Task-universal sentence embeddings from learning natural language inference

Yao Liu, Jeha Yang, Katherine Yu, Alex Kolchinski

{yoliiu, jeha, yokater, yakech}@stanford.edu

Abstract

In this project, we show how fixed-dimension sentence embeddings from models trained on Stanford Natural Language Inference (SNLI) dataset [1] and a new dataset we generated, derived from Stanford Question Answering (SQaD) dataset [2], could be transferred into many other semantic tasks, especially tasks with little training data. The SNLI baseline is based on Conneau et al. [3]. We see following models to learn our sentence embeddings and compare the results on both SNLI and transfer task.

- Siamese model with BLSTM as encoder, followed by MLP as classifier [4].
- Some variants of siamese model with different encoders: Transformer [5], 2 layers BLSTM.
- Decomposable attention model [6].

Siamese for generating sentence embeddings

Siamese models are far not the best-performing models on the two transfer tasks due to using downtuned word-by-word attention; however, we need to use siamese training for to produce generic word embeddings since sharing the same encoder parameters allows the single encoder to learn from all sentences in the training data and at inference on a single sentence, we do not have a target sentence (we cannot use seq2seq attention). We tried the following encoder architectures with Siamese training:

- BiLSTM+LSTM [5], Shared BiLSTM-LSTM
- Transformer [5]
- Transformer [6]

Figure 1: SNLI Siamese architecture (left), and ClassIfQa architecture (right). We try to multilinmken these two tasks.

Multitask training

We (unconventionally) train each of the two tasks for many consecutive steps (e.g., 1000 steps): SNLI (40 steps ClassIfQa) and find that each task can recover and improve on its previous loss has very quickly (within 100 batches) after the other task’s turn (the reason for this initially was that we wanted to compare in-situ progress against a reference single-task learning curve). It seems empirically important to use different batch sizes for the two tasks: 64 for SNLI and 128 for ClassIfQa [5]; thus, we generally take smaller steps for the two tasks.

SOS Methods

We also tried models with joint attention between sentences such as decomposable attention model [5]. The main aim here is to verify that although these models could achieve good performance in SNLI dataset as shown below, it might not be a good choice for learning transferrable representations. Our results were within 2-3 points of published results.

Table 1: Performance of models on SNLI and ClassIfQa datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>SNLI (test)</th>
<th>ClassIfQa (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer</td>
<td>86.8%</td>
<td>66.4%</td>
</tr>
<tr>
<td>Transformer [5]</td>
<td>86.7%</td>
<td>66.2%</td>
</tr>
<tr>
<td>Transformer [6]</td>
<td>86.6%</td>
<td>66.0%</td>
</tr>
</tbody>
</table>

Discussion and future works

- Siamese models with BLSTM encoder from [1] achieved best performance in some tasks, with Multitask training derived from SQaD dataset did well in other tasks.
- Multitask training seems promising, considering we only use models LSTM with max-pooling. This might imply that multitask training could avoid learning features that heavily depend on SNLI dataset or SNLI structure.
- Combination of max-pool and mean-pool could help many encoders achieve better transfer performance.
- Decomposable attention model does well on SNLI dataset but could not learn good transferable embeddings because they rely on inter-sentence attention.

For future work, we would further develop the multitask model. We would try adding additional training data such as MultiNLI, QQP, Question Pair with different word embeddings (torchify, train our own) since they are very important to this model. We also want to work on a decomposition analysis on the best of transfer Multitask tasks so we can understand performance for important task characteristics like sentence length, OOV rate, and relative size of the training data.

References