AAAI 2020 (Oral)

Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment

Di Jin (MIT),* **Zhijing Jin (HKU),*** Joey Tianyi Zhou (A*STAR), Peter Szolovits (MIT)

ArXiv Link

Speaker: Zhijing Jin

Introduction:

B.Eng. (1st Class Honor) University of Hong Kong

Research Scientist Intern, Amazon Shanghai AI Lab

Collaborators:

Publications:

1. EMNLP-IJCNLP 2019: IMaT: Unsupervised Text Attribute Transfer via Iterative Matching and Translation.
2. NAACL 2019: GraphIE: A Graph-Based Framework for Information Extraction.
3. AAHPM 2020: Deep Natural Language Processing to Identify Symptom Documentation in Clinical Notes for Patients with Heart Failure Undergoing Cardiac Resynchronization Therapy
4. (Submitted to IJCAI 2020) Unsupervised Domain Adaptation for Neural Machine Translation with Iterative Back Translation.
5. (Submitted to ACL 2020) Hooks in the Headline: Learning to Generate Headlines with Controlled Styles.

1-Sentence Takeaway Message

Most NLP Models are **very weak against paraphrases.**

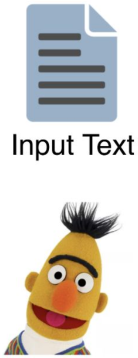
We should promote learnings that captures the **real casual relationships** in data.

Adversarial training is also a good choice.

Text Classification and Natural Language Inference (NLI)

Classification

Classify the text according to their attributes (e.g. sentiment, news category, authenticity)



Input Text

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

Negative!

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

NLI

Recognize the entailment in sentence pairs. (Three labels: entailment, contradiction, and neutral)

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping

Recent Advances in Text Classification and NLI

		Classification	NLI
BERT	2018	BERT, XLNet, RoBERTa	BERT, XLNet, RoBERTa
Re-emergence of RNN- and CNN-based Methods	2013	RNN, DCNN, RCNN, C-LSTM, Tree-LSTM	InferSent, ESIM
Traditional methods	1998	NBOW, MNB, SVM	BoW, MANLI aligner, DepTree aligner, NatLog, Categorical Calculus

Performance on Datasets

- Classification accuracy (%) on 7 datasets show really high performance of NN

Classification	WordCNN	WordLSTM	BERT
AG	92.5	93.1	94.6
Fake	99.9	99.9	99.9
MR	79.9	82.2	85.8
IMDB	89.7	91.2	92.2
Yelp	95.2	96.6	96.1
NLI	**InferSent**	**ESIM**	**BERT**
SNLI	84.6	88.0	90.7
MultiNLI	71.1/71.5	76.9/76.5	83.9/84.1

Neural Network models has (seemingly) conquered classification tasks

Overall good performance on NLI, especially BERT

Performance on Datasets

- Classification accuracy (%) on 7 datasets show really high performance of NN

Does high accuracies (e.g. 90+%) mean that NN conquers the two tasks as humans do?

Classification

A

F

IM

Ye

NLI

SNLI	84.		
MultiNLI	71.1/71.5		

dels has

ed

erall good performance on NLI, especially BERT

Motivation (Adversarial Attacking)

Sentence classification task: We ask the model "Is this a *positive* or *negative* review?".

Adversarial Text Generator

Input Text

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

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

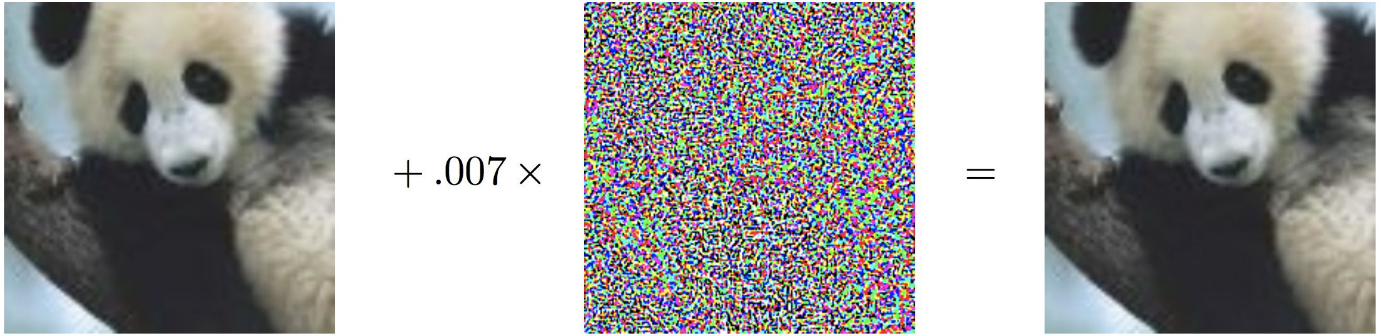
Negative!

