Multi-CLS BERT: An Efficient Alternative to Traditional Ensembling

Haw-Shiuan Chang*   Ruei-Yao Sun*   Kathryn Ricci*   Andrew McCallum

amazon | science

UMassAmherst | Manning College of Information & Computer Sciences

BERT Classifier

- Problem
  - A small text classification task
  - Unstable BERT’s performance
- What About?
  - Ensembling
- But …
  - Costly

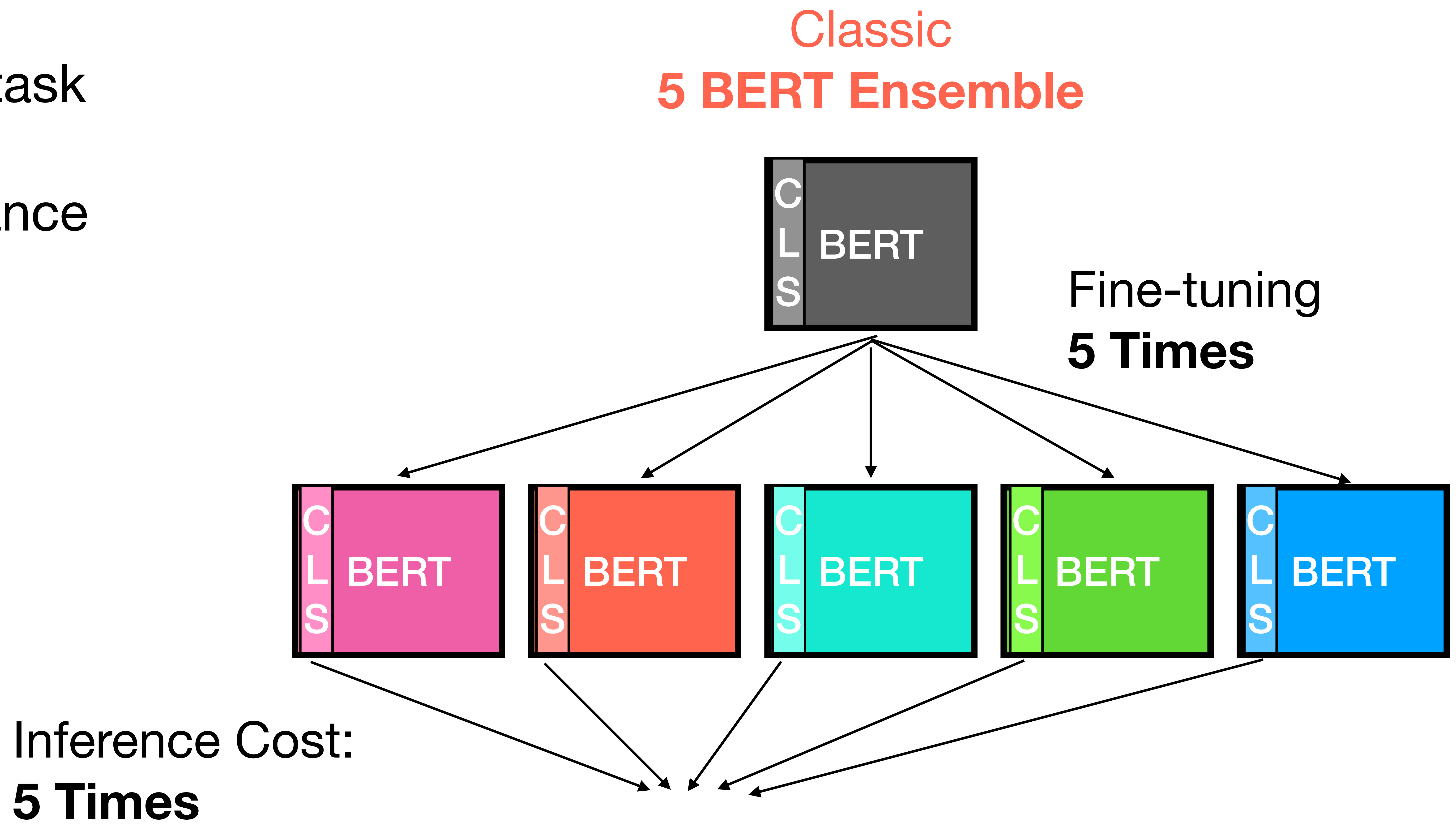
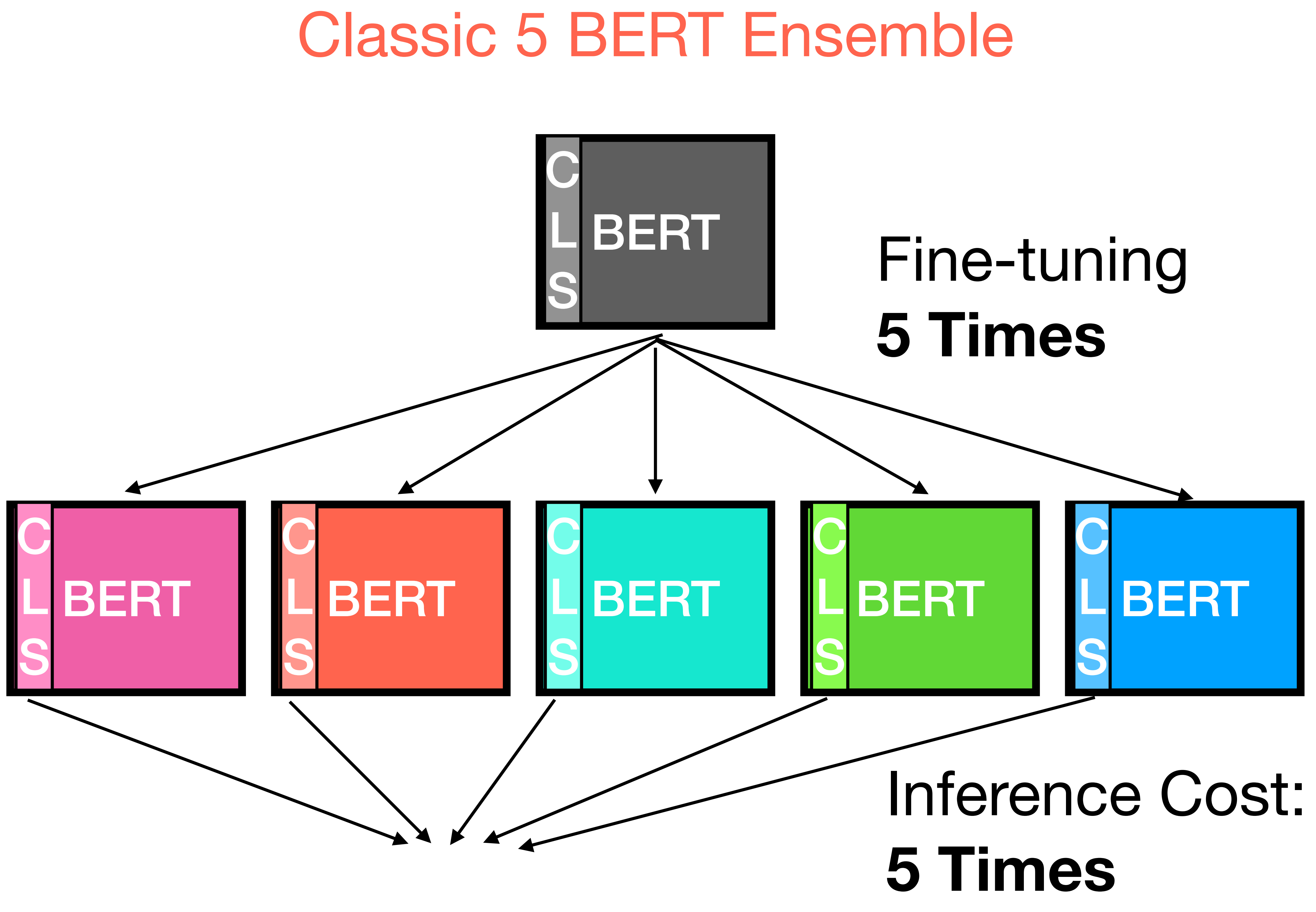


