My primary research interests lie in making machine learning safe and trustworthy for deployment in safety-critical areas and high-stakes applications. To achieve this, my current research focus is on adversarial machine learning. Although there has been an important progress in this field, the majority of the work has been on small norm attacks (Szegedy et al., 2014; Goodfellow, Shlens, & Szegedy, 2015; Madry, Makelov, Schmidt, Tsipras, & Vladu, 2019; Papernot et al., 2016; Carlini & Wagner, 2017) and effective defenses such as adversarial training have been computationally expensive (Xie, Wu, van der Maaten, Yuille, & He, 2019; Kannan, Kurakin, & Goodfellow, 2018). Furthermore, it is important that deep learning models across a variety of domains apart from computer vision are also adversarially robust. I'm excited to pursue these research problems during my PhD in order **to ensure that machine learning systems are provably adversarially robust in a wide variety of settings, and to lower the barriers for making them adversarially robust.** In addition to these research goals, I am also engaged in service to the community. I am a **co-organizer** of the **Trustworthy ML Initiative**, which aims to lower entry barriers into fields associated with trustworthy ML.

Going beyond conventional small-norm attacks For the past six months, I have been working with Dr. **Pin-Yu Chen** as an external student at **MIT-IBM Watson AI Lab**. Together we have been exploring the robustness spectrum beyond conventional adversarial attack techniques and small norm attacks, with the goal to make deployed ML systems truly robust to a wide range of perturbations. During this time I have worked on developing two such unconventional methods to generate adversarial images for Google Cloud's Optical Character Recognition and various object detection APIs. We proposed a new class of adversarial examples: *Simple Transparent Adversarial Examples*. Contrary to present methods of creating adversarial examples, these examples are much more straightforward to craft, even for those without any machine learning knowledge. **This class of examples demands attention due to the ease of use.** The recent surge in the research of creating adversarial attacks has provided great insights into evaluating robustness, but **the current research has not focused on more simple, yet dangerous, unconventional methods that pose huge security concerns.** Considering such simple attacks and going beyond typical small norm attacks (Brown et al., 2018) is **one of the necessary trends for the future of adversarial machine learning** in order to build more inclusive and broad robust defenses. We uncovered several interesting results from this work which resulted in a **preprint** https://jaydeepborkar.github.io/preprint.pdf. For instance, in figure 1, Google's Cloud Vision API misclassified one of the perturbed image of a gun as a hat, which was perturbed by adding transparent patches using a publicly available online tool.

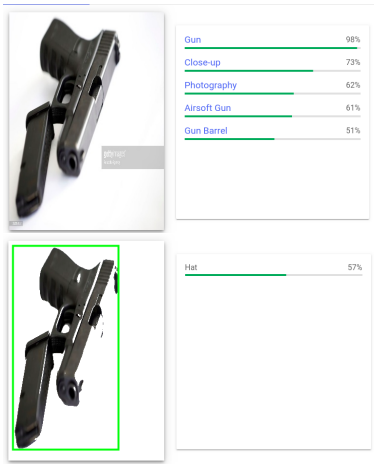

Figure 1: Google Cloud Vision API misclassified perturbed image of a gun as a hat

Computationally cheaper ways of adversarial training Considering the scale at which machine learning models are used to address diverse problems around the world by various small-scale startups and companies, I have often pondered — **how do these small organizations ensure effective adversarial robustness for their products?** Effective defenses such as adversarial training are computationally very expensive ((Xie et al., 2019) use 128 V100s and (Kannan et al., 2018) use 53 P100s for targeted adversarial training), and thus organizations who don't have adequate funding might not be able to afford this level of computation. **In an effort to make such robustness accessible to all, making adversarial training affordable, or exploring cheaper, yet equally robust, methods should be a priority.** (Shafahi et al., 2019) shows some promising methods in this direction. (Ilyas et al., 2019) show that after disentangling robust and non-robust features, we can achieve an adversarially robust model by just standard training (without any adversarial training). However, they use an already adversarially trained robust classifier to disentangle these features. As a possible research direction, I'm interested in whether these same results can be achieved without using an already adversarially robust classifier. If successful, this method could turn out to be an **alternative to make classifiers adversarially robust at no extra cost than standard training.**

Ensuring adversarial robustness of deep learning models across various domains Because deep learning models are deployed across a variety of domains (e.g., vision, language, voice), it is important that

none of the domains are neglected when testing for robustness. There has been a huge amount of work to evaluate the robustness of computer vision and even natural language systems, but still not enough has been done in the audio/voice domain. This is particularly due to the reason that crafting adversarial samples for voice systems is different from vision and language, and difficult due to difference in model architecture. Domain specific Systemization of Knowledge (SoK) papers that analyse the past methods and motivate future work can play a promising role here. One such paper that inspired me to think about this problem is (Abdullah, Warren, Bindschaedler, Papernot, & Traynor, 2020). With the inability to get white-box access to voice-processing systems such as Alexa, a possible future research direction is to query such systems with gibberish inputs in an effort to perform a black-box model-extraction attack. This method has been used to better understand language models (Krishna, Tomar, Parikh, Papernot, & Iyyer, 2020) and I would like to implement this method to experiment on voice-processing systems.

