final_results

April 11, 2025

[1]:
```
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt
from tensorflow.keras.applications.mobilenet_v2 import preprocess_input
```

[2]:
```
# Hyperparameters
BATCH_SIZE = 64
IMG_SIZE = 96  # Upscale CIFAR-10 images (32x32) to 96x96 for MobileNetV2
AUTOTUNE = tf.data.AUTOTUNE
```

[3]:
```
def resize_and_preprocess(image, label):
    image = tf.cast(image, tf.float32)
    image = tf.image.resize(image, [IMG_SIZE, IMG_SIZE])
    image = preprocess_input(image)
    return image, label
```

[4]:
```
# Load CIFAR-10 test dataset
(_, _), (x_test, y_test) = tf.keras.datasets.cifar10.load_data()
y_test = np.squeeze(y_test)
```

```
Downloading data from https://www.cs.toronto.edu/~kriz/cifar-10-python.tar.gz
170498071/170498071              4s
0us/step
```

[5]:
```
model = tf.keras.models.load_model("adversarially_trained_model.keras")
```

[6]:
```
#preprocessing data
test_dataset = tf.data.Dataset.from_tensor_slices((x_test, y_test))
test_dataset = test_dataset.map(resize_and_preprocess,␣
 ↪num_parallel_calls=AUTOTUNE)
test_dataset = test_dataset.batch(BATCH_SIZE).prefetch(AUTOTUNE)
```

[7]:
```
loss, accuracy = model.evaluate(test_dataset)
loss, accuracy
```

```
157/157              11s 34ms/step -
accuracy: 0.9865 - loss: 0.0487
```

[7]: (0.04580976814031601, 0.9868999719619751)

[8]:
```
@tf.function
def batched_fgsm_attack(images, labels, epsilon=0.01):
    with tf.GradientTape() as tape:
        tape.watch(images)
        predictions = model(images, training=False)
        loss = tf.keras.losses.sparse_categorical_crossentropy(labels,␣
↪predictions)
    gradients = tape.gradient(loss, images)
    adv_images = images + epsilon * tf.sign(gradients)
    adv_images = tf.clip_by_value(adv_images, -1, 1)
    return adv_images
```

[9]:
```
@tf.function
def batched_pgd_attack(images, labels, epsilon=0.01, alpha=0.005, num_iter=10):
    adv_images = tf.identity(images)

    for _ in tf.range(num_iter):
        with tf.GradientTape() as tape:
            tape.watch(adv_images)
            predictions = model(adv_images, training=False)
            loss = tf.keras.losses.sparse_categorical_crossentropy(labels,␣
↪predictions)
        gradients = tape.gradient(loss, adv_images)
        adv_images = adv_images + alpha * tf.sign(gradients)

        # Project perturbation
        perturbation = tf.clip_by_value(adv_images - images, -epsilon, epsilon)
        adv_images = tf.clip_by_value(images + perturbation, -1, 1)

    return adv_images
```

[10]:
```
def deepfool_attack(image, num_classes=10, overshoot=0.0000001, max_iter=1):
    image = tf.convert_to_tensor(image, dtype=tf.float32)
    perturbed_image = tf.identity(image)

    # Get original prediction and label
    with tf.GradientTape() as tape:
        tape.watch(perturbed_image)
        logits = model(tf.expand_dims(perturbed_image, axis=0))[0]
    orig_label = tf.argmax(logits)

    r_tot = tf.zeros_like(image)
    i = 0

    while i < max_iter:
```

```
        with tf.GradientTape(persistent=True) as tape:
            tape.watch(perturbed_image)
            logits = model(tf.expand_dims(perturbed_image, axis=0))[0]

        current_label = tf.argmax(logits)
        if current_label != orig_label:
            break

        # Compute gradients for all class logits
        gradients = []
        for k in range(num_classes):
            with tf.GradientTape() as tape2:
                tape2.watch(perturbed_image)
                logit_k = model(tf.expand_dims(perturbed_image, axis=0))[0, k]
            grad_k = tape2.gradient(logit_k, perturbed_image)
            gradients.append(grad_k)
        gradients = tf.stack(gradients)

        # Compute minimal perturbation
        f_orig = logits[orig_label]
        perturbs = []
        for k in range(num_classes):
            if k == orig_label:
                continue
            w_k = gradients[k] - gradients[orig_label]
            f_k = logits[k] - f_orig
            norm_w = tf.norm(tf.reshape(w_k, [-1])) + 1e-8
            pert_k = tf.abs(f_k) / norm_w
            perturbs.append((pert_k, w_k))

        # Choose the closest decision boundary
        perturbs.sort(key=lambda x: x[0])
        pert_k, w_k = perturbs[0]

        # Compute minimal directional perturbation (no sign scaling)
        r_i = (pert_k * w_k) / (tf.norm(w_k) + 1e-8)
        r_tot += r_i

        # Apply accumulated perturbation with small overshoot
        perturbed_image = image + (1 + overshoot) * r_tot
        perturbed_image = tf.clip_by_value(perturbed_image, -1, 1)

        i += 1

    return perturbed_image
```

