My PAL BERT: Using Projected Attention Layers and Additional Fine-Tuning Strategies to Improve BERT's Performance on Downstream Tasks

Colin Sullivan, Abhishek Kumar

Stanford CS224N Default Project, Department of Computer Science

Overview

Problem:

- How can we fine-tune a pre-trained language model on multiple downstream tasks?

Approach:

- We are focused on improving BERT's performance on three downstream tasks: sentiment analysis, paraphrase detection (para), and semantic text similarity (STS)
- We implement seven additions to BERT which are listed in the timeline on the right

Findings:

- Our best model, which we call BP1, consists of PALs, Annealed Sampling, and Unsupervised SimCSE

Overall Findings

Architecture	Additions	Score			
		SST	Para	STS	Overall
Baseline BERT	None	.361	.763	.224	.579
BERT + PALs	None	.424	**.863**	.840	.736
	1. Annealed Sampling (AS)	.490	.861	.866	.761
	BP1 Model: 1. AS 2. SimCSE	**.516**	.862	**.866**	**.770**
	1. AS 2. SimCSE 3. Add. Datasets	.446	.859	.865	.746
	1. AS 2. SimCSE 3. Add. Datasets 4. LR warmup/decay	.466	.860	.860	.752
	1. Annealed Sampling 2. SimCSE 3. Add. Datasets 4. LR warmup/decay 5. RL Layer	.442	.861	.853	.743

References

[1] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR, 2019.

[2] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer, 2019.

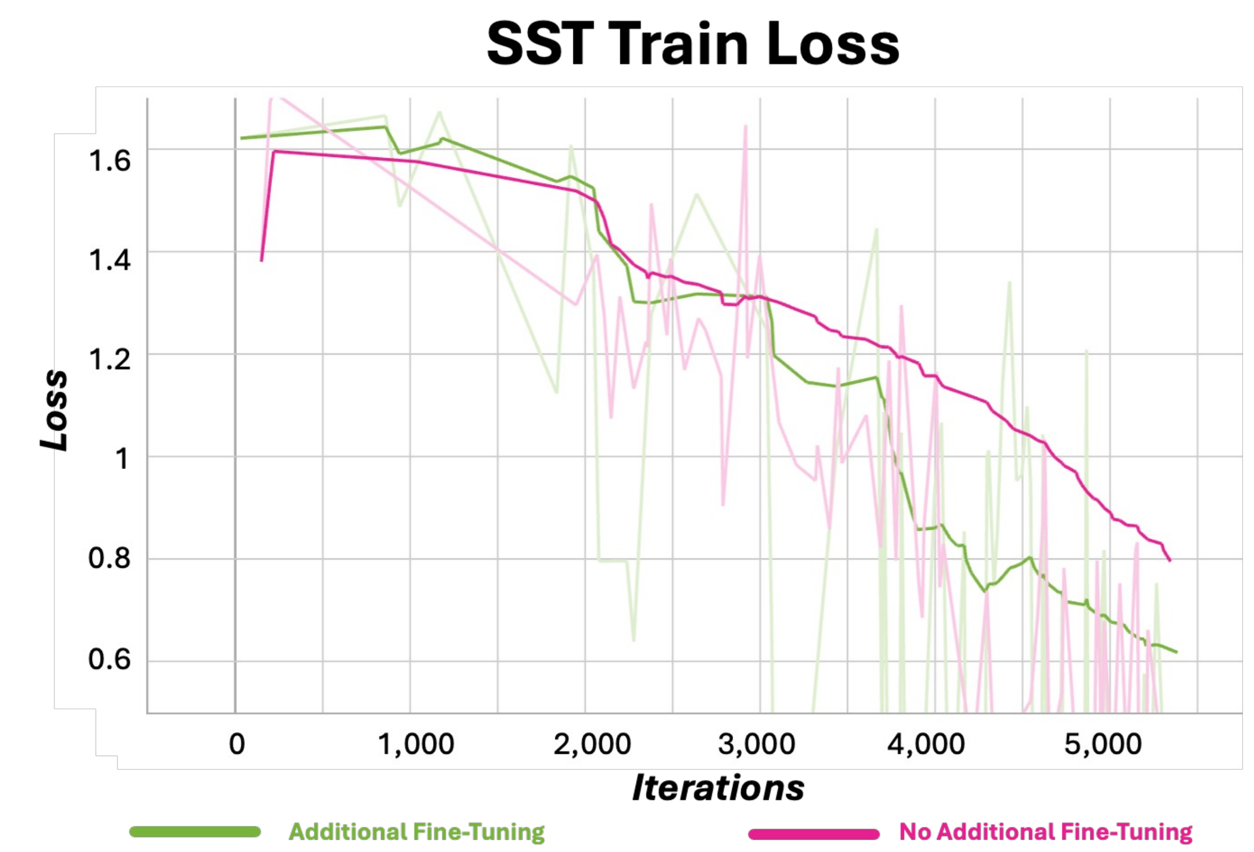
[3] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. CoRR, abs/2104.08821.

