Feedback

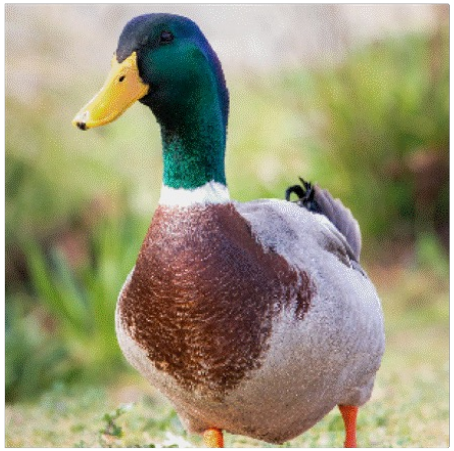
- Slides before lecture
- Notes before lecture

Adversarial

Eric Wong

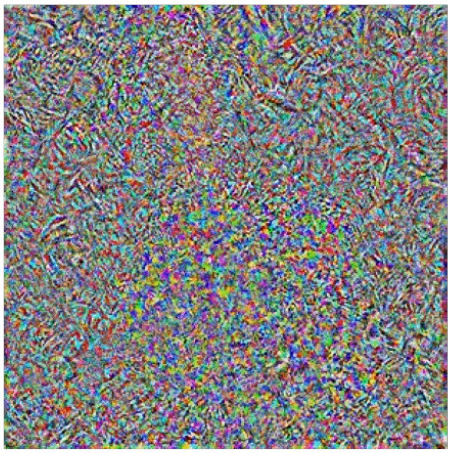
9/8/2022

Noise attack

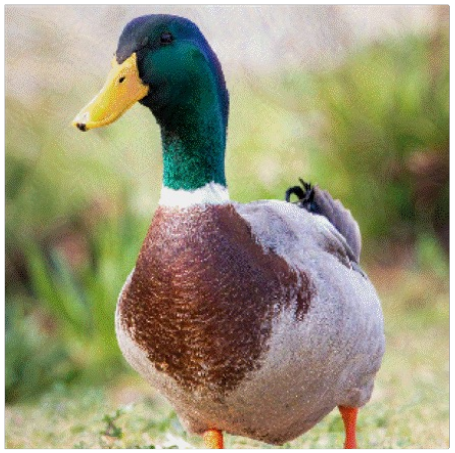


“Duck”

