HEMLOCK

POISON YOUR IMAGES

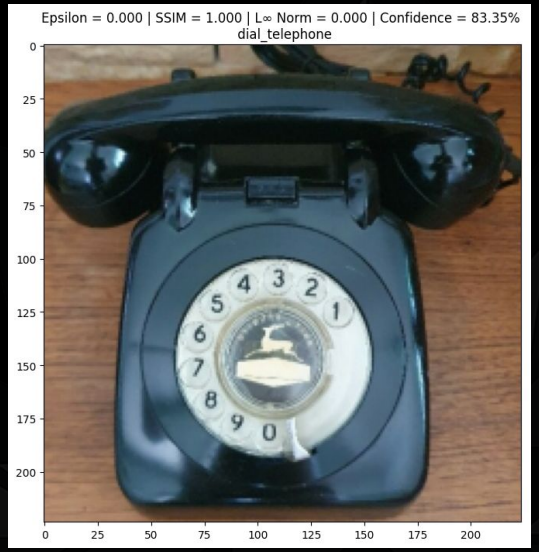
HEMLOCK: THE FIGHT AGAINST AI

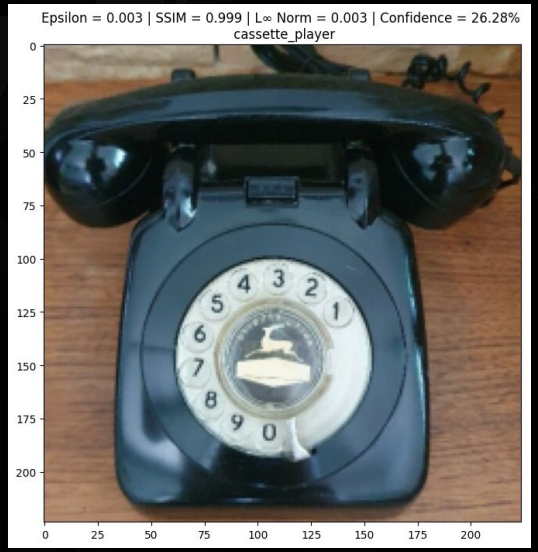
Hemlock is a concept of an interface that allows users to automatically apply poisoning techniques to their content. This app allows social media influencers to protect their work from being replicated or used by big tech companies to train AI models

OVERALL APPROACH

- Decided to use the existing MobileNetV2 image classification model, & TensorFlow and Keras libraries.
- Experimented with different epsilon values to evaluate the model.
- Adversarial Attack Methods: Fast Gradient Sign Method, Projected Gradient Descent Method, and Carlini & Wagner Approach.

IMAGES BEFORE AND AFTER



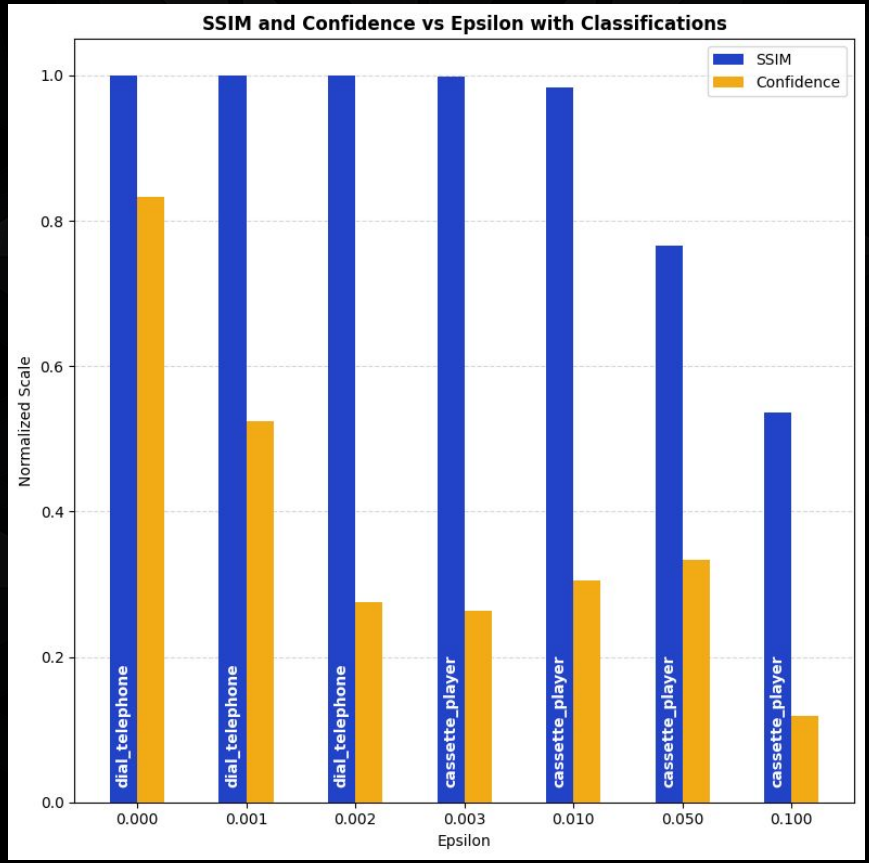


dial_telephone + perturbations = cassette_player

PERFORMANCE METRICS

Structural Similarity Index (SSIM) is a comparison of image structure, brightness and contrast.

Confidence shows how sure the AI is about the identification.



DEMO

```python
# import dependencies
import tensorflow as tf
import matplotlib as mpl
import matplotlib.pyplot as plt
from pathlib import Path
from PIL import Image
from rembg import remove
import numpy as np
import io
import tensorflow_hub as hub

######################################################################
# # Helper function to preprocess the image so that it can be inputted in MobileNetV2
# def preprocess(image):
#   image = tf.cast(image, tf.float32)
#   image = tf.image.resize(image, (224, 224), method=tf.image.ResizeMethod.AREA)
#   image = tf.keras.applications.mobilenet_v2.preprocess_input(image)
#   image = image[None, ...]
#   return image

######################################################################
# # Helper function to extract labels from probability vector
# def get_imagenet_label(probs):
#   return decode_predictions(probs, top=1)[0][0]

######################################################################
# Helper function to remove the background from an image tensor
def remove_background(image, background_color=(255, 255, 255)):
    """
    Removes the background from a TensorFlow image tensor, crops to a square bounding box,
    and replaces the background with a specified color.

    Args:
        image (tf.Tensor): Input image tensor.
        background_color (tuple): RGB color for the new background (default is white).

    Returns:
```

PHASE 2

1. Add more loss functions to the model
2. Experiment with other attack methods
3. Calculate Attack Success Rate
4. Create UI

Q&A

APPENDIX

https://github.com/Jeff-Oliver/hemlock_image_poisoner.git