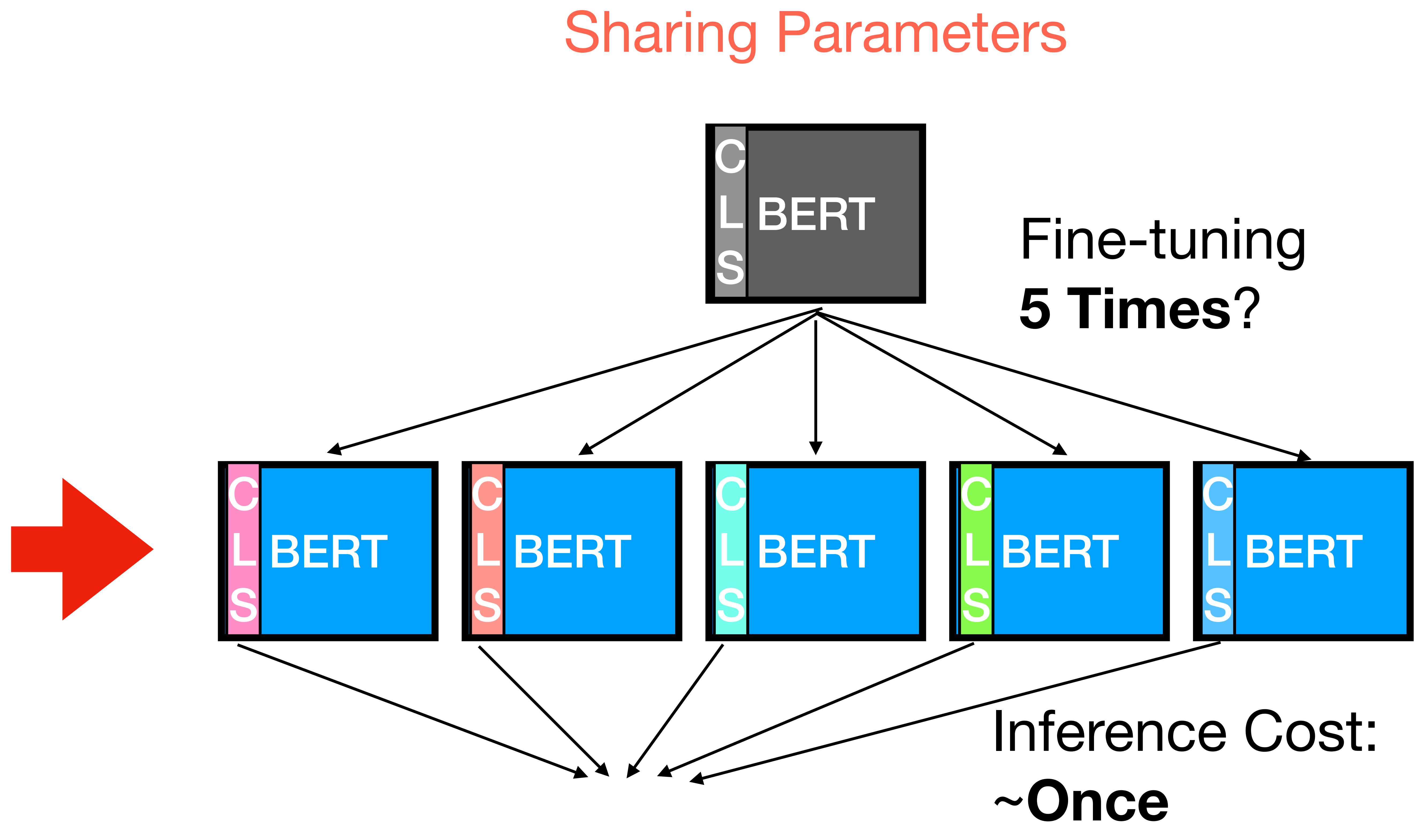
Figure comes from: https://www.flickr.com/photos/gforsythe/50639435271/

Can We Make Ensembling Almost as Efficient as the Single Model?

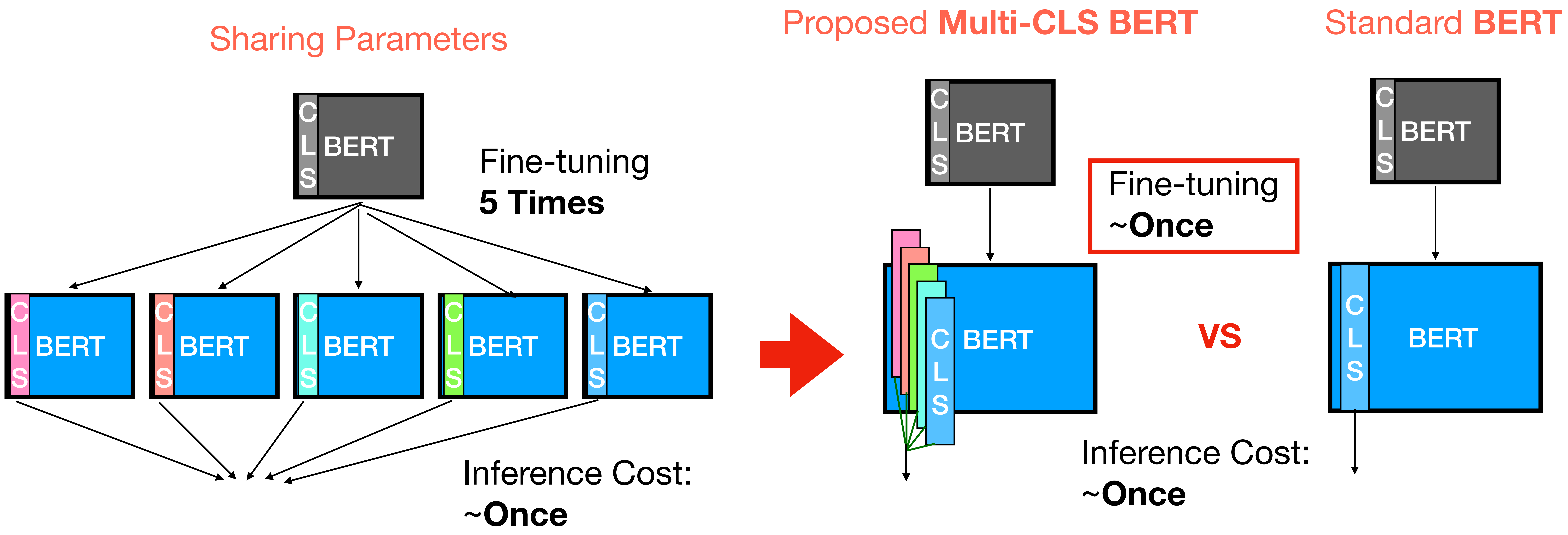
Yes !

