More than BERT: oLMpics on diverse language models

Kevin Zhao
Mentored by Vladislav Lialin, Namratar Shivagunde, Anna Rumshisky
Transfer Learning

- Pre-BERT: most of the model is trained from scratch
- BERT: pre-trained on vast amounts of generic text

![Bar chart showing GLUE scores for different models.](chart.png)

- **GLUE Score**
  - **CBOW**: 58.6
  - **BiLSTM+CoVe**: 62.9
  - **BiLSTM+CoVe+Attn**: 63.1
  - **BiLSTM**: 64.2
  - **BiLSTM+Attn**: 65.6
  - **BERT_Large**: 80.5

---

Untrained Models vs. BERT Performance
Transfer Learning

- **Pre-BERT**: most of the model is trained from scratch
- **BERT**: pre-trained on vast amounts of generic text
- **Post-BERT**: pre-training is one of the pillars of NLP
- The number of pre-training methods has rocketed
Outline

- Transformers
  - Architecture and attention
  - Models
- oLMpics
  - Overview
  - Evaluation methods
  - Task results
- Attention
  - Attention norms
  - Patterns
- Conclusion
Transformers

Architecture

Attention

- Tokens interact directly
- Query, Key, Value
- Multiple heads

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]
Transformers

Architecture

Attention

- Tokens interact directly
- Query, Key, Value
- Multiple heads

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]
Transformer Models

Differences:
- Architecture
- Size
- Pre-training objective
- Pre-training data

Encoder
- BERT
- RoBERTa
- ALBERT
- DistilBERT

Encoder + Decoder
- BART
- T5

Decoder
- GPT
# oLMpics Overview

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Example Question</th>
<th>Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Comparison</td>
<td>A 41 year old person age is [MASK] than a 42 year old person.</td>
<td>younger, older</td>
</tr>
<tr>
<td>Always Never</td>
<td>A lizard [MASK] has a wing.</td>
<td>often, rarely, <strong>never</strong>, sometimes, always</td>
</tr>
<tr>
<td>Object Comparison</td>
<td>The size of a nail is usually much [MASK] than the size of a fork.</td>
<td>smaller, larger</td>
</tr>
<tr>
<td>Antonym Negation</td>
<td>It was [MASK] a fracture, it was really a break.</td>
<td>not, <strong>really</strong></td>
</tr>
<tr>
<td>Taxonomy Conjunction</td>
<td>A ferry and a biplane are both a type of [MASK].</td>
<td>airplane, <strong>craft</strong>, boat</td>
</tr>
<tr>
<td>Property Conjunction</td>
<td>What is related to vertical and is related to honest?</td>
<td><strong>straight</strong>, trustworthy, steep</td>
</tr>
<tr>
<td>Encyclopedic Composition</td>
<td>Where is the headquarters of the company that Giovanni Agusta established located?</td>
<td>Varese, Pisa, Reggio Calabria</td>
</tr>
<tr>
<td>Multi-hop Composition</td>
<td>When comparing a 21 year old, 15 year old, and 19 year old, the [MASK] is oldest.</td>
<td>third, <strong>first</strong>, second</td>
</tr>
</tbody>
</table>

**Blue:** MLM  
**Orange:** QA
oLMpics Evaluation Methods

◎ MLM
◎ QA
◎ Modification for GPT2

Autoregressive Method:
Sequence: “1 3 0 5”
Sequence Probability:

```
(0 1 2 3 4 5)  (0 1 2 3 4 5)  (0 1 2 3 4 5)  (0 1 2 3 4 5)
```

Head (Linear Layer)
Encoder #24
Encoder #2
Encoder #1
Contextualized Embeddings
Hidden Embeddings
Prediction Logits

