Project #8
Adversarial Robustness in Machine Learning Models

Business Understanding Report

Members:
Nischay Uppal - 25%
Patrick Florendo - 25%
Rishabh Pandey - 25%
Supriya Dixit - 25%
Sai Coumar - 0%

1. DETERMINE BUSINESS OBJECTIVES

The first objective of the analyst is to thoroughly understand, from a business perspective, what the customer really wants to accomplish. Often the customer has many competing objectives and constraints that must be properly balanced. The analyst's goal is to uncover important factors at the beginning of the project that can influence the final outcome. A likely consequence of neglecting this step would be to expend a great deal of effort producing the correct answers to the wrong questions.

1.1. Background

Collate the information that is known about the organization's business situation at the start of the project. These details not only serve to more closely identify the business goals to be achieved but also serve to identify resources, both human and material, that may be used or needed during the course of the project.

Organization

- [Optional] Develop organizational charts identifying divisions, departments, and project groups. The chart should also identify managers' names and responsibilities
- [Optional] Identify key persons in the business and their roles
- Identify an internal sponsor (financial sponsor and primary user/domain expert)
- [Optional] Indicate if there is a steering committee and list members
- Identify the business units which are affected by the data mining project (e.g., Marketing, Sales, Finance)

Professor Guang Lin and Radjeep Haldar are the internal sponsor domain experts for the project for the analysis of adversarial attacks on machine learning models.

This project spans Purdue’s Machine Learning research unit.

Problem area

- Identify the problem area (e.g., marketing, customer care, business development, etc.)
- Describe the problem in general terms
- [Optional] Check the current status of the project (e.g., Check if it is already clear within the business unit that a data mining project is to be performed, or whether data mining needs to be promoted as a key technology in the business)
- Clarify prerequisites of the project (e.g., What is the motivation of the project? [Optional] Does the business already use data mining?)
- [Optional] If necessary, prepare presentations and present data mining to the business
- [Optional] Identify target groups for the project result (e.g., Are we expected to deliver a report for top management or an operational system to be used by naive end users?)

- Identify the users' needs and expectations

The problem area is within the intersection of generative AI and adversarial robustness.

This project deals with the problem that image recognition models are currently vulnerable to adversarial attacks — images that have been imperceptibly modified specifically so that they become misclassified.

It is important to ensure that image recognition models are robust enough to repel adversarial attacks before their use becomes more widespread in the case of malicious threat actors attempting to take advantage of vulnerable systems trained on poisoned data. The models and attacks will be generated from already existing image sets, such as MNIST.

Haldar and Professor Lin expect to create adversarial models that can be used to explain why current models are so vulnerable to perturbations in the data and to train more robust models that can withstand the attacks.

Current solution

- Describe any solution currently used to address the problem
- Describe the advantages and disadvantages of the current solution and the level to which it is accepted by the users

Currently, many well-documented adversarial algorithms can be used to generate poisoned data that will trick image classification models.

However, many aspects of these adversarial models are still unexplored, such as how much they hurt the accuracy of current models, how effective models are after training with the adversarial data, and most importantly, why they have such an adverse effect on current models despite the noise not being visible to the naked eye.

1.2. Business Objectives

Describe the customer's primary objective, from a business perspective. In addition to the primary business objective, there are typically a large number of related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor, while a secondary business objective might be to determine whether lower fees affect only one particular segment of customers.

- Informally describe the problem to be solved
- Specify all business questions as precisely as possible

- [Optional] Specify any other business requirements (e.g., the business does not want to lose any customers)
- Specify expected benefits in business terms

Beware of setting unattainable goals—make them as realistic as possible.

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.

1.3. Business Success Criteria

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and readily measurable, such as reduction of customer churn to a certain level, or general and subjective, such as "give useful insights into the relationships." In the latter case, be sure to indicate who would make the subjective judgment.

- Specify business success criteria (e.g., Improve response rate in a mailing campaign by 10 percent and sign-up rate by 20 percent)
- Identify who assesses the success criteria Each of the success criteria should relate to at least one of the specified business objectives.

