Using Transformers for Automatic Short Answer Grading (ASAG)

Midterm Presentation

Period: 20.11.2019—20.05.2020

Yantao Shi
Yantao.shi.thomas@gmail.com

Supervisor
Anna Marie Filighera

KOM – Multimedia Communications Lab
Technical University of Darmstadt
Prof. Dr.-Ing. Ralf Steinmetz (Director)
Dept. of Electrical Engineering and Information Technology
Dept. of Computer Science (adjunct Professor)
www.KOM.tu-darmstadt.de

18. Februar 2020

Overview

- Background Knowledge
- Target Dataset
- Related Models & Motivation
- Targets and Approaches of this work
- Next steps

Automatic Short Answer Grading

"Automatic short answer grading (ASAG) is the task of assessing short natural language responses to objective questions using computational methods."[BGS15]

[BGS15] Burrows, S., Gurevych, I. & Stein, B. The Eras and Trends of Automatic Short Answer Grading. *Int J Artif Intell Educ* **25,** 60–117 (2015).

Transformer

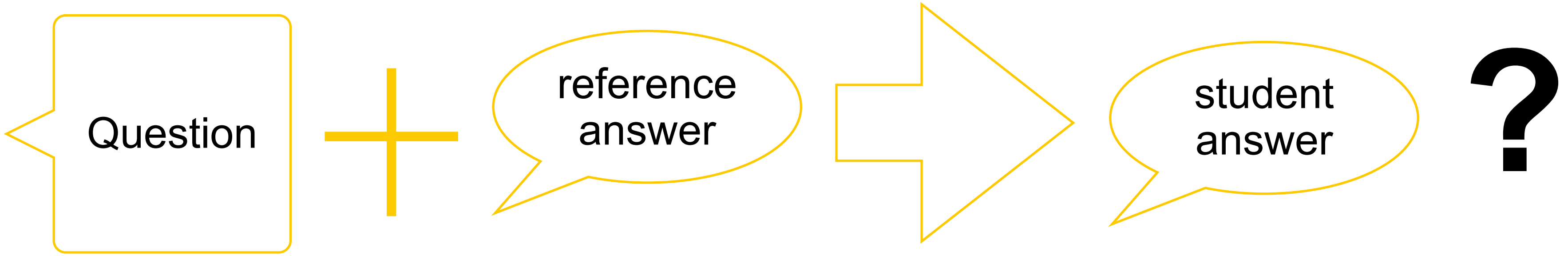
- Attentional mechanism
- Embedding matrix of each word
- All words were calculated by matrix operation
- Feedforward neural network to get a new representation

Advantages

- Take context information into account
- Attention can be achieved in one step of matrix calculation- more efficient

Target Dataset SemEval-2013

Recognizing Textual Entailment Challenge at Semantic Evaluation 2013(SemEval) workshop



