DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

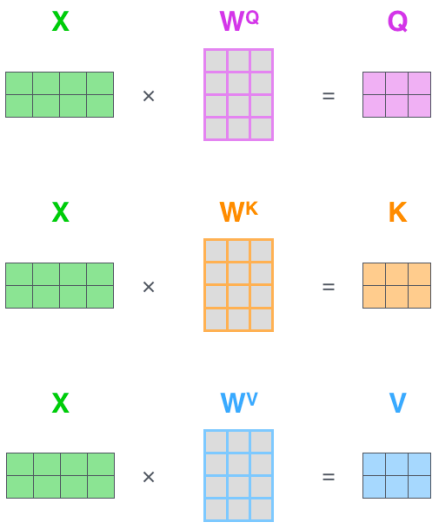
QUARTZ
Quantum Information Access and Retrieval Theory

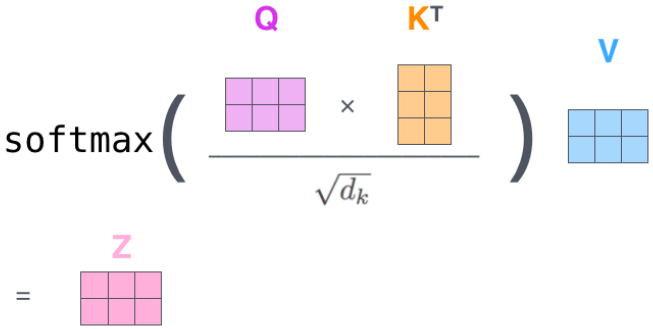
On Position Embeddings in BERT

Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, Jakob Grue Simonsen

University of Padua, Huawei Noah's Ark Lab, University of Copenhagen

Transformer





Z = FFN(MHA(FFN(MHA(x))))

Encoding word features

how to encode features, e.g. Word, Position, Segment ?

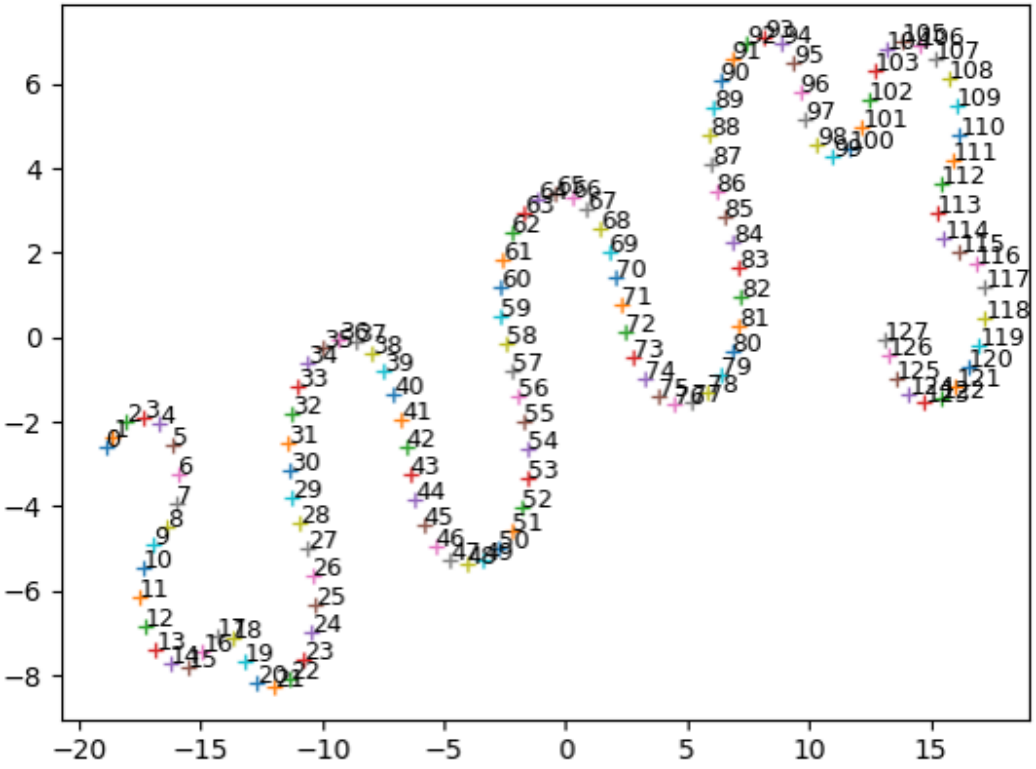
X = WE + PE + SE + ?