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.

Haldar and Professor Lin will assess the success of the project based on the new models' abilities to resist adversarial attacks.

2. ASSESS SITUATION

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and in developing the project plan.

2.1. Inventory of resources

List the resources available to the project, including personnel (business and data experts, technical support, data mining experts), data (fixed extracts, access to live warehoused or operational data), computing resources (hardware platforms), and software (data mining tools, other relevant software).

Hardware resources

- Identify the base hardware
- Establish the availability of the base hardware for the data mining project
- [Optional] Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project
- [Optional] Identify the hardware available for the data mining tool to be used (if the tool is known at this stage)

Hardware resources available for the project include access to the scholar computer cluster. Specifically, we have access to the G, H, and H-MIG sub-clusters. The G sub-cluster has 4 nodes, 16 cores per node, and 192GB of memory per node. The H sub-cluster has 2 nodes, 64 cores per node, and 512GB memory per node. The H-MIG sub-cluster has 2 nodes, 64 cores per node, and 512GB per node.

All team members have user-level account access to the cluster.

Sources of data and knowledge

- Identify data sources
- Identify type of data sources (online sources, experts, written documentation, etc.)
- [Optional] Identify knowledge sources
- [Optional] Identify type of knowledge sources (online sources, experts, written documentation, etc.)
- Check available tools and techniques
- Describe the relevant background knowledge (informally or formally)

For our research, we will be using the following pre-processed datasets: MNIST, CIFAR 10, and SVHN. The MNIST dataset is a large collection of handwritten digits. It has a training set of 60,000 examples and a test set of 10,000 examples. The CIFAR 10 dataset contains 60,000

32x32 color images in 10 different classes. These classes range from different animals to vehicles. The SVHN dataset contains 600,000 32x32 color images of printed digits (from 0 to 9) cropped from pictures of house number plates.

The computer cluster provides a Linux remote desktop, a Jupyter notebook server, and a Python environment.

Jupyter can be used without any reliance on Linux knowledge or experience. The relevant background knowledge includes CNN knowledge (provided in CS37300), basic PyTorch, and Computer Vision Knowledge.

Personnel sources

- Identify project sponsor (if different from internal sponsor as in Section 1.1)
- [Optional] Identify system administrator, database administrator, and technical support staff for further questions
- Identify market analysts, data mining experts, and statisticians, and check their availability
- Check availability of domain experts for later phases

2.2. Requirements, assumptions, and constraints

List all requirements of the project, including schedule of completion, comprehensibility, and quality of results and security, as well as legal issues. As part of this output, make sure that you are allowed to use the data. List the assumptions made by the project. These may be assumptions about the data, which can be verified during data mining, but may also include non-verifiable assumptions related to the project. It is particularly important to list the latter if they will affect the validity of the results. List the constraints made on the project. These constraints might involve lack of resources to carry out some of the tasks in the project in the time required, or there may be legal or ethical constraints on the use of the data or the solution needed to carry out the data mining task.

Requirements

- [Optional] Specify target group profile
- [Optional] Capture all requirements on scheduling
- Capture requirements on comprehensibility, accuracy, deployability, maintainability, and repeatability of the data mining project and the resulting model(s)
- Capture requirements on security, legal restrictions, privacy, reporting, and project schedule

The model should be deployed on the cluster machines and should be reproducible. It should not use any other additional data except for the official dataset provided by MNIST, CIFAR 10, and SVHN.

Since this is publicly available data, there are no privacy concerns associated with this project. This project should be completed in the next ~2 months.

Assumptions

- Clarify all assumptions (including implicit ones) and make them explicit (e.g., to address the business question, a minimum number of customers with age above 50 is necessary)
- List assumptions on data quality (e.g., accuracy, availability)
- [Optional] List assumptions on external factors (e.g., economic issues, competitive products, technical advances)
- [Optional] Clarify assumptions that lead to any of the estimates (e.g., the price of a specific tool is assumed to be lower than $1,000)
- List all assumptions regarding whether it is necessary to understand and describe or explain the model (e.g., how should the model and results be presented to senior management/sponsor)

The project assumes the numerical and pictorial data provided will be a good dataset to train robust models based on adversarial attacks.

