Final Project Proposal

Please make this document anonymous. Your team name should be anonymous.

Team name: ***Pixel Pioneers***

- **What is your project idea?**
  Our project aims to generate adversarial examples to break current computer vision systems. We mainly focus on improving the transferability of the generated adversarial example and its robustness.

- **What is the socio-historical context that this project lives in?**
  The project is set in an era of rapid AI advancements, rising cybersecurity concerns, ethical discussions on AI use, regulatory developments, and strong academic-industry collaboration, underscoring the urgent need for secure, ethically developed AI technologies.

- **Please list three stakeholders that your project could impact, and describe how it could impact them.**
  Our project impacts tech companies and AI developers by revealing system vulnerabilities, pushing for stronger, more secure AI models. Consumers benefit from increased AI safety and trust, though misuse poses privacy risks. Regulatory bodies may use our findings to refine AI governance, enhancing security and ethical standards.

- —

- **What are the skills of the team members? Conduct a skill assessment!**
  Team member 1: Proficient in using the Python language, has developed neural network models using the PyTorch framework. Additionally, has participated in image classification tasks and possesses a certain understanding of neural network principles.
  Team member 2: Successfully developed a runnable CNN neural network using Python. It includes convolutional layers and pooling layers. Additionally, able to explain the principles of this neural network.
  Team member 3: Know the adversarial machine learning techniques, including how to craft adversarial examples that can fool AI models without being detected by humans or other machines. Able to critically evaluate the strengths and weaknesses of AI models and to think creatively about how to test and improve them.

- **What data will you use?**
  We will choose 1000 images of different classes from ILSVRC 2012 validation set.

- **What software/hardware will you use?**
  We will use the PyTorch framework to build our algorithm. For the hardware, we will apply the Oscar computing resource.

- **Who will do what?**
  Team member 1: Focuses on the development and optimization of machine learning models. This includes designing adversarial attack strategies to test the robustness of computer vision systems.
  Team member 2: Specializes in data handling, processing, and analysis. Ensures that the data used for training and testing is of high quality and diversity to effectively challenge AI models.
  Team member 3: Bring expertise in cybersecurity and ethical hacking to anticipate how adversarial examples could be used maliciously and to devise strategies to mitigate these risks.

- **How will you know whether you have made progress? What will you measure?**
  Progress in generating adversarial examples for computer vision systems can be effectively measured by tracking how frequently these examples mislead targeted models and assessing their ability to compromise multiple models across various architectures.

- **What technical problems do you foresee or have?**
  Creating adversarial examples that are effective not just on a specific model but across a wide range of models can be difficult. Achieving high transferability requires deep understanding of different model architectures and their vulnerabilities, which can be complex given the diversity of existing AI systems.

- **Is there anything that we can do to help? E.G., resources, equipment.**
  We may need to access the computing resources (GPU).