Improving Natural Language Understanding with Computation-Efficient Retrieval Representation Fusion

香港城市大學 City University of Hong Kong

哈爾濱工業大學(深圳) HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

Shangyu Wu, Ying Xiong, Yufei Cui, Xue Liu, Buzhou Tang, Tei-wei Kuo, Chun Jason Xue

City University of Hong Kong, Harbin Institute of Technology, Shenzhen, MILA, McGill University

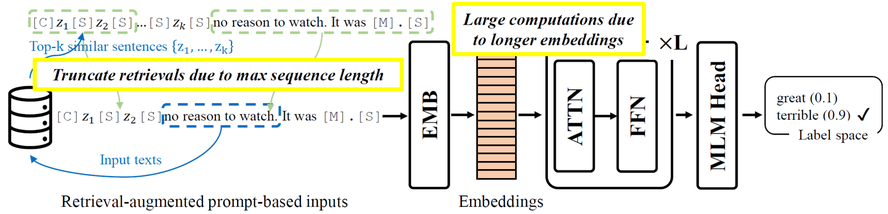
National Taiwan University, Mohamed bin Zayed University of Artificial Intelligence

Mila McGill University 國立臺灣大學 National Taiwan University MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Background

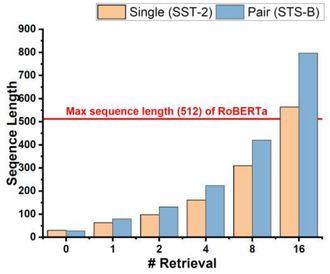
Retrieval-based augmentations in non-knowledge-intensive tasks, such as text classification, are still challenging.

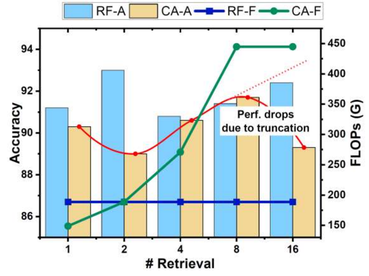
Existing works achieved the SOTA performance on NKI tasks by concatenating retrievals.



Limitations:

1. *Limited retrievals can be added due to the max sequence length of models.*
2. *Concatenating more retrievals results in a longer input sequence, thus leading to large computations during the attention mechanism.*





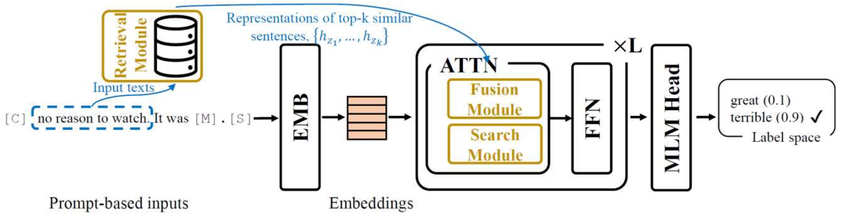
Related Work

1. *Few-shot learning: LM-BFF, DART*
2. *Retrieval-augmentation: RETRO, Atlas, RAG*
3. *Neural architecture search: DARTS, ProxylessNAS*

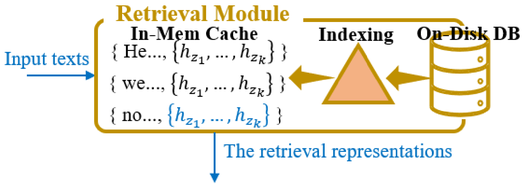
Source code
https://github.com/luffy06/ReFusion

Methodology

Intuition: Directly fusing retrieval representations into transformer-based models in a computation-efficient way.

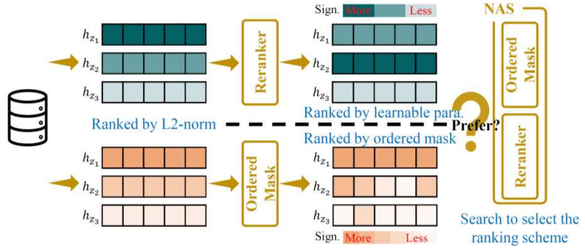


The Online Retrieval Module



- *Query encoder (e.g., BERT).*
- *Task-agnostic retriever*
  - *Indexing (e.g., FAISS).*
  - *Compressed key-value database.*

The Retrieval Fusion Module



- *Reranking the Retrievals*
  - *Learnable Reranker:* $R = \{r_1, \dots, r_k\}$
  - $r_i = \frac{\exp(r_i)}{\sum_j \exp(r_j)}$
  - $h_{y_{[cls]}} = h_{x_{[cls]}} + \frac{1}{k}\sum r_i \cdot h_{z_i}$
- *Ordered Mask over Retrieval Representations*
  - *Ordered Mask on dim-d:* $V^d = \{v_1^d, \dots, v_k^d\}$
  - $c^d \sim \textbf{Gumbel}(\beta, \tau)$
  - $v_i^d = 1 - \textbf{cumsum}(c^d)$
  - $h_{y_{[cls]}} = h_{x_{[cls]}} + \frac{1}{k}\sum v_i^d \cdot h_{z_i}^d$