[11]:
```
def get_test_dataset():
    # Load CIFAR-10 test dataset and preprocess
    (_, _), (x_test, y_test) = tf.keras.datasets.cifar10.load_data()
    y_test = np.squeeze(y_test)
    ds = tf.data.Dataset.from_tensor_slices((x_test, y_test))
    ds = ds.map(resize_and_preprocess, num_parallel_calls=AUTOTUNE)
    ds = ds.batch(BATCH_SIZE).prefetch(AUTOTUNE)
    return ds
```

[12]:
```
clean_ds = get_test_dataset()
model.compile(loss='sparse_categorical_crossentropy', metrics=['accuracy'])
```

[13]:
```
def build_adversarial_dataset_fast(dataset, attack_fn, attack_name="FGSM"):
    adv_images_all = []
    adv_labels_all = []

    print(f"\nBuilding {attack_name} dataset...")

    for images, labels in dataset:
        adv_images = attack_fn(images, labels)
        adv_images_all.append(adv_images)
        adv_labels_all.append(labels)

    adv_images_all = tf.concat(adv_images_all, axis=0)
    adv_labels_all = tf.concat(adv_labels_all, axis=0)

    adv_ds = tf.data.Dataset.from_tensor_slices((adv_images_all,␣
↪adv_labels_all))
    return adv_ds.batch(BATCH_SIZE).prefetch(AUTOTUNE)
```

[14]:
```
def build_adversarial_dataset_deepfool(attack_fn, name="DeepFool",␣
↪max_samples=500, num_classes=10):
    adv_images = []
    adv_labels = []

    print(f"\nGenerating {name} adversarial dataset (max {max_samples} samples).
↪..")
    sample_count = 0

    for images, labels in clean_ds:
        for img, label in zip(images, labels):
            # Pass a fixed number of classes instead of the label value.
            adv_img = attack_fn(img, num_classes)
            adv_images.append(adv_img.numpy())
            adv_labels.append(int(label.numpy()))
            sample_count += 1
```

```
            if sample_count >= max_samples:
                break
        if sample_count >= max_samples:
            break

    adv_images = np.array(adv_images)
    adv_labels = np.array(adv_labels)

    ds = tf.data.Dataset.from_tensor_slices((adv_images, adv_labels))
    ds = ds.batch(BATCH_SIZE).prefetch(AUTOTUNE)
    return ds
```

[15]:
```
def evaluate_model_on_dataset(dataset, name="Dataset"):
    y_true, y_pred = [], []
    total_loss = 0.0
    total_samples = 0

    loss_fn = tf.keras.losses.SparseCategoricalCrossentropy()
    for batch_images, batch_labels in dataset:
        preds = model(batch_images, training=False)
        loss = loss_fn(batch_labels, preds).numpy()
        pred_classes = tf.argmax(preds, axis=1).numpy()
        y_true.extend(batch_labels.numpy())
        y_pred.extend(pred_classes)
        total_loss += loss * len(batch_labels)
        total_samples += len(batch_labels)

    accuracy = np.mean(np.array(y_true) == np.array(y_pred))
    avg_loss = total_loss / total_samples
    correct = sum(np.array(y_true) == np.array(y_pred))
    incorrect = total_samples - correct

    print(f"\n{name} Evaluation:")
    print(f"  Total Samples: {total_samples}")
    print(f"  Accuracy: {accuracy:.4f}")
    print(f"  Loss: {avg_loss:.4f}")
    print(f"  Correct Predictions: {correct}")
    print(f"  Incorrect Predictions: {incorrect}")
    return accuracy, avg_loss
```

