Fine-tuning minBERT for Various Downstream Tasks

Longling Tian,[1] Siqi Wang[2]

[1]Institute of Computational and Mathematical Engineering, Stanford University
[2]Department of Statistics, Stanford University

Stanford
CS224N

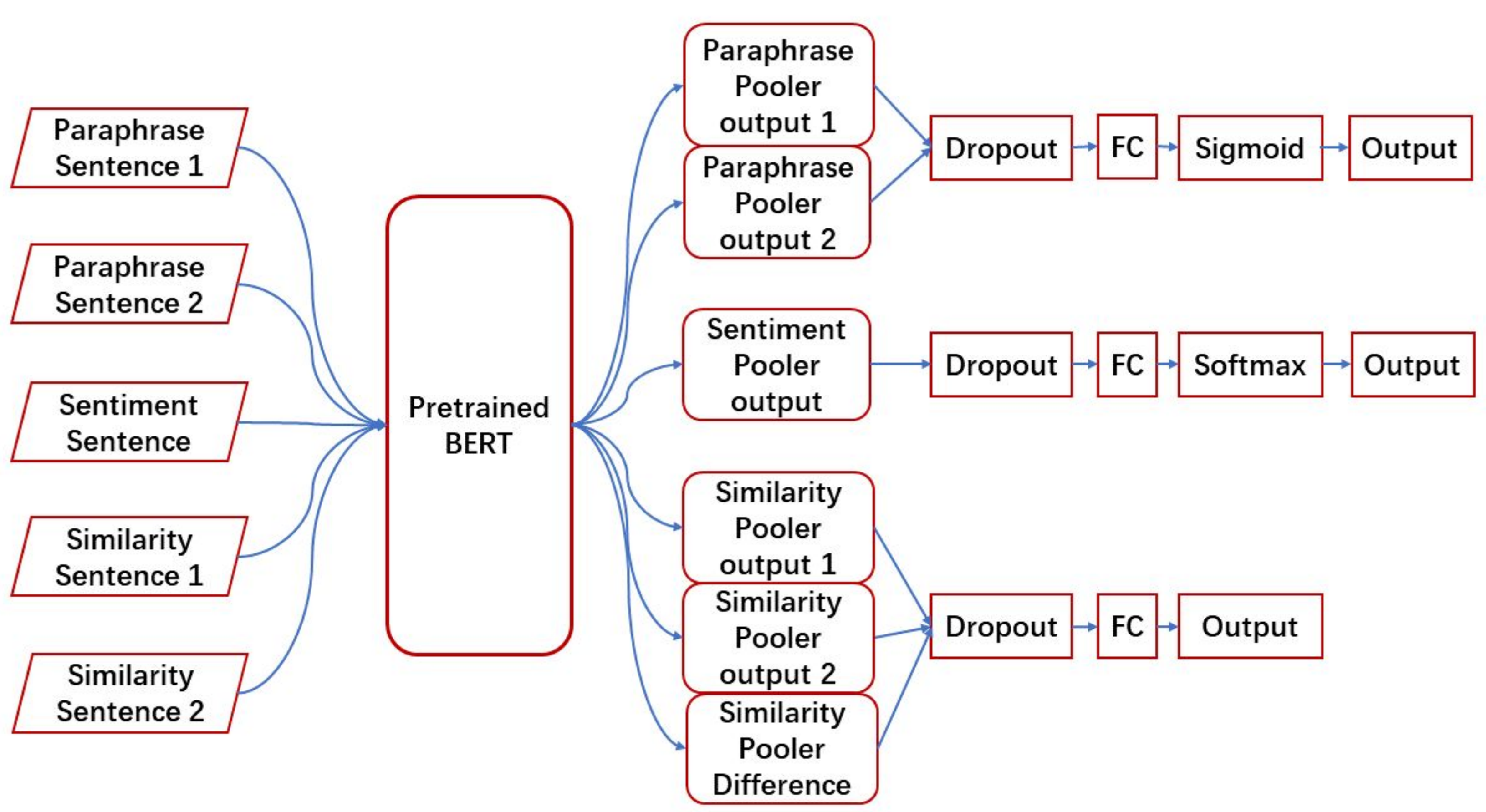
Problem

- It is important for BERT to perform well across multiple tasks:
  - **Efficiency**: save time and computational resources by using the same pre-trained embeddings for various tasks
  - **Generalization**: quickly adapt to new tasks and domains by fine-tuning the model on task-specific datasets
- However, *how* to make BERT perform well is a challenging topic:
  - Different tasks require different architectures & hyperparameters
  - Conflicting objectives

