How Vulnerable are Pre-trained Language Models?

Hannah Chen, Yangfeng Ji, David Evans (University of Virginia)

Motivation

Existing works on adversarial attacks in the text domain have shown great success by crafting semantically and syntactically similar adversarial examples. However, most of them evaluated the attack on small and shallow models that contains only a few feedforward or LSTM layers. While many current state-of-the-art NLP models can be easily constructed by fine-tuning pre-trained language models like ELMo, Open AI GPT, and BERT, we would like to know whether these pre-trained models are also susceptible to adversarial examples. This work studies the adversarial robustness of the pre-trained language models, specifically the BERT models, and compare the results with shallow LSTM models. We also demonstrate the transferability of the adversarial examples that generated for both models.

films
It is sad what they are letting into ~~film~~ festivals these days. I had to sit through over twenty minutes of this dreary short

that wasn't funny at all to get a good seat for a feature film that I wanted to see at a local film festival. The festival

horrifying phenomenal delighted
planners paired this ~~horrible~~ short with a ~~great~~ ~~feature~~. I am just ~~glad~~ the ~~feature~~ was good, otherwise I would have not
horrendous trait characteristics

been a very happy camper! For a comedy short film it got no laughs. The title says it all.

	Negative → Positive	
LSTM	85.46%	73.26%
BERT	98.44%	78.25%

Adversarial Robustness Against Genetic Attacks

Model		Test Accuracy (%)		Attack Success Rate (%)	% of Modifications	
		>=50%	>=70%		Mean	Median
LSTM	v1	88.74	85.02	93.5	11.0	9.6
	v2	89.74	83.53	36.75	12.8	12.5
BERT	v1	88.33	85.86	45.5	10.7	10.8
	v2	88.82	82.69	27.0	13.7	13.0

Note: (1) we count the attack as successful if the model predict with >=70% confidence score for the target label. (2) v1 was validated and attacked on the same set of data, and v2 used two different sets of data to validate and perform attacks.

Genetic Attack (Alzantot et al., EMNLP 2018): Search & replace words with K nearest neighbors in the GloVe embedding space.

What we've found:

1. BERT appears to be more robust than the LSTM model and requires more amount of perturbations to attack successfully.
2. v2 models have less confidence in predicting labels, which may be the reason why the attack success rates decrease significantly.

Shallow vs. Deep Pre-training

Pre-trained Word Embeddings	Pre-trained Language Models
Low dimensional word representations	High dimensional & contextualized word representations
Fail to disambiguate word meanings in different context settings	More robust to word substitutions that do not perfectly fit within the context

Adversarial Examples

Hilarious **film**, I had a great time watching it. the star (Cuneyt Arkin, sometimes credited as Steve Arkin) is a popular actor from turkey. He has played in lots of tough-guy roles, epic - sword films, and romances. It was fun to see him with an international cast and some real lousy **looking** pair of gloves. If I remember it was also dubbed in English which made things even more funnier. (kinda like seeing John Wayne speak Turkish).

Original

Prediction: **Positive (Confidence=99.22%)**
Target: Negative (Confidence=0.78%)

movies
Hilarious ~~film~~, I had a great time...

Prediction: **Positive (Confidence=91.06%)**
Target: Negative (Confidence=8.94%)

researching
...real lousy ~~looking~~ pair of gloves...

Prediction: **Positive (Confidence=92.28%)**
Target: Negative (Confidence=7.72%)

Substitute Both Words

movies
Hilarious ~~film~~, I had a great time... **+** researching ...real lousy ~~looking~~ pair of gloves...

Prediction: Positive (Confidence=26.67%)
Target: **Negative (Confidence=73.3%)** Flipped!

Future Work

Limitations of Genetic Attacks:

- Inefficient heuristic search
- Replace one word at a time → semantic drift
- Cannot find all combination of word substitutions

Future directions:

- Design more efficient and stronger attacks against contextualized word embeddings and pre-trained language models
- Evaluate with formal verifications that can provide certificate against all possible perturbations
- Utilize the open source pre-trained language models to perform transfer attacks

Transferability

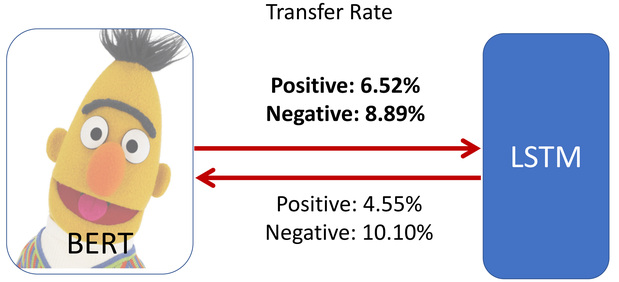

Figure shows the results tested on v1 models,. v2 models have similar results, but lower transfer rate.

NSF Center for Trustworthy Machine Learning
https://ctml.psu.edu/