KoreALBERT: Pretraining a Lite BERT Model for 🇰🇷 Korean Language Understanding

Paper ID : 1510

Hyunjae Lee, Jaewoong Yoon, Bonggyu Hwang, Seongho Joe, Seungjai Min, Youngjune Gwon

SAMSUNG SDS

Motivation

Emerging of Pretrained Language Model(PLM)

- Many state-of-the-art results on natural language processing benchmarks have been accomplished by PLM, mainly BERT and BERT-derieved models

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
2	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	90.7	97.4	91.2	94.5	49.1
3	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
4	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	91.0	96.6	90.9	94.5	51.7
5	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1

< Super-GLUE leader board(December, 2020) >

Motivation

Emerging of Pretrained Language Model(PLM)

- Many state-of-the-art results on natural language processing benchmarks have been accomplished by PLM, mainly BERT and BERT-derieved models

Lack of Korean Language Model

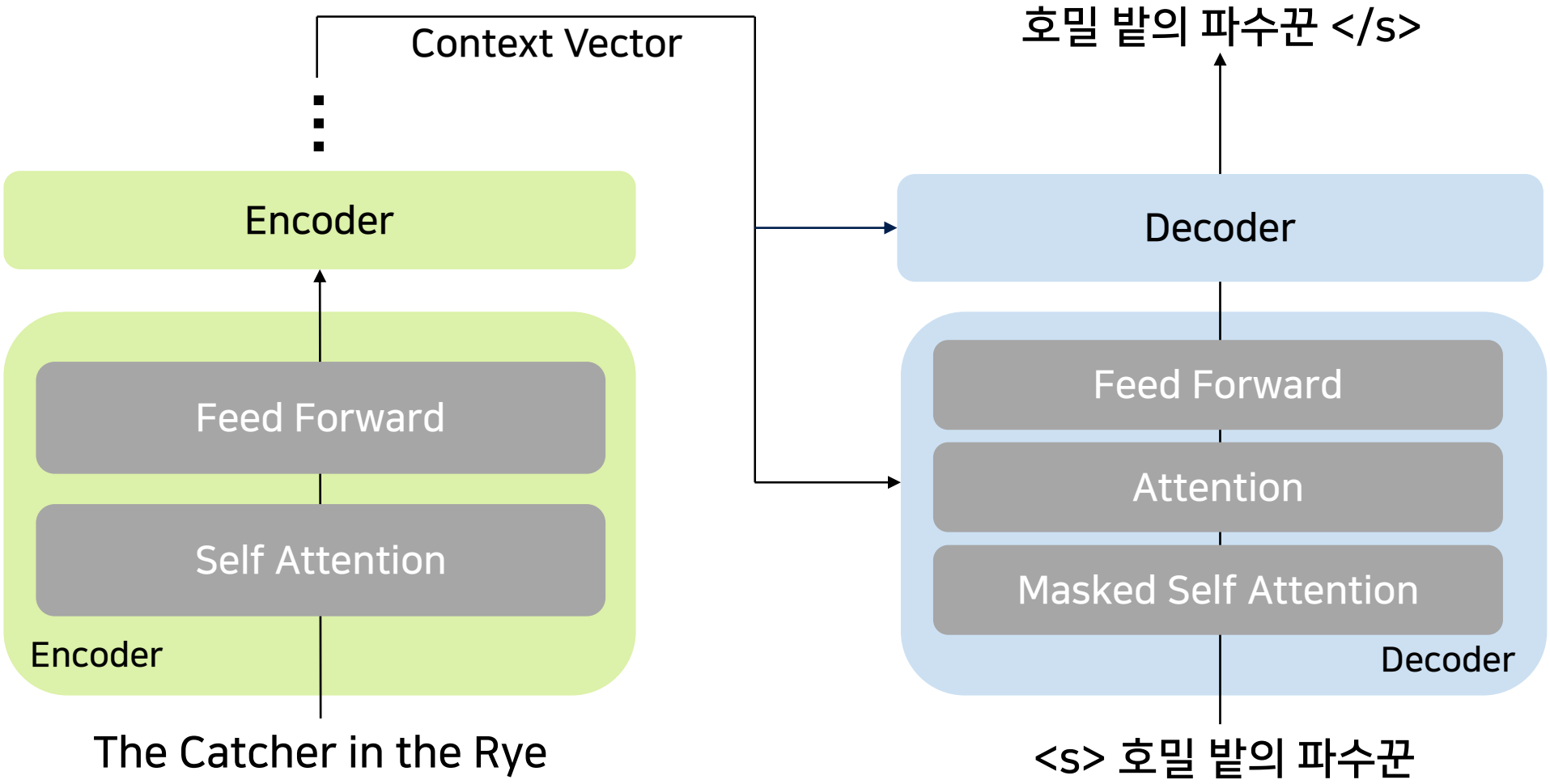
- Most of research in PLM has been limited to the English language

