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

ILLINOIS
Computer Science
GRAINGER COLLEGE OF ENGINEERING

COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH.

Microsoft

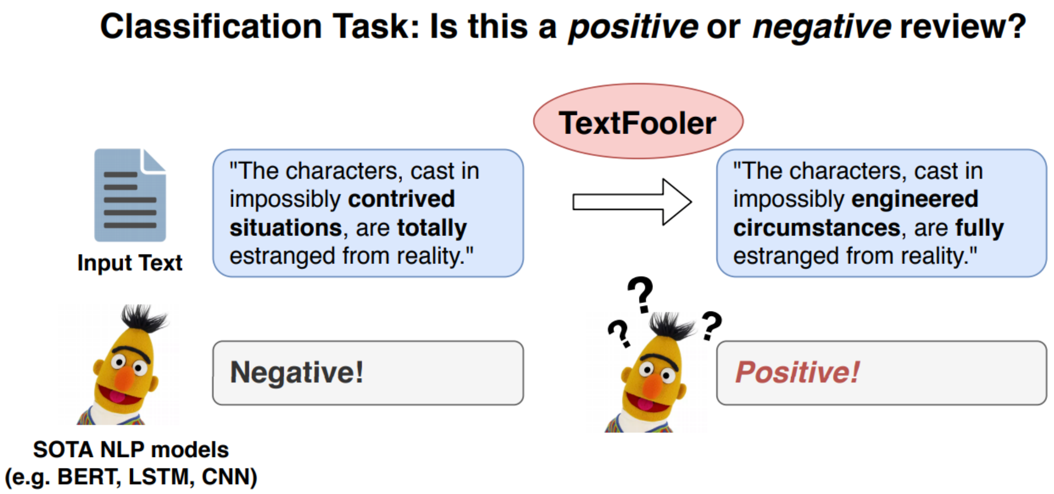
Adversarial Attacks in NLP

- 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 .*

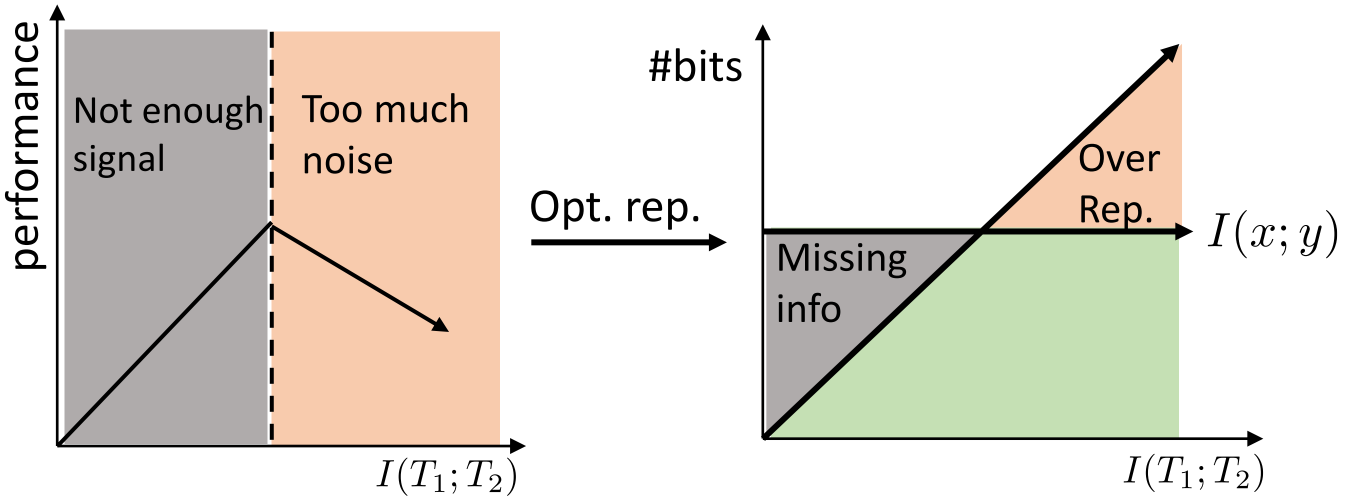
QA Prediction: Jonathan Powell → Donald Trump

- Adversarial examples for classification tasks [2]



[1] Wang, Boxin, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang and Bo Li. "T3: Tree-Autoencoder Constrained Adversarial Text Generation for Targeted Attack." EMNLP (2020).
[2] Jin, Di, Zhijing Jin, Joey Tianyi Zhou and Peter Szolovits. "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment." *AAAI* (2020).

Understanding ML Robustness from the Information-Theoretic Perspectives



[1] Tian et al. What makes for good views for contrastive learning. NeurIPS (2020)

Robust Representation Learning for LM

Task Objective

- Goals:
  - Maximize the mutual information between representation T and label Y
  - Minimize the mutual information between input X and representation T

Information Bottleneck Regularizer

  - Maximize the mutual information between local “robust” feature T_{k_j} and global feature Z

IB Regularizer

- Information Bottleneck (IB) As a Regularizer

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

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

$$I(Y;T) - \beta I(X;T) \geq I(Y;T) - n\beta \sum_{i=1}^{n} I(X_i;T_i). \tag{7}$$