Positive!

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

Why is NLP Adversarial Attack hard?

Q1: Why can't we borrow from CV?



$$\boldsymbol{x} + .007 \times \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)) = \boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$$

"panda" 57.7% confidence | "nematode" 8.2% confidence | $\boxed{\text{“gibbon”}}$ 99.3 % confidence

By adding an unnoticeable perturbation, "panda" is classified as "gibbon".
(Image Credit: (Goodfellow et al., 2014b))

Why is NLP Adversarial Attack hard?

Q1: Why can't we borrow from CV?

	CV	**NLP**
Input Type	Continuous	Discrete
Perceivable?	Unperceivable	Perceivable
Semantic?	Semantic-less	Semantic-sensi tive

Gradient-based adversarial attacks → Invalid characters/ word sequences

Q2: What have other NLP people done?

- Word 2018;

-> Unnatural sentences

- What about edited text?
- What about BERT attacks?

Problem Formulation

Problem Formulation

Given a corpus of N sentences $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$, and a corresponding set of N labels $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$, we have a pre-trained model $F : \mathcal{X} \rightarrow \mathcal{Y}$, which maps the input text space $\mathcal{X}$ to the label space $\mathcal{Y}$.

For a sentence $X \in \mathcal{X}$, a valid adversarial example X_{adv} should conform to the following requirements:

$$F(X_{\text{adv}}) \neq F(X), \text{ and } \mathrm{Sim}(X_{\text{adv}}, X) \geq \epsilon, \qquad (1)$$

where $\mathrm{Sim} : \mathcal{X} \times \mathcal{X} \rightarrow (0, 1)$ is a similarity function and ϵ is the minimum similarity between the original and adversarial examples. In the natural language domain, $\mathrm{Sim}(\cdot)$ is often a semantic and syntactic similarity function.

No access to model parameters!

Method (TextFooler)

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

...	impossibly	contrived	situations	,	...
1. Importance Ranking	0.1	0.3	0.4	0.01	

Leave-one-out Method:

situations

0.4

X = The characters, cast in impossibly contrived situations, are totally estranged from reality.

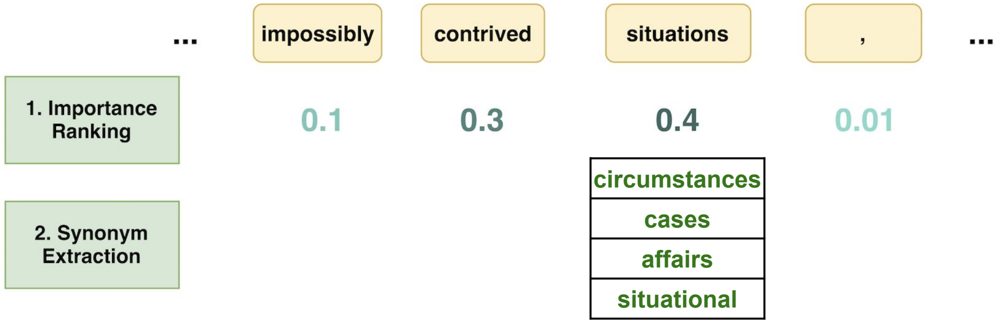
$\mathbf{X}_{\backslash situations}$ = The characters, cast in impossibly contrived , are totally estranged from reality.

$F_Y(X)$: the prediction score of X for the Y label (gold label)

$I_{sitations} = F_Y(X) - F_Y(X_{\backslash sitations}) = 0.4$

Method (TextFooler)

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



$$cos(Embedding(\text{situations}), Embedding(\text{x})) > \delta$$

Method (TextFooler)

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

	...	impossibly	contrived	situations	,	...
1. Importance Ranking		0.1	0.3	0.4	0.01	
2. Synonym Extraction				circumstances cases affairs situational		
3. Part-of-Speech Checking				circumstances cases affairs ~~situational~~		
4. Semantic Scoring				circumstances cases ~~affairs~~		

$$Sim(X, X')=cos(USE(X), USE(X'))> \epsilon$$

Method (TextFooler)

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

	...	impossibly	contrived	situations	,	...
1. Importance Ranking		0.1	0.3	0.4	0.01	
2. Synonym Extraction				circumstances cases affairs situational		
3. Part-of-Speech Checking				circumstances cases affairs ~~situational~~		
4. Semantic Scoring				circumstances cases		

X = The characters, cast in impossibly contrived situations, are totally estranged from reality.