[16]:
```
fgsm_ds = build_adversarial_dataset_fast(clean_ds, lambda x, y:␣
 ↪batched_fgsm_attack(x, y, epsilon=0.01), attack_name="FGSM")
pgd_ds = build_adversarial_dataset_fast(clean_ds, lambda x, y:␣
 ↪batched_pgd_attack(x, y, epsilon=0.01, alpha=0.005, num_iter=10),␣
 ↪attack_name="PGD")
deepfool_ds = build_adversarial_dataset_deepfool(deepfool_attack,␣
 ↪name="DeepFool", max_samples=200)
```

```
Building FGSM dataset…

Building PGD dataset…

Generating DeepFool adversarial dataset (max 200 samples)…
```

[17]:
```python
def load_original_model():
    return tf.keras.models.load_model("adversarially_trained_model.keras")
```

[18]:
```python
def get_gaussian_kernel(size=3, sigma=1.0):
    """Creates a 2D Gaussian kernel."""
    x = tf.range(-size // 2 + 1, size // 2 + 1, dtype=tf.float32)
    x = tf.exp(-(x**2) / (2 * sigma**2))
    kernel_1d = x / tf.reduce_sum(x)
    kernel_2d = tf.tensordot(kernel_1d, kernel_1d, axes=0)
    kernel_2d = kernel_2d / tf.reduce_sum(kernel_2d)
    return kernel_2d[:, :, tf.newaxis, tf.newaxis]  # Shape: (H, W,␣
 ↪in_channels=1, out_channels=1)

def apply_gaussian_blur(x, sigma):
    """Applies Gaussian blur using depthwise convolution."""
    kernel = get_gaussian_kernel(size=3, sigma=sigma)
    channels = tf.shape(x)[-1]
    kernel = tf.tile(kernel, [1, 1, channels, 1])  # Make kernel channel-wise
    x = tf.nn.depthwise_conv2d(x, kernel, strides=[1, 1, 1, 1], padding='SAME')
    return x
```

[19]:
```python
def inference_input_transformation(
    x,
    apply_bitdepth=True,
    bits=4,
    apply_noise=True,
    noise_std=0.05,
    apply_jpeg=True,
    jpeg_quality=75,
    apply_blur=True,
    blur_sigma=0.5
):
    """
    Apply input transformations: quantization, noise, JPEG compression, and␣
 ↪blur.

    Args:
        x (Tensor): Input tensor in [0,1].
    """
    if apply_bitdepth:
```

```
        levels = 2 ** bits
        x = tf.round(x * (levels - 1)) / (levels - 1)

    if apply_noise:
        noise = tf.random.normal(tf.shape(x), mean=0.0, stddev=noise_std,␣
↪dtype=x.dtype)
        x = x + noise

    if apply_jpeg:
        def jpeg_fn(img):
            img_uint8 = tf.image.convert_image_dtype(img, tf.uint8)
            encoded = tf.io.encode_jpeg(img_uint8, quality=jpeg_quality)
            decoded = tf.io.decode_jpeg(encoded)
            return tf.image.convert_image_dtype(decoded, tf.float32)
        x = tf.map_fn(jpeg_fn, x)

    if apply_blur:
        x = apply_gaussian_blur(x, sigma=blur_sigma)

    x = tf.clip_by_value(x, 0.0, 1.0)
    return x
```

