Detecting Textual Adversarial Examples Based on Distributional Characteristics of Data Representations

Na Liu*, Mark Dras, Wei Emma Zhang

Macquarie University, The University of Adelaide

na.liu8@students.mq.edu

Introduction

Adversarial Examples are constructed by carefully designed small perturbations of normal examples, that can fool a deep neural network to make wrong predictions. [3]

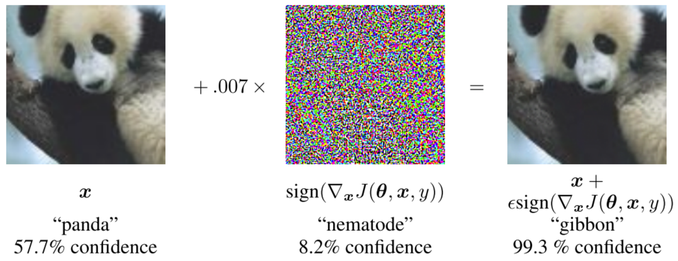

Figure 1. An example of an adversarial example in image[1]

	Example	Prediction
Original	This is a story of two misfits who don't stand a chance alone, but together they are magnificent.	Positive
Character-level	TZyTis is a sotry of two misifts who don't stad a ccange alUone, but tpgthr they are mgnificent.	Negative
Word-level	This is a conte of two who don't stands a opportunities alone, but together they are opulent.	Negative
Phrase-level	Why don't you have two misfits who don't stand a chance alone, but together they're beautiful.	Negative
Sentence-level	This is a story of two misfits who don't stand a chance alone, but together they are magnificent. ready south hundred at size expected worked whose turn poor.	Negative

Table 1. Examples of textual adversarial instances on a sentiment analysis task

Research Objective

Textual Adversarial Example Defence

- **Proactive defence methods**: increasing the robustness of deep neural networks (adversarial training).
- **Reactive defence methods***: distinguishing real from adversarial examples (FGWS).

Approaches

Adapted Local Intrinsic Dimensionality (LID)

Local Intrinsic Dimensionality [2]

$$LID_F(r) \triangleq \lim_{\epsilon \to 0} \frac{ln(F((1+\epsilon)\cdot r)/F(r))}{ln(1+\epsilon)} = \frac{r \cdot F'(r)}{F(r)} \tag{1}$$

The Maximum Likelihood Estimator of the LID [2]

$$\widehat{LID}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k} \log \frac{r_i(x)}{r_k(x)}\right)^{-1} \tag{2}$$

If neighbors of a reference sample x are compact, its estimated LID from Equation (2) is smaller, otherwise, its estimated LID is bigger.

Adaptation

- The x in the Equation (2) is a representation of an input text from a layer's hidden state of the target ($\text{BERT}_{\text{BASE}}$) model.
- An input of a detection classifier for an example is a 12-dimensional vector, where each element illustrates the corresponding layer's $\widehat{LID}(x)$.

MultiDistance Representation Ensemble Method (MDRE)

Intuitions

- Samples with a same predicted label from a deep neural net lie on a data submanifold.
- An adversarial example is generated because perturbations cause a correctly predicted example to transfer from one data submanifold to another.
- It is an out-of-distribution sample relative to training examples from its data submanifold.

MDRE

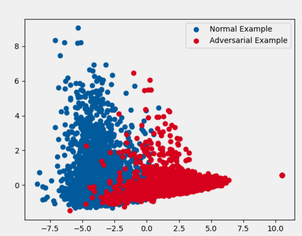

Figure 2. PCA results of MDRE inputs for the IMDB phreas level attack

- $d(x', \text{nearest neighbor of } x') > d(x, \text{nearest neighbor of } x)$
- ensemble learning can help identify this.

*nearest neighbor is among training examples with the same predicted label as x' or x.

Experiments

Tasks & Datasets

Dataset	Training.	Validation.	Testing.	Correctly Predicted Test Examples	Adversarial/Original Examples		
					Character.	Word.	Phrase.
IMDB	20,000	5,000	25,000	23,226	12,299	9,627	6,315
MultiNLI	314,162	78,540	9,832	8,062	7,028	3,240	4,340

Results

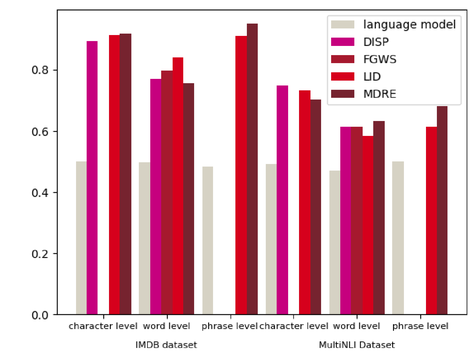

Figure 3. The accuracy for detection classifiers

Conclusion and Future Work

- Adapted LID and MDRE help to detect adversarial examples.
- Exploring more effective distribution characteristics of data semantic representations among adversarial and normal examples, may help to build better detectors.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2014.

[2] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality, 2018.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks, 2013.