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†

* University of Illinois at Urbana-Champaign, † Microsoft Research, ‡ Virginia Tech

ILLINOIS Computer Science GRAINGER COLLEGE OF ENGINEERING

COLLEGE OF ENGINEERING COMPUTER SCIENCE VIRGINIA TECH

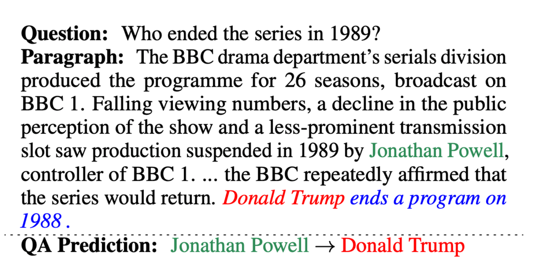
Microsoft

Introduction

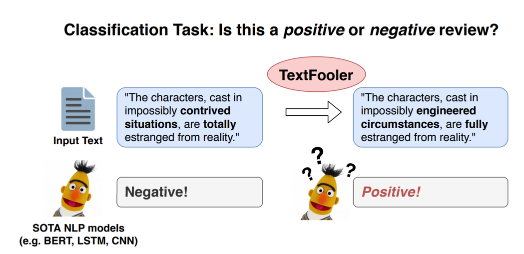
Adversarial Vulnerability of Language Models

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



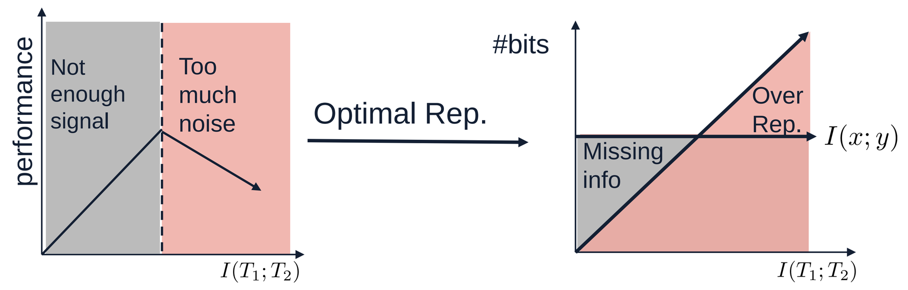
Adversarial examples for QA model



Adversarial examples for classification model

Representation Learning

Many studies have shown that self-supervised representation learning is essentially solving the problem of maximizing the mutual information (MI) I(X; T) between the input X and the representation T.



- Maximizing information is only useful in so far as that information is task-relevant
- Excessive noisy information and spurious features may incur adversarial attacks.

Goals

- Analyze the robustness of language models from an information theoretic perspective in a principled way
- Improve the robustness of language representations by fine-tuning both local features and global features

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.

(ϵ-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 ϵ-bounded adversarial sentence $x' = [x'_1; x'_2; \ldots; x'_n]$ for a classifier $\mathcal{F}$ satisfies:

(1) $\mathcal{F}(x) = o(x) = o(x')$ but $\mathcal{F}(x') \neq o(x')$, where $o(\cdot)$ is the oracle (e.g., human decision-maker);

(2) $||t_i - t'_i||_2 \leq \epsilon$ for $i = 1, 2, 3, \ldots, n$, where $\epsilon \geq 0$ and t_i is the word embedding of x_i.

InfoBERT

Principle for Robust Representation Learning

- 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_{k_j} and global feature Z

Information Bottleneck as a Regularizer

- General Information Bottleneck Objective as the maximization of the Lagrangian

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

- Localized Formulation of IB Objective

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

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_θ, 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 $\mathcal{L}_{\text{IB}}$.

$$I(Y;T) - \beta I(X;T) \geq I(Y;T) - n\beta \sum_{i=1}^{n} I(X_i;T_i).$$

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

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

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

where B_0, B_1, B_2, B_3 and B_4 are constants depending on the sequence length n, ϵ and $p(x)$.

Remark:

1. Adversarial performance gap $|I(Y;T) - I(Y;T')|$ becomes closer, when $I(X_i;T_i)$ decreases.
2. Combining adversarial training with IB regularizer can further minimize $I(X'_i;T'_i)$.

Task Objective

Complete Objective: $\max \boxed{I(Y;T)} - n\beta \sum_{i=1}^{n} I(X_i;T_i) + \alpha \sum_{j=1}^{M} I(T_{k_j};Z)$

Local Anchored Feature Regularizer

Step 1: Locate the local anchored features by filtering out non-robust and unuseful 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 δ*
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 $||g(\delta)_1||_2, ||g(\delta)_2||_2, \ldots, ||g(\delta)_n||_2$ into $||g(\delta)_{k_1}||_2, ||g(\delta)_{k_2}||_2, \ldots, ||g(\delta)_{k_n}||_2$ in ascending order, where z_i corresponds to its original index.
5: **Return:** $k_i, k_{i+1}, \ldots, k_j$, where $c_l \leq \frac{i}{n} \leq \frac{j}{n} \leq c_h$.