how to encode sequential feature?

$$PE'_{2k}(\cdot, pos) = \sin(pos/10000^{2k/d_{model}})$$
$$PE'_{2k+1}(\cdot, pos) = \cos(pos/10000^{2k/d_{model}})$$

how to better encode sequential feature? [Vaswani et.al and Wang et.al]

Fully-learnable PE (after T-SNE)



It seems that there are some clear patterns !!!

Some assumed properties

(to be examined)

Monotonicity: neighboring positions are embedded closer than faraway ones;

e.g, 1 is closer to 2 than 3, 4…

Translation invariance: distances of two arbitrary m-offset position vectors are identical;

distance(1,2) = distance (2,3)

Symmetry: the metric (distance) itself is symmetric. Especially no further info could be provided.

distance(1,2) = distance (2,1)

To understand sinusoidal APE

$$A_{x,y} = \langle \vec{x}, \vec{y} \rangle = \text{sum}\left(\begin{bmatrix} \sin(\omega_1 x) \\ \cos(\omega_1 x) \\ \cdots \\ \sin(\omega_{\frac{D}{2}} x) \\ \cos(\omega_{\frac{D}{2}} x) \end{bmatrix} \odot \begin{bmatrix} \sin(\omega_1 y) \\ \cos(\omega_1 y) \\ \cdots \\ \sin(\omega_{\frac{D}{2}} y) \\ \cos(\omega_{\frac{D}{2}} y) \end{bmatrix} \right) = \text{sum}\left(\begin{bmatrix} \sin(\omega_1 x)\sin(\omega_1 y) \\ \cos(\omega_1 x)\cos(\omega_1 y) \\ \cdots \\ \sin(\omega_{\frac{D}{2}} x)\sin(\omega_{\frac{D}{2}} y) \\ \cos(\omega_{\frac{D}{2}} x)\cos(\omega_{\frac{D}{2}} y) \end{bmatrix} \right) = \sum_{i=0}^{\frac{D}{2}} \cos(\omega_i (x-y))$$

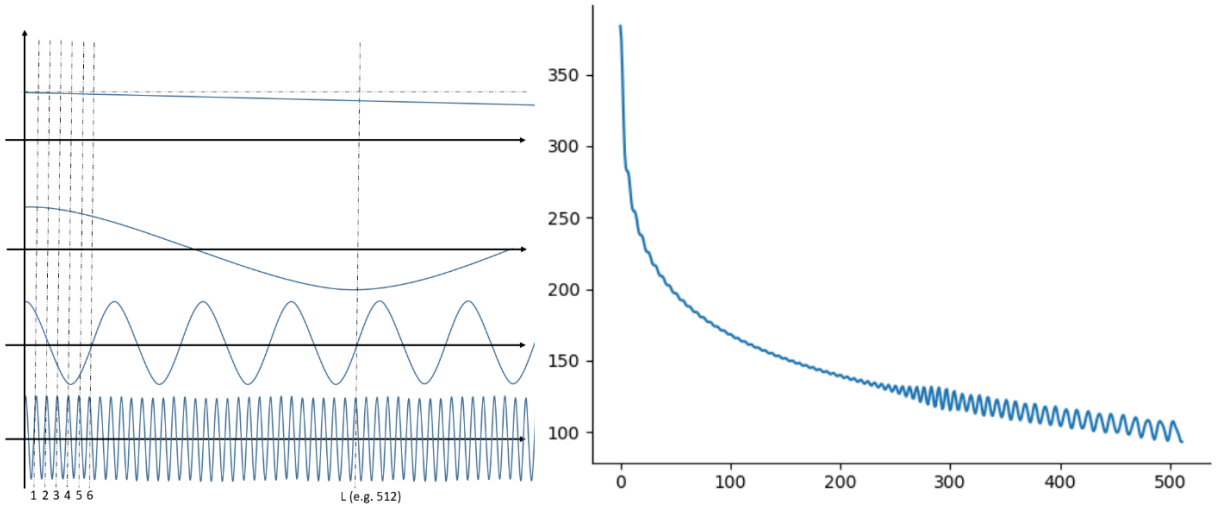
It satisfies **translation invariance and symmetry**