Sharing the BERT Encoder



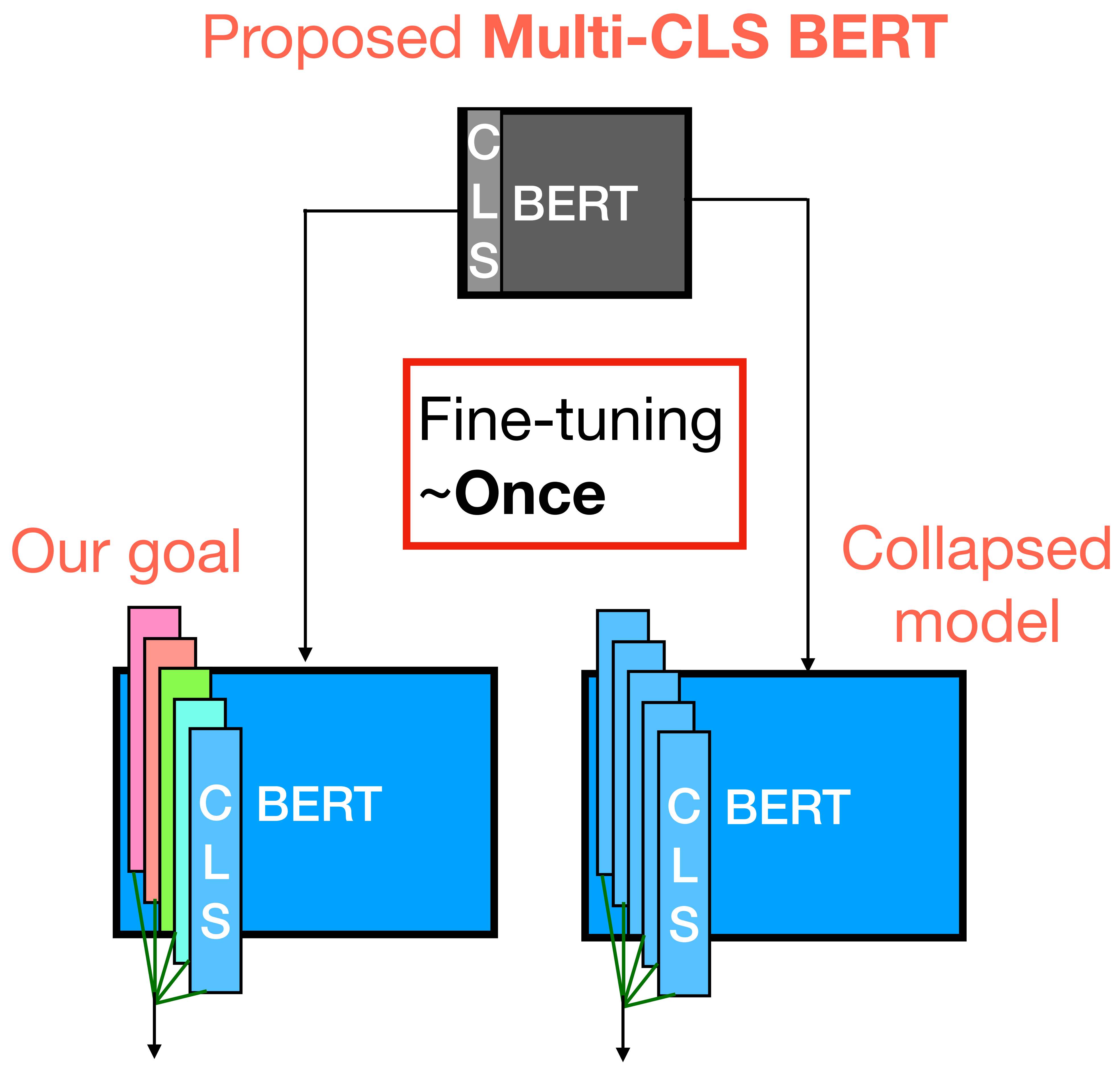


Fine-tuning only Once!



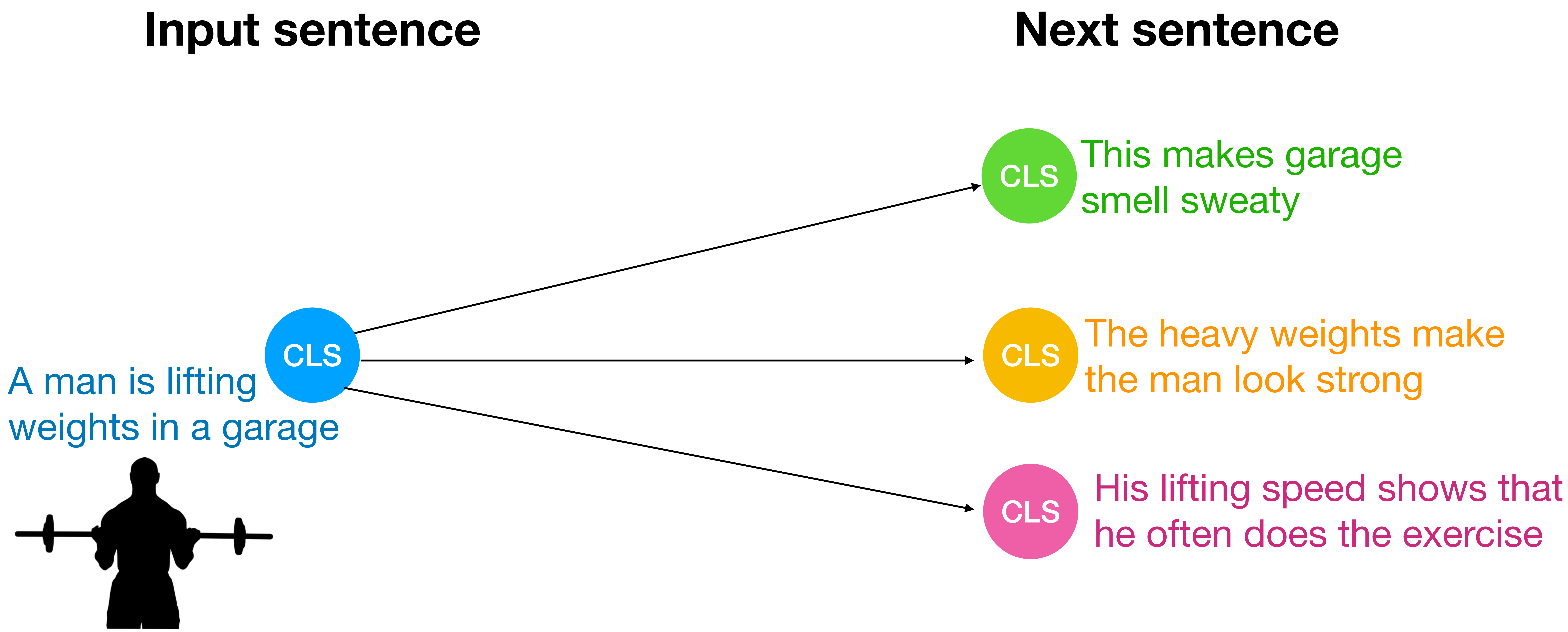
Goal and Challenge

- Our goal
  - Aggregate the contextualized word embeddings differently
- Challenge
  - CLS embeddings are often identical
    - After seeing the same training samples