Multi-lingual Pre-trained Model is not sufficient

- Google has released BERT multilingual model (M-BERT) pre-trained using 104 different languages including the Korean
- In most cases, Mono-lingual Model outperforms Multi-lingual Model

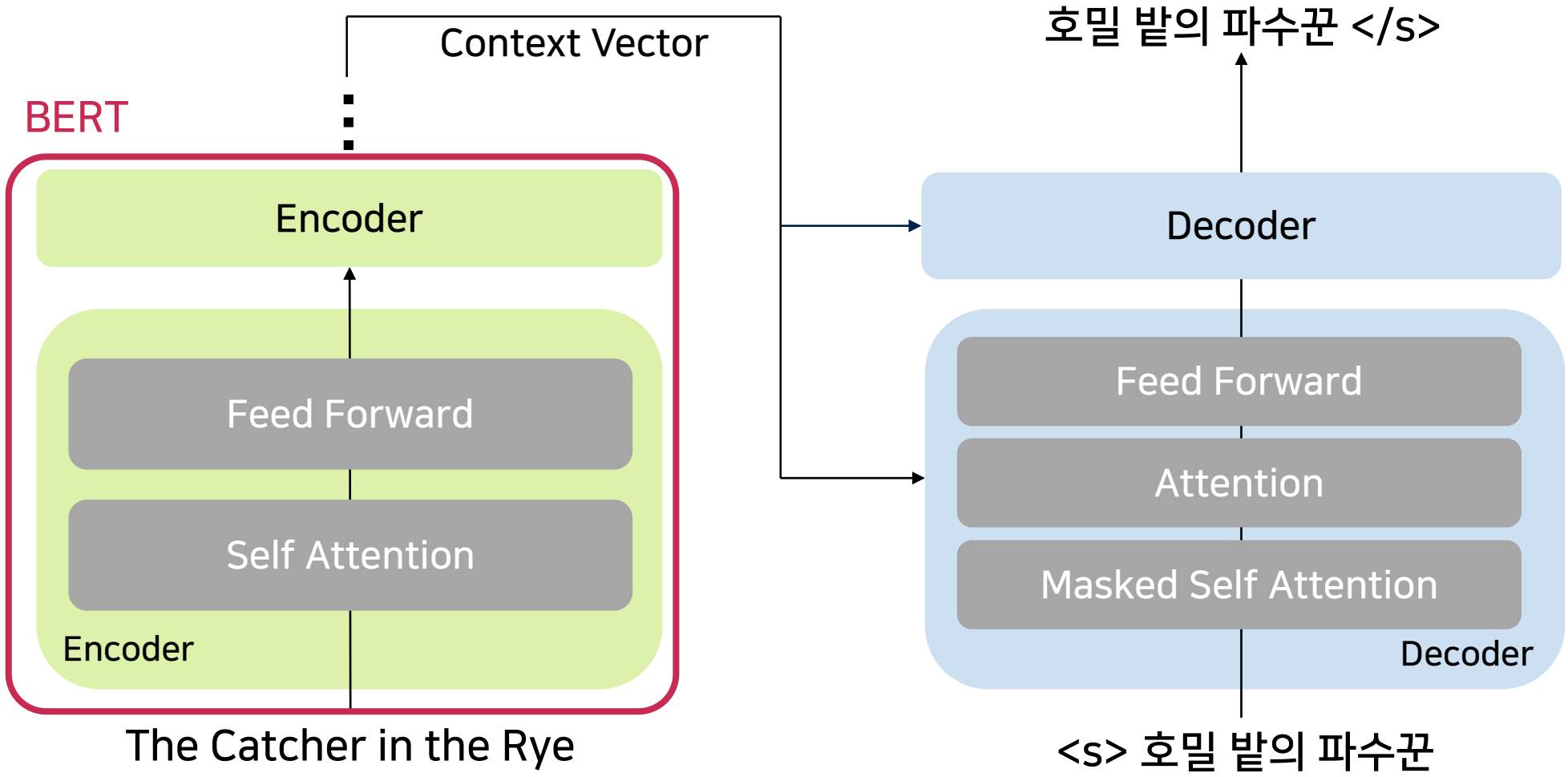
Background

Transformers



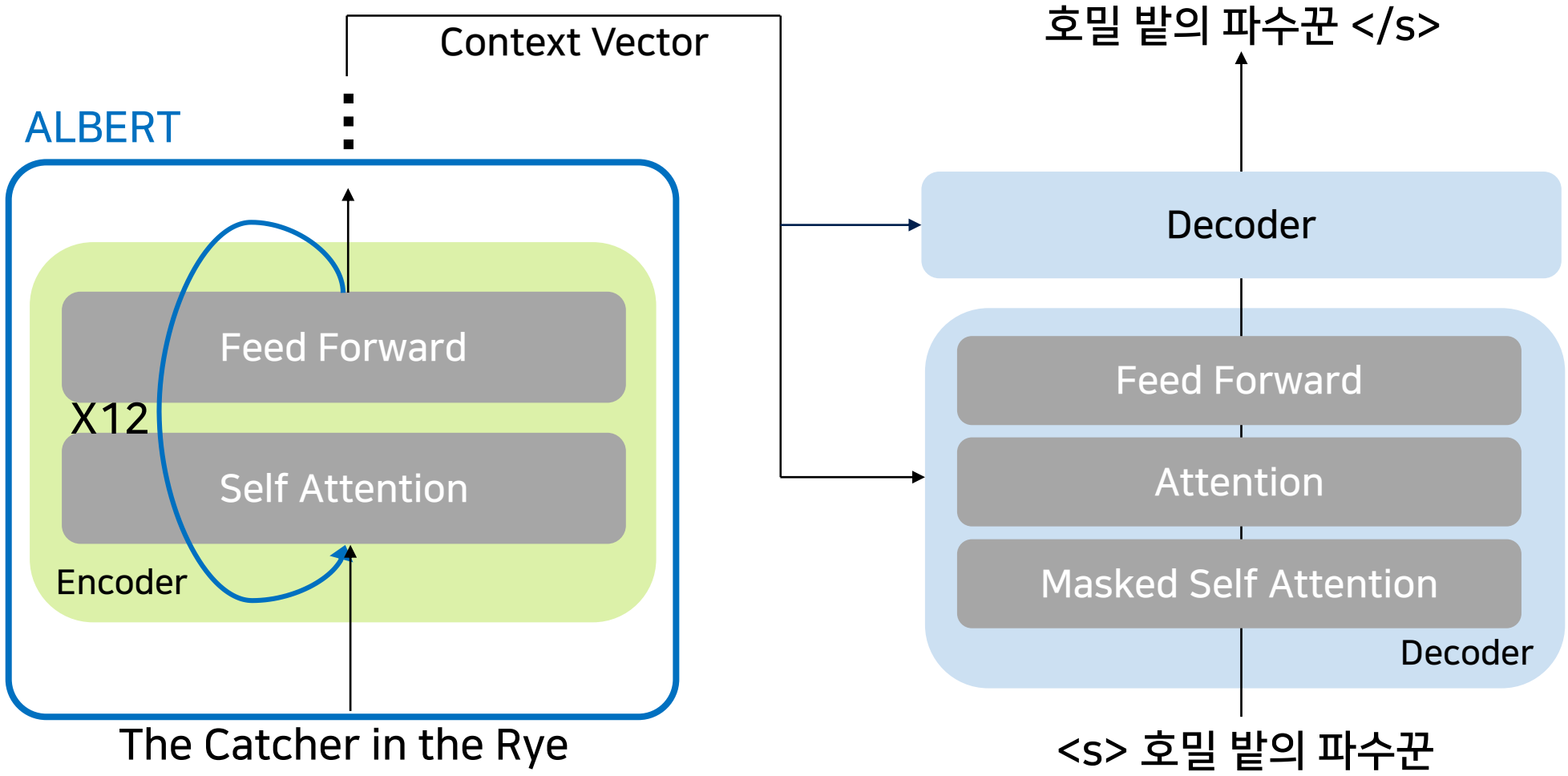
Background