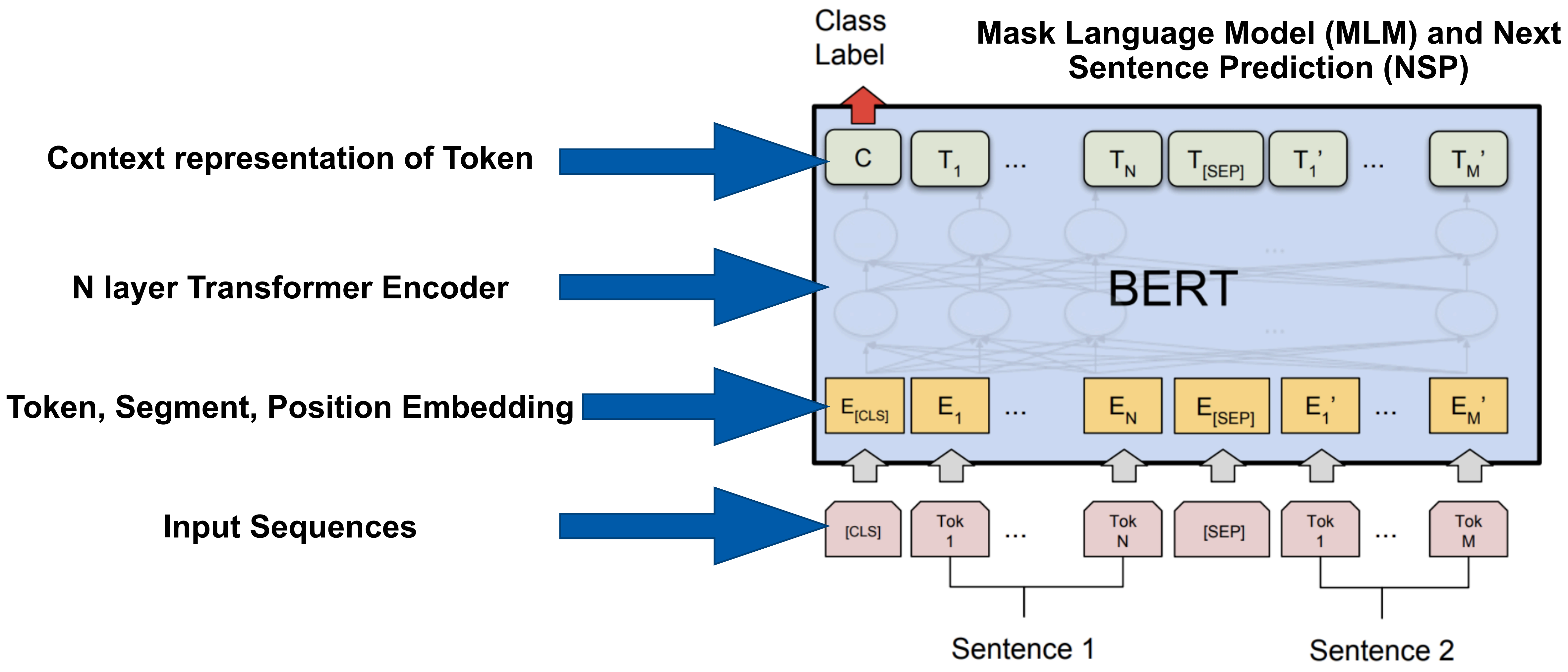
Transformer for ASAG

On the traget dataset of SemEval-2013, Transformer has up to 10% absolute improvement in macro-average-F1 over state-of-the-art(non-transformer) results.