X' = The characters, cast in impossibly contrived circumstances, are totally estranged from reality.

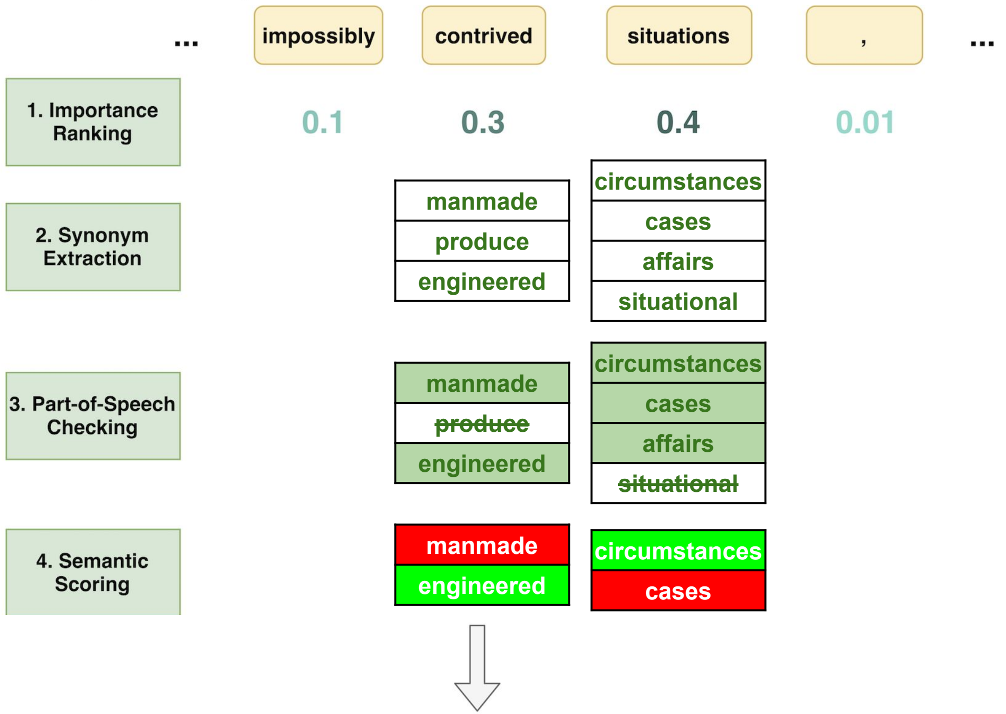
X'' = The characters, cast in impossibly contrived cases, are totally estranged from reality.

Sim(X, X') v.s. *Sim*(X, X'')

F(X') v.s. *F*(X'')

Method (TextFooler)

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



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

Datasets

- Datasets: We use 5 classification datasets, and 2 text entailment datasets
- Samples to attack: 1000 samples were randomly selected from the test set of each dataset

Task	Dataset	Train	Test	Avg Len
Classification	AG's News	30K	1.9K	43
	Fake News	18.8K	2K	885
	MR	9K	1K	20
	IMDB	25K	25K	215
	Yelp	560K	38K	152
Entailment	SNLI	570K	3K	8
	MultiNLI	433K	10K	11

Table 1: Overview of the datasets.

5 Classification Datasets

To study the robustness of our model, we use text classification datasets with various properties, including **news topic classification**, **fake news detection**, and **sentence**- and **document**-level **sentiment analysis**, with average text length ranging from **tens** to **hundreds of words**.

AG's News (AG)

Yelp Polarity (Yelp)

MR

IMDB

Fake News Detection (Fake)

Target models to attack

- Classification: WordCNN (Kim 2014), WordLSTM, BERT (Devlin et al. 2018)
- Entailment: InferSent (Conneau et al. 2017), ESIM (Chen et al. 2016), and fine-tuned BERT

	WordCNN	WordLSTM	BERT
AG	92.5	93.1	94.6
Fake	99.9	99.9	99.9
MR	79.9	82.2	85.8
IMDB	89.7	91.2	92.2
Yelp	95.2	96.6	96.1
	InferSent	**ESIM**	**BERT**
SNLI	84.6	88.0	90.7
MultiNLI	71.1/71.5	76.9/76.5	83.9/84.1

Table 2: Original accuracy of target models on standard test sets.

Neural Network models has (seemingly) conquered classification tasks

BERT scores very high on NLI

Results: Automatic Evaluation

Q: Given the black-box model, what can we measure?

