Your 598 ID: 5107

Title of paper: "Audio Adversarial Examples: Targeted Attacks on Speech Attacks" by Carlini and Wagner.
What is their primary result? The primary result of the paper is to demonstrate that the use of neural networks in audio/speech recognition tasks is vulnerable to adversarial attacks; that is, attacks where the main objective of the attacker is to make the neural network classify an instance x similar to a natural instance y as any target t chosen by the attacker.
Why is this important? This is important because the use of neural networks in speech recognition tasks is more and more relevant to our everyday lives. From Google's Assistant, to Apple's Siri, and Amazon's Alexa—these applications account for
What are their key ideas? They formulate the adversarial attack problem as an optimization problem: Given a natural example x and a target phrase t, solve

$$\begin{aligned} &\text{minimize } \mathrm{dB}_x(\delta), \\ &\text{such that } C(x+\delta) = t \quad \text{for } x+\delta \in [-M, M], \end{aligned}$$

where dB_x is a measure (in decibels) of the magnitude of the distortion δ relative to x, and M is the maximum intensity of the sound. Due to the nonlinearity of the problem, traditional Gradient Descent techniques do not work on this problem. The authors resolve this by minimizing the following reformulation:

$$\text{minimize } \mathrm{dB}_x(\delta) + cl(x+\delta, t),$$

where l is the loss function, and it is constructed such that $l(x', t) \leq 0$ if, and only if, $C(x') = t$. The authors use the CTC loss as the loss function and further formulate an optimization problem to address the oscillatory nature of the previous optimization problem; i.e., they solve

$$\begin{aligned} &\text{minimize } |\delta|_2^2 + cl(x+\delta, t), \\ &\text{such that } \mathrm{dB}_x(\delta) \leq \tau \end{aligned}$$

for sufficiently large τ.
What are the limitations, either in performance or applicability?
What might be an interesting next step based on this work? The authors based their adversarial attacks on a model where the attacker knows the inner workings of the neural network, a so-called "whitebox" setting. It

would be interesting to extend the work to a blackbox setting, where the attacker has no knowledge of the inner workings of the network.
How did they train and evaluate it? The authors test their attacks on Mozilla's implementation of DeepSpeech.
Did they implement something?
Grader's 598 ID: