GradMask: Gradient-Guided Token Masking for Textual Adversarial Example Detection

Han Cheol Moon, Shafiq Joty, and Xu Chi

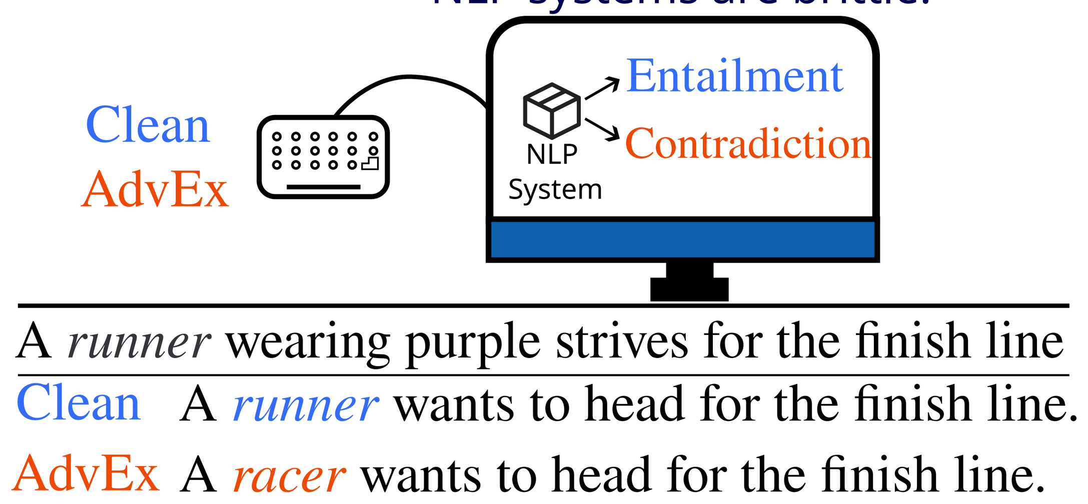
Nanyang Technological University, Salesforce Research, and SIMTECH

Abstract

- We present **GradMask, a simple adversarial example detection scheme for natural language processing (NLP) models**.
- Use gradient signals to detect adversarially perturbed tokens in an input sequence and occludes such tokens by a masking process
- **Improved detection performance** and **an interpretation of its decision**

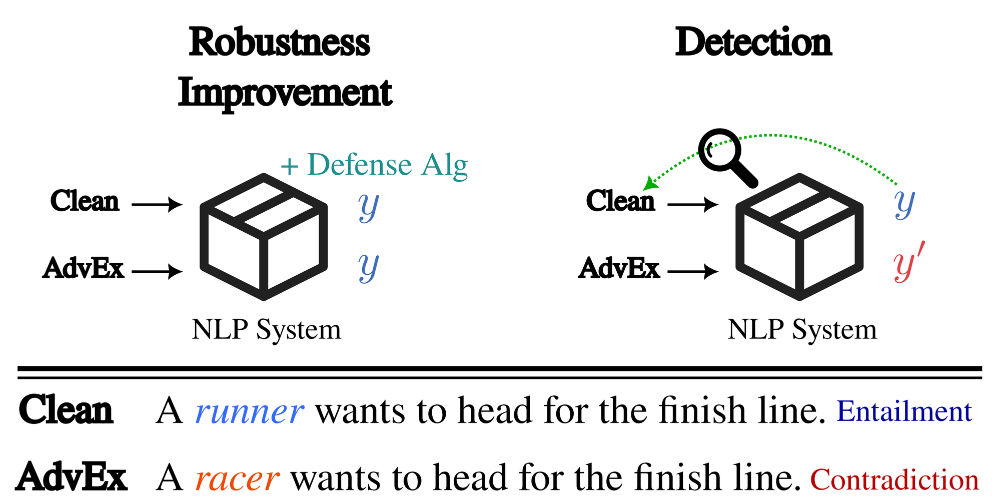
Adversarial Attack in NLP

NLP systems are brittle!