Original Accuracy — Original Accuracy

% Perturbed Words

Query Number

Results: Automatic Evaluation (Classification)

- **Original Accuracy:** Model accuracy before attacking
- **After-Attack Accuracy:** Model accuracy on adversarial examples
- **% Perturbed Words:** Percentage of words in text that are replaced
- **Semantic Similarity:** Similarity score between original text and adversary using USE
- **Query Number:** Number of queries sent to the model

	WordCNN					WordLSTM					BERT				
	MR	IMDB	Yelp	AG	Fake	MR	IMDB	Yelp	AG	Fake	MR	IMDB	Yelp	AG	Fake
Original Accuracy	78.0	89.2	93.8	91.5	96.7	80.7	89.8	96.0	91.3	94.0	86.0	90.9	97.0	94.2	97.8
After-Attack Accuracy	2.8	0.0	1.1	1.5	15.9	3.1	0.3	2.1	3.8	16.4	11.5	13.6	6.6	12.5	19.3
% Perturbed Words	14.3	3.5	8.3	15.2	11.0	14.9	5.1	10.6	18.6	10.1	16.7	6.1	13.9	22.0	11.7
Semantic Similarity	0.68	0.89	0.82	0.76	0.82	0.67	0.87	0.79	0.63	0.80	0.65	0.86	0.74	0.57	0.76
Query Number	123	524	487	228	3367	126	666	629	273	3343	166	1134	827	357	4403
Average Text Length	20	215	152	43	885	20	215	152	43	885	20	215	152	43	885

Results: Automatic Evaluation (NLI)

- Up to -85.4% accuracy change after our attack

	InferSent		ESIM		BERT	
	SNLI	MultiNLI (m/mm)	SNLI	MultiNLI (m/mm)	SNLI	MultiNLI (m/mm)
Original Accuracy	84.3	70.9/69.6	86.5	77.6/75.8	89.4	85.1/82.1
After-Attack Accuracy	3.5	6.7/6.9	5.1	7.7/7.3	4.0	9.6/8.3
% Perturbed Words	18.0	13.8/14.6	18.1	14.5/14.6	18.5	15.2/14.6
Semantic Similarity	0.50	0.61/0.59	0.47	0.59/0.59	0.45	0.57/0.58
Query Number	57	70/83	58	72/87	60	78/86
Average Text Length	8	11/12	8	11/12	8	11/12

Results: Human Evaluation

Q1: Given the human resources, what do we want to measure?

Grammar, Classification accuracy, Semantic Similarity

Q2: How do we measure them?

Grammar

Sentence 1 -> **Score**
Sentence 2 -> **Score**
...
shuffle!

Classification

Sentence 1 -> **Label**
Sentence 2 -> **Label**
...
shuffle!

Semantic Similarity

shuffle!
Sent 1a &1b -> **Same?**
Sent 2a &2b -> **Same?**
...

Q3: What else should we take care of?

Results: Human Evaluation

- Grammar: We ask human annotators to rate Grammaticality on a Likert of 1-5, and calculate avg_score_attacked / avg_score_original
- Classification Accuracy: Let human annotate the adversarial examples and compare with original labels
- Semantic Similarity: Let human judge whether the adversarial example is semantically the same as the original sentence, and calculate the percentage

	MR (WordLSTM)	SNLI (BERT)
Grammar	95%	95%
Classification Accuracy	92%	85%
Semantic Similarity	91%	86%

Qualitative results

Grammar	95%	95%
Classification	92%	85%
Semantic	91%	86%

Movie Review (Positive (POS) ↔ Negative (NEG))	
Original (Label: NEG)	The characters, cast in impossibly ***contrived situations***, are ***totally*** estranged from reality.
Attack (Label: POS)	The characters, cast in impossibly ***engineered circumstances***, are ***fully*** estranged from reality.
Original (Label: POS)	It cuts to the ***knot*** of what it actually means to face your ***scares***, and to ride the ***overwhelming* metaphorical wave** that life wherever it takes you.
Attack (Label: NEG)	It cuts to the ***core*** of what it actually means to face your ***fears***, and to ride the ***big* metaphorical wave** that life wherever it takes you.
SNLI (Entailment (ENT), Neutral (NEU), Contradiction (CON))	
Premise	Two small boys in blue soccer uniforms use a wooden set of steps to wash their hands.
Original (Label: CON)	The boys are in band ***uniforms***.
Adversary (Label: ENT)	The boys are in band ***garment***.
Premise	A child with wet hair is holding a butterfly decorated beach ball.
Original (Label: NEU)	The ***child*** is at the ***beach***.
Adversary (Label: ENT)	The ***youngster*** is at the ***shore***.