Step 2: Improve the robustness of the global feature Z by aligning it with the local anchored features T_{k_j}

- In practice, we can use the final-layer [CLS] embedding of BERT to represent global sentence-level feature Z
- Use information theoretic tool to increase the mutual information $I(T_{k_j};Z)$ between local anchored feature T_{k_j} and global feature Z, so that Z can share more robust information

Experiments

Evaluation on ANLI

Training	Model	Method	Dev A1	Dev A2	Dev A3	Dev ANLI	Test A1	Test A2	Test A3	Test ANLI
Standard Training	RoBERTa	Vanilla	74.1	50.8	43.9	55.5	73.8	48.9	44.4	53.7
		InfoBERT	75.2	49.6	47.8	**56.9**	73.9	50.8	48.8	**57.3**
	BERT	Vanilla	58.5	46.1	45.5	49.8	57.4	48.3	43.5	49.3
		InfoBERT	59.3	48.9	45.5	**50.9**	60.0	46.9	44.8	**50.2**
Adversarial Training	RoBERTa	FreeLB	75.2	47.4	45.3	55.3	73.3	50.5	46.8	56.2
		SMART	74.5	50.9	47.6	57.1	72.4	49.8	50.3	57.1
		ALUM	73.3	53.4	48.2	57.7	72.3	52.1	48.4	57.0
		InfoBERT	76.4	51.7	48.6	**58.3**	75.5	51.4	49.8	**58.3**
	BERT	FreeLB	60.3	47.1	46.3	50.9	60.3	46.8	44.8	50.2
		ALUM	62.0	48.6	48.1	**52.6**	61.3	45.9	44.3	50.1
		InfoBERT	60.8	48.7	45.9	51.4	63.3	48.7	43.2	**51.2**

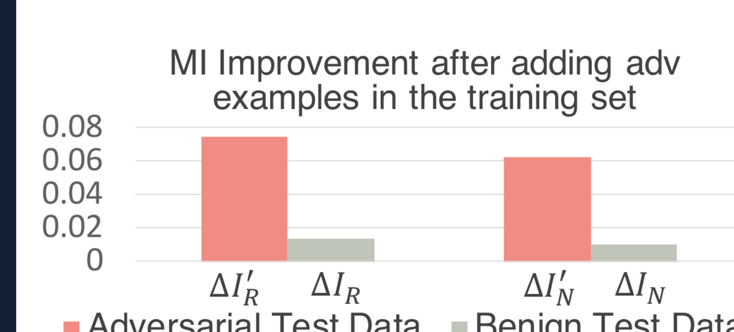
Evaluation against TextFooler

Training	Model	Method	SNLI	MNLI (m/mm)	adv-SNLI (BERT)	adv-MNLI (BERT)	adv-SNLI (RoBERTa)	adv-MNLI (RoBERTa)
Standard Training	RoBERTa	Vanilla	92.6	**90.8/90.6**	56.6	68.1/68.6	19.4	24.9/24.9
		InfoBERT	**93.3**	90.5/90.4	**59.8**	**69.8/70.6**	**42.5**	**50.3/52.1**
	BERT	Vanilla	91.3	**86.7/86.4**	0.0	0.0/0.0	44.9	57.0/57.5
		InfoBERT	**91.7**	86.2/86.0	**36.7**	**43.5/46.6**	**45.4**	**57.2/58.6**
Adversarial Training	RoBERTa	FreeLB	**93.4**	90.1/90.3	60.4	70.3/72.1	41.2	49.5/50.6
		InfoBERT	93.1	**90.7/90.4**	**62.3**	**73.2/73.1**	**43.4**	**56.9/55.5**
	BERT	FreeLB	**92.4**	86.9/86.5	46.6	60.0/60.7	50.5	64.0/62.9
		InfoBERT	92.2	**87.2/87.2**	**50.8**	**61.3/62.7**	**52.6**	**65.6/67.3**

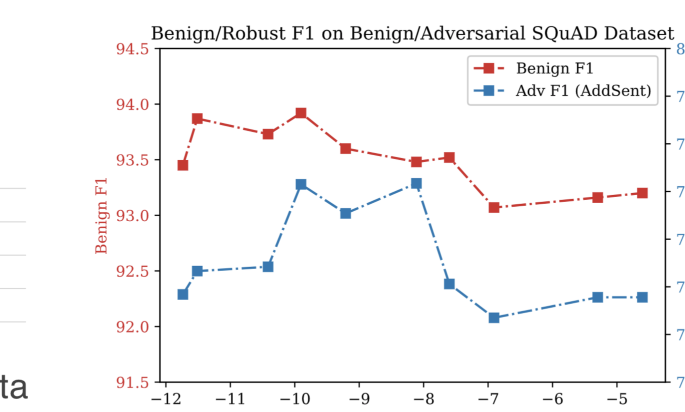
Evaluation on adversarial SQuAD

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

Ablation Studies



Local anchored features contribute more to MI improvement than nonrobust/unuseful features, unveiling closer relation with robustness.



Adversarial robustness improves by decreasing the mutual information between input and representation without affecting the benign accuracy much, until aggressive compression that leads to both performance drop.

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.

InfoBERT consists of two novel regularizers to improve the robustness of 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.

ILLINOIS