Evaluating Protein Transfer Learning with TAPE

Roshan Rao*
UC Berkeley
roshan_rao@berkeley.edu

Nicholas Bhattacharyya*
UC Berkeley
nick_bhat@berkeley.edu

Neil Thomas*
UC Berkeley
nthomas@berkeley.edu

Yan Duan
covariant.ai
rocky@covariant.ai

Xi Chen
covariant.ai
peter@covariant.ai

John Canny
UC Berkeley
canny@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@berkeley.edu

Yun S. Song
UC Berkeley
yss@berkeley.edu

Presented by: Rahmatullah Roche

COMPUTER SCIENCE
VIRGINIA TECH.
Motivation

- Database of protein sequence growing exponentially
- Total number of sequence doubling each year
- Unlabeled sequence contains significant evolutionary information
- Can NLP extract the information?
Contribution: benchmarking

- Self supervised learning from unlabeled dataset
- Task assessing protein embedding (TAPE)
- Systematically evaluated semi-supervised learning on protein sequences
- 5 biologically relevant supervised task
- Hypothesis: multiple tasks are required to accurately benchmark any method
- Performance assessment of
  - Recurrent-based model
  - Convolution-based model
  - Attention-based model
  - Semi-supervised models
Background

- Protein terminology
  - \((x_1, x_2, x_3, \ldots, x_L)\) fixed alphabet for amino acids

- Protein sequence alignments
  - Query → Database → Alignment

- Semi-supervised learning
  - Leverage information from both labeled and unlabeled data
Related Works

- Kernel-based pretraining for homology detection
- NLP-based techniques for transfer learning
- VAE to predict functional impact in mutations
- Transfer learning in protein ss and contact prediction
- Not rigorously benchmarked to assess the comparisons
Dataset

• Unlabeled sequence dataset
  • Pfam database of 31M protein domains
  • Training and test dataset split: 95%/5%

• Supervised datasets
  • Five biologically relevant downstream tasks
  • Dataset ranges in size 8k-50k for training
Tasks

- Self-supervised:
  - Next token prediction
  - Mask token prediction
- Downstream tasks:
  - Protein SS
  - Protein contact map
  - Protein homology detection
  - Fluorescence
  - Stability
Tasks

Figure 1:

(a) Secondary Structure
(b) Contact Prediction
(c) Remote Homology

Figure 2:

(a) Fluorescence
(b) Stability
Losses

• Two self-supervised losses for NLP task
  • Next-token prediction
  • Masked-token prediction

• Protein specific loss
  • Further supervised pretraining of models
  • Supervised pretraining on contact prediction and remote homology detection can improve secondary structure prediction (Beplar et al)
Architectures for Downstream tasks

- **LSTM**
  - Two 3-layer LSTMs with 1024 hidden units corresponding to the forward and backward language models

- **Transformer**
  - 12-layer transformer
  - Each layer hidden size 512 units and 8 attention head
  - 38M parameters

- **ResNet**
  - 35 residual blocks
  - Each containing 2 conv. Layer with 256 filters
  - Kernel size 9, dilation rate 2
Architectures for Downstream tasks (cont.)

• Bepler et al.
  • Two 3-layer LSTMs with 512 hidden units corresponding to the forward and backward language models

• Alley et al.
  • Unidirectional mLSTM
  • 1900 hidden units
Architectures for Downstream tasks: baseline

- Secondary structure: NetSerf2.0
  - Two convolution layers followed by two bidirectional LSTM followed by a linear output layer

- Contact prediction architecture: Similar to RaptorX-contact
  - 30 residual blocks having 2 conv. Layers each
  - 64 filter for each conv. layer

- Remote homology and protein engineering architecture
  - Predict attention value for each position of sequence to compute attention-weighted mean embedding
  - Followed by 512 hidden unit dense layer
  - Followed by relu and linear activation
Results: language modeling

