InfoBERT: Improving Robustness of Language Models from An Information Theoretic Perspective

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, Jingjing Liu
Adversarial Attacks in NLP

- Adversarial examples for **QA** [1]

**Question:** Who ended the series in 1989?

**Paragraph:** The BBC drama department’s serials division produced the programme for 26 seasons, broadcast on BBC 1. Falling viewing numbers, a decline in the public perception of the show and a less-prominent transmission slot saw production suspended in 1989 by Jonathan Powell, controller of BBC 1. ... the BBC repeatedly affirmed that the series would return. *Donald Trump ends a program on 1988.*

**QA Prediction:** Jonathan Powell → Donald Trump

- Adversarial examples for **classification tasks** [2]

**Classification Task:** Is this a *positive* or *negative review*?

**Input Text:** "The characters, cast in impossibly contrived situations, are totally estranged from reality."

**TextFooler:** "The characters, cast in impossibly engineered circumstances, are fully estranged from reality."

**SOTA NLP models** (e.g. BERT, LSTM, CNN)

**Negative!**

**Positive!**

Understanding ML Robustness from the Information-Theoretic Perspectives

Robust Representation Learning for LM

• Goals:
  
  - Maximize the mutual information between representation $T$ and label $Y$
  
  - Minimize the mutual information between input $X$ and representation $T$
  
  - Maximize the mutual information between local “robust” feature $T_{kj}$ and global feature $Z$
IB Regularizer

- Information Bottleneck (IB) As a Regularizer

$$\max L_{IB} = I(Y; T) - \beta I(X; T)$$
IB Regularizer

• Information Bottleneck (IB) As a Regularizer

\[ \mathcal{L}_{\text{IB}} = I(Y; T) - \beta I(X; T) \]

• Localized Information Bottleneck Formulation

\[ \mathcal{L}_{\text{LIB}} := I(Y; T) - n \beta \sum_{i=1}^{n} I(X_i; T_i). \]

**Theorem 3.1.** (Lower Bound of \( \mathcal{L}_{\text{IB}} \)) Given a sequence of random variables \( X = [X_1; X_2; \ldots; X_n] \) and a deterministic feature extractor \( f_\theta \), let \( T = [T_1; \ldots; T_n] = [f_\theta(X_1); f_\theta(X_2); \ldots; f_\theta(X_n)] \). Then the localized formulation of IB \( \mathcal{L}_{\text{LIB}} \) is a lower bound of \( \mathcal{L}_{\text{IB}} \) (Eq. (1)), i.e.,

\[ I(Y; T) - \beta I(X; T) \geq I(Y; T) - n \beta \sum_{i=1}^{n} I(X_i; T_i). \] (7)
IB Regularizer

- Relationship between Adversarial Performance Gap and Mutual Information between input $X$ and representation $T$

**Theorem 3.2.** (Adversarial Robustness Bound) For random variables $X = [X_1; X_2; \ldots; X_n]$ and $X' = [X'_1; X'_2; \ldots; X'_n]$, let $T = [T_1; T_2; \ldots; T_n] = [f_\theta(X_1); f_\theta(X_2); \ldots; f_\theta(X_n)]$ and $T' = [T'_1; T'_2; \ldots; T'_n] = [f_\theta(X'_1); f_\theta(X'_2); \ldots; f_\theta(X'_n)]$ with finite support $\mathcal{T}$, where $f_\theta$ is a deterministic feature extractor. The performance gap between benign and adversarial data $|I(Y; T) - I(Y; T')|$ is bounded above by

$$|I(Y; T) - I(Y; T')| \leq B_0 + B_1 \sum_{i=1}^{n} \sqrt{|\mathcal{T}|}(I(X_i; T_i))^{1/2} + B_2 \sum_{i=1}^{n} |\mathcal{T}|^{3/4}(I(X_i; T_i))^{1/4}$$

$$+ B_3 \sum_{i=1}^{n} \sqrt{|\mathcal{T}|}(I(X'_i; T'_i))^{1/2} + B_4 \sum_{i=1}^{n} |\mathcal{T}|^{3/4}(I(X'_i; T'_i))^{1/4},$$

(8)

where $B_0, B_1, B_2, B_3$ and $B_4$ are constants depending on the sequence length $n$, $\epsilon$ and $p(x)$. 
IB Regularizer Verification

- Adversarial robustness (i.e., the testing accuracy on adversarial examples) increases, as $\beta$ increases and $I(X; T)$ becomes lower.