BERT(Bidirectional Encoder Representation from Transformers)



Background

ALBERT(A Lite BERT)



Related Work

Language-Specific BERT

※ Image Source :https://www.freepik.com/

BERT For Korean

SKT KoBERT

- BERT Architecture, 92M Parameters
- 320M words from Korean Wiki and News
- SentencePiece Tokenizer, Vocab 8K

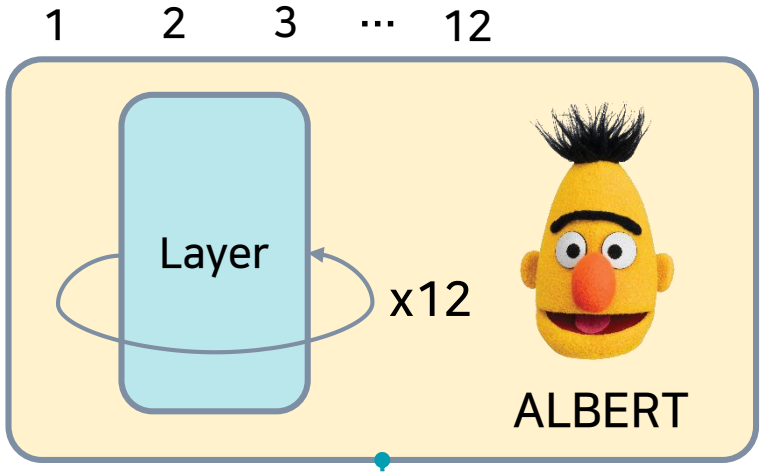
ETRI BERT

- BERT Architecture, 110M Parameters
- 4.7B morphemes
- Morpheme Based Tokenizer, Vocab 30K

No Avaliable ALBERT practice

Method

Korean + ALBERT = KoreALBERT



Masked Token Prediction
Word Order Prediction
Sentence Order Prediction

Corpus

- Korean Wiki, news, book review(43GB)
- General Domain, Written expressions
- SestencePiece Tokenizer, 32K of vocab size