Table 1: Language modeling metrics: Language Modeling Accuracy (Acc), Perplexity (Perp) and Exponentiated Cross-Entropy (ECE)

<table>
<thead>
<tr>
<th></th>
<th>Random Families</th>
<th></th>
<th>Heldout Families</th>
<th></th>
<th>Heldout Clans</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Perp</td>
<td>ECE</td>
<td>Acc</td>
<td>Perp</td>
<td>ECE</td>
</tr>
<tr>
<td>Transformer</td>
<td>0.45</td>
<td>8.89</td>
<td>6.01</td>
<td>0.35</td>
<td>11.77</td>
<td>8.87</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.40</td>
<td>8.89</td>
<td>6.94</td>
<td>0.24</td>
<td>13.03</td>
<td>12.73</td>
</tr>
<tr>
<td>ResNet</td>
<td>0.41</td>
<td>10.16</td>
<td>6.86</td>
<td>0.31</td>
<td>13.19</td>
<td>9.77</td>
</tr>
<tr>
<td>Bepler et al. [11]</td>
<td>0.28</td>
<td>11.62</td>
<td>10.17</td>
<td>0.19</td>
<td>14.44</td>
<td>14.32</td>
</tr>
<tr>
<td>Alley et al. [12]</td>
<td>0.32</td>
<td>11.29</td>
<td>9.08</td>
<td>0.16</td>
<td>15.53</td>
<td>15.49</td>
</tr>
<tr>
<td>Random</td>
<td>0.04</td>
<td>25</td>
<td>25</td>
<td>0.04</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>
## Results: downstream tasks

### Table 2: Results on downstream supervised tasks

<table>
<thead>
<tr>
<th>Method</th>
<th>Structure</th>
<th></th>
<th>Evolutionary</th>
<th></th>
<th>Engineering</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>Contact</td>
<td>Homology</td>
<td>Fluorescence</td>
<td>Stability</td>
<td></td>
</tr>
<tr>
<td>No Pretrain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer</td>
<td>0.70</td>
<td>0.32</td>
<td>0.09</td>
<td>0.22</td>
<td>-0.06</td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td>0.71</td>
<td>0.19</td>
<td>0.12</td>
<td>0.21</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>ResNet</td>
<td>0.70</td>
<td>0.20</td>
<td>0.10</td>
<td>-0.28</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>Pretrain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer</td>
<td>0.73</td>
<td>0.36</td>
<td>0.21</td>
<td>0.68</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td>0.75</td>
<td>0.39</td>
<td><strong>0.26</strong></td>
<td>0.67</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>ResNet</td>
<td>0.75</td>
<td>0.29</td>
<td>0.17</td>
<td>0.21</td>
<td><strong>0.73</strong></td>
<td></td>
</tr>
<tr>
<td>Bepler et al. [11]</td>
<td>0.73</td>
<td>0.40</td>
<td>0.17</td>
<td>0.33</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>Alley et al. [12]</td>
<td>0.73</td>
<td>0.34</td>
<td>0.23</td>
<td>0.67</td>
<td><strong>0.73</strong></td>
<td></td>
</tr>
<tr>
<td>Baseline features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-hot</td>
<td>0.69</td>
<td>0.29</td>
<td>0.09</td>
<td>0.14</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Alignment</td>
<td><strong>0.80</strong></td>
<td><strong>0.64</strong></td>
<td>0.09</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

### Diagrams

- (a) True Contacts
- (b) LSTM
- (c) LSTM Pretrain
- (d) One Hot
- (e) Alignment
Discussions

• Alignment-based input currently outperforms self-supervised featurization
  • All state-of-the-art methods use alignment-based input features
  • Can pertaining along with alignment-based input improve performance?
• Multiple tasks are required to appropriately benchmark a given model
  • Transformer performs worst in ss and contact prediction but best in fluorescence and stability tasks
• A challenge for future research in self-supervised learning
  • Create models for protein specific tasks or generalized tasks?