For **monotonicity**, we need to check its first order derivative

$$\sum_{i=1}^{D/2} -\omega_i \sin(\omega_i m)$$

To understand frequencies

$$A_{x,y} = \langle \vec{x}, \vec{y} \rangle = \text{sum}\left(\begin{bmatrix} \sin(\omega_1 x) \\ \cos(\omega_1 x) \\ \cdots \\ \sin(\omega_{\frac{D}{2}} x) \\ \cos(\omega_{\frac{D}{2}} x) \end{bmatrix} \odot \begin{bmatrix} \sin(\omega_1 y) \\ \cos(\omega_1 y) \\ \cdots \\ \sin(\omega_{\frac{D}{2}} y) \\ \cos(\omega_{\frac{D}{2}} y) \end{bmatrix} \right) = \text{sum}\left(\begin{bmatrix} \sin(\omega_1 x)\sin(\omega_1 y) \\ \cos(\omega_1 x)\cos(\omega_1 y) \\ \cdots \\ \sin(\omega_{\frac{D}{2}} x)\sin(\omega_{\frac{D}{2}} y) \\ \cos(\omega_{\frac{D}{2}} x)\cos(\omega_{\frac{D}{2}} y) \end{bmatrix} \right) = \sum_{i=0}^{\frac{D}{2}} \cos(\omega_i(x-y))$$



(a) Examples of some cosine functions

(b) $\phi(m)$, a sum of cosine functions with frequencies $\omega_i = (1/10000)^{2i/D}$.

The more far, the less similar for PEs

For RPE

- Since It directly parameterize relative distance, it by definition satisfies translation invariance
- P(-m) is different with p(+m), it does not satisfy symmetry