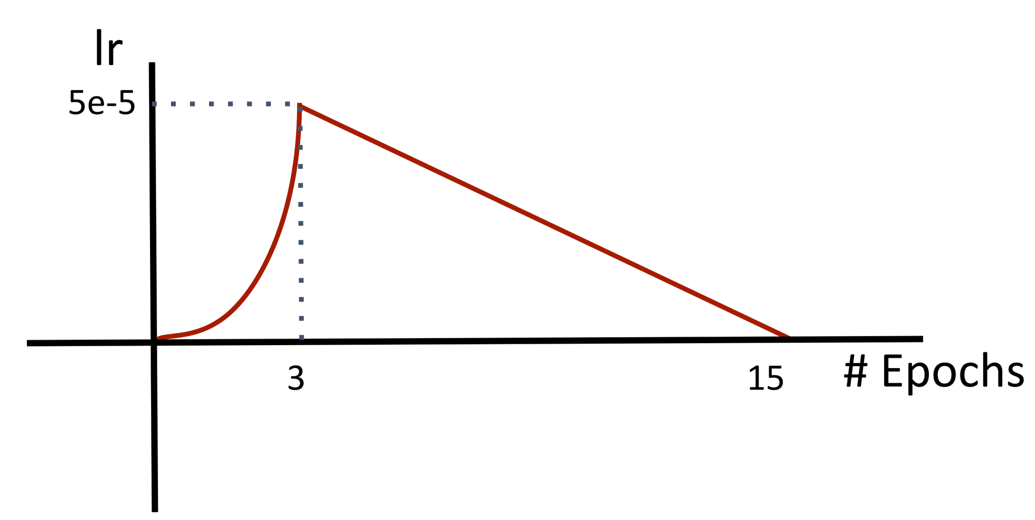
CogBlog.https://web.colby.edu/cogblog/2020/12/05/why-you-should-stop-multitaskingright-now/