Architecture

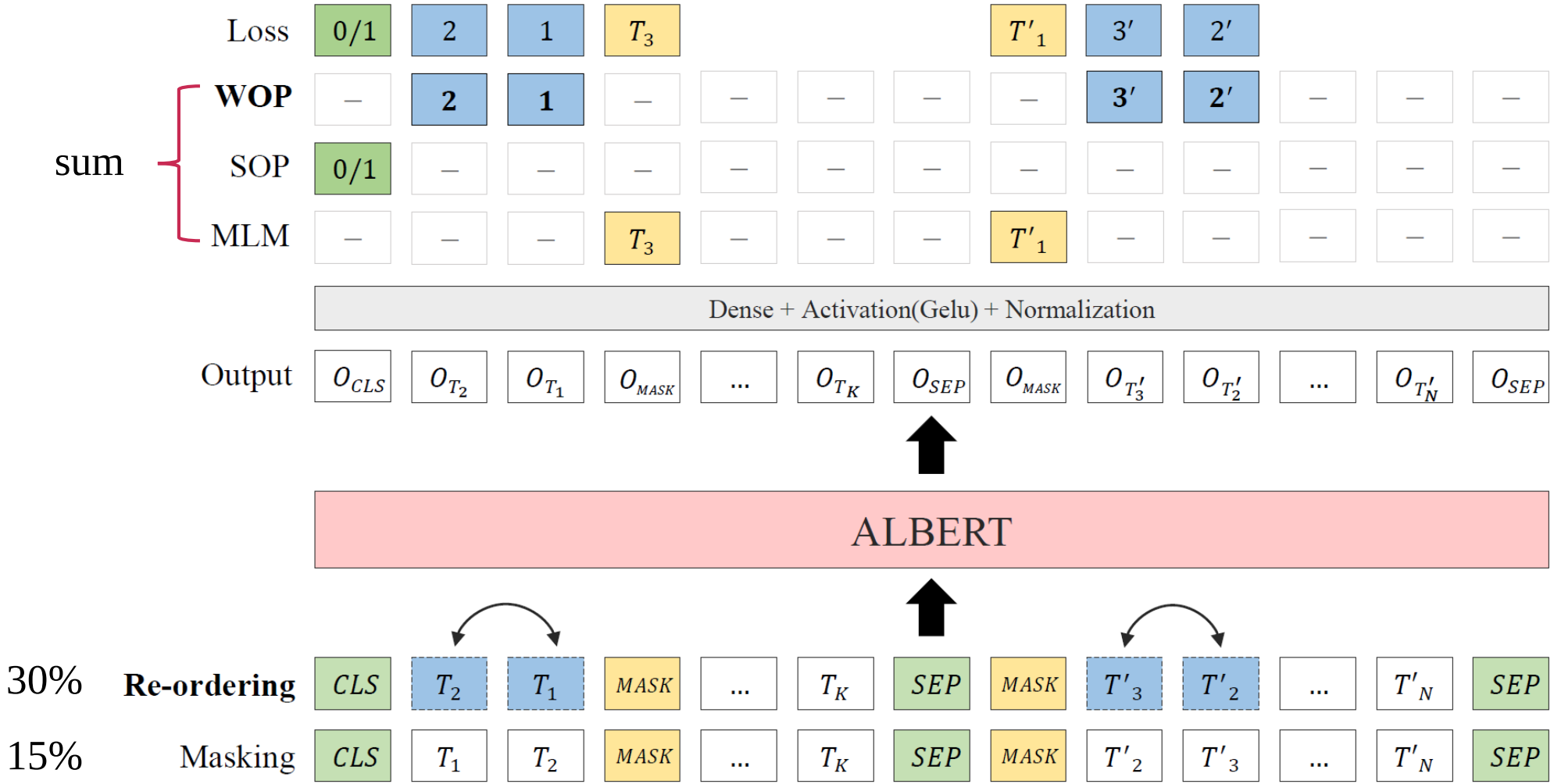
- Same architecture with ALBERT
- Size Comparison :
  BERT Base 110M, ALBERT Base 12M

Training Objectives

- We introduce an additional training objective, Word Order Prediction(WOP)