Existing PEs

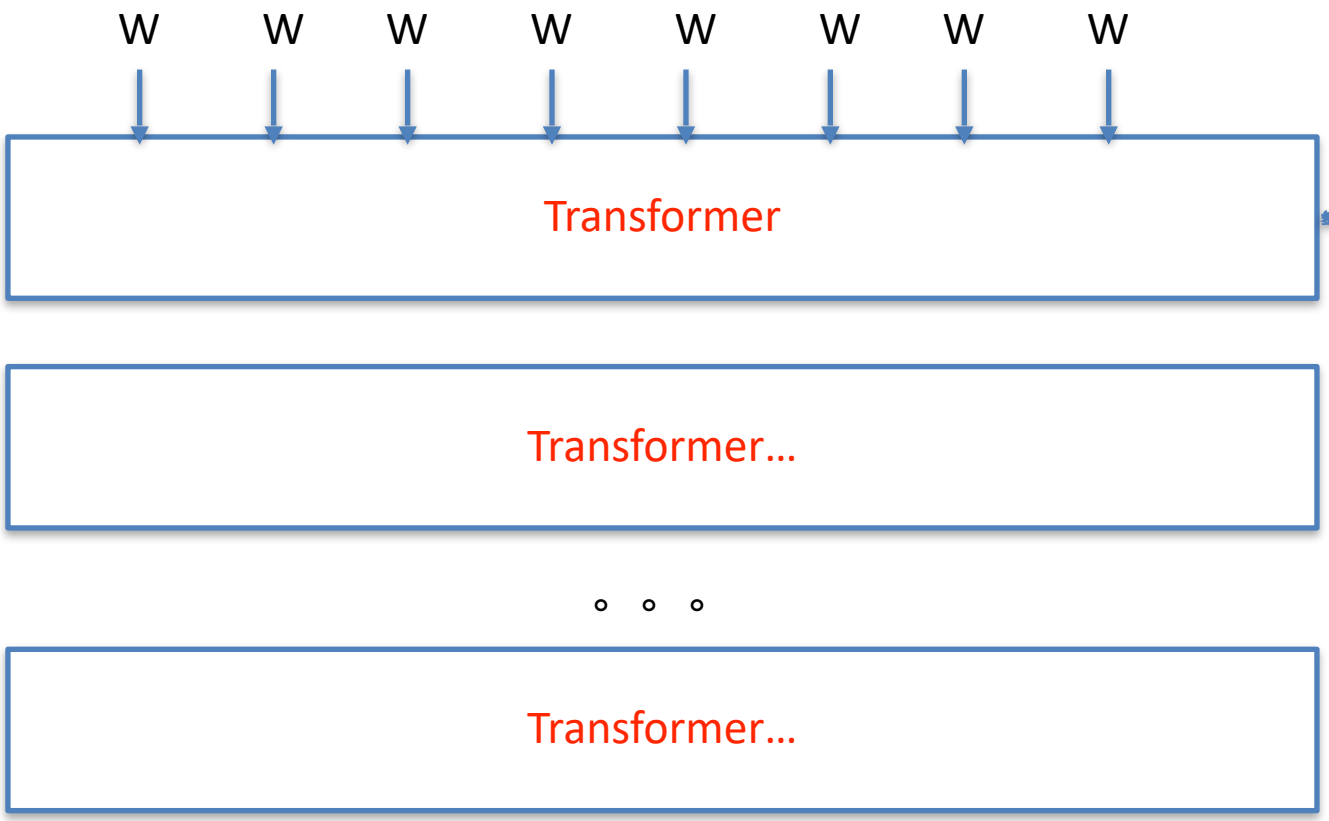
PEs	formulation	parameter scale
fully learnable APE (Gehring et al., 2017)	$P_x \in \mathbb{R}^D$	$L \times D$
fixed sinusoidal APE (Vaswani et al., 2017)	$P(x) = [\cdots, \sin(\omega_i x), \cos(\omega_i x), \cdots]^T$; $\omega_i = (1/10000)^{2i/D}$	0
learnable sinusoidal APE	$P(x) = [\cdots, \sin(\omega_i x), \cos(\omega_i x) \cdots]^T$; $\omega_i \in \mathbb{R}$	$\frac{D}{2}$
fully learnable RPE (Shaw et al., 2018)	$P_x \in \mathbb{R}^D$	$L \times D$
fixed sinusoidal RPE (Wei et al., 2019)	$P(x) = [\cdots, \sin(\omega_i x), \cos(\omega_i x), \cdots]^T$; $\omega_i = (1/10000)^{2i/D}$	0
learnable sinusoidal RPE	$P(x) = [\cdots, \sin(\omega_i x), \cos(\omega_i x), \cdots]^T$; $\omega_i \in \mathbb{R}$	L

Be either fully-learnable or functional parameterised

Pre-training setting

- Replace the PE component and train with 5 more epochs (128 length) and 2 more epoch (512 length)
    - Dataset: wiki and books (16G raw text)
    - Task: Mask word prediction and next word prediction