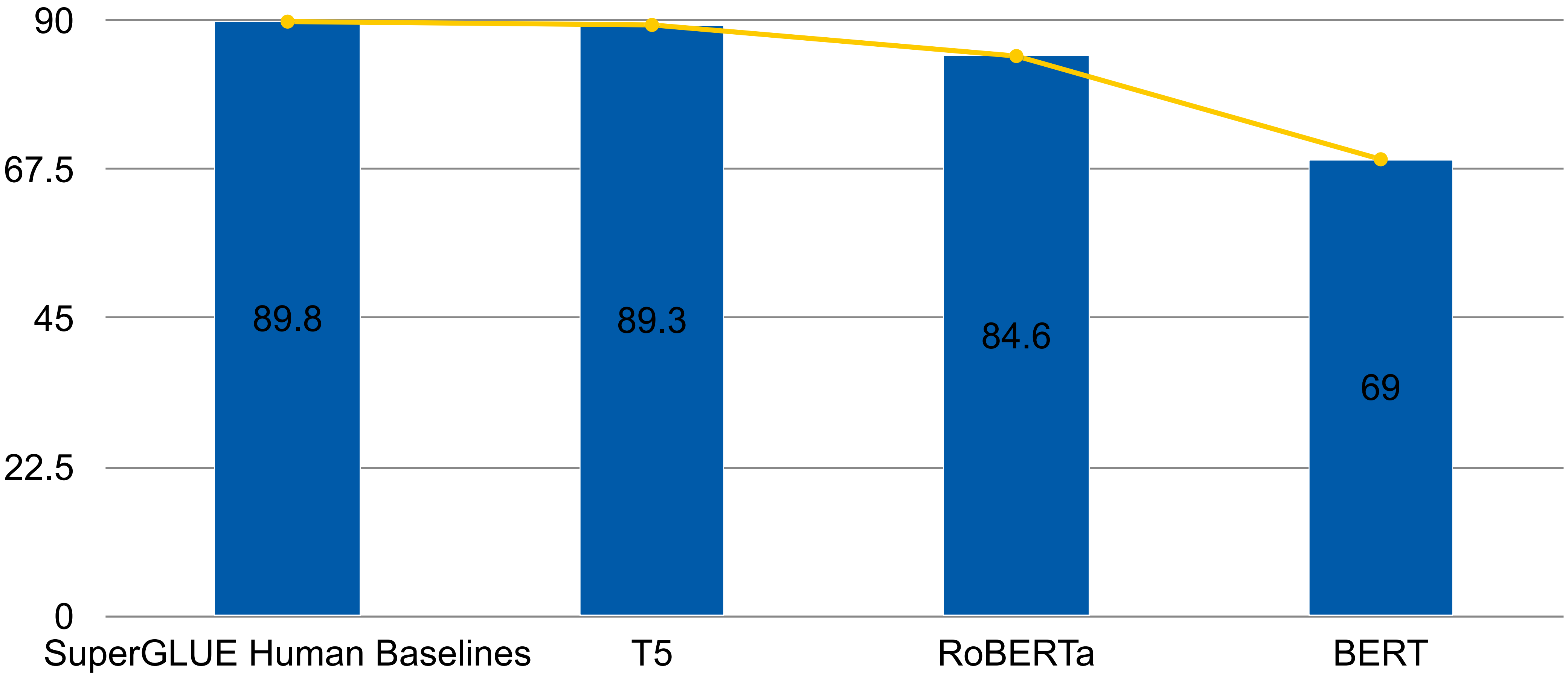
Non-Transformer method vs Transformer method on SemEval-2013

	Unseen answer			Unseen question			Unseen domain		
	acc	M-F1	W-F1	acc	M-F1	W-F1	acc	M-F1	W-F1
Saha et al. (feature encoding method)	71.8	66.6	71.4	61.4	49.1	62.8	63.2	47.9	61.2
Bert-base (State-of-the-art)	**75.0**	**72.0**	**75.8**	**65.3**	**57.5**	**64.8**	**63.8**	**57.9**	**63.4**

Bidirectional Encoder Representations from Transformers (BERT)



State-of-the-Art Transformer Model



SuperGLUE Leaderboard as of Feb 2020. Note: CB evaluation is done via F1 score / accuracy

Motivation

BERT v.s. RoBERTa v.s. T5

	BERT	**RoBERTa**	**T5**
Size(Millions)	Base: 110 Large: 340	Base: 110 Large: 340	Base: 220 Large: 770
Training Time	Base: 8* V100*12days Large: 64 TPU Chips*4days	Large: 1024*V100*1day; 4-5times more than BERT	Not mentioned
Data	16GB BERT data	160GB(16GB BERT data+additional)	750GB C4 data

Target: Better, Smaller, Faster

Better: Improve the acc/W-F1/M-F1 scores on the target dataset SemEval-2013.
Smaller: Reduce the number of parameters.
Faster: Reduce the fine-tuning time.

acc/W-F1/M-F1

Fine-tuning time

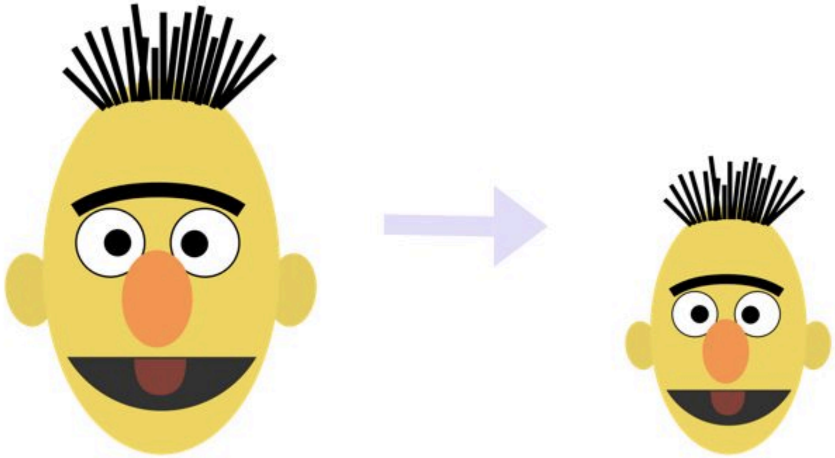
Number of parameters

A Lite BERT (ALBERT)