Figure 2: Benign/robust F1 score on benign/adversarial QA datasets. Models are trained on the benign SQuAD dataset with different $\beta$. 
Robust Representation Learning for LM

- Goals:
  - Maximize the mutual information between representation $T$ and label $Y$
  - Minimize the mutual information between input $X$ and representation $T$
  - Maximize the mutual information between local “robust” representation $T_{k,j}$ and global representation $Z$
Local Anchored Feature Extraction

- Use adversarial attack to determine the local “robust” features

**Algorithm 1 - Local Anchored Feature Extraction.** This algorithm takes in the word local features and returns the index of local anchored features.

1: **Input:** Word local features $t$, upper and lower threshold $c_h$ and $c_l$
2: $\delta \leftarrow 0$  // Initialize the perturbation vector $\delta$
3: $g(\delta) = \nabla_{\delta} \ell_{\text{task}}(q_\psi(t + \delta), y)$  // Perform adversarial attack on the embedding space
4: Sort the magnitude of the gradient of the perturbation vector from $||g(\delta)_1||_2, ||g(\delta)_2||_2, ..., ||g(\delta)_n||_2$ into $||g(\delta)_{k_1}||_2, ||g(\delta)_{k_2}||_2, ..., ||g(\delta)_{k_n}||_2$ in ascending order, where $z_i$ corresponds to its original index.
5: **Return:** $k_i, k_{i+1}, ..., k_j$, where $c_l \leq \frac{i}{n} \leq \frac{j}{n} \leq c_h$.

- Increase the Mutual Information between local anchored features $T_{k,j}$ and global features $Z$

$$\max_{j=1}^{M} \sum_{j=1}^{M} I(T_{k,j}; Z)$$
Why local robust features are helpful

- Evaluate the difference of mutual information between global and local features for models before and after adv training.

- From the mutual information change, local anchored features are indeed more aligned with the global representation after adv training, which leads to a more robust model.
Complete Version

\[
\max I(Y; T) - n\beta \sum_{i=1}^{n} I(X_i; T_i) + \alpha \sum_{j=1}^{M} I(T_{k_j}; Z)
\]

- The first term uses the standard task objective (e.g., Maximum Log Likelihood)
- The second term uses CLUB [1] to calculate the upper bound
- The last term uses InfoNCE as the lower bound

Experiments

• Evaluation against Different Adversarial Attacks
  • Natural Language Inference (NLI)
    • ANLI
    • TextFooler
  • Question Answering
    • adv-SQuAD
## Evaluation of Model Robustness (I) - ANLI

<table>
<thead>
<tr>
<th>Training</th>
<th>Model</th>
<th>Method</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A1</td>
<td>A2</td>
</tr>
<tr>
<td>RoBERTa Standard</td>
<td>Vanilla</td>
<td>74.1</td>
<td>50.8</td>
<td>43.9</td>
</tr>
<tr>
<td></td>
<td>InfoBERT</td>
<td>75.2</td>
<td>49.6</td>
<td>47.8</td>
</tr>
<tr>
<td></td>
<td>Vanilla</td>
<td>58.5</td>
<td>46.1</td>
<td>45.5</td>
</tr>
<tr>
<td></td>
<td>InfoBERT</td>
<td>59.3</td>
<td>48.9</td>
<td>45.5</td>
</tr>
<tr>
<td>Adversarial</td>
<td>FreeLB</td>
<td>75.2</td>
<td>47.4</td>
<td>45.3</td>
</tr>
<tr>
<td></td>
<td>SMART</td>
<td>74.5</td>
<td>50.9</td>
<td>47.6</td>
</tr>
<tr>
<td></td>
<td>ALUM</td>
<td>73.3</td>
<td>53.4</td>
<td>48.2</td>
</tr>
<tr>
<td></td>
<td>InfoBERT</td>
<td>76.4</td>
<td>51.7</td>
<td>48.6</td>
</tr>
<tr>
<td></td>
<td>FreeLB</td>
<td>60.3</td>
<td>47.1</td>
<td>46.3</td>
</tr>
<tr>
<td></td>
<td>ALUM</td>
<td>62.0</td>
<td>48.6</td>
<td>48.1</td>
</tr>
<tr>
<td></td>
<td>InfoBERT</td>
<td>60.8</td>
<td>48.7</td>
<td>45.9</td>
</tr>
</tbody>
</table>

