Alexander Khosrowshahi
William Stone
Ross Williams
Saswata Majumder
Prof. Ewing
CSCI1470
4/25/2025

Team Yell — Check-in #3

Project Title:
Team Yell: Adversarial Attacks on OpenAI's Whisper ASR Capabilities

Team Members (Name, Login):
Alexander Khosrowshahi (akhosrow)
William Stone (wlstone)
Saswata Majumder (smajum14)
Ross Williams (rjwillia)

GitHub Link

Check-in Questions:

Introduction: What problem are you trying to solve and why?

We are attempting to poison speech-to-text models using adversarial learning. Our main motivation for this project is improving digital privacy by (hopefully) being able to provide a model that can generate an audio filter that tricks speech-to-text models into transcribing your speech incorrectly. To do this, we plan to introduce an adversarial perturbation model trained against OpenAI's Whisper, a leading open-source audio transcription model. Our motivation is in personal privacy—allowing secure audio communication protected from automated transcription, protecting intellectual property in orated art mediums, and generally giving users more discretion over how their speech is used in model training. We approach the model as both a black box and white box, attacking it through means of adversarial black box optimization using evolutionary strategies and white box gradient ascent.

Challenges: What has been the hardest part of the project you've encountered so far?

Our project has gone through many iterations throughout the research process. The realization we would have to use a gradient-less method for the black box attack was a stirring one, and the research to get to our current methods was substantial. Though we feel fairly confident in the project's current place and path forward to get our training working, our main concern is compute intensity. Evolutionary strategies, the method we plan to use, seems to work best with very small perturbations over many epochs. Given Whisper has a nontrivial overhead per-run and each epoch can involve multiple runs, we are very concerned about if we'll be able to train in time.

Insights: Are there any concrete results you can show at this point?

We're just bordering on concrete results as of now. Our model does seem to move in the direction of producing higher word error rates in training, but we have some odd bugs we need to iron out regarding the training vs. the actual results of running a sample through our perturbation model and having whisper transcribe it again.

- How is your model performing compared with expectations?

Currently our model is not performing as well as we hoped it would, but we think a lot of this is up to having to run longer training sessions to see how it will change over time.

Plan: Are you on track with your project?

Yes, we feel we are on track, but also that we will need to do some heavy lifting in the next few days. We've fully implemented our model, two different implementations of the ES attack, in depth preprocessing and utility functions, etc. so we're in a good spot materially. Like already mentioned, the main question is in speed and effectiveness of training.

- What do you need to dedicate more time to?

Once again, training. We also need to run the open-source attack, though we expect this to be very simple.

- What are you thinking of changing, if anything?

We might test out some more variants on ES such as NES or adversarial bandits, but these are really either stretch goals or last resorts if ES/CMAES prove ineffective.