- **Factorized embedding parameterization:**
The large word embedding matrix is decomposed into two small matrices, thus significantly reducing the number of parameters.

- **Cross-layer parameter sharing:**
Reduce the number of parameters by sharing parameters between layers. This technique prevents the number of parameters from increasing as the depth of the network increases.

- **A Sentence-order prediction (SOP) was proposed to replace NSP.**



ALBERT v.s. BERT

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.5/83.3	80.3/77.3	84.1	91.7	68.3	82.1	17.7x
	large	334M	92.4/85.8	83.9/80.8	85.8	92.2	73.8	85.1	3.8x
	xlarge	1270M	86.3/77.9	73.8/70.5	80.5	87.8	39.7	76.7	1.0
ALBERT	base	12M	89.3/82.1	79.1/76.1	81.9	89.4	63.5	80.1	21.1x
	large	18M	90.9/84.1	82.1/79.0	83.8	90.6	68.4	82.4	6.5x
	xlarge	59M	93.0/86.5	85.9/83.1	85.4	91.9	73.9	85.5	2.4x
	xxlarge	233M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	1.2x

The effect of controlling for training time, BERT-large vs ALBERT-xxlarge configurations.[LCG+19]

Burrows, S., Gurevych, I. & Stein, B. The Eras and Trends of Automatic Short Answer Grading. *Int J Artif Intell Educ* **25,** 60–117 (2015).

