BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning

Asa Cooper Stickland and Iain Murray
University of Edinburgh

Background: BERT

Our model builds on BERT (Devlin et al., 2018), a powerful (and big) sentence representation model.

Background: BERT

Our model builds on BERT (Devlin et al., 2018), a powerful (and big) sentence representation model.

Based off the ‘transformer’ architecture, with the key component self-attention.

BERT is trained on large amounts of text from the web (think: all of English wikipedia).

This model can be fine-tuned on tasks with a text input.

Best paper award at NAACL, 238 citations since 11/10/2018, SOTA on many tasks.

Our Approach

BERT is a huge model (approx. 100 or 300 million parameters), we don’t want to store many different versions of it.

Motivations: Mobile devices, web scale apps.

Can we do many tasks with one powerful model?

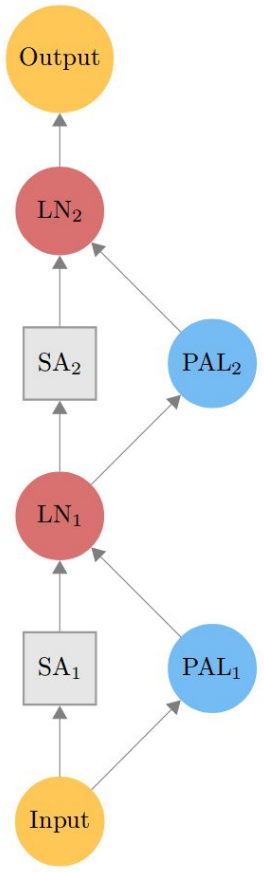
Our Approach

We consider multi-task learning on the GLUE benchmark (Wang et al, 2018), and we want the model to share most parameters but have some task-specific ones to increase flexibility.

We concentrate on <1.13× 'base' parameters.

Where should we add parameters?

What form should they take?



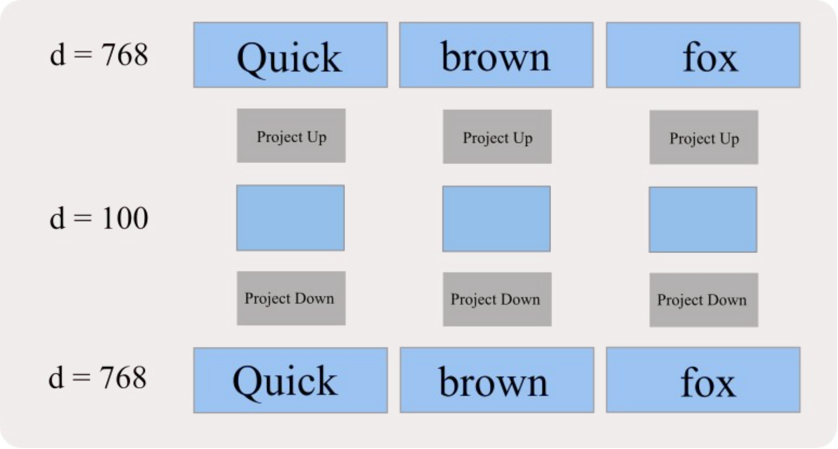
Adapters: Basics

$$\mathbf{h}^{l+1} = \mathrm{LN}(\mathbf{h}^l + \mathrm{SA}(\mathbf{h}^l) + \mathrm{TS}(\mathbf{h}^l))$$

We can add a simple linear projection down from the normal model dimension **d_m** to **d_s**:

$$\mathrm{TS}(\mathbf{h}) = V^D g(V^E \mathbf{h})$$

V^E projects down to d_s, we apply function g(), then **V^D** projects back up to d_m.

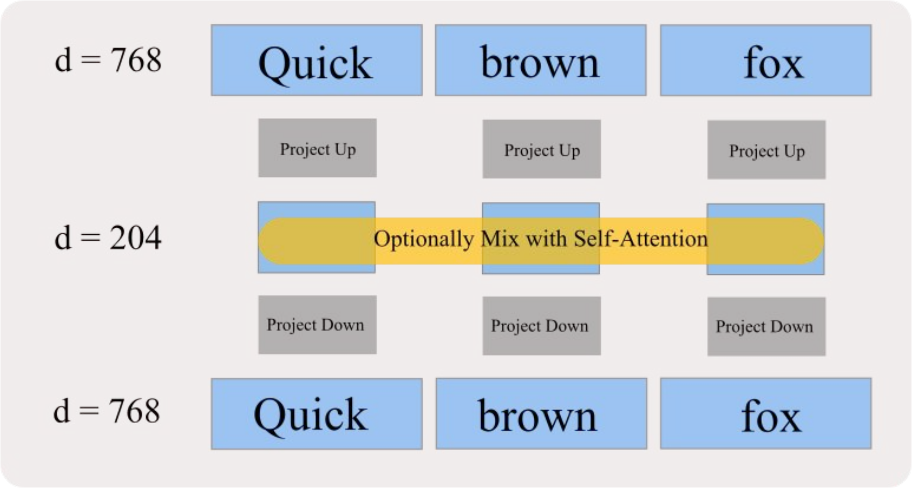


Adapters: PALs

V^E projects down to d_s, we apply function g(), then V^D projects back up to d_m.

$$\mathrm{TS}(\mathbf{h}) = V^D g(V^E \mathbf{h})$$

Our PALs method **shares V^D and V^E across all layers,** so we have the ‘budget’ to make function g() be self-attention.



Experiments

METHOD	PARAMS	MNLI-(M/MM) 392K	QQP 363K	QNLI 108K	SST-2 67K	COLA 8.5K	STS-B 5.7K	MRPC 3.5K	RTE 2.5K	AV.
BERT-BASE	8×	84.6/83.4	89.2/71.2	90.1	93.5	52.1	85.8	84.8/88.9	66.4	79.6
SHARED	1.00×	84.0/83.4	88.9/70.8	89.3	93.4	51.2	83.6	81.3/86.7	76.6	79.9
TOP PROJ. ATTN.	1.10×	84.0/83.2	88.8/71.2	89.7	93.2	47.1	85.3	83.1/87.5	75.5	79.6
PALs (204)	1.13×	84.3/83.5	89.2/71.5	90.0	92.6	51.2	85.8	84.6/88.7	76.0	**80.4**

Thanks!

Contact me @AsaCoopStick on Twitter, or email a.cooper.stickland@ed.ac.uk.

Our paper is on Arxiv, and it's called ‘BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning’.

Our poster is on Wednesday at 6:30 pm, Pacific Ballroom #258.