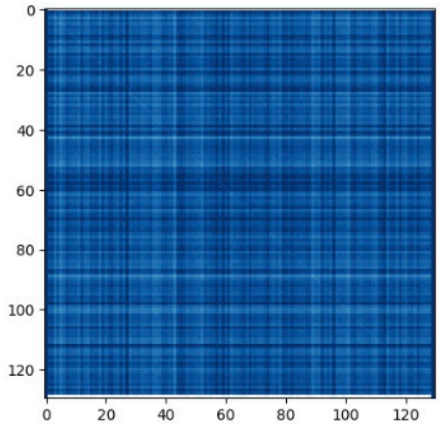
Probing test



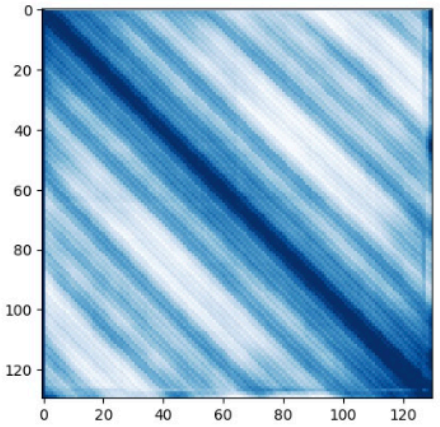
Probe the
first-layer attention

Take the attention weights in the first layers

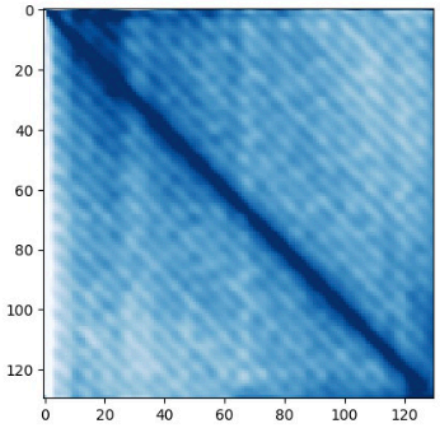
Probing test



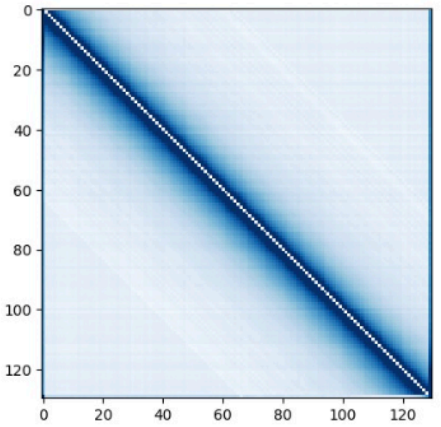
(a) BERT without PE



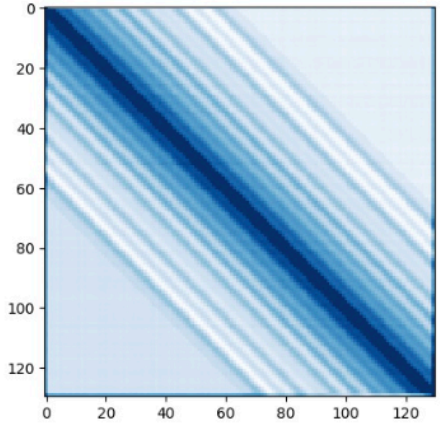
(b) fully-learnable APE



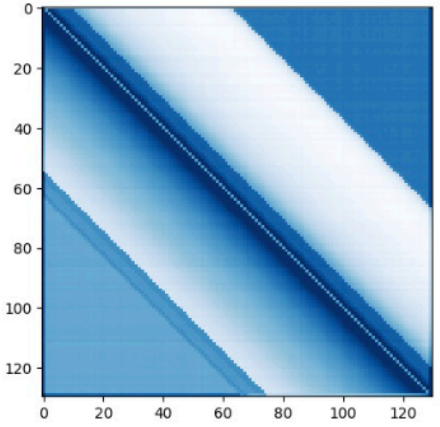
(c) learnable sin. APE



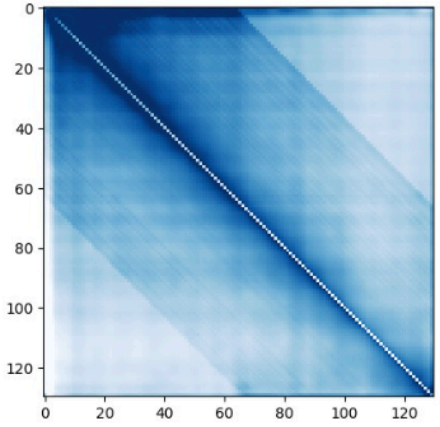
(d) fully-learnable RPE



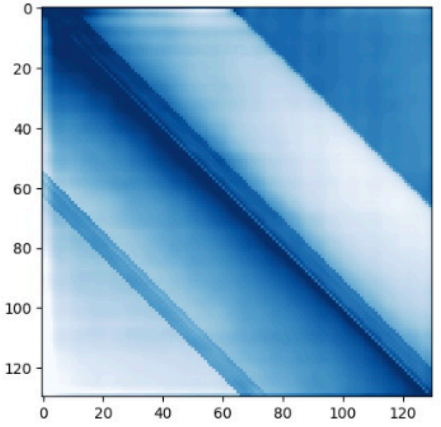
(e) fixed sin. RPE



(f) learnable sin. RPE



(g) learnable sin. APE + fully-learnable RPE



(h) learnable sin. APE + learnable sin. RPE

Probing test

PEs	monotonicity		translation invariance		symmetry	direction
	all offsets	first 20 offsets	w/ [CLS]	w/o [CLS]		balance
BERT without PE	0.5430	0.1393	0.9497	0.9939	0.0005	1.0136
BERT-style APE	0.2461	0.0208	0.5030	**0.0143**	0.0012	1.1940
fixed sin. APE	0.1937	**0.0190**	0.2552	0.2143	0.0010	**1.0266**
learnable sin. APE	**0.1936**	0.0237	**0.0653**	0.0378	**0.0004**	1.0281
fully-learnable RPE	0.1576	**0.0048**	0.1178	**0.0007**	0.0007	1.1930
fixed sin. RPE	**0.1273**	0.0054	**0.0924**	0.0020	**0.0007**	**1.1565**
learnable sin. RPE	0.3157	0.0057	0.1397	0.0038	0.0014	1.3223
BERT-style APE + fully-learnable RPE	0.1993	**0.0071**	0.2601	**0.0059**	0.0009	1.1971
BERT-style APE + fixed sin. RPE	**0.1579**	0.0143	**0.1376**	0.0072	**0.0007**	**1.1302**
BERT-style APE+ learnable sin. RPE	0.2364	0.0158	0.2334	0.0088	0.0014	1.3804
learnable sin. APE + fully-learnable RPE	0.1248	0.0065	0.0487	0.0238	0.0007	1.1196
learnable sin. APE + fixed sin. RPE	**0.0746**	**0.0040**	**0.0243**	**0.0168**	**0.0007**	**1.0773**
learnable sin. APE + learnable sin. RPE	0.1796	0.0052	0.0399	0.0252	0.0027	1.6722