Experiments

- One BERT does three tasks

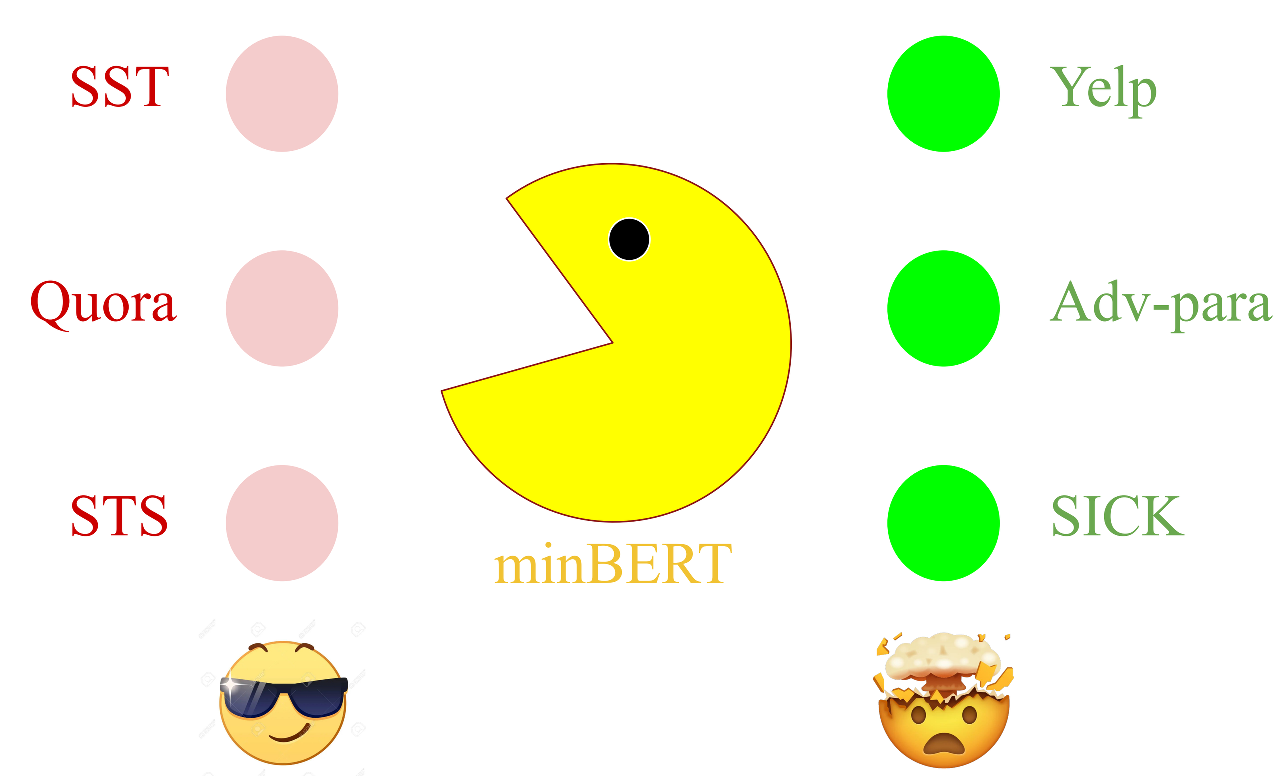


- Learning rate warmup & decay

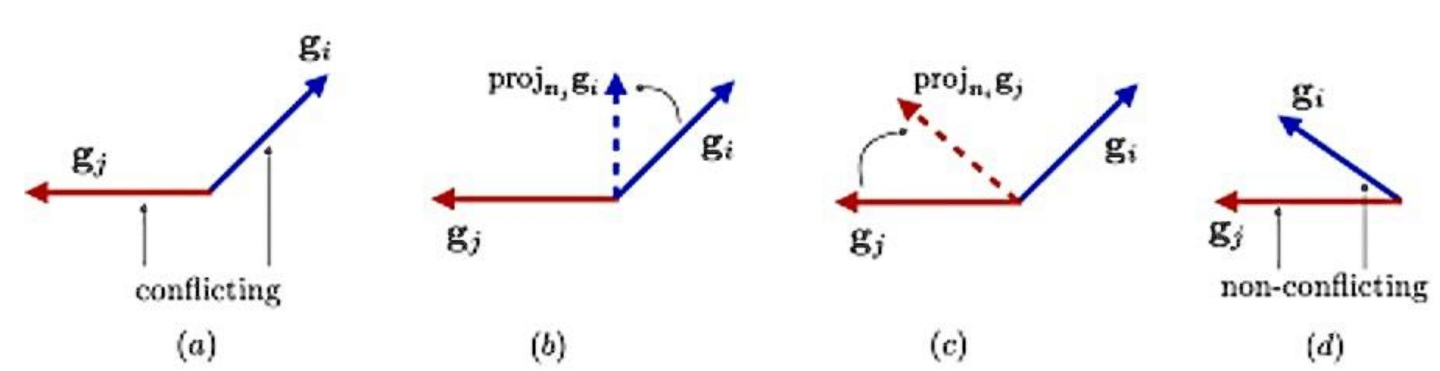


Liu et.al, ICLR 2020

- Additional Datasets



- Gradient Surgery to remove conflicting gradients



Yu et.al, NeurIPS 2020

Result & Analysis

Finetuning Technique	Overall	Sentiment Classification	Paraphrase	Similarity
minBERT	/	0.528	/	/
Baseline	0.511	0.479	0.781	0.274
Cosine Similarity	0.546	0.503	0.722	0.412
Diff	0.615	**0.520**	**0.795**	0.529
Diff + 2-dense-layer	0.605	0.497	0.787	0.531
Diff + additional data	0.582	0.462	0.782	0.502
Grad-surg + diff	**0.648**	0.513	0.776	**0.655**
Grad-surg + diff + add-data	0.638	0.498	0.777	0.639
Test set	0.640	0.510	0.775	0.634

🥇Congrats to **grad-surg + diff layer**!

🙏The chosen one made its way to the test set, got **64%** overall accuracy

🤕Cos-sim and additional data failed

😲Baseline model already did well on paraphrase and sentiment classification task

📈 **25%** performance boost after adding in difference - need to explicitly tell the model what's the goal

💪 Another **12%** increase in similarity task after using gradient surgery

- Conflicting issue is severe in similarity task
- ~1% drop in the other two tasks😦: grad-surg may take off useful information

🤔The way from **27.4%** to **65.5%**: hyperparameter tuning only contributed to **4%**

🤫Double dense layer performed worse: **easy** model sometimes does it all

Future Steps

- Perform gradient surgery on similarity task only
- Modify loss function to penalize more on similarity task
- Grab more data, more **similar** data
- Use data more on pretrain than fine-tune

References

1. Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn.Gradient surgery for multi-task learning, 2020
2. Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. arXiv preprint, arXiv:1908.03265, 2019.