Houdini: Fooling Deep Structured Prediction Models

Ravi Raju

12-28-2018

1 Motivation

This paper introduces a framework called Houdini for producing adverserial examples not exclusive to classification. Some of the problems that are posed in the introduction is that some of these attacks are based on gradients, which may be a combinatorial problem to attack. These are focusing on structured prediction tasks.

2 Solution

In the abstract attacks with Houdini achieve higher success rate than those based on the traditional surrogates used to train these models. We cannot train directly on the loss as it so nondifferential so a surrogate is used, this is called Houdini. What it actually is a product of a stochastic margin (look on pg 4) and the loss function.

3 Results

Experiments were conducted for human pose estimation, semantic segmentation, and speech recognition. Basically the perturbation added masked the persence of the humans in the picture. Perturbation screwed with correctly cutting objects; maybe that has something to do with the boundary?

4 Discussion/Takeaway

Honestly, my understanding of segmentation and speech recognition is too poor to really be able to properly appreciate this paper. Maybe if I look at attacks on these systems when I get more background knowledge, I will revisit this paper.