It is assumed that the data set provides enough data for our models to learn. We also assume that a general model for the MNIST + SVHN, and CIFAR predictions can be learned from past data.

An objective evaluation score will be used to measure the performance of different models to automate finding the best model and provide an easy way to rank them. Our deliverable will contain accuracy scores for models trained on multiple categories of adversarial attacks. The presented models should be straightforward, brief, and easy to understand, and the results of the models should contain the evaluation which follows the definition of the evaluation score.

Constraints

- Check general constraints (e.g., legal issues, budget, timescales, and resources)
- Check access rights to data sources (e.g., access restrictions, password required)
- Check technical accessibility of data (operating systems, data management system, file or database format) • Check whether relevant knowledge is accessible
- [Optional] Check budget constraints (fixed costs, implementation costs, etc.)

The 3 primary datasets are only sources provided, and we should not look for any external data source. We need to make sure to use the data within ethical boundaries.

The tools and hardware should be equipped to handle CSV files and analyze more than 100,000 images and save the required outputs smoothly.

All the relevant images are accessible through PyTorch libraries!

2.3. Risks and contingencies

List the risks, that is, the events that might occur, impacting schedule, cost, or result. List the corresponding contingency plans: what action will be taken to avoid or minimize the impact or recover from the occurrence of the foreseen risks.

Identify risks

- [Optional] Identify business risks (e.g., competitor comes up with better results first)
- [Optional] Identify organizational risks (e.g., department requesting project doesn't have funding for the project)
- [Optional] Identify financial risks (e.g., further funding depends on initial data mining results)
- Identify technical risks
- Identify risks that depend on data and data sources (e.g., poor quality and coverage)

Since this is a widely-used open-source/standardized dataset, there typically shouldn't be problems with missing data values. If there happen to be any missing values (or fewer data values for a particular class), it will result in underfitting the model for that particular class.

Develop contingency plans

- Determine conditions under which each risk may occur
- Develop contingency plans

There are no conditions under which our dataset might be affected! Hypothetically, if it did happen, we could use another open-source dataset instead.

2.4. Terminology

Compile a glossary of terminology relevant to the project. This should include at least two components: (1) A glossary of relevant business terminology, which forms part of the business understanding available to the project (2) A glossary of data mining terminology, illustrated with examples relevant to the business problem in question

- [Optional] Check prior availability of glossaries; otherwise begin to draft glossaries
- Talk to domain experts to understand their terminology
- Become familiar with the business terminology

AT: Adversarial Training

FGSM: Fast-signed gradient max

PGD: Projected-gradient-descent

CW: Carlini and Wagner

White Box: attacks that require access to the trained model

Black Box: attacks that don't require access to the trained model

Transferrable attacks: depend on transferability; require access to the training dataset

Score-based attacks: only rely on the predicted scores

Decision-based attacks: direct attacks that solely rely on the final decision of the model

MNIST, CIFAR 10, SVHN: Datasets utilized in our research

2.5. Costs and benefits

Prepare a cost-benefit analysis for the project, comparing the costs of the project with the potential benefits to the business if it is successful

- [Optional] Estimate costs for data collection
- [Optional] Estimate costs of developing and implementing a solution
- Identify benefits (e.g., improved customer satisfaction, ROI, and increase in revenue)
- [Optional] Estimate operating costs
- [Optional] Remember to identify hidden costs, such as repeated data extraction and preparation, changes in workflows, and time required for training.

Since this is more of a research-based project, there is no immediate financial incentive for completing it. Some business applications of this research are in the security field: identity theft, automated driving, bypassing spam filters, etc.

Costs: Computational Resources, Power (used for training significantly large models), Time
Benefit: Increased Security of Models

Successful implementation will improve model robustness against adversarial attacks to make more secure systems to prevent general data poisoning that can have broader consequences.

