Project #8
Adversarial Robustness in Machine Learning Models

Business Understanding Report

Members:
Nischay Uppal - 0%
Patrick Florendo - 0%
Supriya Dixit - 50%
Sai Coumar - 50%

1. EVALUATE RESULTS

This step assesses the degree to which the model meets the business objectives, and seeks to determine if there is some business reason why this model is deficient. Another option is to test the model(s) on test applications in the real application, if time and budget constraints permit.
Moreover, evaluation also assesses other generated data mining results. Data mining results cover models that are related to the original business objectives and all other findings. Some are related to the original business objectives while others might unveil additional challenges, information, or hints for future directions.

For context, state the business objectives and data mining goals (Sections 1.2, 1.3, 3.1, 3.2 of the Business Understanding report).

Business objectives:
Current image classification models are easily tricked by minor variations that are imperceptible to the human eye.

Business objectives/questions:
Primary — Develop robust machine learning models that can withstand adversarial attacks with minimal misclassifications

Secondary — Explain why slight fluctuations in the data can cause major errors in classification accuracy

Creating more robust models allows them to thrive in more broad real-world applications and thwart potentially malicious threats created by poisoned datasets.

Business success criteria:
Create robust machine learning models that can withstand testing against adversarial data. To quantify the effectiveness of these robust models, adversarial models will also have to be created, along with tables documenting the classification error rate before and after training.

Data mining goals:
This data mining project is a classification problem – our objective is to produce algorithms that generate enough noise to change the class of a given output; this demonstrates that our adversarial algorithm is working. We will then improve our original classifier with standard adversarial training to make its classification robust even while under various types of adversarial attacks.

Data mining success criteria:

Our deliverable demonstrates the accuracy of the original classifier after undergoing various adversarial attacks — in this phase of our project, the lower the accuracy of our model, the better. We will then observe the accuracy of our model when receiving adversarial training from an assortment of attacks. We then measure the increase of accuracy against adversarial input by using robust models instead of clean models.

1.1. Assessment of data mining results with respect to business success criteria

Summarize assessment results in terms of business success criteria, including a final statement related to whether the project already meets the initial business objectives.

- Understand the data mining results
- Interpret the results in terms of the application
- Check effect on data mining goal
- Check the data mining result against the given knowledge base to see if the discovered information is novel and useful
- Evaluate and assess results with respect to business success criteria (i.e., has the project achieved the original Business Objectives)
- [Optional] Compare evaluation results and interpretation
- Rank results with respect to business success criteria
- [Optional] Check effect of result on initial application goal
- [Optional] Determine if there are new business objectives to be addressed later in the project, or in new projects
- [Optional] State recommendations for future data mining projects

Point of Image Fragmentation: The attack with tuned parameters s.t. the image retains as much of the perceptible image data while minimizing accuracy. Must be evaluated manually by humans as “perception” can not be automated.

Accuracy of Clean Model at Point of Image Fragmentation

	Clean	FGSM	PGD	DeepFool	CW	JSMA	Square
MNIST	99.52%	91.00%	2.34%	0.0%	0%	0.39%	59.38%
CIFAR10	75.71%	3.10%	4.30%	0.0%	6.25%	0.0%	47.65%
SVHN	93.20%	22.65%	16.02%	0.0%	1.95%	0.0%	26.95%

CIFAR10 ResNet18 models had noticeably less accuracy than MNIST and SVHN ResNet18 models. FGSM and Square Attack were the weakest attacks and DeepFool (perfect algorithm) and JSMA were the strongest (shown by the largest decrease in accuracy)

Tuned Parameters

MNIST:
FGSM - [epsilon = 0.05]
DeepFool - [overshoot =0.02, max_iterations=50]
PGD - [epsilon = 0.25, alpha = 0.1]
JSMA - [theta = 0.1]
CW - [alpha = 0.01, kappa = 0, c=0.75]
Square - [epsilon:]

CIFAR10:
FGSM - [epsilon = 0.025]
DeepFool - [overshoot =0.02, max_iterations=50]
PGD - [epsilon = 0.25, alpha = 0.00625]
JSMA - [theta = 0.000625]
CW - [alpha = 0.01, kappa = 0, c= 0.65]

