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 $<1.13 \times$ ‘base’ parameters.

Where should we add parameters?

What form should they take?
Adapters: Basics

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

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

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

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

$V^E$ 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 $d_m$.

$$TS(h) = V^D g(V^E h)$$

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Params</th>
<th>MNL1-(m/mm) 392k</th>
<th>QQP 363k</th>
<th>QNLI 108k</th>
<th>SST-2 67k</th>
<th>CoLA 8.5k</th>
<th>STS-B 5.7k</th>
<th>MRPC 3.5k</th>
<th>RTE 2.5k</th>
<th>Av.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-BASE</td>
<td>8x</td>
<td>84.6/83.4</td>
<td>89.2/71.2</td>
<td>90.1</td>
<td>93.5</td>
<td>52.1</td>
<td>85.8</td>
<td>84.8/88.9</td>
<td>66.4</td>
<td>79.6</td>
</tr>
<tr>
<td>Shared</td>
<td>1.00x</td>
<td>84.0/83.4</td>
<td>88.9/70.8</td>
<td>89.3</td>
<td>93.4</td>
<td>51.2</td>
<td>83.6</td>
<td>81.3/86.7</td>
<td>76.6</td>
<td>79.9</td>
</tr>
<tr>
<td>Top Proj. Attn.</td>
<td>1.10x</td>
<td>84.0/83.2</td>
<td>88.8/71.2</td>
<td>89.7</td>
<td>93.2</td>
<td>47.1</td>
<td>85.3</td>
<td>83.1/87.5</td>
<td>75.5</td>
<td>79.6</td>
</tr>
<tr>
<td>PALs (204)</td>
<td>1.13x</td>
<td>84.3/83.5</td>
<td>89.2/71.5</td>
<td>90.0</td>
<td>92.6</td>
<td>51.2</td>
<td>85.8</td>
<td>84.6/88.7</td>
<td>76.0</td>
<td><strong>80.4</strong></td>
</tr>
</tbody>
</table>
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.