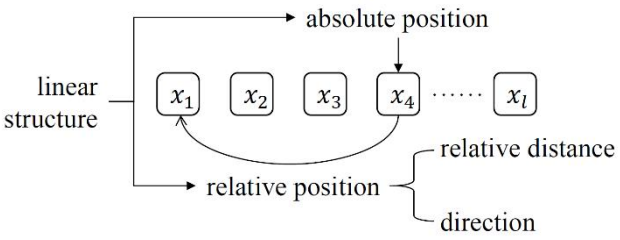
Bidirectional Transformer with Absolute-position Aware Relative Position Encoding for Encoding Sentences

Le Qi, Yu Zhang, Ting Liu

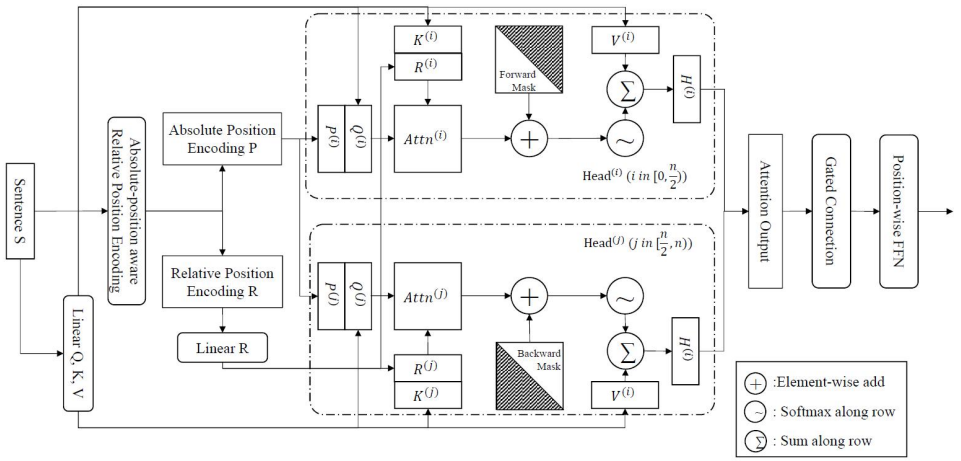
Frontiers of Computer Science, DOI: 10.1007/s11704-022-0610-2

Problems & Ideas

- Problems of Transformers on Encoding Sentences:
  - Transformers are weak in modeling the linear structure of sentences.
  - Existing methods cannot combine all the linear structural information organically, containing:
    - The absolute position of tokens
    - The relative distance and direction between tokens
- Ideas:
  - Model the absolute position and relative distance together through the absolute position aware relative position encoding
  - Model the direction between tokens through the bidirectional mask strategy



The composition of sentence linear structure



The overview of our BiAR-Transformer

Main Contributions

- Contributions:
  - We propose a novel bidirectional transformer with absolute-position aware relative position encoding (BiAR-Transformer) for enhancing the ability of Transformers in utilizing linear structural information to better model sentences.
  - In order to model the relative distance bias and the semantic bias related to the position of tokens in the attention calculation process, we propose an absolute-position aware relative position encoding (A-RPE), which combines the absolute position of tokens and the relative distance between tokens together.
  - In order to capture the directional information between tokens which is not aware by the A-RPE, we propose a bidirectional mask strategy, which applies the forward and backward masks simultaneously to the multi-head attention based on the A-RPE.

Table 2 Comparison results with encoders training from scratch.

Model	Dim	$\|\theta\|$	SNLI	SST-5	En-De
Transformer [14]	300	-	82.2	50.4	-
Transformer [1]	-	-	-	-	27.3
DiSAN [6]	600	2.4m	85.6	51.7	-
TreeLSTM [15]	150	316K	-	51.0	-
Star-Transformer [14]	300	-	86.0	52.9	-
PSAN [16]	300	2.0m	86.1	-	-
Distance-based SAN [17]	1200	4.7m	86.3	-	-
DRCN [18]	-	5.6m	86.5	-	-
HBMP [19]	600	22m	86.6	-	-
Transformer w/rpe [20]	-	-	-	-	26.8
Transformer w/rec. pe [21]	-	-	-	-	28.3
Transformer w/reorder. pe [22]	-	-	-	-	28.2
SRAR [23]	300	3.2m	86.8	52.6	28.2
BiAR-Transformer	600	1.8m	**86.9***	**53.2***	28.2*

Table 3 Experimental results comparing with BERT-base.

Model	SNLI	QQP	MNLI-m/mm	SST-5
BERT-base	85.7	89.6	75.6/75.3	56.1
Transformer*$_{bert}$	86.4	89.8	76.0/75.7	56.5
BiAR-Transformer$_{bert}$	**87.5**	**90.1**	**76.4/76.5**	**57.1**

Experimental results.