The Architecture Search Module

- *Search Space*
  - *No Fusion.*
  - *Fusion with Reranker.*
  - *Fusion with Ordered Mask.*
- *Searching Details*
  - *Architectural weights:* $\alpha = \{\alpha_1, \dots, \alpha_l\}$
  - *Forward a search module*

$$\hat{o}(h) = \sum_i \frac{\exp(\alpha_i)}{\sum_j \exp(\alpha_j)} o_i(h)$$

Experimental Results

Experimental Setting

- *15 NKI tasks,*
  - *8 taks from GLUE;*
  - *SNLI, SST-5, MR, CR, MNLI, MNLI-mm, Subj, and TREC.*
- *16-shot learning.*
- *Seeds: 13, 21, 42, 87, 100.*
- *Backbone model: RoBERTa-large.*

Main Results

Methods	SST-2	SST-5	MR	CR	MPQA	SUBJ	TREC	CoLA	Avg-S
LM-BFF	$92.7_{0.9}$	$47.4_{2.5}$	$87.0_{1.2}$	$90.3_{1.0}$	$84.7_{2.2}$	$91.2_{1.1}$	$84.8_{5.1}$	$9.3_{7.3}$	73.4
DART	$\mathbf{93.5}_{0.5}$	-	$\mathbf{88.2}_{1.0}$	$\mathbf{91.8}_{0.5}$	-	$90.7_{1.4}$	$87.1_{3.8}$	-	-
KPT	$90.3_{1.6}$	-	$86.8_{1.8}$	$88.8_{3.7}$	-	-	-	-	-
CA-512	$91.3_{1.4}$	$46.7_{1.1}$	$85.1_{1.4}$	$88.3_{1.7}$	$76.9_{2.8}$	$88.0_{1.9}$	$82.2_{4.4}$	$7.4_{3.3}$	70.7
ReFusion	$93.4_{0.6}$	$\mathbf{49.8}_{1.4}$	$87.9_{1.1}$	$91.7_{0.3}$	$\mathbf{86.7}_{1.1}$	$\mathbf{92.5}_{0.8}$	$\mathbf{90.3}_{3.7}$	$\mathbf{11.4}_{4.1}$	**75.5**
Methods	**MNLI**	**MNLI-m**	**SNLI**	**QNLI**	**RTE**	**MRPC**	**QQP**	**Avg-P**	**Avg-all**
LM-BFF	$68.3_{2.3}$	$70.5_{1.9}$	$77.2_{3.7}$	$64.5_{4.2}$	$69.1_{3.6}$	$74.5_{5.3}$	$65.5_{5.3}$	69.9	71.8
DART	$67.5_{2.6}$	-	$75.8_{1.6}$	$66.7_{3.7}$	-	$\mathbf{78.3}_{4.5}$	$67.8_{3.2}$	-	-
KPT	$61.4_{2.1}$	-	-	$61.5_{2.8}$	-	-	$\mathbf{71.6}_{2.7}$	-	-
CA-512	$66.2_{1.0}$	$67.8_{1.3}$	$71.6_{2.2}$	$66.9_{3.2}$	$66.6_{3.1}$	$73.5_{6.9}$	$64.0_{1.9}$	68.1	69.5
ReFusion	$\mathbf{69.3}_{1.5}$	$\mathbf{70.9}_{1.5}$	$\mathbf{80.6}_{1.4}$	$\mathbf{73.0}_{1.1}$	$\mathbf{70.9}_{2.3}$	$77.0_{3.6}$	$68.9_{3.3}$	**72.9**	**74.3**

The results of LM-BFF, DART refer to their original paper. The results of KPT refer to Chen et al. (2022). The numbers are the average results. The subscript numbers are the standard deviation results.

Ablation Study

Methods	MPQA	SUBJ	TREC	SNLI	QNLI	RTE
Roberta-Large	$83.6_{2.5}$	$90.3_{2.8}$	$83.8_{5.2}$	$73.5_{5.2}$	$65.0_{3.0}$	$64.1_{2.0}$
Reranker	$84.2_{2.2}$	$91.3_{1.3}$	$85.0_{4.2}$	$74.3_{4.6}$	$68.8_{1.4}$	$65.6_{3.1}$
Ordered Mask	$83.3_{1.9}$	$90.8_{1.4}$	$83.0_{5.8}$	$74.9_{4.0}$	$68.3_{1.4}$	$65.8_{3.1}$
NAS with Reranker	$86.9_{1.3}$	$92.4_{1.3}$	$\mathbf{90.8}_{2.5}$	$80.3_{1.9}$	$\mathbf{73.5}_{1.8}$	$69.2_{2.4}$
NAS with Ordered Mask	$\mathbf{87.0}_{1.5}$	$92.4_{0.7}$	$90.7_{3.0}$	$80.3_{1.3}$	$73.0_{1.0}$	$70.4_{2.5}$
ReFusion	$86.7_{1.1}$	$\mathbf{92.5}_{0.8}$	$90.3_{3.7}$	$\mathbf{80.6}_{1.4}$	$73.0_{1.1}$	$\mathbf{70.9}_{2.3}$

Full Training Set

Methods	SST-2	SST-5	MR	CR	MPQA	SUBJ	TREC	CoLA	RTE
LM-BFF	95.0	58.7	90.8	89.4	87.8	97.0	97.4	62.6	80.9
ReFusion	**95.6**	**61.0**	**92.3**	**91.4**	**84.4**	**97.1**	**97.6**	**62.8**	**85.2**