Feature Aggregation in Zero-Shot Cross-Lingual Transfer Using Multilingual BERT

Beiduo Chen [1] Wu Guo [1] Quan Liu [2] Kun Tao [2]

[1]NERC-SLIP, University of Science and Technology of China [2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research

Introduction

- **Task** zero-shot cross-lingual transfer
- **Object** pre-trained multilingual BERT (mBERT)
- **Issue** the mainstream methods to solve the cross-lingual downstream tasks are always using the last transformer layer's output of mBERT as the representation of linguistic information
- **Contribution**
  - We prove that the output of layers before the last layer can provide supplementary information to the last layer of mBERT for different zero-shot cross-lingual downstream tasks. The optimal dynamic equilibrium between cross-lingual capability and language-structured ability of mBERT is discussed.
  - We design a feature aggregation module based on an attention mechanism to fuse information from two transformer layers.
  - Experimental results on four cross-lingual downstream datasets show that our method improves the performance of mBERT on all tasks compared to the baseline and is generalized in various situations.

System Architecture

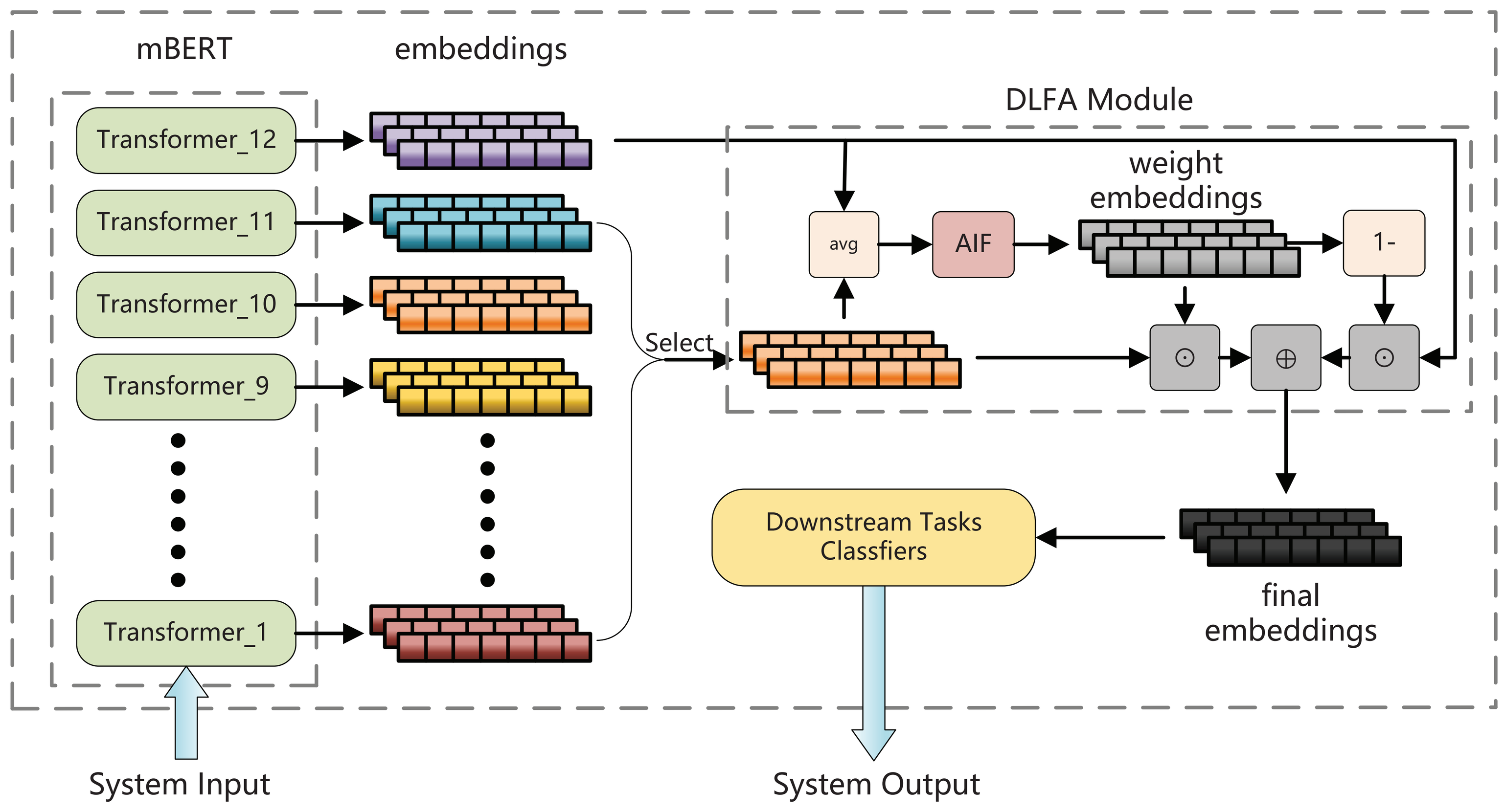

Figure 1. The overall structure of the proposed system.