Method

KoreALBERT Pre-training architecture



Evaluation

6 Evaluation Tasks

Task	Dataset	Input Example(Translated in English)	Prediction
Sentiment Analysis	NSMC	“I give it a MILLION STARS!!!”	1(Positive)
Phraphrasing Detection	PD	“Katz was born in Sweden in 1947 and moved to New York City at the age of 1.” “Katz was born in 1947 in Sweden and moved to New York at the age of one.”	1(Positive)
Sentence Similarity	KorSTS	“A main is eating food.” “A man is eating something”	0.84 (Spearman)
Natural Language Inference	KorNLI	“You don’t have to stay there” “You can leave.”	Entailment
Reading Comprehension	KorQuAD1.0	Pair(Paraphrase, Question)	Start/End Index of Answer
Named Entity Recognition	NER	“I went to the Changwon Univ. to watch the Daedongjae”	Changwon Univ/LOC_B, Daedongjae/EVT_B

Evaluation

Compatibility of Word Order Prediction (WOP)

- WOP hardly hurts the performance of MLM or SOP
- WOP has been proved to improve performance of downstream tasks when it is added to other two objectives

TABLE II
Experimental results on downstram tasks performance comparing between different combination of pretraining objectives

Objectives	MLM acc	SOP acc	WOP acc	KorNLI acc	KorSTS spearman	NSMC acc	PD acc	NER acc	KorQuAD1.0 f1
MLM + SOP	35.3	79.8	-	76.4	75.6	**88.6**	92.9	80.7	89.5
MLM + SOP + WOP	35.1	79.1	80.7	**76.9**	**76.6**	88.4	**93.2**	**91.2**	**89.8**
MLM + WOP	35.6	-	84.0	76.8	73.3	88.5	92.3	81.0	89.3

* avg. score of 10 different seeds(50K steps)

Evaluation

- **Final Results on 6 Downstream tasks**
  - KoreALBERT consistently **outperforms multi and monolingual baselines** on 6 downstream NLP tasks while having **much fewer parameters**

TABLE III
Experimental results on downstream tasks and model parameters

Model	Params (M)	Speedup	KorNLI acc	KorSTS spearman	NSMC acc	PD acc	NER acc	KorQuAD1.0 f1	Avg.
Multilingual BERT	172	1.0	76.8	77.8	87.5	91.1	80.3	86.5	83.3
XLM-R Base	270	0.5x	80.0	79.4	**90.1**	92.6	**83.9**	92.3	86.4
KoBERT	92	1.2x	78.3	79.2	**90.1**	91.1	82.1	90.3	85.2
ETRI BERT	110	-	79.5	80.5	88.8	93.9	82.5	94.1	86.6
KoreALBERT Base	12	5.7x	79.7	81.2	89.6	93.8	82.3	92.6	86.5
KoreALBERT Large	18	1.3x	**81.1**	**82.1**	89.7	**94.1**	83.7	**94.5**	**87.5**

Evaluation

- **Final Results on 6 Downstream tasks**
  - KoreALBERT consistently **outperforms multi and monolingual baselines** on 6 downstream NLP tasks while having **much fewer parameters**

TABLE III
Experimental results on downstream tasks and model parameters

Model	Params (M)	Speedup	KorNLI acc	KorSTS spearman	NSMC acc	PD acc	NER acc	KorQuAD1.0 f1	Avg.
Multilingual BERT	172	1.0	76.8	77.8	87.5	91.1	80.3	86.5	83.3
XLM-R Base	270	0.5x	80.0	79.4	**90.1**	92.6	**83.9**	92.3	86.4
KoBERT	92	1.2x	78.3	79.2	**90.1**	91.1	82.1	90.3	85.2
ETRI BERT	110	-	79.5	80.5	88.8	93.9	82.5	94.1	86.6
KoreALBERT Base	12	5.7x	79.7	81.2	89.6	93.8	82.3	92.6	86.5
KoreALBERT Large	18	1.3x	**81.1** +4.3	**82.1** +4.3	89.7 +2.2	**94.1** +3	83.7 +3.4	**94.5** +8	**87.5**

Conclusion

- We have introduced KoreALBERT, **a pre-trained monolingual ALBERT** model for Korean language understanding.

- We have proposed a **word order prediction** loss, a new training objective, which is compatible with the original MLM and SOP objectives of ALBERT.

- Further questions and ideas, Contact us : h8.lee@samsung.com