In addition to adversarial ML, I am also interested in other measures of trustworthiness such as privacy and fairness. Various works such as (Carlini, Liu, Úlfar Erlingsson, Kos, & Song, 2019; Buolamwini & Gebru, 2018) have motivated me to study these fields. I am currently working with **Prof. Hima Lakkaraju** (Harvard University) on a research project related to fairness. This project has provided me the opportunity to explore the direction of making fair predictors robust under dataset shifts. Inspired by Joy Buolamwini's work, I presented an independent poster on AI Bias — AI, why you ain't fair? : Understanding AI Bias at PyCon India 2019 to spread awareness about this issue. Participating in discussions at various venues such as Towards Trustworthy ML workshop, Participatory Approaches to Machine Learning workshop, and USENIX'20 have given me helpful insights about some important problems in this field. One such discussion was on the trade-off between adversarial robustness and accuracy (Raghunathan, Xie, Yang, Duchi, & Liang, 2020; Tsipras, Santurkar, Engstrom, Turner, & Madry, 2019) at Towards Trustworthy ML workshop. I am also passionate about using Computer Science for social good. In March 2020, I built the platform **COVID Letters** along with my friend to help people combat anxiety and stay hopeful during COVID-19.

How I got interested in Adversarial ML and challenges faced My journey in adversarial ML began in the third year of my undergraduate degree when, in reading papers by Percy Liang's group at Stanford, I stumbled upon their paper (Jia & Liang, 2017). I began seeking out related research and found myself questioning the assumptions and coming up with possible extensions for the papers I was reading. **I found myself enjoying this process, which motivated me to get involved in research.**
In the process of pursuing research opportunities, I faced significant challenges. There were no experts in this subject at my college, and it was challenging to independently grapple with the difficult concepts presented in the papers I was reading. However, **I was dedicated to learning the material**, and sought out online courses and supplementary material that would help me do so. An additional challenge was that my undergraduate college in India had no research environment. To get involved and pursue my passions, **I had no other options but to seek out external research collaborations myself.** In April 2019, I began cold-emailing researchers whose work I had enjoyed reading, including questions and possible extensions to their work. Months passed by and I did not find any collaborations. Often it felt as if I was traveling a dark tunnel which would lead me to a dead end. But my interest to pursue research kept me going. After **emailing more than 50 various researchers for over 12 months**, I finally found my first collaborator when I met Dr. Pin-Yu Chen virtually at ICLR in April 2020.
Though challenging, this process of finding a collaborator taught me more than any course I have ever taken. I realized that I enjoyed working independently and pushing myself through uncertainties. **It taught me to not give up on what I believe in and to keep going with patience and perseverance.** These experiences have given me a dedication and commitment to learning, and have prepared me to embark on a PhD journey. I would be overjoyed to have the opportunity to continue to grow at UVA.

Why get a PhD? My career goal is to work on research questions in Trustworthy Machine Learning, with the eventual hope of leading my own research team. **Drawing from my own journey and experiences, I am also committed to creating a nonprofit organization in India to make starting in research more accessible for students from institutions with less resources.** I believe that a PhD would provide me the opportunity to explore the research questions I'm excited about, and develop a network of mentors and collaborators, which would take me closer to my goals.

Working at UVA I want to study at UVA because it has amazing groups working on Trustworthy Machine Learning. I'm very interested in working with **Prof. David Evans** at the **Security Research Group** on security and privacy for ML. I have enjoyed reading the paper "Hybrid Batch Attacks: Finding Black-box Adversarial Examples with Limited Queries". My interests also fall in line with **Prof. Yuan Tian**.

*

References

Abdullah, H., Warren, K., Bindschaedler, V., Papernot, N., & Traynor, P. (2020). *Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems.*

Brown, T. B., Carlini, N., Zhang, C., Olsson, C., Christiano, P., & Goodfellow, I. (2018). *Unrestricted adversarial examples.*

Buolamwini, J., & Gebru, T. (2018, 23–24 Feb). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), (Vol. 81, pp. 77–91). New York, NY, USA: PMLR. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

Carlini, N., Liu, C., Úlfar Erlingsson, Kos, J., & Song, D. (2019). *The secret sharer: Evaluating and testing unintended memorization in neural networks.*

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)* (p. 39-57). doi: doi:10.1109/SP.2017.49

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples.*

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 125–136). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf

Jia, R., & Liang, P. (2017). *Adversarial examples for evaluating reading comprehension systems.*

Kannan, H., Kurakin, A., & Goodfellow, I. (2018). *Adversarial logit pairing.*

Krishna, K., Tomar, G. S., Parikh, A. P., Papernot, N., & Iyyer, M. (2020). *Thieves on sesame street! model extraction of bert-based apis.*

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019). *Towards deep learning models resistant to adversarial attacks.*

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 ieee european symposium on security and privacy (euros p)* (p. 372-387). doi: doi:10.1109/EuroSP.2016.36

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., & Liang, P. (2020). *Understanding and mitigating the tradeoff between robustness and accuracy.*

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., ... Goldstein, T. (2019). *Adversarial training for free!*

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks.*

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). *Robustness may be at odds with accuracy.*

Xie, C., Wu, Y., van der Maaten, L., Yuille, A., & He, K. (2019). *Feature denoising for improving adversarial robustness.*