Attentional Information Fusion

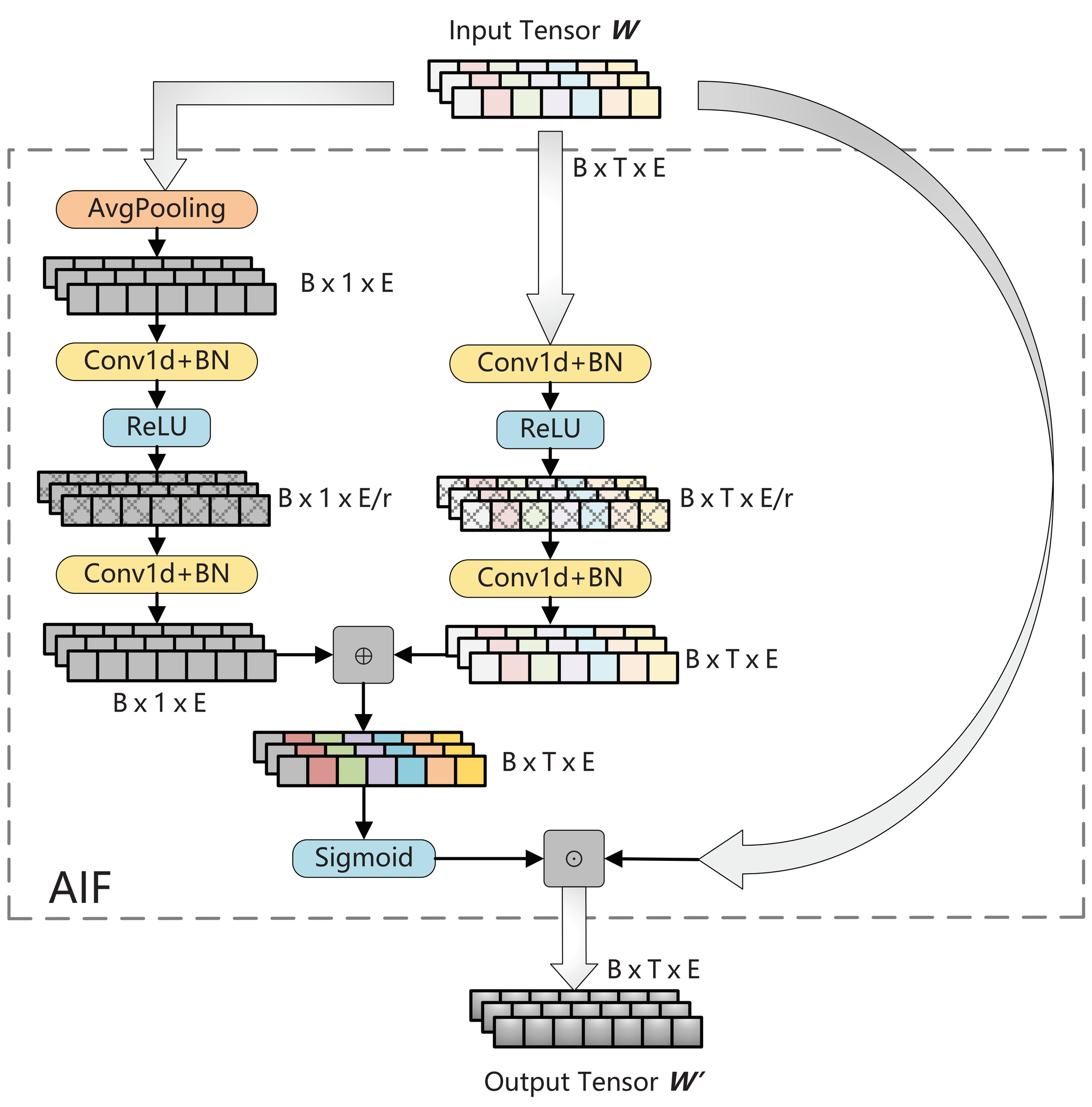

Figure 2. The architecture of the proposed attentional information fusion (AIF) module. The AIF extracts global and local information via two branches and element-wisely multiplies the result with the input tensor