[20]:

```
class TransformedModel(tf.keras.Model):
    def __init__(
        self,
        base_model,
        apply_bitdepth=True,
        bits=4,
        apply_noise=True,
        noise_std=0.05,
        apply_jpeg=True,
        jpeg_quality=75,
        apply_blur=True,
        blur_sigma=0.5
    ):
        super().__init__()
        self.base_model = base_model
        self.apply_bitdepth = apply_bitdepth
        self.bits = bits
        self.apply_noise = apply_noise
        self.noise_std = noise_std
        self.apply_jpeg = apply_jpeg
        self.jpeg_quality = jpeg_quality
        self.apply_blur = apply_blur
        self.blur_sigma = blur_sigma

    def call(self, inputs, training=False):
```

```
        # Convert from [-1, 1] to [0, 1] before transformation
        inputs = (inputs + 1.0) / 2.0

        transformed = inference_input_transformation(
            inputs,
            apply_bitdepth=self.apply_bitdepth,
            bits=self.bits,
            apply_noise=self.apply_noise,
            noise_std=self.noise_std,
            apply_jpeg=self.apply_jpeg,
            jpeg_quality=self.jpeg_quality,
            apply_blur=self.apply_blur,
            blur_sigma=self.blur_sigma
        )

        # Convert back to [-1, 1] for model input
        transformed = transformed * 2.0 - 1.0
        return self.base_model(transformed, training=training)
```

[21]:
```
# all transformations applied
model = TransformedModel(
    load_original_model(),
    apply_bitdepth=True, bits=4,
    apply_noise=True, noise_std=0.05,
    apply_jpeg=True, jpeg_quality=75,
    apply_blur=True, blur_sigma=0.5
)
```

[22]:
```
evaluate_model_on_dataset(clean_ds, name='Clean + Transformed_default')
```

```
Clean + Transformed_default Evaluation:
  Total Samples: 10000
  Accuracy: 0.4337
  Loss: 2.2827
  Correct Predictions: 4337
  Incorrect Predictions: 5663
```

[22]:
```
(np.float64(0.4337), np.float32(2.2826812))
```

[23]:
```
evaluate_model_on_dataset(fgsm_ds, name="FGSM + Transformed_default")
```

```
FGSM + Transformed_default Evaluation:
  Total Samples: 10000
  Accuracy: 0.4156
  Loss: 2.3821
  Correct Predictions: 4156
```

```
    Incorrect Predictions: 5844
```

[23]: (np.float64(0.4156), np.float32(2.3821476))

[24]:
```python
evaluate_model_on_dataset(pgd_ds, name="PGD + Transformed_default")
```

```
PGD + Transformed_default Evaluation:
  Total Samples: 10000
  Accuracy: 0.4134
  Loss: 2.3758
  Correct Predictions: 4134
  Incorrect Predictions: 5866
```

[24]: (np.float64(0.4134), np.float32(2.3758323))

[25]:
```python
evaluate_model_on_dataset(deepfool_ds, name="DeepFool + Transformed_default")
```

```
DeepFool + Transformed_default Evaluation:
  Total Samples: 200
  Accuracy: 0.1650
  Loss: 5.1539
  Correct Predictions: 33
  Incorrect Predictions: 167
```

[25]: (np.float64(0.165), np.float32(5.1538787))

[26]:
```python
# no bitdepth
model = TransformedModel(
    load_original_model(),
    apply_bitdepth=False,  # Bitdepth reduction disabled
    apply_noise=True, noise_std=0.05,
    apply_jpeg=True, jpeg_quality=75,
    apply_blur=True, blur_sigma=0.5
)
```

[27]:
```python
evaluate_model_on_dataset(clean_ds, name='Clean + Transformed_no_bitdepth')
```

```
Clean + Transformed_no_bitdepth Evaluation:
  Total Samples: 10000
  Accuracy: 0.4913
  Loss: 1.9947
  Correct Predictions: 4913
  Incorrect Predictions: 5087
```

[27]: (np.float64(0.4913), np.float32(1.9946779))

[28]:
```
evaluate_model_on_dataset(fgsm_ds, name="FGSM +␣
↪Transformed_default_no_bitdepth")
```

```
FGSM + Transformed_default_no_bitdepth Evaluation:
  Total Samples: 10000
  Accuracy: 0.4563
  Loss: 2.1446
  Correct Predictions: 4563
  Incorrect Predictions: 5437
```

[28]:
```
(np.float64(0.4563), np.float32(2.144608))
```

[29]:
```
evaluate_model_on_dataset(pgd_ds, name="PGD + Transformed_default_no_bitdepth")
```

```
PGD + Transformed_default_no_bitdepth Evaluation:
  Total Samples: 10000
  Accuracy: 0.4613
  Loss: 2.1215
  Correct Predictions: 4613
  Incorrect Predictions: 5387
```

