Pit One Against Many: Leveraging Attention-head Embeddings for Parameter-efficient Multi-head Attention

Huiyin Xue, Nikolaos Aletras[1]

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

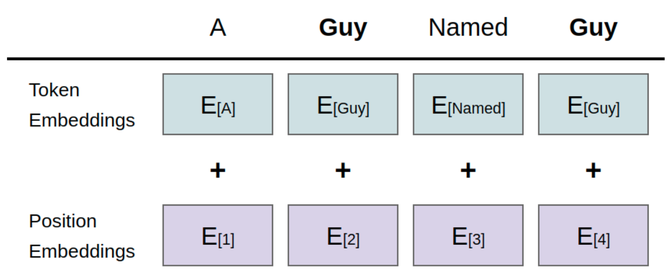
University of Sheffield

INTUITION

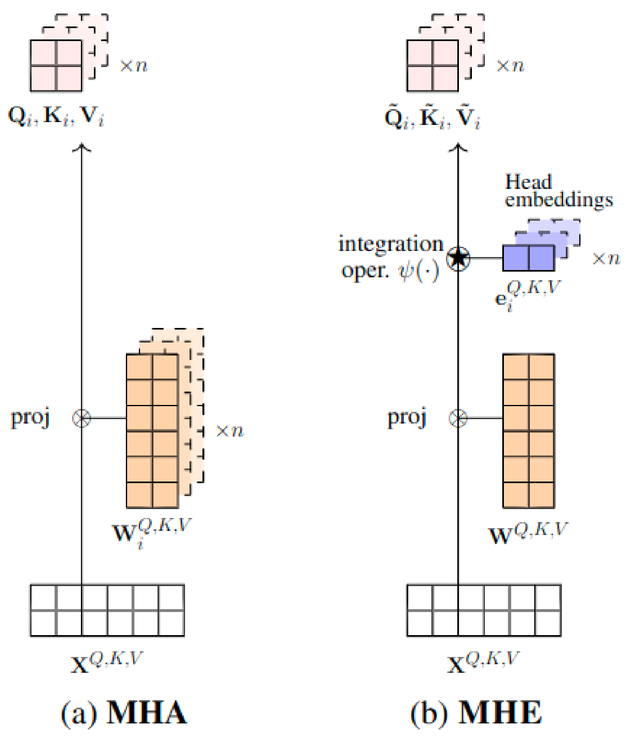
- Multi-head attention (MHA) module leads to a large memory footprint that increases with the number of layers and attention heads per layer in pre-trained language models.
- Scaling MHA can hinder researchers with limited access to computing resources to participate in advancing the field.
  This results in inequalities, where only a privileged few can actively contribute, potentially impeding diversity and inclusivity.
- **Goal:** Make pre-trained language models scale parameter-efficient with the number of layers and attention heads.
- **Approach:** Multiple Head Embeddings Attention (MHE) uses shared projection matrices across heads that are modified by corresponding embedding heads.
- **Contributions:** MHE achieves high predictive performance retention ratio (i.e. 92.9~98.7%) to MHA on several downstream tasks while being $(3n^2 - 3n)d^2 - 3nd$ smaller for a single attention sublayer with n attention heads of d hidden dimension.

METHODS

INSPIRATION



MHA –> MHE



APPROACHES

Multi-head Attention (MHA)

$$\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i = \mathbf{X}\mathbf{W}_i^{Q,K,V}$$
$$\mathbf{H}_i = \text{Att}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$$
$$= \text{SoftMax}(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_h}})\mathbf{V}_i$$

Multiple Head Embeddings Attention (MHE)

- Seed Projection Matrix

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{X}\mathbf{W}^{Q,K,V}$$

- Attention Head Embeddings

$$\tilde{\mathbf{Q}}_i, \tilde{\mathbf{K}}_i, \tilde{\mathbf{V}}_i = \psi(\mathbf{Q}; \mathbf{K}; \mathbf{V}, \mathbf{e}_i^{Q,K,V})$$
$$\mathbf{H}_i = \text{Att}(\tilde{\mathbf{Q}}_i, \tilde{\mathbf{K}}_i, \tilde{\mathbf{V}}_i)$$

- Queries, Keys and Values with Head Embeddings

For

$$\mathbf{A} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$$
$$\mathbf{b} \in \{\mathbf{e}^Q, \mathbf{e}^K, \mathbf{e}^V\}$$

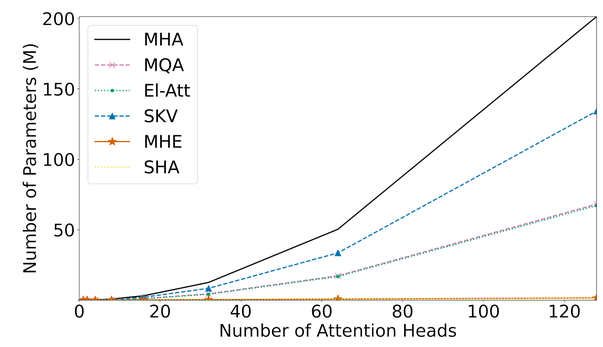
MHE-Add :
$$\psi(\mathbf{A}, \mathbf{b}) := \mathbf{A} + \mathbf{b}$$

