HashFormers: Towards Vocabulary-independent Pre-trained Transformers

Huiyin Xue, Nikolaos Aletras[1]

1 Department of Computer Science, University of Sheffield, United Kingdom

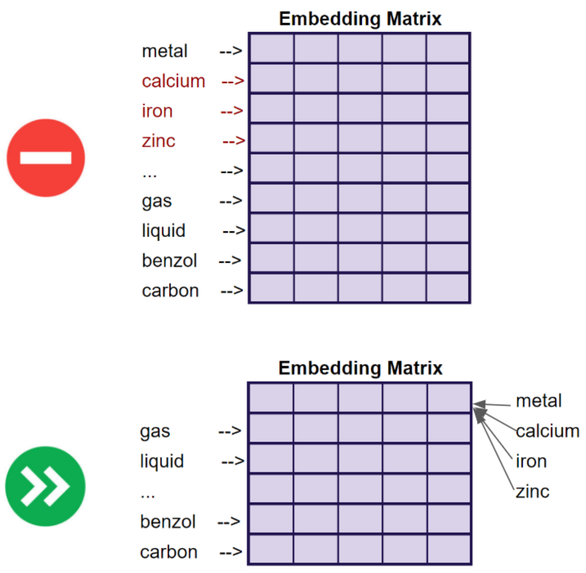
University of Sheffield

INTUITION

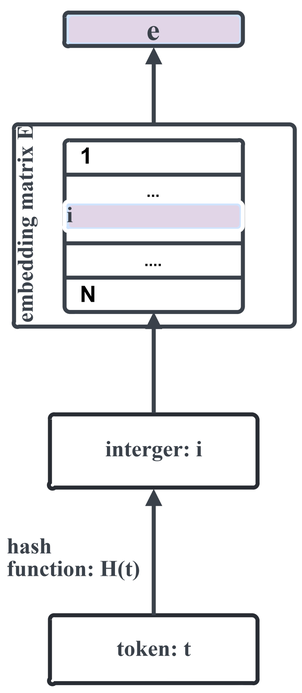
- Transformer-based pre-trained language models are vocabulary-dependent, mapping by default each token to its corresponding embedding.
- However... This one-to-one mapping results into embedding matrices that occupy a lot of memory (i.e. millions of parameters) and grow linearly with the size of the vocabulary.
- **Goal:** Make pre-trained transformers to support an unlimited vocabulary given a substantially smaller fixed-sized embedding matrix by reducing the memory footprint.
- **Approach:** HashFormers use a many-to-one mapping between tokens and embeddings using hashing functions to group together multiple tokens.
- **Contributions:** Our most efficient HashFormer variant achieves a negligible performance degradation (0.4% on GLUE) using only 99.1K parameters for representing the embeddings compared to 12.3-38M parameters of state-of-the-art models.

METHODS

MANY-TO-ONE MAPPING!



HASHFORMERS-EMB



HASHING APPROACHES

Message-Digest Hashing (HashFormers-MD)

- Use a Message-Digest (MD5) hash function to map each token to its 128-bits output, $v = \mathcal{H}(t)$.
- Does not require any feature extraction step.
- Return mostly different hashes for tokens with morphological or semantic similarity.

MD5('play') = 077f244def8a70e5ea758bd8352fcd8178
MD5('plays') = 4a258d930b7d3409982d727ddbb4ba88

Locality-Sensitive Hashing (HashFormers-LSH)

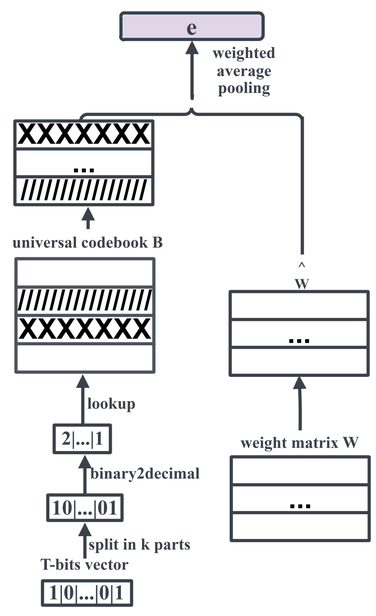
- Use random projections to hash similar tokens into the same indices with high probability.
- Extract n-grams features for Locality-Sensitive Hashing.
- Assign tokens with similar morphology to the same hash encoding.

LSH('play') = LSH('plays')

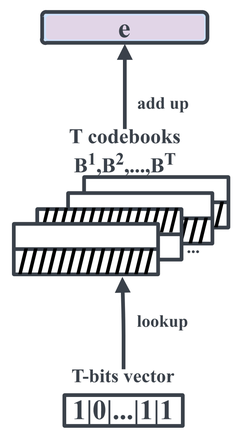
Further Compression of the Embedding Space

Manipulated Word Detection
S+R
as **Pre-training Objective**

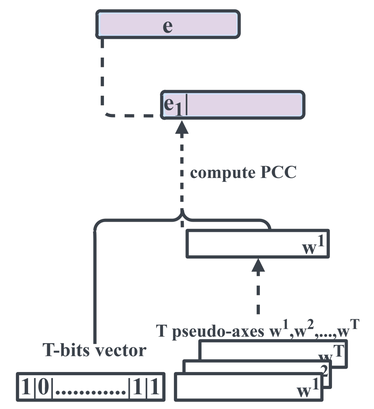
HASHFORMERS-POOL



HASHFORMERS-ADD



HASHFORMERS-PROJ



RESULTS

Model	Token	#Params(Emb)	#Params(Total)	Infer Speed-up Rate	GLUE (dev) Avg.
BERT-MLM	subword	38.6M	124.6M	1.0	79.5
BERT-S+R	subword	38.6M	124.6M	1.0	79.6(+0.1)
CANINE-C	unicode	12.3M	121.0M	0.6x	70.4(-9.1)
ProFormer	word	322.6K	15.1M	-	51.8(-27.7)
HashFormers-MD (Ours)					
Emb (50K)	word	38.6M	124.6M	1.0x~2.4x	79.9(+0.4)
Emb (1K)	word	797.2K	86.8M	-	66.8(-12.7)
Pool	word	797.2K	86.8M	1.0x~2.3x	75.3(-4.2)
Add	word	197.4K	86.2M	1.0x~2.4x	75.7(-3.8)
Proj	word	99.1K	86.1M	1.0x~2.6x	75.7(-3.8)
HashFormers-LSH (Ours)					
Emb (50K)	word	38.6M	124.6M	1.0x~2.4x	76.0(-3.5)
Emb (1K)	word	797.2K	86.8M	-	64.2(-15.3)
Pool	word	797.2K	86.8M	1.0x~2.3x	78.9(-0.6)
Add	word	197.4K	86.2M	1.0x~2.4x	78.9(-0.6)
Proj	word	99.1K	86.1M	1.0x~2.6x	79.1(-0.4)

e-mail: {hxue12, n.aletras}@sheffield.ac.uk

EMNLP 2022

Presented at the the 2022 Conference on Empirical Methods in Natural Language Processing in Abu Dhabi on 7-11 Dec 2022