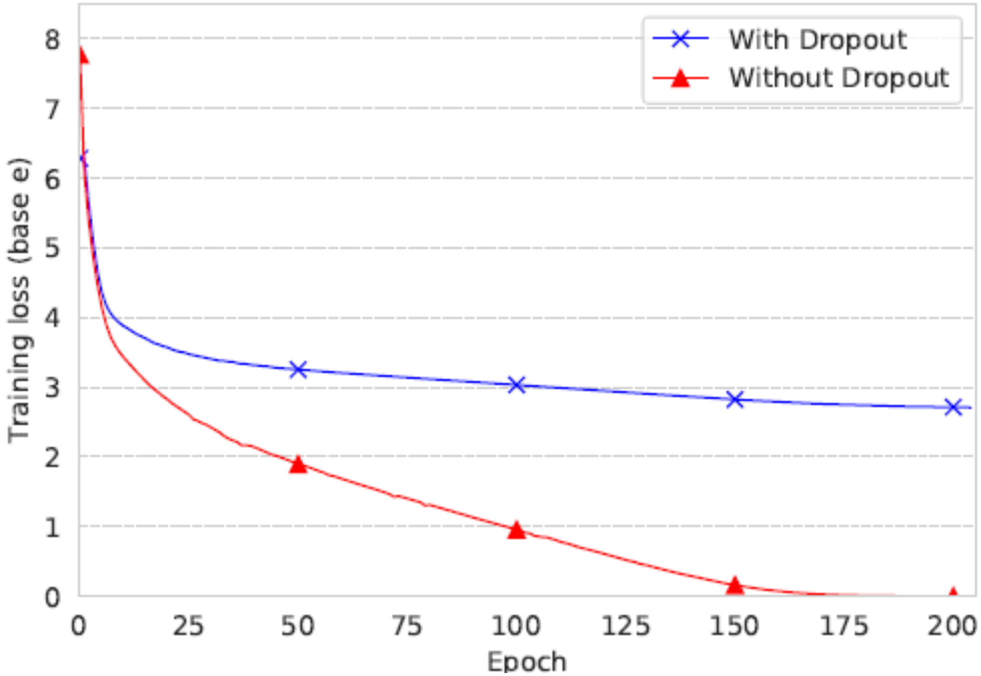
Qualitative Analysis - 2

- Simple vs Neural Representation
  - Interpolating an n-gram model with a Transformer LM



Qualitative Analysis - 3

- Implicit vs Explicit Memory (model parameters vs datastore)
  - Training a Transformer LM with no dropout
  - Interpolating the memorizing LM with the original LM
- Improved by just 0.1 – compared to 1.9 from kNN-LM



Discussion