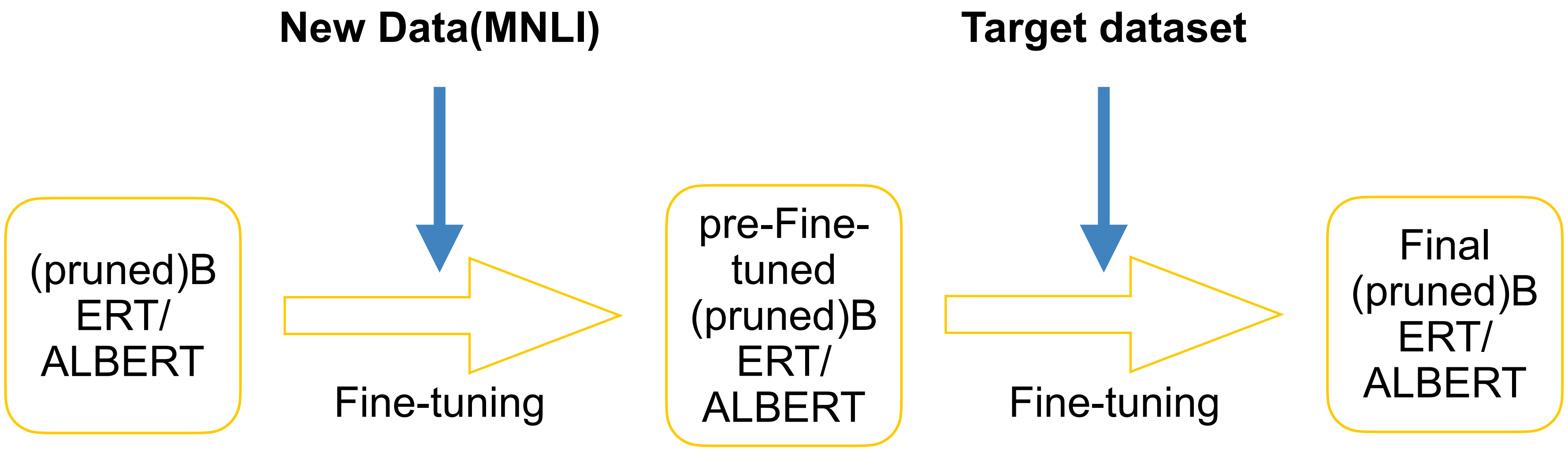
ALBERT v.s. BERT

BERT vs ALBERT on SemEval-2013

	Unseen answer			Unseen question			Unseen domain		
	acc	M-F1	W-F1	acc	M-F1	W-F1	acc	M-F1	W-F1
ALBERT-base	74.2	68.6	74.5	62.6	48.9	63.7	**66.5**	**59.0**	**67.4**
Bert-base (State-of-the-art)	**75.0**	**72.0**	**75.8**	**65.3**	**57.5**	**64.8**	63.8	57.9	63.4

Better: Using More Data

MNLI(Multi-Genre Natural Language Inference): 3-way Classification Dataset.



Better: Result of Using More Data

BERT vs ALBERT on SemEval-2013

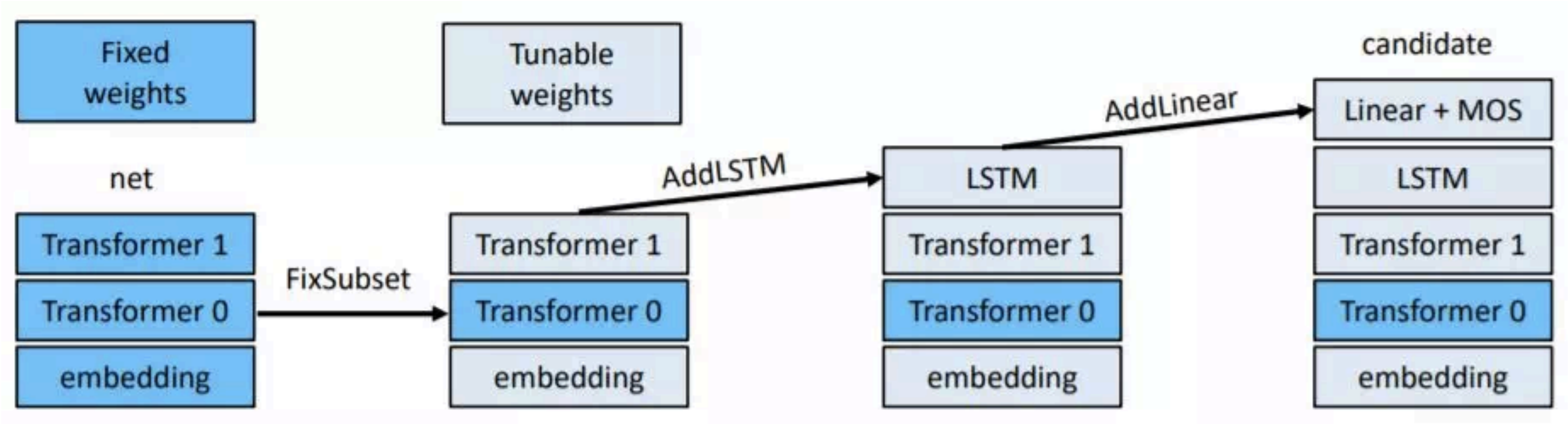
	Unseen answer			Unseen question			Unseen domain		
	acc	M-F1	W-F1	acc	M-F1	W-F1	acc	M-F1	W-F1
ALBERT-base	74.2	68.6	74.5	62.6	48.9	63.7	66.5	59.0	**67.4**
Bert-base (State-of-the-art)	75.0	72.0	**75.8**	65.3	57.5	64.8	63.8	57.9	63.4
Albert-pruned-v1+mnli(2e-5_90.3)	**75.5**	**72.8**	75.6	**68.7**	**58.8**	**69.2**	**66.7**	**60.7**	67.2

Next Step

Next Step

Better: Add LSTM + Mixture of Softmaxes(MoS)