MHE-Mul :
$$\psi(\mathbf{A}, \mathbf{b}) := \mathbf{A} \odot (\mathbf{b} + 1)$$

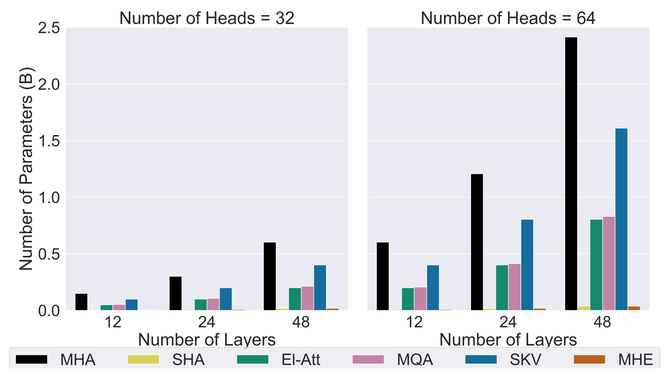
THEORETICAL MEMORY COMPLEXITY

Attention	Complexity	#Params	#Params (+)
SHA	$O(n)$	$3d^2n$	0
MHA	$O(n^2)$	$3d^2n^2$	$(3n^2 - 3n)d^2$
EL-ATT	$O(n^2)$	d^2n^2	$(n^2 - 3n)d^2$
MQA	$O(n^2)$	$d^2n^2 + 2d^2n$	$(n^2 - n)d^2$
SKV	$O(n^2)$	$2d^2n^2$	$(2n^2 - 3n)d^2$
MHE (ours)			
-ADD	$O(n)$	$3d^2n + 3dn$	$3nd$
-MUL	$O(n)$	$3d^2n + 3dn$	$3nd$

#PARAMS WHILE SCALING #HEADS



SCALING #HEADS AND #LAYERS



RESULTS

PERFORMANCE OF ENCODER-ONLY MODELS

Attention	#params	GLUE Acc	GLUE PRR	GLUE PEoP	SUPERGLUE Acc	SUPERGLUE PRR	SUPERGLUE PEoP	SQUAD v1.1 Acc	SQUAD v1.1 PRR	SQUAD v1.1 PEoP	SQUAD v2.0 Acc	SQUAD v2.0 PRR	SQUAD v2.0 PEoP
SHA	8.85M	79.2	96.7	-	67.1	95.1	-	82.5	93.1	-	67.6	87.8	-
MHA	28.32M	81.9	100.0	0.02	70.5	100.0	0.02	88.6	100.0	0.03	77.0	100.0	0.06
EL-ATT	14.16M	80.3	98.0	0.02	69.5	98.5	0.06	86.5	97.6	0.08	72.2	93.8	0.11
MQA	15.34M	81.3	99.2	0.04	69.3	98.2	0.04	86.7	97.9	0.07	74.8	97.1	0.15
SKV	21.23M	**81.4**	**99.4**	0.02	**69.9**	**99.1**	0.03	**88.1**	**99.4**	0.05	**75.9**	**98.6**	0.09
MHE-ADD	8.88M	80.4	98.2	4.92	69.1	97.9	9.44	83.7	94.5	4.65	71.8	93.2	19.88
MHE-MUL	8.88M	80.6	98.3	**5.53**	69.6	98.7	**12.07**	85.9	97.0	**13.19**	72.3	93.9	**22.25**

ROBUSTNESS TO SCALING (RESULTS OF ENCODER-DECODER MODELS ON WMT-14 ENGLISH TO GERMAN)

	N	d_m	h	d_h	p_{drop}	#Steps	#Params(M) MHA	#Params(M) MHE-MUL	BLEU MHA	BLEU MHE-MUL	PRR
BASE	12	512	8	64	0.1	100K	18.87	6.52	24.8	23.6	95.0
	12	512	16	32	0.1	100K	18.87	5.63	25.1	22.9	91.5
	12	512	4	128	0.1	100K	18.87	8.29	24.7	23.6	95.3
4L	8	512	8	64	0.1	100K	12.58	4.34	23.9	22.4	93.6
8L	16	512	8	64	0.1	100K	25.17	8.69	25.3	24.3	96.0
12H	12	768	12	64	0.15	100K	42.47	13.31	25.7	24.2	94.2
BIG	12	1024	16	64	0.3	300K	75.50	22.47	26.5	24.8	93.6

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

Presented at the 2023 Conference on Empirical Methods in Natural Language Processing in Singapore on 6-10 Dec 2023

EMNLP 2023