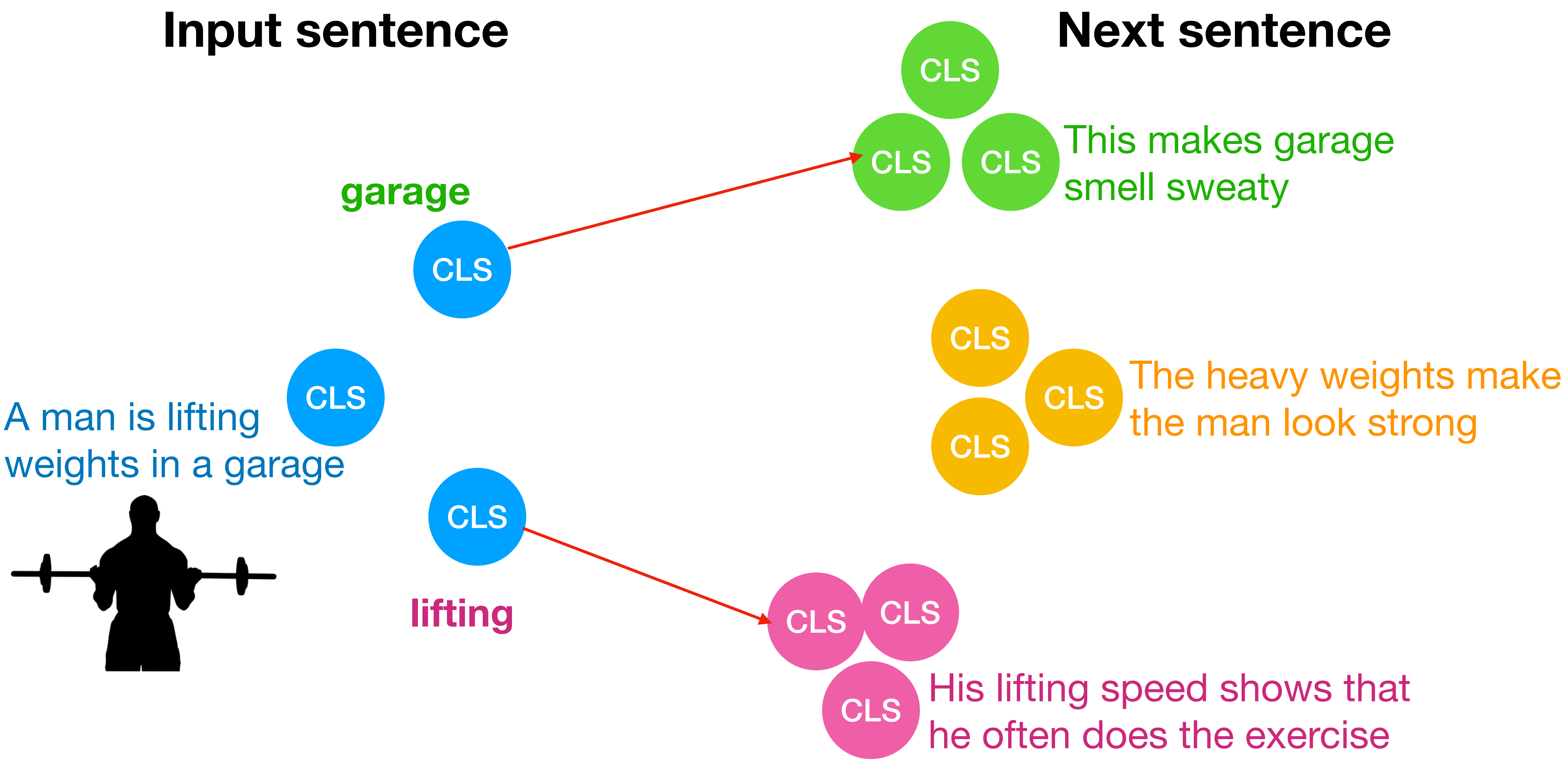


Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. 2019. Deep ensembles: A loss landscape perspective. *ArXiv preprint*, abs/1912.02757