[29]:
```
(np.float64(0.4613), np.float32(2.1214905))
```

[30]:
```
evaluate_model_on_dataset(deepfool_ds, name="DeepFool +␣
↪Transformed_no_bitdepth")
```

```
DeepFool + Transformed_no_bitdepth Evaluation:
  Total Samples: 200
  Accuracy: 0.1750
  Loss: 5.0621
  Correct Predictions: 35
  Incorrect Predictions: 165
```

[30]:
```
(np.float64(0.175), np.float32(5.062115))
```

[31]:
```
# no noise
model = TransformedModel(
    load_original_model(),
    apply_bitdepth=True, bits=4,
    apply_noise=False,    # Noise addition disabled
    apply_jpeg=True, jpeg_quality=75,
    apply_blur=True, blur_sigma=0.5
)
```

[32]:
```
evaluate_model_on_dataset(clean_ds, name='Clean + Transformed_no_noise')
```

```
Clean + Transformed_no_noise Evaluation:
  Total Samples: 10000
  Accuracy: 0.9017
  Loss: 0.2887
  Correct Predictions: 9017
  Incorrect Predictions: 983
```

[32]: (np.float64(0.9017), np.float32(0.28868285))

[33]:
```
evaluate_model_on_dataset(fgsm_ds, name="FGSM + Transformed_no_noise")
```

```
FGSM + Transformed_no_noise Evaluation:
  Total Samples: 10000
  Accuracy: 0.7851
  Loss: 0.7013
  Correct Predictions: 7851
  Incorrect Predictions: 2149
```

[33]: (np.float64(0.7851), np.float32(0.7012873))

[34]:
```
evaluate_model_on_dataset(pgd_ds, name="PGD + Transformed_no_noise")
```

```
PGD + Transformed_no_noise Evaluation:
  Total Samples: 10000
  Accuracy: 0.8125
  Loss: 0.5893
  Correct Predictions: 8125
  Incorrect Predictions: 1875
```

[34]: (np.float64(0.8125), np.float32(0.58929026))

[35]:
```
evaluate_model_on_dataset(deepfool_ds, name="DeepFool + Transformed_no_noise")
```

```
DeepFool + Transformed_no_noise Evaluation:
  Total Samples: 200
  Accuracy: 0.3000
  Loss: 4.1993
  Correct Predictions: 60
  Incorrect Predictions: 140
```

[35]: (np.float64(0.3), np.float32(4.1993437))

[36]:
```
# no jpeg compression
model = TransformedModel(
```

```
    load_original_model(),
    apply_bitdepth=True, bits=4,
    apply_noise=True, noise_std=0.05,
    apply_jpeg=False,      # JPEG compression disabled
    apply_blur=True, blur_sigma=0.5

)
```

[37]:
```
evaluate_model_on_dataset(clean_ds, name='Clean + Transformed_no_jpeg')
```

```
Clean + Transformed_no_jpeg Evaluation:
  Total Samples: 10000
  Accuracy: 0.3874
  Loss: 2.3156
  Correct Predictions: 3874
  Incorrect Predictions: 6126
```

[37]: (np.float64(0.3874), np.float32(2.3156402))

[38]:
```
evaluate_model_on_dataset(fgsm_ds, name="FGSM + Transformed_no_jpeg")
```

```
FGSM + Transformed_no_jpeg Evaluation:
  Total Samples: 10000
  Accuracy: 0.3612
  Loss: 2.4278
  Correct Predictions: 3612
  Incorrect Predictions: 6388
```

[38]: (np.float64(0.3612), np.float32(2.4277833))

[39]:
```
evaluate_model_on_dataset(pgd_ds, name="PGD + Transformed_no_jpeg")
```

```
PGD + Transformed_no_jpeg Evaluation:
  Total Samples: 10000
  Accuracy: 0.3578
  Loss: 2.4544
  Correct Predictions: 3578
  Incorrect Predictions: 6422
```

[39]: (np.float64(0.3578), np.float32(2.4544492))

[40]:
```
evaluate_model_on_dataset(deepfool_ds, name="DeepFool + Transformed_no_jpeg")
```

```
DeepFool + Transformed_no_jpeg Evaluation:
```

```
  Total Samples: 200
  Accuracy: 0.1250
  Loss: 5.6194
  Correct Predictions: 25
  Incorrect Predictions: 175
```

