Hey, Alexa: Sorry I fooled you

February 24 2020, by Rachel Gordon

Credit: CC0 Public Domain

A human can likely tell the difference between a turtle and a rifle. Two years ago, Google's AI wasn't so sure. For quite some time, a subset of computer science research has been dedicated to better understanding how machine learning models handle these "adversarial" attacks, which are inputs deliberately created to trick or fool machine learning

algorithms.

While much of this work has focused on speech and images, recently, a team from MIT's Computer Science and Artificial Intelligence Laboratory tested the boundaries of text. They came up with "TextFooler," a general framework that can successfully attack natural language processing (NLP) systems—the types of systems that let us interact with our Siri and Alexa voice assistants—and "fool" them into making the wrong predictions.

One could imagine using TextFooler for many applications related to internet safety, such as email spam filtering, hate speech flagging, or "sensitive" political speech text detection—which are all based on text classification models.

"If those tools are vulnerable to purposeful adversarial attacking, then the consequences may be disastrous," says Di Jin, MIT Ph.D. student and lead author on a new paper about TextFooler. "These tools need to have effective defense approaches to protect themselves, and in order to make such a safe defense system, we need to first examine the adversarial methods."

TextFooler works in two parts: altering a given text, and then using that text to test two different language tasks to see if the system can successfully trick machine learning models.

The system first identifies the most important words that will influence the target model's prediction, and then selects the synonyms that fit contextually. This is all while maintaining grammar and the original meaning to look "human" enough, and until the prediction is altered.

Then, the framework is applied to two different tasks—text classification, and entailment, (which is the relationship between text

fragments in a sentence), with the goal of changing the classification or invalidating the entailment judgment of the original models.

In one example, TextFooler's input and output were:

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

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

In this case, when testing on an NLP model, it gets the example input right, but then gets the modified input wrong.

In total, TextFooler successfully attacked three target models, including "BERT," the popular open-source NLP model. It fooled the target models with an accuracy of over 90 percent to under 20 percent, by changing only 10 percent of the words in a given text. The team evaluated success on three criteria: changing the model's prediction for classification or entailment, if it looked similar in meaning compared with the original example to a human reader, and lastly if the text looked natural enough.

The researchers note that while attacking existing models is not the end goal, they hope that this work will help more abstract models generalize to new, unseen data.

"The system can be used or extended to attack any classification-based NLP models to test their robustness," says Jin. "On the other hand, the generated adversaries can be used to improve the robustness and generalization of deep learning models via adversarial training, which is a critical direction of this work."

More information: Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. arXiv:1907.11932 [cs.CL] arxiv.org/abs/1907.11932