Pretrained Transformers Improve Out-of-Distribution Robustness

Dan Hendrycks*, Xiaoyuan Liu*, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, Dawn Song

UC Berkeley
Shanghai Jiao Tong University
University of Chicago
Out-of-Distribution Robustness
Out-of-Distribution Robustness

In reality, test distribution will not match training

Image Credit: Aleksander Mądry
Out-of-Distribution Robustness

Image Credit: Aleksander Mądry
Out-of-Distribution Robustness

Two Goals:
- Generalize
- Detect
Our Paper’s Goal

- How robust are current NLP models?
Our Paper’s Goal

- How robust are current NLP models?
- Why might transformers be brittle?
  - high accuracy != high robustness [Hendrycks and Dietterich, 2019]
  - use superficial dataset patterns [Gururangan et al. 2018]
Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets
Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Sentiment Analysis
Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Sentiment Analysis

American → Chinese, Italian, and Japanese
Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

**Sentiment Analysis**
- American → Chinese, Italian, and Japanese

**Semantic Similarity**
- Headlines → Images
Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

**Sentiment Analysis**
- American → Chinese, Italian, and Japanese

**Semantic Similarity**
- Headlines → Images

**Reading Comprehension**
- CNN → DailyMail
Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Sentiment Analysis
  - American → Chinese, Italian, and Japanese

Semantic Similarity
  - Headlines → Images

Reading Comprehension
  - CNN → DailyMail

Textual Entailment
  - Telephone → Letters
Pretrained Transformers are More Robust

Semantic Textual Similarity (STS-B) Generalization

Pearson Correlation (%)

- IID Data (Images)
- OOD Data (MSRvid)

<table>
<thead>
<tr>
<th>Model</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. BoW</td>
<td>40</td>
</tr>
<tr>
<td>Avg. w2v</td>
<td>60</td>
</tr>
<tr>
<td>ConvNet w2v</td>
<td>80</td>
</tr>
<tr>
<td>LSTM w2v</td>
<td>80</td>
</tr>
<tr>
<td>BERT Base</td>
<td>90</td>
</tr>
<tr>
<td>BERT Large</td>
<td>90</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>90</td>
</tr>
</tbody>
</table>
Pretrained Transformers are More Robust

**IMDb Sentiment Classifier Generalization**

- **IID Data (IMDb)**
- **OOD Data (SST-2)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc (IID)</th>
<th>Acc (OOD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. BoW</td>
<td>85</td>
<td>75</td>
</tr>
<tr>
<td>Avg. w2v</td>
<td>82</td>
<td>70</td>
</tr>
<tr>
<td>ConvNet w2v</td>
<td>90</td>
<td>88</td>
</tr>
<tr>
<td>LSTM w2v</td>
<td>88</td>
<td>85</td>
</tr>
<tr>
<td>BERT Base</td>
<td>92</td>
<td>90</td>
</tr>
<tr>
<td>BERT Large</td>
<td>95</td>
<td>93</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>97</td>
<td>95</td>
</tr>
</tbody>
</table>
Bigger Models Are Not Always Better
Bigger Models Are Not Always Better

SST-2 Model Size vs. Accuracy Drop

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy - IMDb Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT base</td>
<td>4</td>
</tr>
<tr>
<td>BERT large</td>
<td>5</td>
</tr>
<tr>
<td>ALBERT base</td>
<td>2</td>
</tr>
<tr>
<td>ALBERT large</td>
<td>8</td>
</tr>
<tr>
<td>ALBERT xlarge</td>
<td>4</td>
</tr>
<tr>
<td>ALBERT xxlarge</td>
<td>4</td>
</tr>
</tbody>
</table>
Pretrained Transformers Are Better OOD Detectors

- softmax probability for scoring anomalies [Hendrycks and Gimpel, 2017]
Pretrained Transformers Are Better OOD Detectors

- softmax probability for scoring anomalies [Hendrycks and Gimpel, 2017]
- feed in OOD inputs and report false alarm rate at 95% recall
Pretrained Transformers Are Better OOD Detectors

Detecting OOD Examples for an SST-2 Sentiment Classifier

False Alarm Rate (%) (Lower Is Better)

Model Type
- Random Detector
- Bag of Words
- Avg. word2vec
- LSTM word2vec
- ConvNet word2vec
- BERT Large

20 NG Multi30K RTE SNLI WMT16 Average
Conclusions

● OOD benchmark for four NLP tasks

● Pretrained Transformers improve OOD generalization

● Pretrained Transformers improve OOD detection

● Further work needed to make models robust

Code + Data and Paper available