+

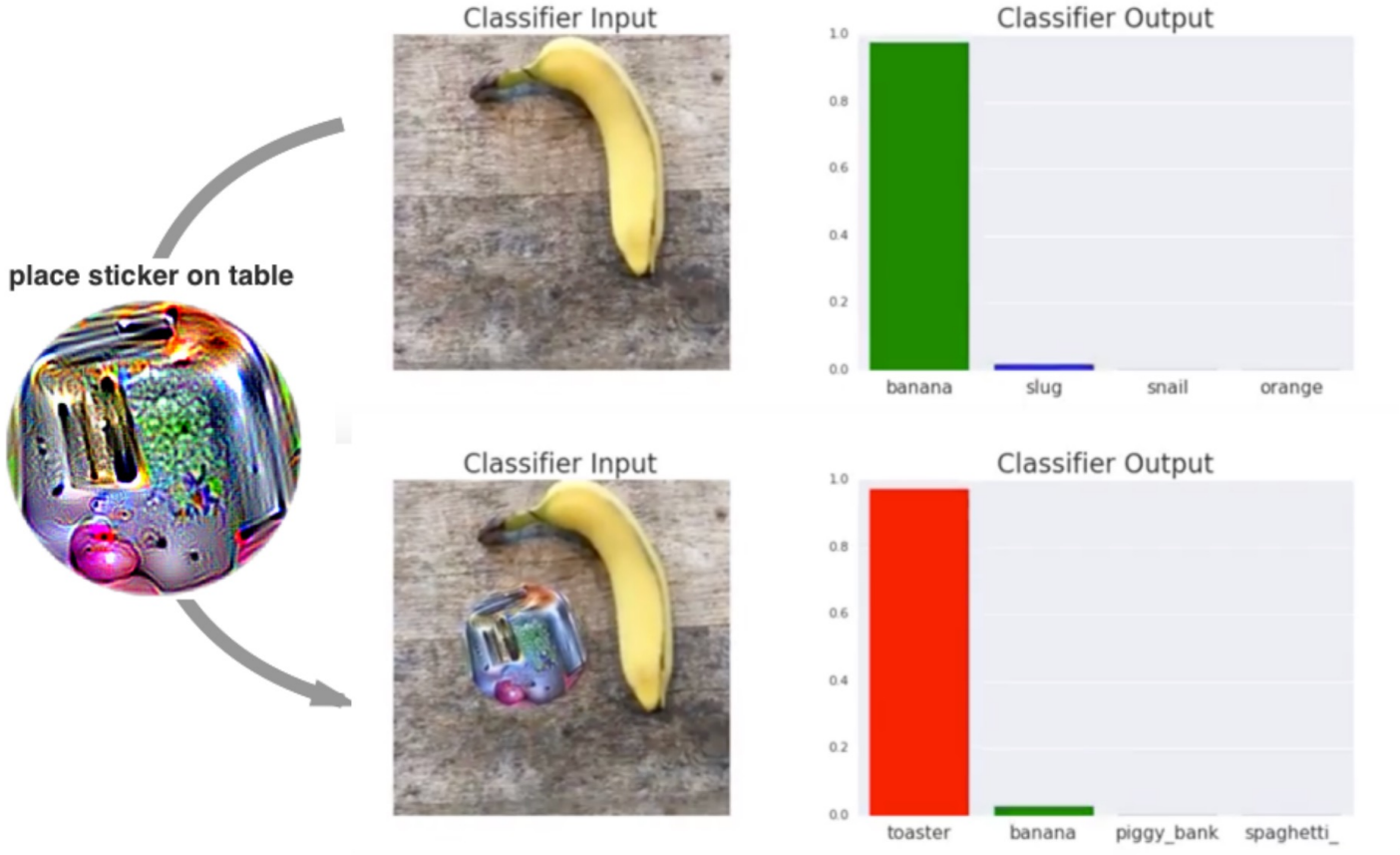


=

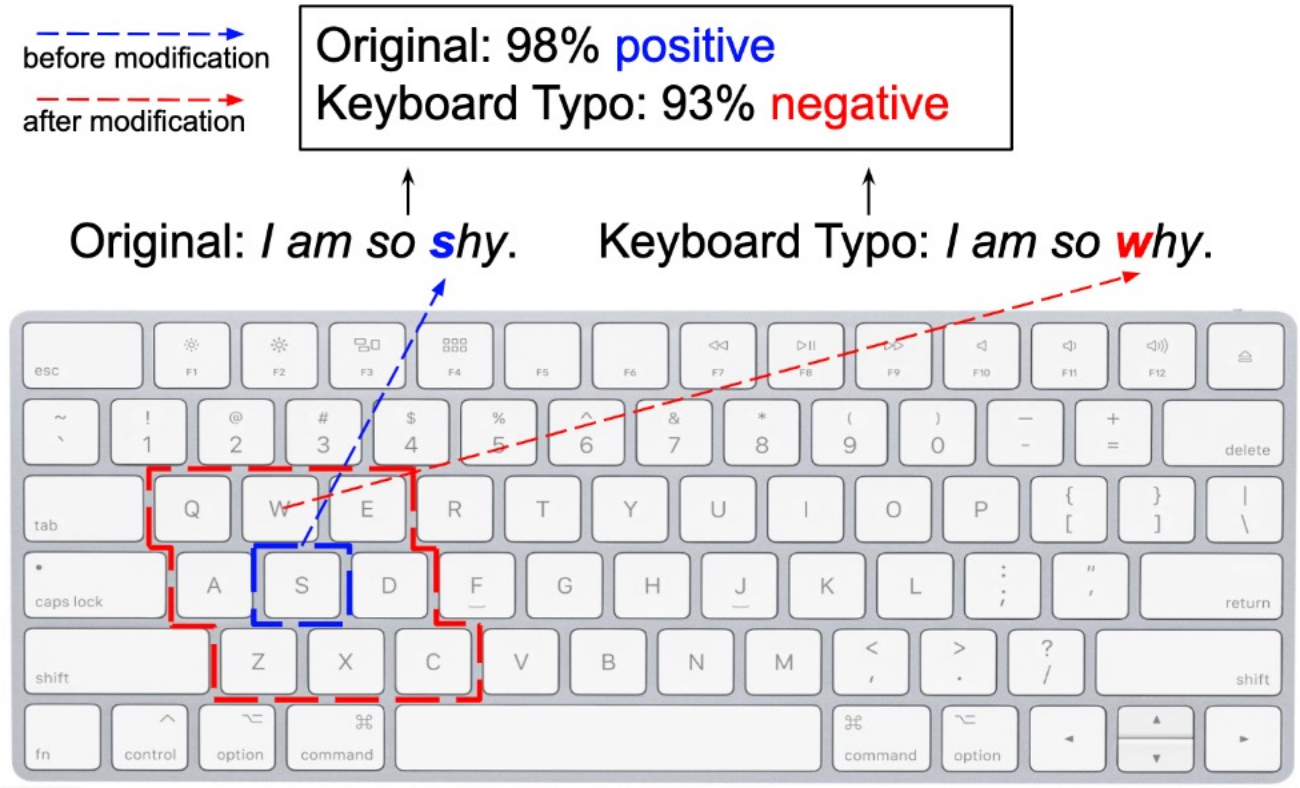


“Hermit crab”

Patch attack



Typos



Sun et al. 2020 “Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT”

Sentences

Label	Sentence
P	I am currently trying to give this company another chance. I have had the same scheduling experience as others have written about. Wrote to them today
N	I am currently trying to give this company another review. I have had the same dental experience about others or written with a name. Thanks to them today

Hsieh et al. 2019 “Natural Adversarial Sentence Generation with Graident based Perturbation”