Pretraining Diversification

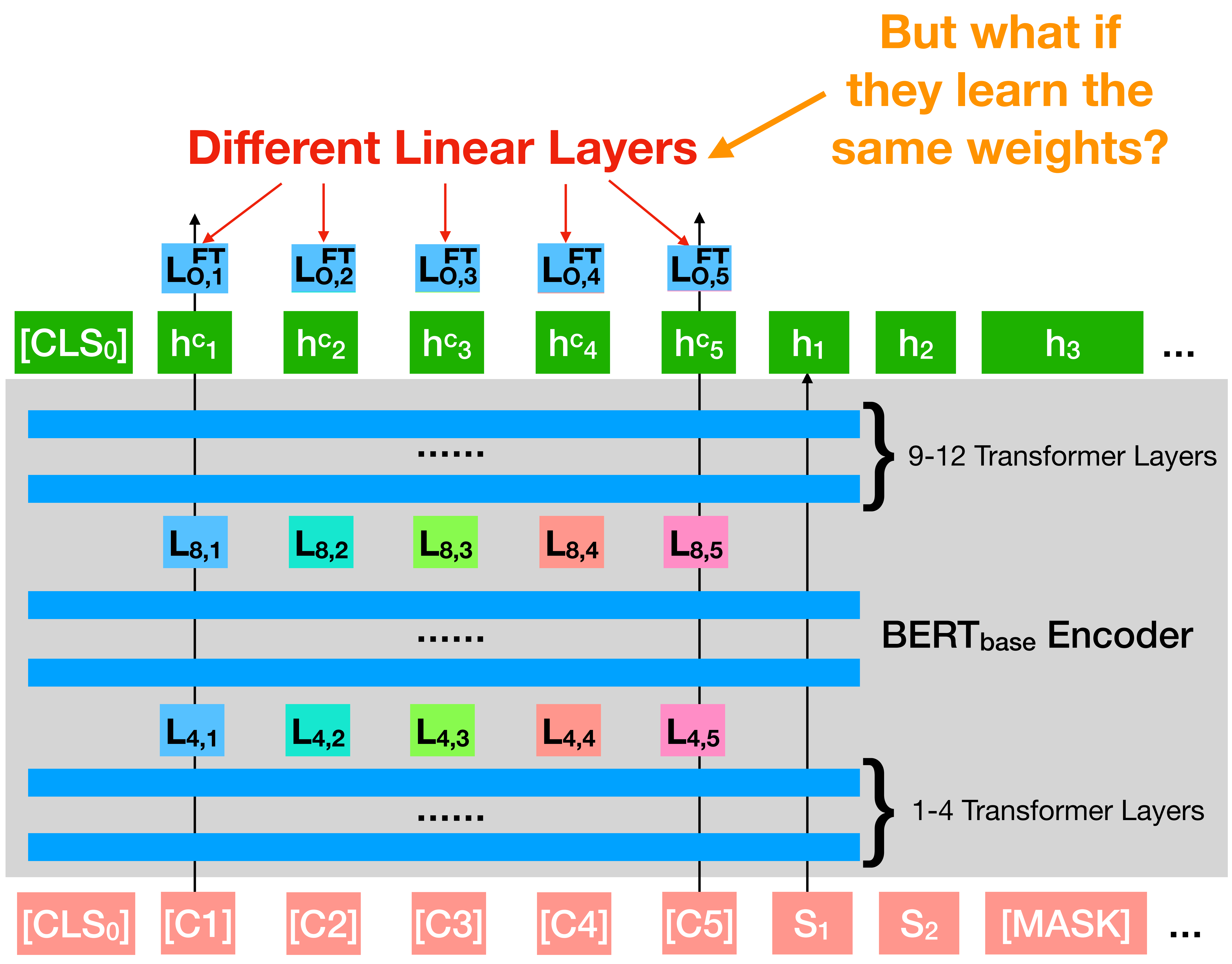


Pretraining Diversification



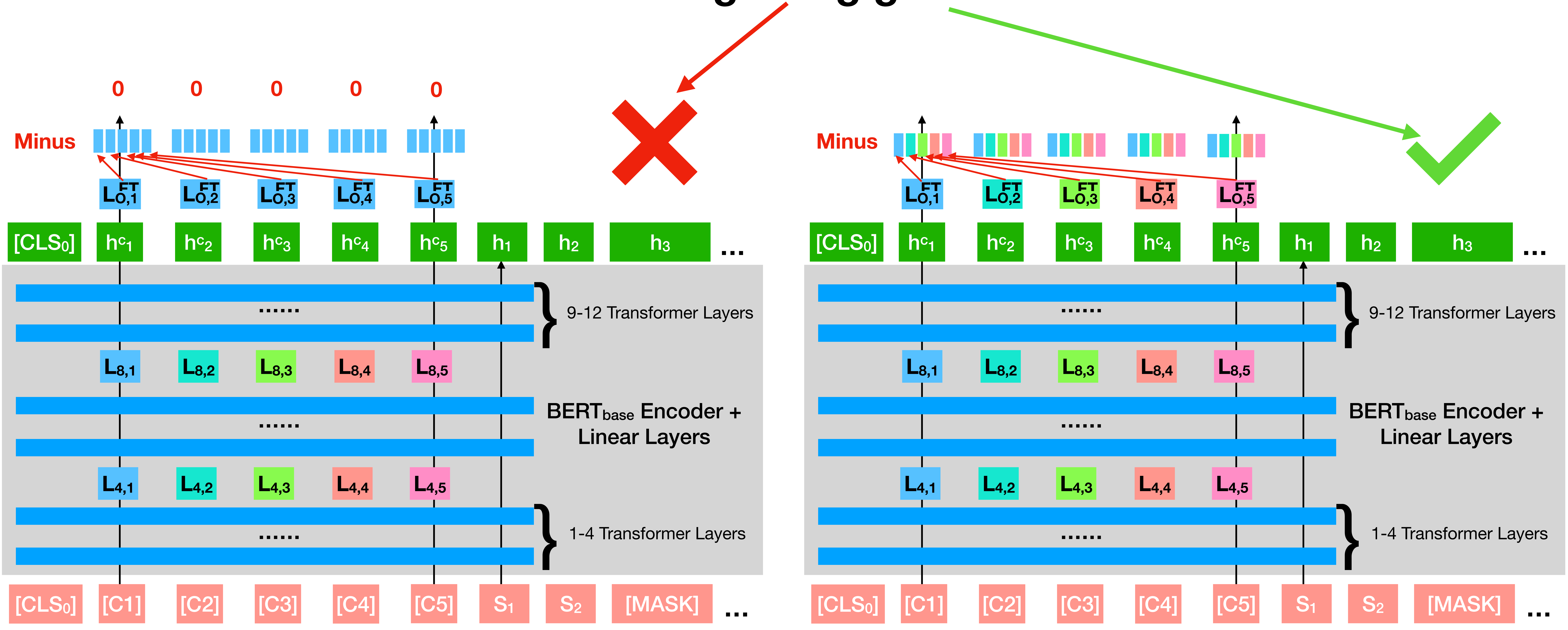
Architecture Diversification

- Insert different linear layers for different CLS tokens
  - The differences of CLS could be stored in the linear weights
  - The parameter increase is relatively small



Fine-tuning Diversification

After fine-tuning using gradient descent



Experiment Settings

- Our main baseline MTL
    - By optimizing the pretraining and fine-tuning methods of a state-of-the-art BERT model (Aroca-Ouellette and Rudzicz, 2020)
- Repeat training 16 times
    - Pretraining 4 times and fine-tuning 4 times
    - Many previous work shows that random seeds are important in GLUE and SuperGLUE

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2021. The multiberts: BERT reproductions for robustness analysis. *ArXiv preprint*, abs/2106.16163
Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv preprint*, abs/2002.06305.
Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021a. Revisiting few- sample BERT fine-tuning. In *9th International Con- ference on Learning Representations, ICLR 2021, Vir- tual Event, Austria, May 3-7, 2021*.
Marius Mosbach, Maksym Andriushchenko, and Diet- rich Klakow. 2021. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. In *9th International Conference on Learn- ing Representations, ICLR 2021, Virtual Event, Aus- tria, May 3-7, 2021*.