Table 2: Robust accuracy on the ANLI dataset. Models are trained on both adversarial and benign datasets (ANLI (training) + FeverNLI + MNLI + SNLI).
<table>
<thead>
<tr>
<th>Training</th>
<th>Model</th>
<th>Method</th>
<th>SNLI</th>
<th>MNLI (m/mm)</th>
<th>adv-SNLI (BERT)</th>
<th>adv-MNLI (BERT)</th>
<th>adv-SNLI (RoBERTa)</th>
<th>adv-MNLI (RoBERTa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RoBERTa</td>
<td>Vanilla</td>
<td>92.6</td>
<td>90.8/90.6</td>
<td>56.6</td>
<td>68.1/68.6</td>
<td>19.4</td>
<td>24.9/24.9</td>
</tr>
<tr>
<td>Standard Training</td>
<td></td>
<td>InfoBERT</td>
<td><strong>93.3</strong></td>
<td>90.5/90.4</td>
<td><strong>59.8</strong></td>
<td><strong>69.8/70.6</strong></td>
<td><strong>42.5</strong></td>
<td><strong>50.3/52.1</strong></td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>Vanilla</td>
<td>91.3</td>
<td>86.7/86.4</td>
<td>0.0</td>
<td>0.0/0.0</td>
<td>44.9</td>
<td>57.0/57.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>InfoBERT</td>
<td><strong>91.7</strong></td>
<td>86.2/86.0</td>
<td><strong>36.7</strong></td>
<td><strong>43.5/46.6</strong></td>
<td><strong>45.4</strong></td>
<td><strong>57.2/58.6</strong></td>
</tr>
<tr>
<td>Adversarial Training</td>
<td>RoBERTa</td>
<td>FreeLB</td>
<td>93.4</td>
<td>90.1/90.3</td>
<td>60.4</td>
<td>70.3/72.1</td>
<td>41.2</td>
<td>49.5/50.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>InfoBERT</td>
<td>93.1</td>
<td>90.7/90.4</td>
<td><strong>62.3</strong></td>
<td><strong>73.2/73.1</strong></td>
<td><strong>43.4</strong></td>
<td><strong>56.9/55.5</strong></td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>FreeLB</td>
<td><strong>92.4</strong></td>
<td>86.9/86.5</td>
<td><strong>46.6</strong></td>
<td><strong>60.0/60.7</strong></td>
<td><strong>50.5</strong></td>
<td><strong>64.0/62.9</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>InfoBERT</td>
<td>92.2</td>
<td>87.2/87.2</td>
<td><strong>50.8</strong></td>
<td><strong>61.3/62.7</strong></td>
<td><strong>52.6</strong></td>
<td><strong>65.6/67.3</strong></td>
</tr>
</tbody>
</table>

Table 3: Robust accuracy on the adversarial SNLI and MNLI(-m/mm) datasets generated by TextFooler based on blackbox BERT/RoBERTa (denoted in brackets of the header). Models are trained on the benign datasets (MNLI+SNLI) only.
### Evaluation of Model Robustness (III) - adv-SQuAD

<table>
<thead>
<tr>
<th>Training</th>
<th>Method</th>
<th>benign</th>
<th>AddSent</th>
<th>AddOneSent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Training</td>
<td>Vanilla</td>
<td>93.5/86.9</td>
<td>72.9/66.6</td>
<td>80.6/74.3</td>
</tr>
<tr>
<td></td>
<td>InfoBERT</td>
<td>93.5/87.0</td>
<td>78.5/72.9</td>
<td>84.6/78.3</td>
</tr>
<tr>
<td>Adversarial Training</td>
<td>FreeLB</td>
<td>93.8/87.3</td>
<td>76.3/70.3</td>
<td>82.3/76.2</td>
</tr>
<tr>
<td></td>
<td>ALUM</td>
<td>-</td>
<td>75.5/69.4</td>
<td>81.4/75.9</td>
</tr>
<tr>
<td></td>
<td>InfoBERT</td>
<td>93.7/87.0</td>
<td>78.0/71.8</td>
<td>83.6/77.1</td>
</tr>
</tbody>
</table>

Table 4: Robust F1/EM scores based on RoBERTa\textsubscript{Large} on the adversarial SQuAD datasets (AddSent and AddOne-Sent). Models are trained on standard SQuAD 1.0 dataset.
Thank you!

Paper

Code
https://github.com/AI-secure/InfoBERT