**Introductory Notes**

Adversarial Vulnerability of Language Models

Deep neural networks are known to be prone to adversarial examples, i.e., the outputs of neural networks can be arbitrarily wrong when human-interpretable adversarial perturbations are added to the inputs.

Textual adversarial attacks typically perform word-level substitution or sentence-level paraphrasing to achieve semantic/utility preservation that seems innocuous to human, while fools NLP models. Recent studies further show that even large-scale pre-trained language models (LM) such as BERT are vulnerable to adversarial attacks.

**Definition of Textual Adversarial Examples**

We mainly focus on the dominant word-level attack as the main threat model, since it

- achieves higher attack success
- is generally less noticeable to human readers than other attacks

Most word-level adversarial attacks constrain word perturbations via the bounded magnitude in the semantic embedding space.

By adapting from Jacobsen et al. (2019), we define the adversarial text examples with distortions constrained in the embedding space:

\[
\|x^\prime - x\|_2 \leq \epsilon
\]

(c-bounded Textual Adversarial Examples) Given a sentence \( x = [x_1, x_2, \ldots, x_n] \), where \( x_i \) is the word at \( i \)-th position, the \( c \)-bounded adversarial sentence \( x^\prime = [x_1, x_2, \ldots, x_n] \) for a classifier \( f \) satisfies:

1. \( f(x) = f(x^\prime) \) but \( f(x^\prime) \neq f(x) \), where \( f(x) \) is the oracle (e.g., human decision-maker).
2. \( \|x_i - x_i^\prime\|_2 \leq c \) for \( i = 1, 2, \ldots, n \), where \( c \geq 0 \) and \( t_i \) is the word embedding of \( x_i \).

**InfoBERT**

**Principle for Robust Representation Learning**

- Maximizing the mutual information between representation \( T \) and label \( Y \)
- Maximizing the mutual information between input \( X \) and representation \( T \)
- Maximizing the mutual information between local "robust" feature \( T_k \) and global feature \( Z \)

**Information Bottleneck as a Regularizer**

- General Information Bottleneck Objective as the maximization of the Lagrangian
- Localized Formulation of IB Objective

**Theorem 3.1 (Lower bound of \( \mathcal{L}_IB \))**

Given a sequence of random variables \( X = \{X_1, X_2, \ldots, X_n\} \) and a deterministic feature extractor \( f_\theta \), let \( T = \{T_1, \ldots, T_n\} = \{f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)\} \) and \( Y = \{y_1, \ldots, y_n\} \), then the localized formulation of IB \( \mathcal{L}_IB \) is a lower bound of \( \mathcal{E}_B \)

\[ \mathcal{L}_IB = I(Y; T) - \beta I(T; Y) \geq -\beta n \sum_{i=1}^n I(T_i; Y_i) \]

**Theorem 3.2 (Adversarial Robustness Bound)** For random variables \( X = \{X_1, X_2, \ldots, X_n\} \) and \( X = \{X_1', X_2', \ldots, X_n'\} \) and \( T = \{T_1, \ldots, T_n\} = \{f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)\} \) and \( T' = \{T_1', \ldots, T_n'\} = \{f_\theta(x_1'), f_\theta(x_2'), \ldots, f_\theta(x_n')\} \), then \( f_\theta \) is a deterministic feature extractor. The performance gap between benign and adversarial data \( I(Y; T) - I(Y; T') \) is bounded above by

\[ \|I(Y; T) - I(Y; T')\|_1 \leq \lambda_0 + \lambda_1 \sum_{i=1}^n \left( \frac{1}{\sqrt{\beta}} I(T_i; Y_i) + \frac{\beta}{\sqrt{\beta}} I(T_i; Y_i) \right) + \lambda_2 \sum_{i=1}^n \left( \frac{1}{\sqrt{\beta}} I(T_i; Y_i) + \frac{\beta}{\sqrt{\beta}} I(T_i; Y_i) \right) \]

where \( \lambda_0, \lambda_1, \lambda_2 \) are constants.

**Remark**

1. Adversarial performance gap \( I(Y; T) - I(Y; T') \) becomes closer when \( I(T_i; Y_i) \) decreases.
2. Combining adversarial training with IB regularizer can further minimize \( I(T_i; Y_i) \).

**Local Anchored Feature Regularizer**

Given a sentence \( x = \{x_1, x_2, \ldots, x_n\} \) and a deterministic feature extractor \( f_\theta \), let \( T = \{T_1, \ldots, T_n\} \), then the localized formulation of IB \( \mathcal{L}_IB \) is a lower bound of the mutual information between representation \( T \) and label \( Y \). In practice, we can use the final-layer [CLS] embedding of BERT to represent global sentence-level feature \( Z \).

**Ablation Studies**

Evaluation on adversarial SQuAD

In this paper, we propose a novel learning framework InfoBERT from an information theoretic perspective to perform robust fine-tuning over pre-trained language models.

**InfoBERT** consists of two novel regularizers to robustify the learned representations:

(a) Information Bottleneck Regularizer, learning to extract the approximated minimal sufficient statistics and denoise the excessive spurious features;
(b) Local Anchored Feature Regularizer, which improves the robustness of global features by aligning them with local anchored features.

**Conclusions**

In this paper, we propose a novel learning framework InfoBERT from an information theoretic perspective to perform robust fine-tuning over pre-trained language models.

**Experiments**

**Evaluation on ANLI**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>56.5</td>
<td>48.7</td>
</tr>
<tr>
<td>SMART</td>
<td>57.5</td>
<td>49.8</td>
</tr>
<tr>
<td>ALBER</td>
<td>58.4</td>
<td>50.6</td>
</tr>
<tr>
<td>ALEUR</td>
<td>59.4</td>
<td>51.6</td>
</tr>
</tbody>
</table>

**Evaluation against TextFooter**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>57.0</td>
<td>49.2</td>
</tr>
<tr>
<td>SMART</td>
<td>58.0</td>
<td>50.8</td>
</tr>
<tr>
<td>ALBER</td>
<td>59.0</td>
<td>51.8</td>
</tr>
<tr>
<td>ALEUR</td>
<td>60.0</td>
<td>52.6</td>
</tr>
</tbody>
</table>

**Evaluation on adversarial SQuAD**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>58.0</td>
<td>50.4</td>
</tr>
<tr>
<td>SMART</td>
<td>59.0</td>
<td>51.4</td>
</tr>
<tr>
<td>ALBER</td>
<td>60.0</td>
<td>52.6</td>
</tr>
<tr>
<td>ALEUR</td>
<td>61.0</td>
<td>53.8</td>
</tr>
</tbody>
</table>