- Can you trust your NLP systems?
- Document forgery: Craft an AdvEx that flips the decision of a screening process of a bank.
- Disease diagnosis: You have to get surgery
- False translation.

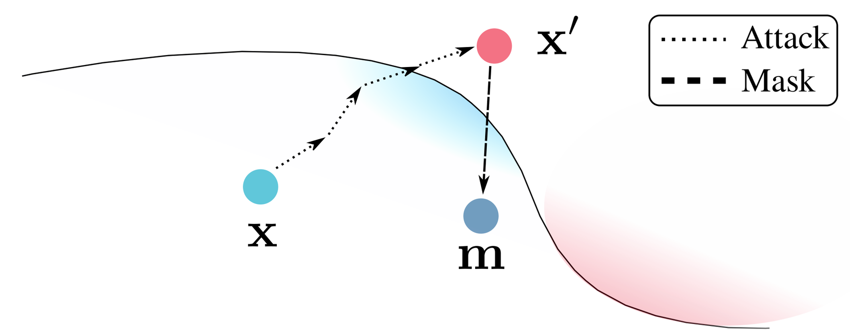
What is adversarial example detection?



Adversarial Example Detection! Pros & Cons

(+) No negative impact on the model performance.
(+) Identify the intention (adversarial or not).
(+) Allow users can take actions (reject or revise) accordingly.
(−) Typically work as a separate module.

Gradient-Guided Textual Adversarial Example Detection Algorithm:



$\mathbf{x}$	... during the 2004 Republican convention in New York were not that compelling to make a ...
$\mathbf{x}'$	... during the 2004 Republican convention in New York were not that compel to make a documentary ...
$\mathbf{m}$	... during the 2004 Republican convention in New York were not that [MASK] to make a [MASK]...

What is so special?

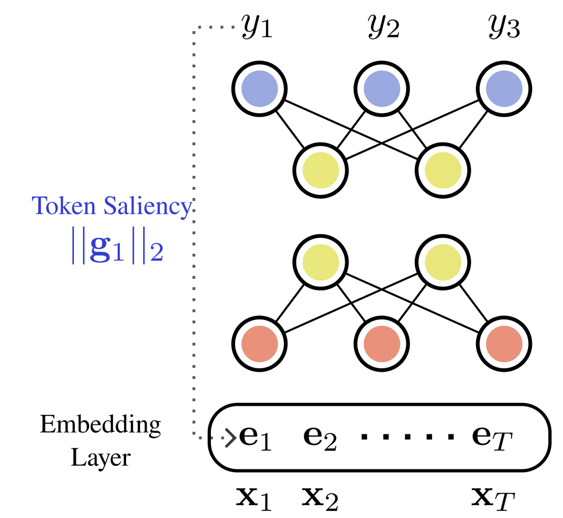
1. Very simple and low computation cost!
2. No assumptions about potential attacks such as word-frequency and synonym substitution sets.
3. No additional sub-modules such as synonym search module and Additional classifiers.
4. Provide interpretation.
5. Significantly low FPR95 scores.
6. Superior performance.

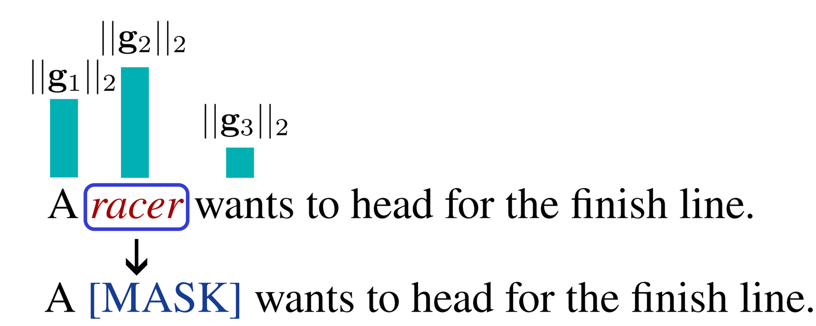
GradMask Works Really Well!