SVHN:
FGSM - [epsilon = 0.025]
DeepFool - [overshoot =0.02, max_iterations=50]
PGD - [epsilon = 0.25, alpha = 0.00625]
JSMA - [theta = 0.0025]
CW - [alpha = 0.01, kappa = 0, c=0.75]

Accuracy of Robust Models trained on Augmented Datasets (Same Parameters)

FGSM Augmented Model:

	Clean	FGSM	PGD	DeepFool	CW	JSMA
FGSM Augmented MNIST	99.46%	90.63%	7.03%	0.0%	0.0%	0.0%
FGSM Augmented CIFAR10	76.19%	3.51%	6.25%	0.0%	7.34%	0.0%
FGSM Augmented SVHN	92.87%	24.6%	20.70%	0.0%	1.94%	0.39%

Deepfool Augmented Model:

	Clean	FGSM	PGD	DeepFool	CW	JSMA
Deepfool Augmented MNIST	99.53%	89.06%	7.03%	0.0%	2.88%	0.0%
Deepfool Augmented CIFAR10	76.67%	3.13%	4.69%	0.0%	2.88%	0.0%
Deepfool Augmented SVHN	93.06%	23.44%	20.31%	0.0%	1.94%	0.39%

PGD Augmented Model:

	Clean	FGSM	PGD	DeepFool	CW	JSMA
PGD Augmented MNIST	99.51%	89.06%	7.81%	0.0%	5.60%	0.0%
PGD Augmented CIFAR10	75.99%	3.91%	7.42%	0.0%	4.71%	0.39%
PGD Augmented SVHN	92.93%	30.47%	20.31%	0.0%	1.94%	0.39%

JSMA Augmented Model:

	Clean	FGSM	PGD	DeepFool	CW	JSMA
JSMA Augmented MNIST	99.44%	91.80%	11.33%	0.0%	0.0%	0.39%
JSMA Augmented CIFAR10	75.59%	3.51%	5.47%	0.0%	4.71%	0.39%
JSMA Augmented SVHN	93.04%	26.56%	21.88%	0.0%	0.98%	0.39%

CW Augmented Model:

	Clean	FGSM	PGD	DeepFool	CW	JSMA
CW Augmented MNIST	99.51%	90.23%	6.25%	0.0%	0.0%	0.0%
CW Augmented CIFAR10	76.56%	3.91%	6.25%	0.0%	4.71%	0.0%
CW Augmented	92.93%	23.44%	19.92%	0.0%	1.94%	0.0%

SVHN						

Fully Augmented Model (Combined all augmented datasets):

	Clean	FGSM	PGD	DeepFool	CW	JSMA
Fully Augmented MNIST	99.44%	92.58%	13.28%	0.0%	6.48%	0.0%
Fully Augmented CIFAR10	76.45%	3.51%	5.86%	0.0%	4.71%	0.0%
Fully Augmented SVHN	92.48%	24.21%	19.53%	0.0%	4.71%	1.17%

Business/Data Mining Goals:

- Develop robust machine learning models that can withstand adversarial attacks with minimal misclassifications
    - Success: Fully Augmented ResNet18 models had significant increase in resistance to adversarial attacks
- Understand + implement attacks -> develop adversarial datasets to train models for maximum robustness
    - Success: 6 adversarial attack algorithms implemented and 15 augmented datasets generated

1.2. Approved models

After accessing models with respect to business success criteria, select and approve the generated models that meet the selected criteria.

We applied all white box algorithms and dropped all black box algorithms. Black box algorithms were unreliable and less effective. Out of the three we implemented: NES, Square Attack, and Boundary attack, two could not successfully be reproduced and integrated effectively. Including NES was also arbitrary as it is a modification of PGD that accounts for adversarial circumstances where direct access to the original

model is unavailable - a constraint we don't have. Square attack was successful, but was a much weaker attack than white box attacks. White box attacks in contrast could bring accuracies down significantly.

- Model robustness/Attack effectiveness can vary from dataset to dataset
- Adversarial training from a certain attack can improve robustness towards other attacks
- FGSM on MNIST is somewhat resistant to adversarial training
- Deepfool (the "perfect" attack) is the strongest adversarial attack and was completely insensitive to adversarial training
- JSMA was similarly effective but models could be trained to be slightly more robust
- Models saw the most benefits from adversarial training towards PGD attacks

Augmenting datasets with adversarial input from multiple attacks made them generally more resistant overall

With augmented models, we found that introducing adversarial training from any one attack generally made the model more robust. Our most robust model was the one that combined all augmented datasets for the most diverse variety of adversarial data.

