Automatic Student Network Search for Knowledge Distillation

1 MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, China

Zhexi Zhang[1], Wei Zhu[2], Junchi Yan[1*], Peng Gao[2] and Guotong Xie[2*]

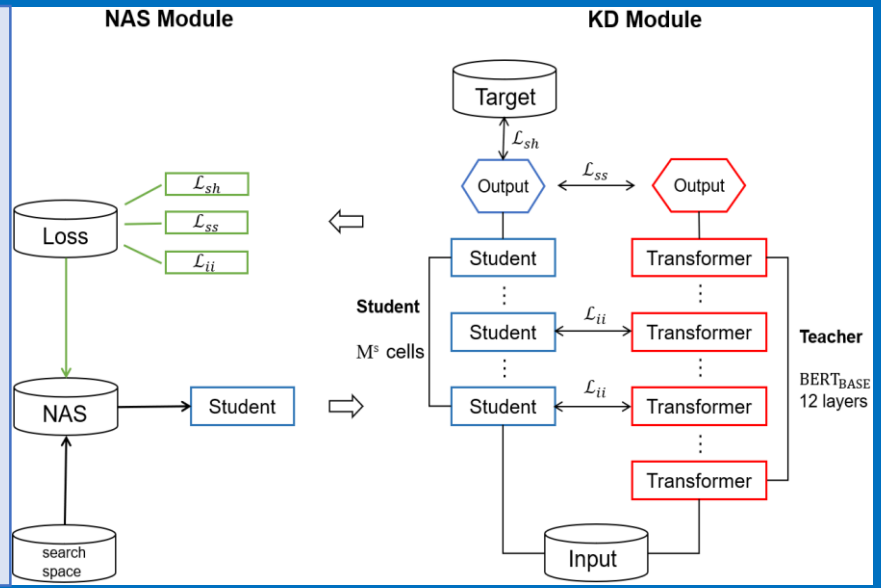
2 Ping An Technology (Shenzhen) Co. Ltd., Shenzhen, China

Introduction

- BERT is a commonly used pretrained language model (PLM) that obtains state-of-the-art results on 11 different natural language processing (NLP) tasks.
- However, BERT contains a large number of parameters and requires vast numbers of computational resources. Concretely, $\text{BERT}_{\text{BASE}}$ has 110 million parameters while $\text{BERT}_{\text{LARGE}}$ has 340 million.
- Knowledge distillation (KD) is a promising compression method and has achieved great success in compressing BERT. KD follows a student-teacher framework where the light-weight student network learns from the teacher.
- However, the student networks in previous KD studies are manually designed. Researchers have tried to compressing BERT into MLP, BiLSTM and network with less Transformer layers. These manually designed students are not optimal since they either still contain redundant parameters or have week representation ability.
- Motivated by the above observations, we propose to **automatically search for a compact student network for compressing BERT using neural architecture search (NAS)**.

NAS-KD: our Framework

As illustrated in the right, NAS-KD is composed of NAS module and KD module. In NAS module, we select and combine the candidate operations in search space to generate a student cell. In KD module, multiple student cells are stacked as the student network in knowledge distillation. The distillation objective is formulated by three different loss functions. The total loss is returned to NAS module to update the parameters and structure of student cell.



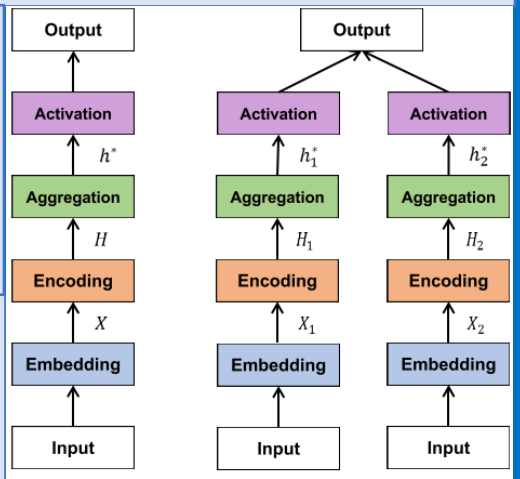
Search Space in NAS Module

Encoding Layer

- LSTM
- Identity
- 4, 8 head attention
- Standard convolution
- Depthwise-separable convolution
- None

Aggregation Layer

- Max pooling
- Average pooling
- Dynamic routing
- Self-attention pooling



(a) basic cell (b) cell structure on sentence-pair tasks

KD Module

Except vanilla KD objective where the student learns from the ouput of teacher, we encourage each student cell to uniformly mimic a corresponding Transformer layer of BERT. The formulations are in the right.

$$\mathcal{L}_{ss} = -\sum_{i=1}^{N}\sum_{j=1}^{C}(\hat{y}^t_{i,j} \cdot \log(\hat{y}^s_{i,j}/t))$$

$$\mathcal{L}_{sh} = -\sum_{i=1}^{N}\sum_{j=1}^{C}(y_i \cdot \log \hat{y}^s_{i,j})$$

$$\mathcal{L}_{ii} = \sum_{i=1}^{N}\sum_{j=1}^{M^s-1} \| \frac{\mathbf{h}^s_{i,j}}{\|\mathbf{h}^s_{i,j}\|_2^2} - \frac{\mathbf{h}^t_{i,\frac{M^t}{M^s}j}}{\|\mathbf{h}^t_{i,\frac{M^t}{M^s}j}\|_2^2} \|_2^2$$

$$\mathcal{L} = \alpha\mathcal{L}_{ss} + (1-\alpha)\mathcal{L}_{sh} + \beta\mathcal{L}_{ii}$$

Method	Par. (w/o Emb)	Par. (total)	Layers	SST-2	MRPC	QQP	MNLI-m	MNLI-mm	QNLI	RTE
$\text{BERT}_{\text{BASE}}$ (Google)	85.2	109	12	93.5	88.9	71.2	84.6	83.4	90.5	66.4
$\text{BERT}_{\text{BASE}}$ (Teacher)	85.2	109	12	93.1	87.7	71.2	83.5	82.8	90.3	66.1
Distilled $\text{BiLSTM}_{\text{SOFT}}$	0.96	10.1	1	90.7	-	68.2	73.0	72.6	-	-
TinyBERT (w/o DA)	4.8	9.7	4	-	82.4	-	80.5	81.0	-	-
$\text{BERT}_{\text{SMALL}}$	4.8	9.7	4	87.6	83.2	66.5	75.4	74.9	84.8	62.6
DistilBERT	28.4	52.2	4	91.4	82.4	68.5	78.9	78.0	85.2	54.1
BERT_3-FT	23.9	45.7	3	86.4	80.5	65.8	74.8	74.3	84.3	55.2
BERT_3-KD	23.9	45.7	3	86.9	79.5	67.3	75.4	74.8	84.0	56.2
BERT_3-PKD	23.9	45.7	3	87.5	80.7	68.1	76.7	76.3	84.7	58.2
NAS-KD_3	**9.4±1.5**	**33.2±1.5**	3	86.9	79.3	67.5	76.1	75.5	83.9	**58.9**
BERT_6-FT	43.2	67.0	6	90.7	85.9	69.2	80.4	79.7	86.7	63.6
BERT_6-KD	43.2	67.0	6	91.5	86.2	70.1	80.2	79.8	88.3	64.7
BERT_6-PKD	43.2	67.0	6	92.0	85.0	70.7	81.5	81.0	89.0	65.5
NAS-KD_6	**18.6±2.9**	**42.4±2.9**	6	**92.2**	**86.3**	70.4	81.0	80.2	88.6	**65.9**