Dataset	Attack	ASR (%)	AUROC (%) ↑		EER (%) ↓		FPR95 (%) ↓	
			FGWS	GM	FGWS	GM	FGWS	GM
IMDb	BAE-R	63.45	66.56	**95.15**	40.51	**7.35**	94.05	**8.90**
	A2T	52.25	84.04	**95.05**	19.11	**8.30**	87.66	**9.80**
	TextFooler	82.44	85.40	**96.40**	17.20	**5.60**	86.52	**6.70**
	PWWS	87.41	90.92	**95.43**	12.03	**7.35**	77.68	**8.60**
AG	BAE-R	15.75	62.59	**83.82**	44.95	**19.15**	94.15	**35.20**
	A2T	13.04	75.09	**83.49**	27.80	**20.38**	90.26	**40.02**
	TextFooler	84.89	89.68	**96.53**	12.30	**5.35**	79.53	**5.60**
	PWWS	65.96	94.74	**95.69**	**6.46**	7.70	50.80	**9.30**
SST-2	BAE-R	58.17	60.08	**79.40**	43.80	**23.30**	94.33	**61.70**
	A2T	20.07	65.57	**78.16**	33.95	**23.07**	93.14	**52.44**
	TextFooler	93.28	74.14	**84.82**	29.05	**17.10**	91.66	**35.40**
	PWWS	85.18	85.25	**85.49**	**16.76**	19.62	82.11	**38.50**
MNLI	BAE-R	64.23	52.77	**69.99**	50.96	**33.80**	95.17	**73.50**
	A2T	49.85	66.34	**69.92**	37.96	**33.95**	92.82	**65.50**
	TextFooler	91.41	70.25	**74.24**	34.35	**29.50**	92.00	**55.40**
	PWWS	83.06	**76.94**	74.15	**27.38**	31.05	88.88	**65.47**

Table 1. Adversarial example detection restuls. GM stands for GradMask.

Gradient-Guided Masking





- Gradient-based attribution analysis.
- Token saliency: L_2-norm of $\mathbf{g}_t$, $||\mathbf{g}_t||_2$.
- $||\mathbf{g}_5||_2 < ||\mathbf{g}_3||_2 < \cdots < \underbrace{||\mathbf{g}_1||_2 < ||\mathbf{g}_2||_2}_{\text{Top}-k \text{ salient tokens}}$

1. Mask K salient tokens $\mathbf{x}' \rightarrow \mathbf{m}$.
2. Measure model confidence change:
   $w = \Big(f_{\boldsymbol{\theta}}(\mathbf{x})_i - f_{\boldsymbol{\theta}}(\mathbf{m})_i\Big)^2$
3. Decision making using an indicator function.:

Main assumptions

1. Masking suspicious tokens drops the model confidence.
   - Adversarial examples are results of sophisticate optimization.
2. NLP models are generally robust to a weak-level of noise.
3. The partial information loss in clean examples can be offset by the overall context of the input text.

Table 2. Statistics of extracted features.

Dataset	K	w-A/Conf-A (Avg±Std)	w-C/Conf-C (Avg±Std)
IMDb	MSP	-/49.58±49.67	-/92.88±13.53
	1	32.48±29.39/-	2.81±12.03/-
	2	53.71±36.92/-	3.84±18.04/-
	3	59.75±34.53/-	4.28±18.85/-
AG	MSP	-/49.43±49.55	-/89.75±15.58
	1	25.11±24.04/-	2.09±11.01/-
	2	47.18±31.39/-	3.32±16.03/-
	3	50.84±30.18/-	3.77±16.79/-

- Confidence on adversarial examples tends to be low.
- High w values:
  - AdvEx are brittle.
- Low w in clean examples:
  - NLP models are generally robust.

Adversarially Perturbed Word Detection