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* $X = [X_1; X_2; ...; X_n]$ *and* $X' = [X'_1; X'_2; ...; X'_n]$*, let* $T = [T_1; T_2; ...; T_n] = [f_\theta(X_1); f_\theta(X_2); ...; f_\theta(X_n)]$ *and* $T' = [T'_1; T'_2; ...; T'_n] = [f_\theta(X'_1); f_\theta(X'_2); ...; f_\theta(X'_n)]$ *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*

$$|I(Y;T) - I(Y;T')| \leq B_0 + B_1 \sum_{i=1}^{n} \sqrt{|\mathcal{T}|}(I(X_i;T_i))^{1/2} + B_2 \sum_{i=1}^{n} |\mathcal{T}|^{3/4}(I(X_i;T_i))^{1/4}$$
$$+ B_3 \sum_{i=1}^{n} \sqrt{|\mathcal{T}|}(I(X'_i;T'_i))^{1/2} + B_4 \sum_{i=1}^{n} |\mathcal{T}|^{3/4}(I(X'_i;T'_i))^{1/4}, \quad (8)$$

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

IB Regularizer Verification

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

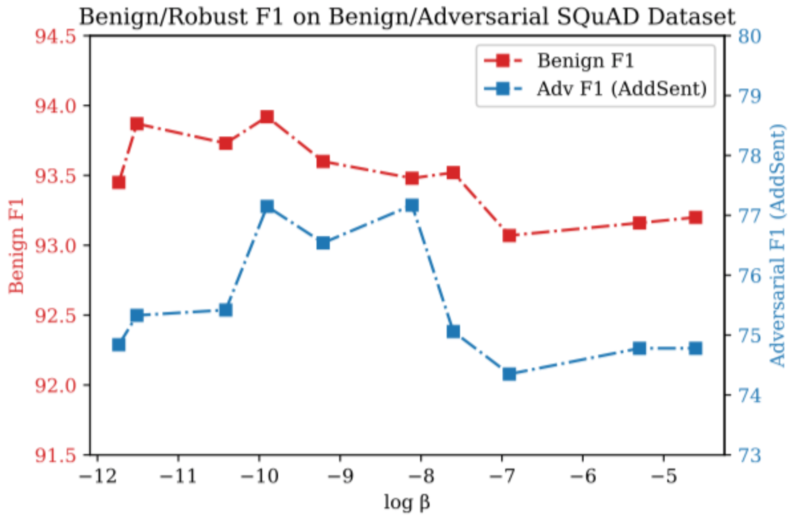


Figure 2: Benign/robust F1 score on benign/adversarial QA datasets. Models are trained on the benign SQuAD dataset with different β.

Robust Representation Learning for LM

- Goals:
  - Maximize the mutual information between representation T and label Y

    Representation Learning

  - Minimize the mutual information between input X and representation T

    Information Bottleneck Regularizer

  - Maximize the mutual information between local “robust” representation T_{k_j} and global representation Z

    Local Anchored Feature Regularizer

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.

Input: Word local features t, upper and lower threshold c_h and c_l

$\delta \leftarrow 0$ *// Initialize the perturbation vector δ*

$g(\delta) = \nabla_\delta \ell_{\text{task}}(q_\psi(t + \delta), y)$ *// Perform adversarial attack on the embedding space*

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.

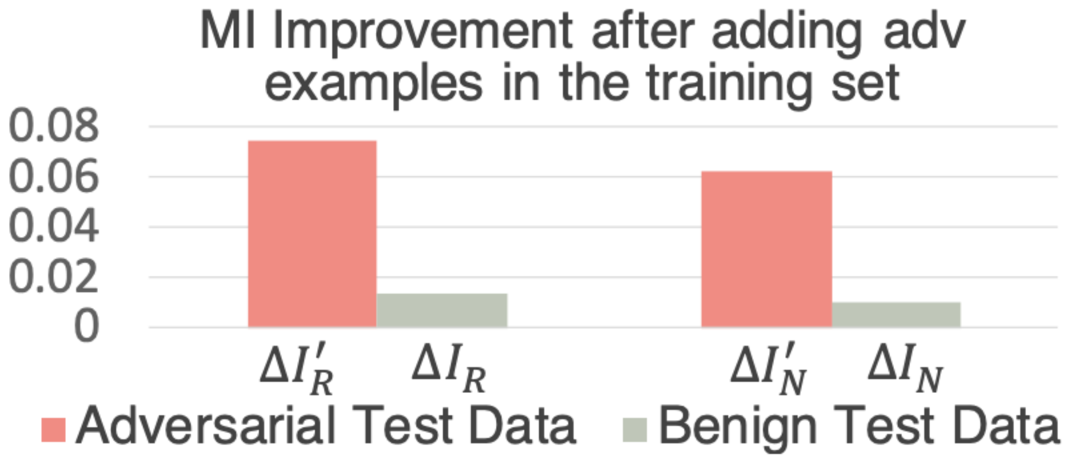
Return: $k_i, k_{i+1}, ..., k_j$, where $c_l \leq \frac{i}{n} \leq \frac{j}{n} \leq c_h$.

- Increase the Mutual Information between local anchored features T_{k_j} and global features Z

$$\max \sum_{j=1}^{M} I(T_{k_j}; Z)$$

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

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

Task objective

Information Bottleneck Regularizer

Local Anchored Feature Regularizer

- 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

[1] Cheng, Pengyu, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan and L. Carin. "CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information." ICML (2020).

Experiments

- Evaluation against Different Adversarial Attacks
  - Natural Language Inference (NLI)
    - ANLI
    - TextFooler
  - Question Answering
    - adv-SQuAD

Evaluation of Model Robustness (I) -ANLI

Training	Model	Method	Dev				Test			
			A1	A2	A3	ANLI	A1	A2	A3	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**

Table 2: Robust accuracy on the ANLI dataset. Models are trained on both adversarial and benign datasets (ANLI (training) + FeverNLI + MNLI + SNLI).

Evaluation of Model Robustness (II) - 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**

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

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**

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

ILLINOIS Computer Science
GRAINGER COLLEGE OF ENGINEERING

Paper

https://arxiv.org/abs/2010.02329

Code

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

Thank you!