Natural Language Understanding

BERT Base could be better than BERT Large

Configuration ↓	Model Name ↓	Model Size ↓	GLUE 100	GLUE 1k	GLUE Full	SuperGLUE 100*	SuperGLUE 1k*	SuperGLUE Full
BERT Base	Pretrained	109.5M	55.71 ± 0.62	71.67 ± 0.15	82.05 ± 0.08	57.18 ± 0.43	61.55 ± 0.37	65.04 ± 0.36
	MTL	109.5M	59.29 ± 0.27	73.26 ± 0.13	83.30† ± 0.07	57.50 ± 0.41	62.94 ± 0.36	66.33 ± 0.33
	Our[illegible]	111.3M	57.84 ± 0.32	[illegible]	[illegible]3.40 ± 0.07	57.31 ± 0.35	63.35 ± 0.18	66.29 ± 0.18
	Ours (K=5, $\lambda = 0$)	118.4M	61.54 ± 0.32	**74.14** ± 0.12	83.41 ± 0.07	58.29 ± 0.33	63.71 ± 0.26	**66.80** ± 0.25
	Ours (K=5, $\lambda = 0.1$)	118.4M	**61.80** ± 0.35	74.10 ± 0.13	**83.47** ± 0.05	58.20 ± 0.31	63.61 ± 0.27	66.74 ± 0.26
	Ours (K=5, $\lambda = 0.5$)	118.4M	60.49 ± 0.35	74.02 ± [illegible]	**83.47** ± 0.08	**58.41** ± 0.38	**63.78** ± 0.25	**66.80** ± 0.24
	Ours [illegible]	118.4M	59.86 ± 0.34	7[illegible] ± 0.14	83.43 ± 0.07	57.84 ± 0.40	63.56 ± 0.22	66.39 ± 0.22
BERT Large	MTL	335.2M	61.39 ± 0.37	75.30 ± 0.27	84.13 ± 0.11	59.03 ± 0.54	65.21 ± 0.38	69.16 ± 0.37
	Ours (K=1)	338.3M	59.19 ± 0.43	75.35 ± 0.21	84.59 ± 0.07	57.35 ± 0.42	64.67 ± 0.43	69.24 ± 0.41
	Ours (K=5, $\lambda = 0$)	350.9M	63.19 ± 0.49	75.73 ± 0.26	84.51 ± 0.05	59.46 ± 0.44	65.43 ± 0.38	69.56 ± 0.31
	Ours (K=5, $\lambda = 0.1$)	350.9M	**64.24** ± 0.40	**76.27** ± 0.12	**84.61** ± 0.08	**59.88** ± 0.43	65.58 ± 0.26	**70.03** ± 0.25
	Ours (K=5, $\lambda = 0.5$)	350.9M	63.02 ± 0.42	75.95 ± 0.10	84.49 ± 0.08	59.42 ± 0.34	**65.84** ± 0.25	69.79 ± 0.25
	Ours (K=5, $\lambda = 1$)	350.9M	62.07 ± 0.45	75.85 ± 0.17	**84.61** ± 0.07	58.74 ± 0.50	65.00 ± 0.29	69.04 ± 0.27

+ 8.9M

+ 2.51

+ 225.7 M

+ 2.1

Aroca-Ouellette, Stéphane, and Frank Rudzicz. "On Losses for Modern Language Models." In *EMNLP*. 2020.

Natural Language Understanding

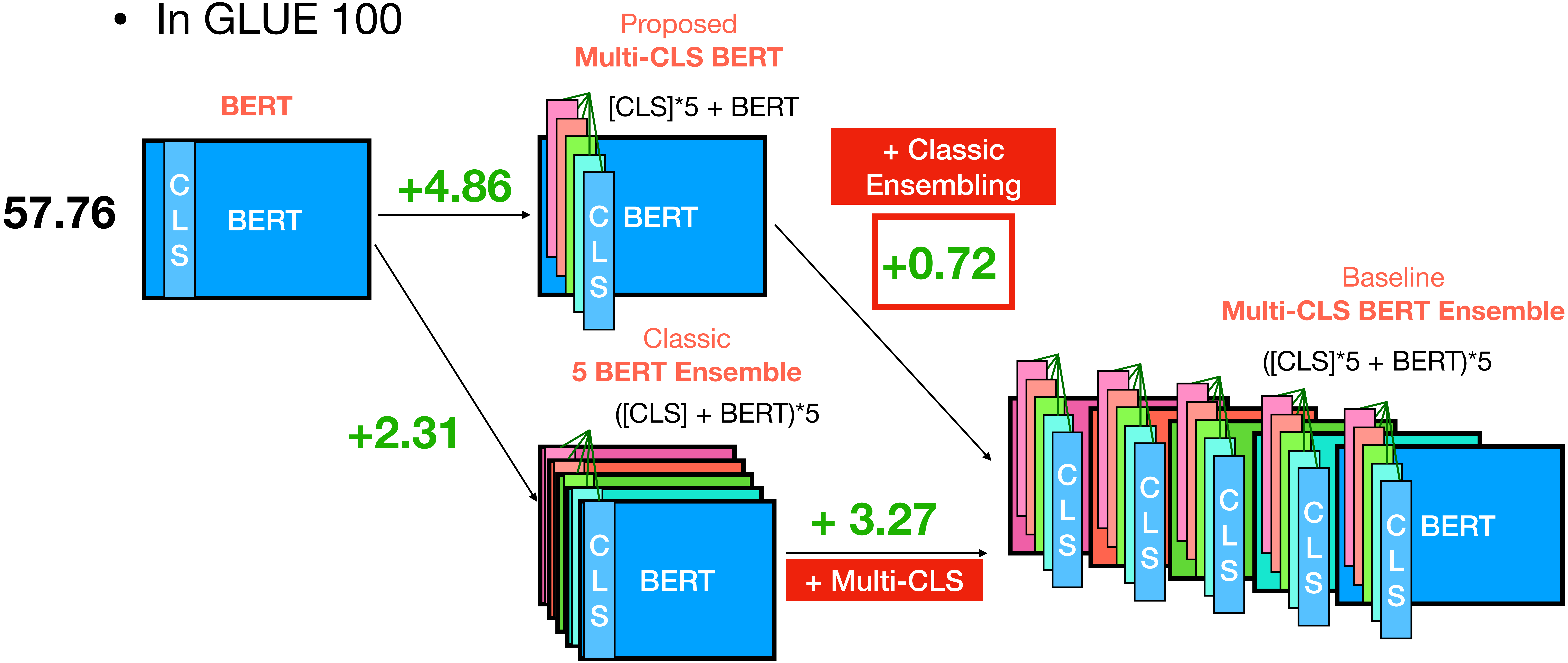
The improvement of BERT Large is usually larger than the improvement of BERT Base

