Compressing Sentence Representation for Semantic Retrieval via Homomorphic Projective Distillation

Xuandong Zhao
UC Santa Barbara
xuandongzhao@cs.ucsb.edu

Zhiguo Yu
Microsoft
zhiguo.yu@microsoft.com

Ming Wu
Microsoft
mingwu@microsoft.com

Lei Li
UC Santa Barbara
leili@cs.ucsb.edu

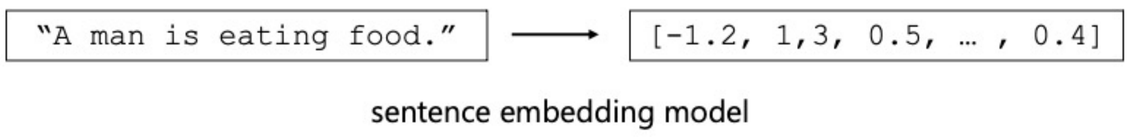
Sentence Embedding

Good sentence representations have wide applications in NLP:

- web search, question answering, knowledge inference, machine translation…

Sentence embedding model:

- Input: sentence
- Output: fixed-length continuous vector



Existing Models are Big!

State-of-the-Art Sentence Embedding Models:

- Sentence-BERT (SBERT) (Reimers and Gurevych, 2019)
- SimCSE (Gao et al., 2021)
- …

Large Model Size

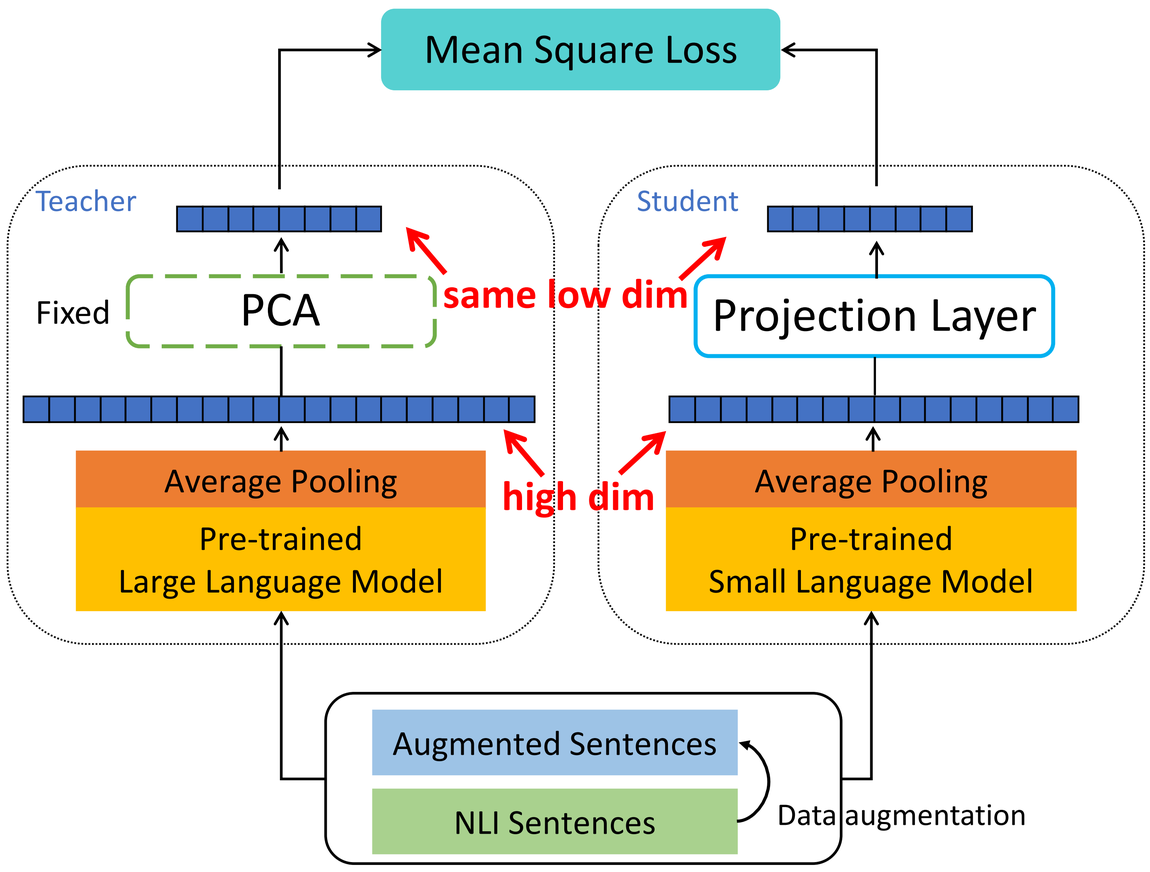
Sentence-BERT$_{base}$: **109M**
SimCSE-RoBERTa$_{large}$: **355M**

Large Representation Size

Sentence-BERT$_{base}$: **768-dim**
SimCSE-RoBERTa$_{large}$: **1024-dim**

Require large memory and suffer from high latency

Homomorphic Projective Distillation



Experiments

Tasks

- Semantic Textual Similarity (STS) Task
- Semantic Retrieval (SR) Task

Data

- SNLI and MNLI dataset
- Data augmentation: WordNet substitution and back translation

Model

- Teacher: SimCSE-MPNet & SimCSE-RoBERTa$_{large}$
- Student: MiniLM & TinyBERT

Results

STS Task (HPD-MiniLM vs SimCSE-RoBERTa$_{large}$)

- **97.7%** Spearman's correlation performance
- **7 times** higher speed
- **6.5%** of parameter

Model	STS Avg.	Size	Dim	Speed
Large models				
SBERT$_{base}$	74.89	109M	768	993
SRoBERTa$_{large}$	76.68	355M	1024	385
SimCSE-MPNet ♠	82.75	109M	768	986
SimCSE-RoBERTa$_{large}$ ♣	83.76	355M	1024	291
Backbone for compact model: TinyBERT				
SimCSE-TinyBERT	78.78	14M	312	2650
+Projection-128	78.12	14M	128	2604
+Whitening-128	78.74	14M	128	2612
HPD-128 (Teacher: ♠)	80.99	14M	128	2608
HPD-128 (Teacher: ♣)	**81.02**	14M	128	2609
Backbone for compact model: MiniLM				
SimCSE-MiniLM	77.16	23M	384	2031
+Projection-128	77.25	23M	128	2022
+Whitening-128	77.28	23M	128	2015
HPD-128 (Teacher: ♠)	81.20	23M	128	2025
HPD-128 (Teacher: ♣)	**81.80**	23M	128	2024

SR Task (HPD-TinyBERT vs SimCSE-RoBERTa$_{large}$)

- Competitive MRR performance
- Reduce the retrieval time by **8.2x**
- Reduce the memory usage by **8.0x**

Model	MRR	Time	Mem
HPD-TinyBERT-128	0.613	63.1	42.61
HPD-TinyBERT-256	0.616	130.4	85.22
HPD-TinyBERT-312	0.615	165.4	103.86
HPD-MiniLM-128	0.610	68.6	42.61
HPD-MiniLM-256	0.615	132.1	85.22
HPD-MiniLM-384	0.612	194.4	127.83
SimCSE-MPNet-768	0.671	385.8	255.66
SimCSE-RoBERTa$_{large}$-1024	0.670	518.0	340.88

References

Nils Reimers and Iryna Gurevych. Sentencebert: Sentence embeddings using siamese bertnetworks. EMNLP 2019

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. EMNLP 2021

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. ArXiv, abs/2103.15316

https://github.com/XuandongZhao/HPD