Annealed Sampling

$$\alpha = 1 - 0.8\frac{\epsilon - 1}{E - 1}$$

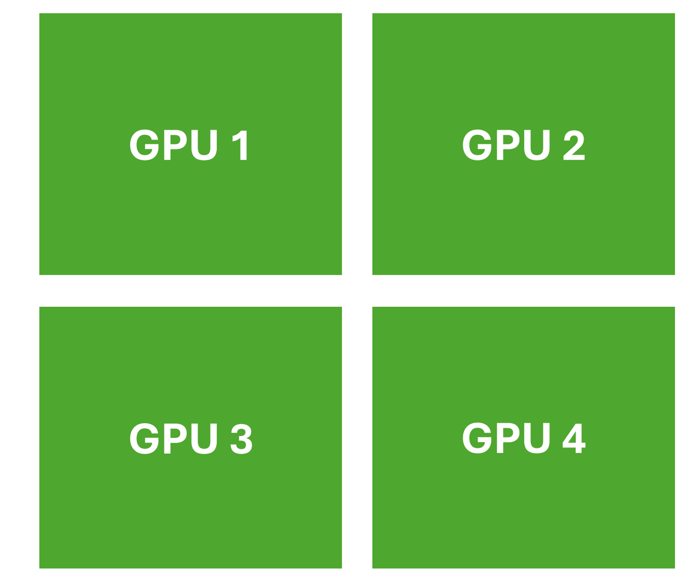
Fine-Tuning on Additional Datasets

- **Stanford Natural Language Inference (SNLI):** 570k English sentence entailment pairs for para
- **CFIMDB:** 2,434 highly polarized movie reviews for sentiment



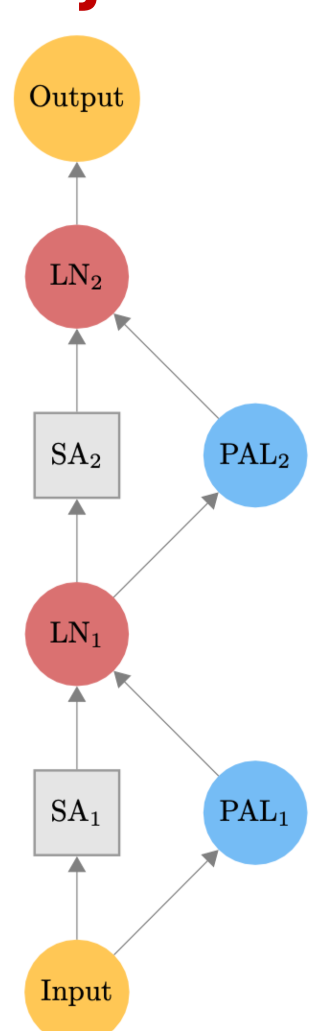
Multi-GPU Training

- Multi-GPU setup with 4 Nvidia T4 GPU.
- Used PyTorch's DistributedDataParallel (DDP) to parallelize our workload.
- The gradients calculated after forward pass are averaged across all GPUs.
- Training and Validation was run using a Multi-GPU setup while the test was run using a single GPU.



PALs → Annealed Sampling → SimCSE → Fine-Tuning on Additional Datasets → Relational Layer and LR Warmup / Decay → Multi-GPU Training

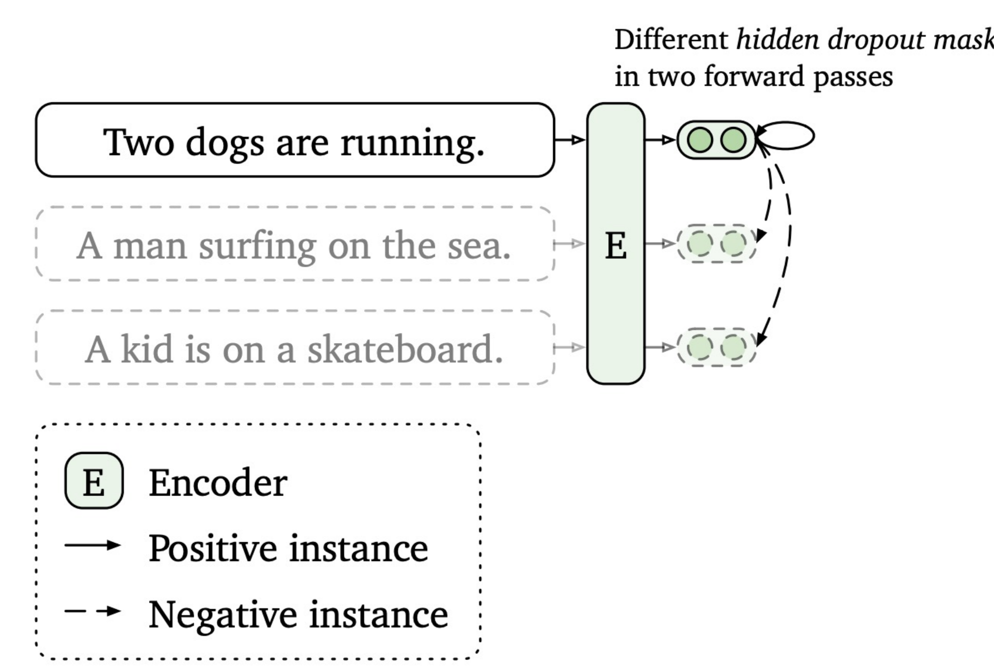
Projected Attention Layers (PALs)



- PALs resulted in massive improvements on the overall score on all 3 downstream tasks.
- Added additional task specific parameters in the model which are trained in addition to the different layers of BERT.
- Adding task specific parameters early on enables the parameters to learn richer representation of signals, as compared to adding them after getting [CLS] embedding from BERT.

Unsupervised Contrastive Learning of Sentence Embeddings (SimCSE)

- SST improvement from 0.490 to 0.516



RL and LR Warmup / Decay

- Increase LR linearly for first 10% of steps, then decrease linearly for the remainder of training to 0