GluonNLP 0.6: Closing the Gap in Reproducible Research with BERT

Eric Haibin Lin  
Mar 19 · 4 min read

BERT (Bidirectional Encoder Representations from Transformers) is arguably the most notable pre-training model in natural language processing (NLP). For instance, BERT lifts the score from 72.8 to 80.5 in the GLUE benchmark for 9 different NLP tasks—this is the biggest recent advancement[6].

Although BERT is exciting, unfortunately there have been no open source implementations that simultaneously

- enable scalable pre-training with GPUs;
- reproduce results on various tasks;
- support model exporting for deployment.

Thus, we release GluonNLP 0.6 to address such pain points by i) pre-training BERT with 8 GPUs in 6.5 days; ii) reproducing multiple natural language understanding results; iii) streamlining deployment.

GluonNLP’s mission is to provide a one stop shop experience for easy prototyping for deep learning models for NLP.

We Pre-train BERT with 8 GPUs in 6.5 Days
In case you live under a rock and have not heard of BERT, here is how it works. It 1) uses stacked bidirectional transformer encoders, 2) learns parameters by masked language modeling and next sentence prediction on large corpora with self-supervision, and 3) transfers these learnt text representations to specific downstream NLP tasks with a small set of labeled data by fine-tuning.

You may wonder: the official BERT repository has released multiple pre-trained models for free from the result of many TPU hours, why should we still care about pre-training BERT? This is because the choice of corpus for pre-training is very important. Like any other transfer learning setting, the model is more likely to perform well if the pre-trained data source is close to the task at hand. For example, pre-training on Wikipedia may not help us do better on tweets due to the difference in language style.

We pre-trained the BERT Base model from scratch. We used an English Wikipedia data dump that contains 2.0 billion words after removing images and tables, and a books corpus dataset which contains 579.5
million words after de-duplication. With mixed precision training and gradient accumulation, the BERT Base model takes 6.5 days using 8 Volta 100 GPUs and achieves the following results on validation sets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GluonNLP</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Research Paraphrase Corpus</td>
<td>88.0</td>
<td>86.4</td>
</tr>
<tr>
<td>The Stanford Sentiment Treebank</td>
<td>93.0</td>
<td>92.7</td>
</tr>
</tbody>
</table>

**We Reproduce BERT Fine-tuning for NLU Tasks with Scripts and Logs Available**

Promoting reproducible research is one of the important goals of GluonNLP. In GluonNLP, we provide both training scripts and logs that replicate state-of-the-art results on RTE[6], MNLI[8], SST-2, MRPC[10], SQuAD 1.1[9] and SQuAD 2.0[11]. Our code is modularized to facilitate BERT on many tasks in one framework.

We report F1 and exact match scores on the validation set for question answering datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BERT Model</th>
<th>GluonNLP</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD 1.1</td>
<td>BERT Base</td>
<td>88.5 81.0</td>
<td>88.5 80.8</td>
</tr>
</tbody>
</table>

Below please find accuracies on validation sets for the following sentence classification tasks with BERT Base model:
<table>
<thead>
<tr>
<th>Dataset</th>
<th>GluonNLP</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Research Paraphrase</td>
<td>88.7</td>
<td>86.4</td>
</tr>
<tr>
<td>Corpus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Stanford Sentiment</td>
<td>93.0</td>
<td>92.7</td>
</tr>
<tr>
<td>Treebank</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**We Streamline BERT Deployment**

With the power of MXNet, we provide BERT model that can be serialized into json format and deployed in in C++, Java, Scala, and many other languages. With float16 support, we see approximately 2 times speed up for BERT inference on GPUs. We are also working on int8 quantization on CPUs.

**How to Get Started?**

To get started with BERT using GluonNLP, visit our tutorial that walks through the code for fine-tuning BERT for sentence classification. You can also check out our BERT model zoo for BERT pre-training scripts, and fine-tuning scripts for SQuAD and GLUE benchmarks.

For other new features added in GluonNLP, please read our release notes. We are also working on BERT distributed training enhancements, and bringing GPT-2, BiDAF[12], QANet[3], BERT for NER/parsing, and many more to GluonNLP.

Happy BERTing with GluonNLP 0.6!

**Acknowledgement**

We thank great contributions from the GluonNLP community:
@havent-jeon @fiercex @kenjewu @imgarylai @TaoLv @Ishitori @szha @astonzhang @cgraywang

**References**


