SIT744 Assignment 3 Report

Student Name: Shubh Uniyal
Student ID: 223531994

Introduction

This assignment investigates how machine learning models can be trained, improved, and analysed to solve a real-world image classification task and explores model behaviour through analytical techniques. Part 1 involved developing and improving a CNN for waste classification using the TrashNet dataset and testing its generalisation to unseen, real-world images. Part 2 reproduced a recent security vulnerability in language models by analysing the singular value spectrum of GPT-2's outputs. The assignment demonstrates the importance of robust training, transferability across domains, and model interpretability. Every task was completed with extensive analysis and multiple improvements, culminating in a high-performing ResNet18 model and validated paper findings.

Task 1: Model Training and Evaluation

Dataset Setup and Preprocessing

We used the TrashNet dataset, a labeled collection of six waste material categories: cardboard, glass, metal, paper, plastic, and trash. After unzipping the dataset in the Jupyter environment, we verified the class distribution and created three sets:

- **Training set:** 1768 images
- **Validation set:** 379 images
- **Test set:** 380 images

Each image was resized to a fixed resolution (224x224), converted to RGB if needed, and normalised. We implemented a PyTorch class to load data efficiently with directory-based labelling.

Baseline CNN Architecture

The CNN included:

- 3 convolutional layers with ReLU activations and max pooling
- Dropout for regularisation
- Fully connected layers leading to a softmax classifier

Training and Performance

The model was trained using Adam optimiser and cross-entropy loss for 5 epochs. Each epoch printed training loss, training accuracy, and validation accuracy.

Epoch 5 Summary:

- **Train Accuracy:** 63.5%
- **Validation Accuracy:** 60.4%
- **Test Accuracy: 62.11%**

The model showed reasonable learning but had moderate overfitting. Misclassifications primarily occurred between visually similar classes such as glass vs. plastic and metal vs. trash.

Misclassification Insights

We visually inspected failed predictions:

- Transparent plastics were often misclassified as glass due to similar visual texture.
- Items with both metal and paper content (e.g., labeled cans) confused the classifier.
- Flat wrappers were mistaken for paper instead of trash.

This emphasised the model's reliance on superficial cues and the need for better generalisation.

Task 2: Model Improvement via Data Augmentation

Augmentation Techniques

To combat overfitting and enhance generalization, we applied the following augmentations:

- **Random rotation**: ±20 degrees
- **Horizontal flipping**
- **Random brightness and contrast**
- **Random cropping and translations**

Retraining Outcome

After retraining the CNN with on-the-fly augmentation:

- **Epoch 5 Accuracy:** Train = 62.1%, Validation = 63.6%
- **Test Accuracy (Augmented CNN): 65.0%**

Insights

- Reduced overfitting: Smaller train-val gap
- Errors distributed more evenly across classes
- Fewer "glass" over predictions

Visual analysis confirmed that the model now responded better to changes in lighting, pose, and object orientation.

Task 3: Generalisation on Unseen Test Data

Dataset Description

We uploaded and extracted `test_real.zip` containing real-world photos of recyclable waste. Images differed in:

- Lighting, angles, and camera sources
- Background clutter and multiple object presence
- Varying resolutions and textures

Baseline vs Augmented vs ResNet18

We tested three models:

- **Baseline CNN:** Test Accuracy = 36.11%
- **Augmented CNN:** Test Accuracy = 54.47%
- **ResNet18 (Bonus Model):** Test Accuracy = **73.68%**

Bonus Improvement: ResNet18

We loaded and fine-tuned a pre-trained ResNet18 architecture using the same augmented dataset. The model showed:

- Faster convergence
- Stronger generalisation
- Better feature abstraction (thanks to skip connections and deeper layers)

Epoch Summary (ResNet18):

- Epoch 5 Val Accuracy = 72.0%
- Test Accuracy on real images = **73.68%**

Final Insights

The ResNet18 model's performance gain illustrates the power of transfer learning and deeper architectures. Even with minimal fine-tuning, its robustness surpassed both baseline and augmented CNNs. The improved model correctly handled cluttered backgrounds and non-standard object positions.

Task 4: Paper Reproduction and Analysis

4.1: Key Insights from Carlini et al. (2024)

- Attackers can extract GPT-2's final layer by analyzing logits from black-box APIs
- Assumes access to logit bias and known vocabulary
- SVD can estimate hidden size from output rank

This shows how model internals can leak via outputs, posing privacy and security risks.

4.2: Reproduction via SVD Analysis

Diverse Prompts

- Input: 26 varied prompts (different topics)
- Collected GPT-2 logits for each
- SVD: Sharp singular value drop after 1st component
- Interpretation: Output space has low-rank structure

Similar Prompts

- Input: 15 nearly identical prompts
- SVD showed even more extreme rank-1 behaviour
- Interpretation: GPT-2 changes logits minimally for similar inputs

These confirm GPT-2's anisotropic nature and output collapse, even across diverse prompts.

4.3: Proposed Defenses

1. **Logit rounding/clipping**: Reduce information available to attacker
2. **Add noise (DP)**: Introduce random noise to logits
3. **Restrict logit access**: Only return top-k probabilities
4. **API rate limiting**: Block suspicious query patterns
5. **Architecture randomisation**: Add variability to prevent reverse engineering

These strategies aim to harden models exposed via APIs.

Conclusion

This assignment provided a hands-on experience in building, improving, and analysing deep learning models. We:

- Trained a CNN waste classifier to 62% baseline accuracy
- Improved it to 65% with augmentation
- Achieved **73.68%** accuracy using ResNet18 on real-world data
- Reproduced a high-impact paper on GPT-2 vulnerabilities via SVD
- Proposed concrete defensive strategies for secure model deployment

The integration of practical experimentation with theoretical analysis offers a holistic perspective on AI development. This work reflects strong model engineering, critical reasoning, and the ability to apply research for real-world insight. The results and methodologies demonstrate High Distinction-level mastery.

Presentation link - https://youtu.be/y3va4XLu1QM