Double Layers Feature Aggregation

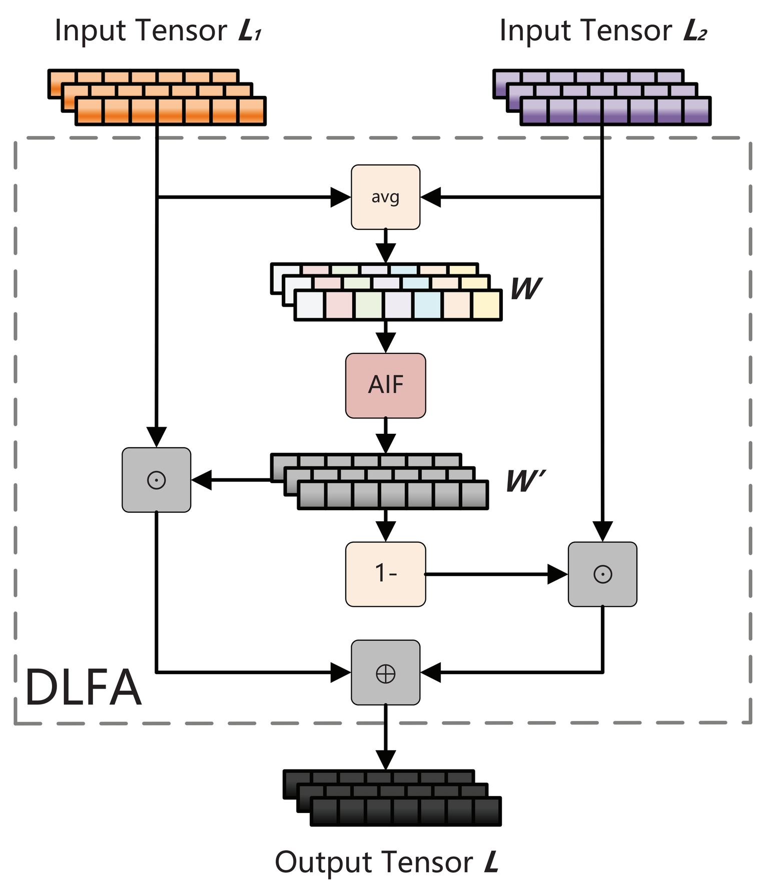

Figure 3. The architecture of the proposed double layers feature aggregation (DLFA) module. The DLFA fuses the information from one selected lower transformer layer of mBERT and that from the last transformer layer.

$$W' = W \odot Sigmoid(W_{Global} \oplus W_{Local}) = AIF(W) \quad (1)$$

$$L = L_1 \odot AIF(\frac{L_1 \oplus L_2}{2}) + L_2 \odot (1 - AIF(\frac{L_1 \oplus L_2}{2})) \quad (2)$$

Main Results on Xtreme Benchmark

Table 1. All results of zero-shot cross-lingual transfer trials for 4 tasks. "D_x" means the system with the DLFA module that fuses the last and the x^{th} transformer layers' output.

Task	XNLI	PAWS-X	NER	POS
Model\Metrics	Acc (%)	Acc (%)	F1 (%)	F1 (%)
baseline	65.40	81.94	62.17	70.28
D_11	**66.91**	83.04	62.27	71.53
D_10	66.55	84.33	62.76	**71.81**
D_9	66.57	84.24	62.43	71.58
D_8	66.90	82.91	**63.34**	71.36
D_7	66.20	83.48	62.63	71.52
D_6	66.75	**84.37**	61.84	71.29
D_5	65.42	82.44	61.66	71.08
D_4	66.14	82.75	62.28	71.26
D_3	66.00	83.81	61.88	71.21
D_2	66.15	83.04	61.34	71.38
D_1	65.85	82.50	61.73	71.18

Best performances on these four tasks are obtained with different fusion layers. Different tasks focus on different aspects of language structure learning ability, resulting in the fusion of different layers. Table 2,3,4 indicate that the information of language structure lies on the upper layers while the lower layers of mBERT are more flexible for cross-lingual transfer.

Analysis and Discussion

Table 2. Results of Classification tasks. The scores in the "avg_enf" column denote the average performance of the subset "enf" (involving "en, de"). The scores in the "avg_noenf" column denote the average performance of the subset "noenf".

XNLI			PAWS-X		
Model	avg_enf	avg_noenf	Model	avg_enf	avg_noenf
baseline	75.40	63.86	baseline	89.85	78.80
D_11	**77.34**	65.31	D_10	**90.30**	81.94
D_8	77.11	**65.34**	D_6	90.00	**82.12**

Table 3. Results of cosine similarity experiment on XNLI and PAWS-X.

XNLI		PAWS-X	
Model	Avg C.S.	Model	Avg C.S.
D_11	0.5689	D_10	0.8548
D_8	**0.6822**	D_6	**0.9461**

Table 4. Several languages' results on PAWS-X(Acc.)

Model\Lang	en	de	fr	es	ko	zh
mBERT	94.0	85.7	87.4	87.0	69.6	77.0
D_10	**94.1**	**86.5**	**88.3**	88.79	75.8	80.3
D_6	93.9	86.0	87.8	**89.09**	**76.1**	**81.2**