Motivation

- Pre-trained language model, such as BERT, has proven to be highly effective for downstream NLP tasks
- However, the high demand for computing resources during model training hinders their application in practice
- Knowledge Distillation (KD) is proven to be useful for model compression in previous work
- We propose **Patient Knowledge Distillation**, which learns knowledge from previous layers of the teacher network, and is more generalizable and effective than vanilla KD

Notations

- **BERT-Teacher**: BERT with 12 or 24 layers fine-tuned on downstream tasks
- **BERT-Student**: Transformer with 3 or 6 layers to be learned from the Teacher and downstream tasks
- **CE-Loss**: Cross-entropy loss
- **DS-Loss**: Distillation loss between teacher’s and student’s soft labels
- **Embedding of [CLS]**: $h_i = [h_{i,1}, h_{i,2}, \ldots, h_{i,k}] = \text{BERT}_t(x_i) \in \mathbb{R}^{k \times d}$
- **PT Loss on [CLS]**: $L_{PT} = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{||h_{i,j}||^2} - \frac{1}{||h_{i,P(t)}(j)||^2}$

Learning Curves on the Training and Dev sets of QNLI and MNLI

- Learning curves on QNLI and MNLI, two large-scale NLI datasets, where the Student network learned with vanilla KD quickly saturates on the dev set, while the proposed Patient-KD starts to plateau only in a later stage

Experimental Results

<table>
<thead>
<tr>
<th>Setting</th>
<th>Teacher</th>
<th>Student</th>
<th>SST-2</th>
<th>MRPC</th>
<th>QQP</th>
<th>MNLI-m</th>
<th>MNLI-mm</th>
<th>QNLI</th>
<th>RTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>BERT_6</td>
<td>94.3</td>
<td>89.2/85.2</td>
<td>70.9/89.0</td>
<td>83.7</td>
<td>82.8</td>
<td>90.4</td>
<td>69.1</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>BERT_{6}[Base]</td>
<td>94.3</td>
<td>88.2/84.3</td>
<td>71.9/89.4</td>
<td>85.7</td>
<td>84.8</td>
<td>92.2</td>
<td>72.8</td>
</tr>
<tr>
<td>#1</td>
<td>BERT_{6}[Teacher]</td>
<td>BERT_{6}[Base]-KD</td>
<td>91.5</td>
<td>86.2/80.6</td>
<td>70.1/88.8</td>
<td>79.7</td>
<td>79.1</td>
<td>88.3</td>
<td>64.7</td>
</tr>
<tr>
<td>#2</td>
<td>BERT_{6}[Teacher]</td>
<td>BERT_{6}[Base]-KD</td>
<td>91.2</td>
<td>86.1/80.7</td>
<td>69.4/88.6</td>
<td>80.2</td>
<td>79.7</td>
<td>87.5</td>
<td>65.7</td>
</tr>
<tr>
<td>#3</td>
<td>BERT_{6}[Teacher]</td>
<td>BERT_{6}[Large]-KD</td>
<td>89.6</td>
<td>79.0/70.0</td>
<td>65.0/86.7</td>
<td>75.3</td>
<td>74.6</td>
<td>83.4</td>
<td>53.7</td>
</tr>
<tr>
<td>#4</td>
<td>BERT_{6}[Teacher]</td>
<td>BERT_{6}[Large]-PKD</td>
<td>89.8</td>
<td>77.8/68.3</td>
<td>67.1/87.9</td>
<td>77.2</td>
<td>76.7</td>
<td>83.8</td>
<td>53.2</td>
</tr>
</tbody>
</table>

- #1 vs. #2: there is no much difference between the Student’s performance when changing teacher from BERT-Large to BERT-Base
- #2 vs. #3: BERT_{6}[Large] Student has 1.6 times more parameters than BERT_{6}[Base], but it performs much worse
- #3 vs. #4: PKD-Skip outperforms KD, which indicates PKD is a generic approach independent of the selection of the Teacher model

Initialization Mismatch

- Ideally, we should use pre-trained 6-layer BERT as initialization
- We are using first 6 layers of BERT-Base and BERT-Large because of computation limitation
- The first six layers of BERT-Large may not be able to capture high-level features, leading to worse KD performance