Configuration ↓	Model Name ↓	Model Size ↓	GLUE 100	GLUE 1k	GLUE Full	SuperGLUE 100*	SuperGLUE 1k*	SuperGLUE Full
BERT Base	Pretrained	109.5M	55.71 ± 0.62	71.67 ± 0.15	82.05 ± 0.08	57.18 ± 0.43	61.55 ± 0.37	65.04 ± 0.36
	MTL	109.5M	59.29 ± 0.27	73.26 ± 0.13	83.30† ± 0.07	57.50 ± 0.41	62.94 ± 0.36	66.33 ± 0.33
	Ours (K=1)	111.3M	57.84 ± [illegible]	73.28 ± [illegible]	83.40 ± [illegible]	57.31 ± [illegible]	63.35 ± [illegible]	66.29 ± [illegible]
			+ 2.51	+ 0.84	+ 0.17	+ 0.70	+ 0.67	+ 0.41
	Ours (K=5, $\lambda = 0$)	118.4M	61.54 ± 0.32	**74.14** ± 0.12	83.41 ± 0.07	58.29 ± 0.33	63.71 ± 0.26	**66.80** ± 0.25
	Ours (K=5, $\lambda = 0.1$)	118.4M	**61.80** ± 0.35	74.10 ± 0.13	**83.47** ± 0.05	58.20 ± 0.31	63.61 ± 0.27	66.74 ± 0.26
	Ours (K=5, $\lambda = 0.5$)	118.4M	60.49 ± 0.35	74.02 ± 0.12	**83.47** ± 0.08	**58.41** ± 0.38	**63.78** ± 0.25	**66.80** ± 0.24
	Ours (K=5, $\lambda = 1$)	118.4M	59.86 ± 0.34	73.75 ± 0.14	83.43 ± 0.07	57.84 ± 0.40	63.56 ± 0.22	66.39 ± 0.22
BERT Large	MTL	335.2M	61.39 ± 0.37	75.30 ± 0.27	84.13 ± 0.11	59.03 ± 0.54	65.21 ± 0.38	69.16 ± 0.37
	Ours (K=1)	338.3M	59.19 ± [illegible]	75.35 ± [illegible]	84.59 ± [illegible]	57.35 ± [illegible]	64.67 ± [illegible]	69.24 ± [illegible]
			+ 2.85	+ 0.97	+ 0.48	+ 0.85	+ 0.37	+ 0.87
	Ours (K=5, $\lambda = 0$)	350.9M	63.19 ± 0.49	75.73 ± 0.26	84.51 ± 0.05	59.46 ± 0.44	65.43 ± 0.38	69.56 ± 0.31
	Ours (K=5, $\lambda = 0.1$)	350.9M	**64.24** ± 0.40	**76.27** ± 0.12	**84.61** ± 0.08	**59.88** ± 0.43	65.58 ± 0.26	**70.03** ± 0.25
	Ours (K=5, $\lambda = 0.5$)	350.9M	63.02 ± 0.42	75.95 ± 0.10	84.49 ± 0.08	59.42 ± 0.34	**65.84** ± 0.25	69.79 ± 0.25
	Ours (K=5, $\lambda = 1$)	350.9M	62.07 ± 0.45	75.85 ± 0.17	**84.61** ± 0.07	58.74 ± 0.50	65.00 ± 0.29	69.04 ± 0.27

Aroca-Ouellette, Stéphane, and Frank Rudzicz. "On Losses for Modern Language Models." In *EMNLP*. 2020.

Multi-CLS vs Ensembling

- In GLUE 100

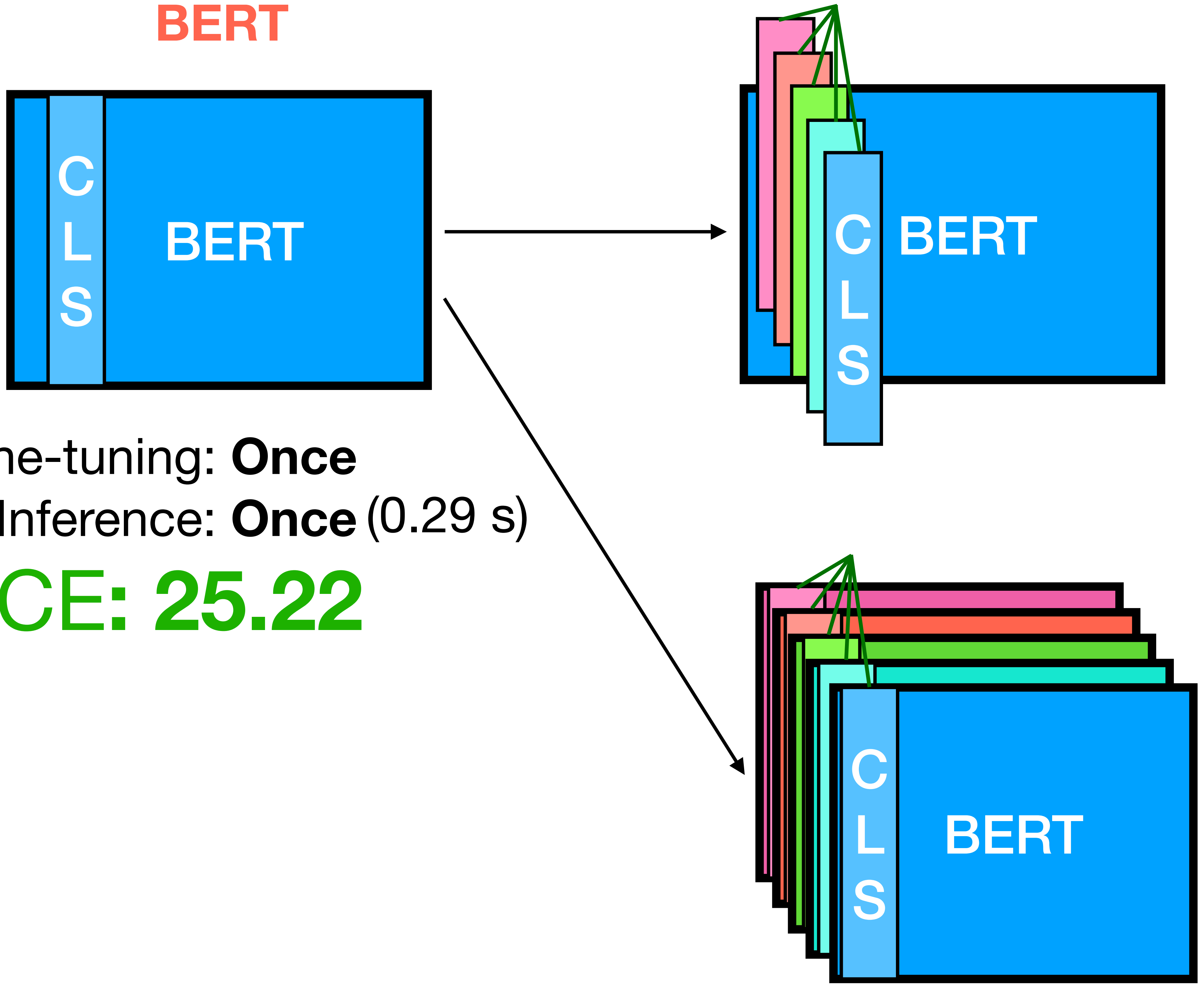


Multi-CLS vs Ensembling

- In GLUE 100, Comparison of expected calibration errors (ECE).