The bigger, the more it violates the properties (monotonicity, translation invariance and symmetry)

Applications

- Document-level classifications (GLUE)
  - Use [CLS] for prediction
- Token-level classifications (SQuAD)
  - Use each token for prediction

GLUE requires that PEs can flexibly deal with CLS and normal positions

Downstream tasks - GLUE

PEs	single sentence		sentence pair							
	CoLA	SST-2	MNLI	MRPC	QNLI	QQP	RTE	STS-B	WNLI	
	acc	acc	acc	F1	acc	F1	acc	spear. cor.	acc	mean $\pm$ std
BERT without PE	39.0	86.5	80.1	86.2	83.7	86.5	63.0	87.4	33.8	76.6 $\pm$ 0.41
fully learnable (BERT-style) APE	60.2	**93.0**	84.8	89.4	88.7	87.8	**65.1**	**88.6**	37.5	82.2 $\pm$ 0.30
fixed sin. APE	57.1	92.6	84.3	89.0	88.1	87.5	58.4	86.9	45.1	80.5 $\pm$ 0.71
learnable sin. APE	56.0	92.8	84.8	88.7	88.5	87.7	59.1	87.0	40.8	80.6 $\pm$ 0.29
fully-learnable RPE	58.9	92.6	84.9	**90.5**	88.9	**88.1**	60.8	88.6	50.4	81.7 $\pm$ 0.31
fixed sin. RPE	60.4	92.2	84.8	89.5	88.8	88.0	62.9	88.1	45.1	81.8 $\pm$ 0.53
learnable sin. RPE	60.3	92.6	**85.2**	90.3	**89.1**	**88.1**	63.5	88.3	49.9	82.2 $\pm$ 0.40
fully learnable APE + fully-learnable RPE	59.8	92.8	85.1	89.6	88.6	87.8	62.5	88.3	**51.5**	81.8 $\pm$ 0.17
fully learnable APE + fixed sin. RPE	59.2	92.4	84.8	89.9	88.8	87.9	61.0	88.3	48.2	81.5 $\pm$ 0.20
fully learnable APE+ learnable sin. RPE	**61.1**	92.8	**85.2**	**90.5**	89.5	87.9	**65.1**	88.2	49.6	**82.5** $\pm$ 0.44
learnable sin. APE + fully-learnable RPE	57.2	92.7	84.8	88.9	88.5	87.8	58.6	88.0	51.3	80.8 $\pm$ 0.44
learnable sin. APE + fixed sin. RPE	57.6	92.6	84.5	88.8	88.6	87.6	63.1	87.4	48.7	81.3 $\pm$ 0.43
learnable sin. APE + learnable sin. RPE	57.7	92.7	85.0	89.6	88.7	87.8	62.3	87.5	50.1	81.4 $\pm$ 0.33