Table 6: Examples of original and adversarial sentences from MR (WordLSTM) and SNLI (BERT) datasets.

Comparison with Benchmark

- **Success Rate:** Percentage of successful attacks
- Up to **25.8%** improvement on the success rate
- Up to **5.0%** improvements on perturbed words

Dataset	Model	Success Rate	% Perturbed Words
IMDB	(Li et al. 2018)	86.7	6.9
	(Alzantot et al. 2018)	97.0	14.7
	Ours	**99.7**	**5.1**
SNLI	(Alzantot et al. 2018)	70.0	23.0
	Ours	**95.8**	**18.0**
Yelp	(Kuleshov et al. 2018)	74.8	-
	Ours	**97.8**	**10.6**

Ablation (Step 1: Word Importance Ranking)

- After removing Step 1 and instead randomly selecting the words to perturb, the after-attack accuracy increases by more than 45% on all three datasets
- In this table, BERT model is used as the target model.

	MR	AG	SNLI
% Perturbed Words	16.7	22.0	18.5
Original Accuracy	86.0	94.2	89.4
After-Attack Accuracy	11.5	12.5	4.0
After-Attack Accuracy (Random)	**68.3**	**80.8**	**59.2**

Ablation (Step 2: Semantic Similarity Constraint)

- In Step 2 of Algorithm 1, for every possible word replacement, we check the semantic similarity, and apply a similarity threshold.
- In this table, BERT model is used as the target model.

original / - Sim.

	MR	IMDB	SNLI	MNLI(m)
After-Attack Accu.	11.5/6.2	13.6/11.2	4.0/3.6	9.6/7.9
% Perturbed Words	16.7/14.8	6.1/4.0	18.5/18.3	15.2/14.5
Query Number	166/131	1134/884	60/57	78/70
Semantic Similarity	0.65/0.58	0.86/0.82	0.45/0.44	0.57/0.56

Ablation (Step 2: Semantic Similarity Constraint)

- In Step 2 of Algorithm 1, for every possible word replacement, we check the semantic similarity, and apply a similarity threshold.
- In this table, BERT model is used as the target model.

Original	like a south of the border melrose ***place***
Adversarial	like a south of the border melrose ***spot***
- Sim.	like a south of the border melrose ***mise***
Original	their computer animated faces are very ***expressive***
Adversarial	their computer animated face are very ***affective***
- Sim.	their computer animated faces are very ***diction***

Transferability

- Transferability of adversarial text: whether adversarial samples curated based on one model can also fool another
- Attacking against BERT shows higher transferability than others
- Transferability in Entailment > Transferability in Classification

		WordCNN	WordLSTM	BERT
IMDB	WordCNN	—	84.9	90.2
	WordLSTM	74.9	—	87.9
	BERT	84.1	85.1	—
		InferSent	**ESIM**	**BERT**
SNLI	InferSent	—	62.7	67.7
	ESIM	49.4	—	59.3
	BERT	58.2	54.6	—

Adversarial Training

- **Af. Acc.:** after-attack accuracy
- **Pert.:** percentage of perturbed words
- **Adv. Training:** add adversarial examples to original data and re-train the model
- We can enhance the robustness of a model to future attacks by training it with the generated adversarial examples

	MR		SNLI	
	Af. Acc.	**Pert.**	**Af. Acc.**	**Pert.**
Original	11.5	16.7	4.0	18.5
+ Adv. Training	**18.7**	**21.0**	**8.3**	**20.1**

Contributions

1. TextFooler: a **simple but strong** baseline for black-box language adversarial attack. Code is at 👉
2. Revealed that on text classification and entailment, most state-of-the-art NLP models (BERT, LSTM, and CNN) are **delicate against simple adversarial attacks**.
3. Successfully degraded BERT's performance by **-74.5% to -90.4%** on five classification datasets and **-80.8% to -85.4%** on two NLI datasets.
4. A comprehensive **four-way automatic** and **three-way human** evaluation.

Takeaway Messages

1. Current NLP models pays attention to peripheral correlations such as specific words and their cooccurrences. Thus they are **weak against paraphrases**.
2. We should promote learnings that captures the **real casual relationships** in data.
3. **Adversarial training** can increase the model robustness.




ArXiv Link of our paper:

Contact Email:
zhijing.jin@connect.hku.hk

Thank you!

Feel free to contact me for idea brainstorming.