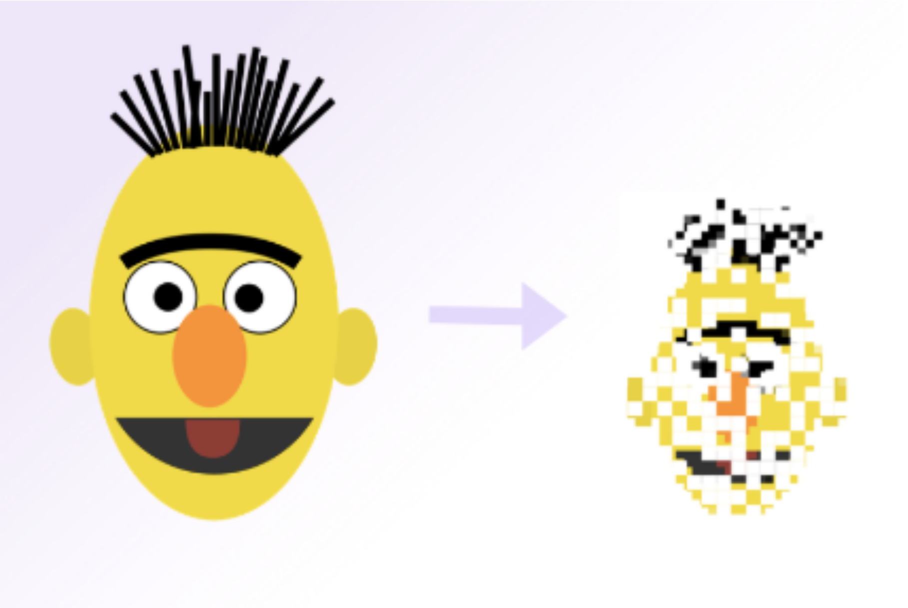
Mixture of Softmaxes(MoS) : Address the problem that the standard Softmax-based language model for word embeddings not good at model natural language.



Structure of Transformer + LSTM + MOS[YDSC17]

[YDSC17] Yang Z, Dai Z, Salakhutdinov R, et al. Breaking the softmax bottleneck: A high-rank RNN language model[J]. arXiv preprint arXiv:1711.03953, 2017.

Smaller & Faster: Head(Layer) pruning



Layer \ Head	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.03	0.07	0.05	-0.06	0.03	**-0.53**	0.09	**-0.33**	0.06	0.03	0.11	0.04	0.01	-0.04	0.04	0.00
2	0.01	0.04	0.10	**0.20**	0.06	0.03	0.00	0.09	0.10	0.04	**0.15**	0.03	0.05	0.04	0.14	0.04
3	0.05	-0.01	0.08	0.09	0.11	0.02	0.03	0.03	-0.00	0.13	0.09	0.09	-0.11	**0.24**	0.07	-0.04
4	-0.02	0.03	0.13	0.06	-0.05	0.13	0.14	0.05	0.02	0.14	0.05	0.06	0.03	-0.06	-0.10	-0.06
5	**-0.31**	-0.11	-0.04	0.12	0.10	0.02	0.09	0.08	0.04	**0.21**	-0.02	0.02	-0.03	-0.04	0.07	-0.02
6	0.06	0.07	**-0.31**	0.15	-0.19	0.15	0.11	0.05	0.01	-0.08	0.06	0.01	0.01	0.02	0.07	0.05

Difference in BLEU score for each head of the encoder's self attention mechanism. Underlined numbers indicate that the change is statistically significant with $p < 0.01$. The base BLEU score is 36.05.

Questions?

Sources

[BGS15] Burrows, S., Gurevych, I. & Stein, B. The Eras and Trends of Automatic Short Answer Grading. Int J Artif Intell Educ 25, 60–117 (2015).

[YDSC17] Yang Z, Dai Z, Salakhutdinov R, et al. Breaking the softmax bottleneck: A high-rank RNN language model[J]. arXiv preprint arXiv:1711.03953, 2017.

[LOG+ 19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre- training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[LCG+19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural informa- tion processing systems, pages 5998–6008, 2017.