The fully-learnable PE performs well, not PE variants significantly outperform it

Downstream tasks - SQuADs

PEs	SQuAD V1.1		SQuAD V2.0	
	F1	EM	F1	EM
BERT without PE	36.47 ± 0.19	24.24 ± 0.33	50.48 ± 0.12	49.30 ± 0.14
fully learnable (BERT-style) APE	89.44 ± 0.08	81.92 ± 0.11	76.43 ± 0.63	73.07 ± 0.63
fixed sin. APE	89.45 ± 0.07	81.93 ± 0.11	76.12 ± 0.48	72.75 ± 0.55
learnable sin. APE	89.65† ± 0.11	82.24† ± 0.17	77.24 ± 0.43	73.93 ± 0.44
fully-learnable RPE	90.50† ± 0.08	83.38 † ± 0.11	79.85† ± 0.27	76.68† ± 0.49
fixed sin. RPE	90.30† ± 0.07	83.24 † ± 0.08	78.76† ± 0.29	75.38† ± 0.28
learnable sin. RPE	90.45† ± 0.11	83.49 † ± 0.14	79.40† ± 0.37	76.14† ± 0.33
fully learnable APE + fully-learnable RPE	90.57† ± 0.04	83.45 † ± 0.10	**80.31**† ± 0.10	76.94† ± 0.20
fully learnable APE + fixed sin. RPE	90.24† ± 0.17	83.06 † ± 0.21	78.74† ± 0.50	75.40† ± 0.52
fully learnable APE+ learnable sin. RPE	89.56 ± 0.28	82.26 † ± 0.30	77.82† ± 0.42	74.51† ± 0.39
learnable sin. APE + fully-learnable RPE	**90.72**† ± 0.13	**83.68** † ± 0.27	80.24 † ± 0.35	**76.98** † ± 0.34
learnable sin. APE + fixed sin. RPE	90.36† ± 0.08	83.25 † ± 0.10	78.81† ± 0.33	75.71† ± 0.28
learnable sin. APE + learnable sin. RPE	90.49† ± 0.14	83.59 † ± 0.14	79.93† ± 0.34	76.69† ± 0.39

RPEs and sin. APEs perform better than the fully learnable PE

Tips of PEs

- Untie [CLS] and PEs for document-level classification
- Use RPE for token-level classification

Correlations between properties and performance

Properties		CoLA	SST-2	MNLI	QQP	GLUE	SQuAD V1.1	SQuAD V2.0
monotonicity	all offsets	0.44	0.43	**0.56**	0.32	0.48	-0.31	-0.27
	first 20 offsets	-0.18	0.44	-0.24	-0.42	-0.21	**-0.91**	-0.86
translation invariance	w/ `[CLS]/[SEP]`	0.48	0.52	0.04	-0.07	0.42	**-0.63**	-0.57
	w/o `[CLS]/[SEP]`	-0.47	0.01	**-0.69**	-0.68	-0.61	-0.51	-0.58
symmetry		0.17	0.24	**0.40**	0.09	0.31	0.15	0.16
direction balance		0.32	0.16	**0.63**	0.35	0.48	0.32	0.37

violating local monotonicity and translation invariance is harmful, while violating symmetry (and direction-balance) is beneficial

- Thanks
- We will further post a blog to explain more details of position embeddings, very soon