BERT

57.76



Fine-tuning: **Once**
Inference: **Once** (0.29 s)
ECE**: 25.22**

Proposed
Multi-CLS BERT

Fine-tuning: **Once**
Inference: **Once** (0.31 s)
ECE**: 15.46**

Classic
5 BERT Ensemble

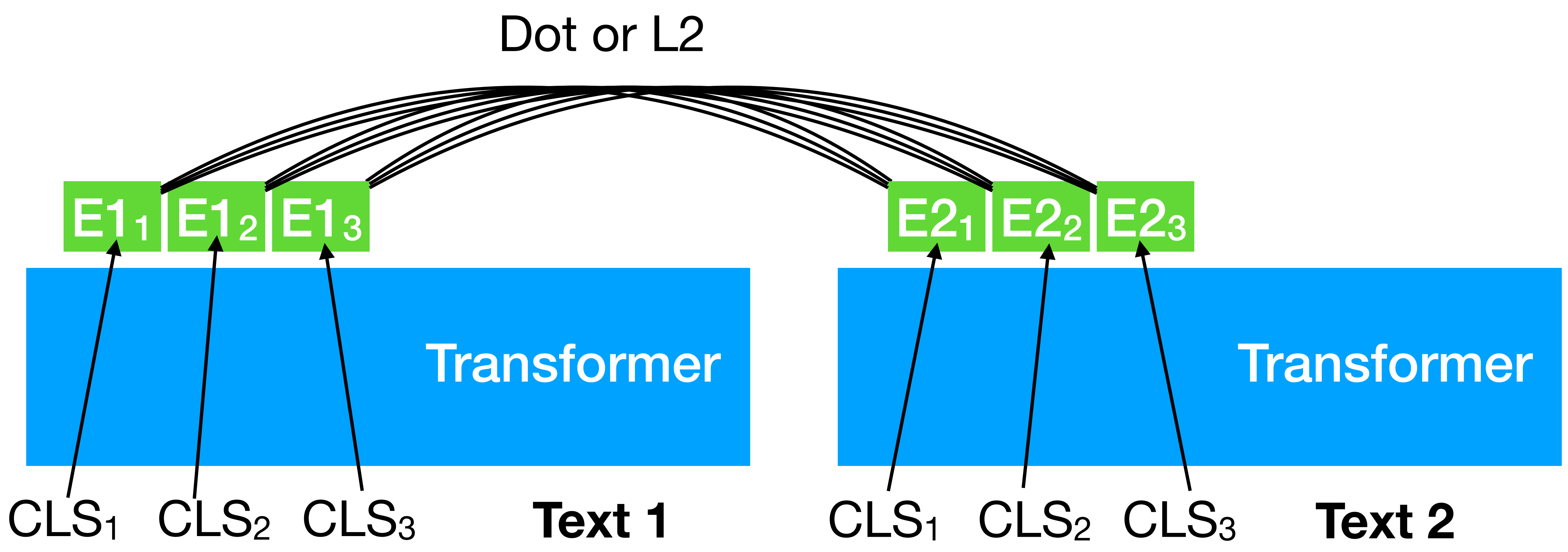
Fine-tuning: **5 Times**
Inference: **5 Times** (1.46 s)
ECE**: 13.85**

Conclusion

- Ensembling BERT almost without extra cost is achievable
- We need some tricks to diversify the multiple CLS hidden states
- Compared to standard ensembling
  - Improve more when the training dataset is small
  - Improve less otherwise

Our Other Work using Multiple Embeddings

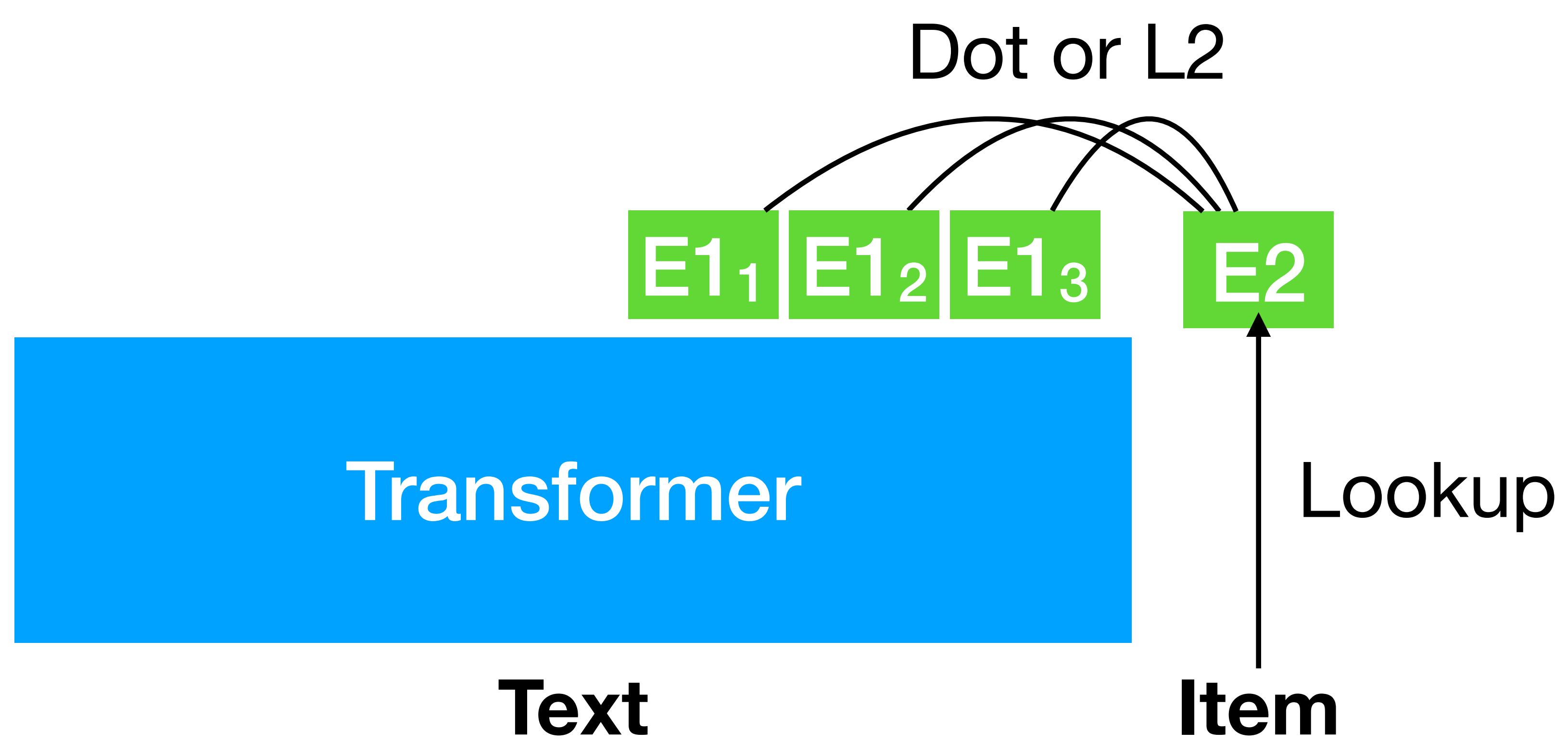
BERT-like LM encoder for NLU



More Accurate and Calibrated

NLI QA IR Sent sim ……

GPT-like LM decoder for NLG



More Factual and Less Repetition

Text Completion Summarization

H.-S. Chang* , Z. Yao*, A. Gon, H. Yu, and A. McCallum, "Revisiting the Architectures like Pointer Networks to Efficiently Improve the Next Word Distribution, Summarization Factuality, and Beyond" ACL Findings 2023

H.-S. Chang, and A. McCallum, "Softmax Bottleneck Makes Language Models Unable to Represent Multi-mode Word Distributions," ACL 2022

H.-S. Chang, "Modeling the Multi-mode Distribution in Self-Supervised Language Models, "PhD Thesis 2022