Masked LM
0 1 2 3 4 5
<table>
<thead>
<tr>
<th>Model</th>
<th>Always</th>
<th>Never</th>
<th>Object</th>
<th>Antonym</th>
<th>Taxonomy</th>
<th>Multi-hop</th>
<th>Encyclopedia</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random Baseline</strong></td>
<td>20</td>
<td>50</td>
<td>50</td>
<td>33.3</td>
<td>33.3</td>
<td>33.3</td>
<td>33.3</td>
<td>33.3</td>
</tr>
<tr>
<td>BERT&lt;sub&gt;base&lt;/sub&gt;</td>
<td>13.3</td>
<td>55.4</td>
<td>53.8</td>
<td>46.7</td>
<td>33.2</td>
<td>56.1</td>
<td><strong>62.6</strong></td>
<td></td>
</tr>
<tr>
<td>BERT&lt;sub&gt;large&lt;/sub&gt;</td>
<td>22.5</td>
<td>52.4</td>
<td>51.0</td>
<td>53.9</td>
<td>33.8</td>
<td>57.1</td>
<td>58.3</td>
<td></td>
</tr>
<tr>
<td>BERT&lt;sub&gt;large WWM&lt;/sub&gt;</td>
<td>10.7</td>
<td>55.6</td>
<td>57.2</td>
<td>46.2</td>
<td>33.8</td>
<td>56.4</td>
<td>60.1</td>
<td></td>
</tr>
<tr>
<td>RoBERT&lt;sub&gt;large&lt;/sub&gt;</td>
<td>13.5</td>
<td><strong>87.4</strong></td>
<td><strong>74.4</strong></td>
<td>45.4</td>
<td>28.0</td>
<td>55.5</td>
<td>55.5</td>
<td></td>
</tr>
<tr>
<td>DistilBERT&lt;sub&gt;base&lt;/sub&gt;</td>
<td>15.0</td>
<td>50.8</td>
<td>50.8</td>
<td>46.9</td>
<td>33.4</td>
<td>53.9</td>
<td>56.2</td>
<td></td>
</tr>
<tr>
<td>ALBERT&lt;sub&gt;large&lt;/sub&gt;</td>
<td>10.7</td>
<td>55.6</td>
<td>57.2</td>
<td>46.2</td>
<td>33.8</td>
<td><strong>57.2</strong></td>
<td>60.2</td>
<td></td>
</tr>
<tr>
<td>BART&lt;sub&gt;large&lt;/sub&gt;</td>
<td>14.3</td>
<td>50.8</td>
<td>53.8</td>
<td>42.6</td>
<td>33.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>T5&lt;sub&gt;large&lt;/sub&gt;</td>
<td>25.7</td>
<td>79.8</td>
<td>59.2</td>
<td>44.2</td>
<td>33.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GPT2</td>
<td><strong>50.1±1.54</strong></td>
<td><strong>50.1±1</strong></td>
<td><strong>52.8±1.93</strong></td>
<td><strong>48.4±1.01</strong></td>
<td>32.2±2.37</td>
<td>32.2</td>
<td>42.9</td>
<td></td>
</tr>
<tr>
<td>GPT2&lt;sub&gt;medium&lt;/sub&gt;</td>
<td>40.8±2.24</td>
<td>49.6±0.92</td>
<td>54.7±2.38</td>
<td>49.1±1.65</td>
<td>29.6±2.12</td>
<td>31.8</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>GPT2&lt;sub&gt;large&lt;/sub&gt;</td>
<td>20.2±1.73</td>
<td>50.4±0.97</td>
<td>50.1±2.68</td>
<td>46.9±1.47</td>
<td>33.5±1.34</td>
<td>47.5</td>
<td>35.2</td>
<td></td>
</tr>
<tr>
<td>UniLM&lt;sub&gt;base&lt;/sub&gt;</td>
<td>15.5±1.49</td>
<td>47.8±1.25</td>
<td>43.5±0.71</td>
<td>-</td>
<td><strong>34.9±0.78</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UniLM&lt;sub&gt;large&lt;/sub&gt;</td>
<td>19.2±2.1</td>
<td>61.12±1.43</td>
<td>50.8±0.77</td>
<td>-</td>
<td>33.1±1.21</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Attention Norms

◎ Attention weights can be useful in understanding what a model looks at

◎ However, more recently attention norms have been shown to be more accurate
  ○ Attention formula can be rearranged
  ○ The norm of this product between the attention weights and transformed value vectors is the “attention norm”
Attention Norm Patterns

Age-Age Pattern

Age-MASK Pattern
Age-Age and Age-MASK Importance

To determine whether heads are important, we compare the effect of disabling the heads to disabling the same amount of random heads.

<table>
<thead>
<tr>
<th>Modification</th>
<th>BERT (20-40)</th>
<th>RoBERTa (20-40)</th>
<th>BERT (40-60)</th>
<th>RoBERTa (40-60)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (Age Comparison)</td>
<td>76.0 (0)</td>
<td>98.6 (0)</td>
<td>36.6 (0)</td>
<td>99.2 (0)</td>
</tr>
<tr>
<td>Age-Age</td>
<td>66.2 (20)</td>
<td>64.2 (20)</td>
<td>32.2 (20)</td>
<td>98.2 (20)</td>
</tr>
<tr>
<td>Age-Mask</td>
<td>67.4 (5)</td>
<td>98.6 (3)</td>
<td>68 (5)</td>
<td>99.2 (3)</td>
</tr>
<tr>
<td>Random (20 heads)</td>
<td>76.3 ± 5.4 (20)</td>
<td>92 ± 8.6 (20)</td>
<td>26.9 ± 5.3 (20)</td>
<td>97.8 ± 1.7 (20)</td>
</tr>
<tr>
<td>Random (5 or 3 heads)</td>
<td>72.7 ± 3.5 (5)</td>
<td>97.0 ± 1.1 (3)</td>
<td>39.9 ± 2.5 (5)</td>
<td>99.1 ± 0.3 (3)</td>
</tr>
</tbody>
</table>

Table 6: Results after disabling heads. The number in parentheses is the number of heads disabled.
Conclusion

- We analyzed the differences between pre-trained models
  - Zero-shot evaluation on oLMpics tasks
    - Different models perform well on different tasks, there’s no clear leader
    - None of the models can solve composition task
  - Hidden representation analysis - attention norms
    - Intuitive features like Age-MASK do not contribute to performance
- Adapted oLMpics zero-shot setup for autoregressive models
Acknowledgements

Thanks to:

▪ My mentors Vladislav Lialin and Dr. Rumshisky for their guidance and support
▪ Namratar Shivagunde for doing the GPT and UniLM experiments, as well as helping develop oLMpics for autoregressive models
▪ Dr. Gerovitch and the PRIMES program for giving me this research opportunity
▪ My parents
Questions?