This thesis examines adversarial examples in machine learning, specifically in the image classification domain. State-of-the-art deep learning models are able to recognize patterns better than humans. However, we can significantly reduce the model's accuracy by adding imperceptible, yet intentionally harmful noise. This work investigates various methods of creating adversarial images as well as techniques that aim to defend deep learning models against these malicious inputs. We choose one of the contemporary defenses and design an attack that utilizes evolutionary algorithms to deceive it. Our experiments show an interesting difference between adversarial images created by evolution and images created with the knowledge of gradients. Last but not least, we test the transferability of our created samples between various deep learning models.