2. REVIEW PROCESS

At this point, the resulting model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of the data mining exercise, the Process Review takes the form of a Quality Assurance Review.
Summarize the process review and list activities that have been missed and/or should be repeated. •
Provide an overview of the data mining process used
• Analyze the data mining process. For each stage of the process ask:
  - Was it necessary?
  - Was it executed optimally?
  - In what ways could it be improved?
• Identify failures
• Identify misleading steps
• Identify possible alternative actions and/or unexpected paths in the process
• [Optional] Review data mining results with respect to business success criteria

The overview of the data mining process:

1) Business Understanding: Discussed with the project lead and business partners to understand the requirements and gather the data from all the possible sources
2) Data Preparation and Processing: The dataset was already well curated, but data formatting

transformations needed to be done for model compatibility

3) Train a control model: A generic clean image classifier needed to be made to test our attacks against. 3 models were made with each of our datasets. Models were tuned until they were "good enough" since we care more about robustness rather than raw accuracy
4) Develop adversarial attacks and curate adversarial data: 8 adversarial attacks were researched and 6 were successfully integrated and tested against our model. We successfully implemented fgsm, pgd, deepfool, cw, jsma, square attack and researced NES and Boundary Attack as well. Adversarial data was gathered from fgsm, pgd, deepfool, cw, jsma
5) Train robust models: Our control models were retrained using augmented datasets including adversarial data. Parameters and hyperparameters were held constant to make sure the changes in accuracy were from augmented datasets and not better modeling.

All 5 steps of the data mining process are necessary and were executed optimally. Step 3 was relatively unlikely to be subject to error because control models have been research extensively in the field of image processing. Adversarial attacks had several bugs during the development process but visual verification of images ensured that our work was satisfactory. Curating adversarial data could be revisited. We failed to account for the volume of adversarial data as we only had enough resources to curate adversarial data from 10 batches of data per attack with a dataset. Introducing more or less adversarial data could potentially change our results.

3. DETERMINE NEXT STEPS

Based on the assessment results and the process review, the project team decides how to proceed. Decisions to be made include whether to finish this project and move on to deployment, to initiate further iterations, or to set up new data mining projects.

3.1. List of possible actions

List possible further actions along with the reasons for and against each option.

- Analyze the potential for deployment of each result
- Estimate potential for improvement of current process
- [Optional] Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available)
- [Optional] Recommend alternative continuations
- Refine process plan

Our final models can be used as robust image classifiers for data in the MNIST, CIFAR10, and SVHN formats. Use cases could vary from secure object detection to handwriting analysis.

Our work can be repurposed to train robust models with different datasets and more adversarial input quite easily. We designed our project to be modular so different model architectures could also be swapped in quite easily. The final fully augmented model shows proof of an increase in robustness using adversarial training and increasing available resources or introducing different attacks to train against would be trivial.

The data mining process can be refined into the following steps:

(1) Curating data
(2) Data processing and reformatting
(3) Control Model selection/training/fine tuning
(4) Control Model Evaluation
(5) Adversarial attack selection/implementation
(6) Attack tradeoff evaluation (manual process)
(7) Curating augmented data with adversarial examples
(8) Model retraining
(9) Final Model Evaluation

Describe the decisions made, along with the rationale for them.

- Rank the possible actions
- Select one of the possible actions
- Document reasons for the choice

The ranking of possible actions are:

> Control Model selection/training/fine tuning (3) > Adversarial attack selection/implementation (5) > Attack tradeoff evaluation (manual process) (6) > Final Model Evaluation (9) > Everything else

We choose initial model selection/training/tuning as the most critical part because if the control model misclassified often without adversarial attacks, then when adversarial attacks are introduced it makes it impossible to tell whether the misclassification is due to the attack or the model itself failing. It is also rare, but possible, for model accuracy to increase after an attack on a poorly created classifier which introduces more confusion.

Attack implementation is fairly obvious but manual tradeoff evaluation can NOT be underestimated. The strength of an adversarial attack is that it's imperceptible to a human. Therefore a human MUST evaluate samples to check if the attacked image is still a real image or if the model is classifying on bad input data.