[40]: (np.float64(0.125), np.float32(5.6193657))

[41]:
```python
# no blur, higher jpeg quality
model = TransformedModel(
    load_original_model(),
    apply_bitdepth=True, bits=4,
    apply_noise=True, noise_std=0.05,
    apply_jpeg=True, jpeg_quality=90,  # Higher JPEG quality
    apply_blur=False,   # Blur disabled
)
```

[42]:
```python
evaluate_model_on_dataset(clean_ds, name='Clean + Transformed__no_blur')
```

```
Clean + Transformed__no_blur Evaluation:
  Total Samples: 10000
  Accuracy: 0.2894
  Loss: 2.9377
  Correct Predictions: 2894
  Incorrect Predictions: 7106
```

[42]: (np.float64(0.2894), np.float32(2.937708))

[43]:
```python
evaluate_model_on_dataset(fgsm_ds, name="FGSM + Transformed_no_blur")
```

```
FGSM + Transformed_no_blur Evaluation:
  Total Samples: 10000
  Accuracy: 0.2784
  Loss: 2.9834
  Correct Predictions: 2784
  Incorrect Predictions: 7216
```

[43]: (np.float64(0.2784), np.float32(2.9833708))

[44]:
```python
evaluate_model_on_dataset(pgd_ds, name="PGD + Transformed_no_blur")
```

```
PGD + Transformed_no_blur Evaluation:
  Total Samples: 10000
  Accuracy: 0.2785
  Loss: 2.9884
```

```
  Correct Predictions: 2785
  Incorrect Predictions: 7215
```

[44]: (np.float64(0.2785), np.float32(2.9884486))

[45]:
```
evaluate_model_on_dataset(deepfool_ds, name="DeepFool + Transformed_no_blur")
```

```
DeepFool + Transformed_no_blur Evaluation:
  Total Samples: 200
  Accuracy: 0.0950
  Loss: 5.6535
  Correct Predictions: 19
  Incorrect Predictions: 181
```

[45]: (np.float64(0.095), np.float32(5.6534715))

[46]:
```
# no noise and no bitdepth
model = TransformedModel(
    load_original_model(),
    apply_bitdepth=False, bits=4, # Disable bitdepth
    apply_noise=False,     # Disable noise
    noise_std=0.05,
    apply_jpeg=True, jpeg_quality=75,
    apply_blur=True,
    blur_sigma=0.5
)
```

[47]:
```
evaluate_model_on_dataset(clean_ds, name='Clean +␣
 ↪Transformed_no_bitdepth_noise')
```

```
Clean + Transformed_no_bitdepth_noise Evaluation:
  Total Samples: 10000
  Accuracy: 0.9428
  Loss: 0.1624
  Correct Predictions: 9428
  Incorrect Predictions: 572
```

[47]: (np.float64(0.9428), np.float32(0.16239214))

[50]:
```
evaluate_model_on_dataset(deepfool_ds, name="DeepFool +␣
 ↪Transformed_no_bitdepth_noise")
```

```
DeepFool + Transformed_no_bitdepth_noise Evaluation:
  Total Samples: 200
  Accuracy: 0.2650
```

```
  Loss: 4.3098
  Correct Predictions: 53
  Incorrect Predictions: 147
```

[50]: (np.float64(0.265), np.float32(4.3098006))

[51]:
```python
evaluate_model_on_dataset(fgsm_ds, name="FGSM + Transformed_no_bitdepth_noise")
```

```
FGSM + Transformed_no_bitdepth_noise Evaluation:
  Total Samples: 10000
  Accuracy: 0.7741
  Loss: 0.6948
  Correct Predictions: 7741
  Incorrect Predictions: 2259
```

[51]: (np.float64(0.7741), np.float32(0.69478834))

[52]:
```python
evaluate_model_on_dataset(pgd_ds, name="PGD + Transformed_no_bitdepth_noise")
```

```
PGD + Transformed_no_bitdepth_noise Evaluation:
  Total Samples: 10000
  Accuracy: 0.8416
  Loss: 0.4605
  Correct Predictions: 8416
  Incorrect Predictions: 1584
```

[52]: (np.float64(0.8416), np.float32(0.46049884))

[ ]:
```python

```