3. **DETERMINE DATA MINING GOALS**

A business goal states objectives in business terminology; a data mining goal states project objectives in technical terms. For example, the business goal might be, "Increase catalog sales to existing customers," while a data mining goal might be, "Predict how many widgets a customer will buy, given their purchases over the past three years, relevant demographic information, and the price of the item."

3.1. Data mining goals

Describe the intended outputs of the project that enable the achievement of the business objectives. Note that these are normally technical outputs.

- Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).

- Specify data mining problem type (e.g., classification, description, prediction, and clustering).

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.

3.2. Data mining success criteria

Define the criteria for a successful outcome to the project in technical terms, for example, a certain level of predictive accuracy or a propensity-to-purchase profile with a given degree of "lift." As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.

- Specify criteria for model assessment (e.g., model accuracy, performance and complexity)
- Define benchmarks for evaluation criteria
- Specify criteria which address subjective assessment criteria (e.g., model explainability and data and marketing insight provided by the model)
  Remember that the data mining success criteria are different than the business success criteria defined earlier

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 with the Carlini & Wagner attack and the Projected Gradient Descent attack. We then measure the TRADES—a metric that describes the tradeoff between model robustness and accuracy—with two optimal parameters β1 and β2

4. PRODUCE PROJECT PLAN

Describe the intended plan for achieving the data mining goals and thereby achieving the business goals.

4.1. Project plan

List the stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. Wherever possible, make explicit the large-scale iterations in the data mining process—for example, repetitions of the modeling and evaluation phases. As part of the project plan, it is also important to analyze dependencies between time schedule and risks. Mark results of these analyses explicitly in the project plan, ideally with actions and recommendations for actions if the risks are manifested.

Although this is the only task in which the project plan is directly named, it nevertheless should be consulted continually and reviewed throughout the project. The project plan should be consulted at minimum whenever a new task is started or a further iteration of a task or activity is begun.

Define the initial project plan [Optional] and discuss the feasibility with all involved personnel

- Combine all identified goals and selected techniques in a coherent procedure that solves the business questions and meets the business success criteria
- [Optional] Estimate the effort and resources needed to achieve and deploy the solution. (It is useful to consider other people's experience when estimating timescales for data mining projects. For example, it is often postulated that 50-70 percent of the time and effort in a data mining project is used in the Data Preparation Phase and 20-30 percent in the Data Understanding Phase, while only 10-20 percent is spent in each of the Modeling, Evaluation, and Business Understanding Phases and 5-10 percent in the Deployment Phase.)
- Identify critical steps
- [Optional] Mark decision points
- [Optional] Mark review points
- [Optional] Identify major iterations

The project duration is January 2023 to May 2023, consisting of 13 weeks of project research and development. The general timeline consists of understanding the different tools and techniques to implement adversarial attacks for the first 5 weeks, followed by adversarial network implementation for the next 5 weeks, and consolidation with documentation and results in the final 3 weeks.

The basic idea of the project is to implement different adversarial attack methods within three general groupings and study their effects on the error rates of adversarial deep learning models. Based on the problem statement, we are studying a classification problem.

A critical step will be to establish a comprehensive set of comparison metrics to evaluate these attacks and identify the characteristics of each.

4.2 Initial Assessment of Tools and Techniques

At the end of the first phase, the project team performs an initial assessment of tools and techniques. Here, it is important to select a data mining tool that supports various methods for different stages of the process, since the selection of tools and techniques may influence the entire project.

- Create a list of selection criteria for tools and techniques (or use an existing one if available)
- Choose potential tools and techniques
- Evaluate appropriateness of techniques
- Review and prioritize applicable techniques according to the evaluation of alternative solutions

Due to the nature of our work, we do not have selection criteria. It is a given that there are a variety of fundamental adversarial attacks that fall into three buckets: white-box attacks, black-box attacks, and score-based attacks, and we seek to implement these methods and observe their effects.

These attacks and the adversarial networks will be implemented using PyTorch with GPU clusters in the Purdue Scholar environment. The underlying development environment is Anaconda. We do not anticipate leveraging cloud-based computing clusters.

The size of the image datasets are as follows: SVHM -> CIFA10 -> MNIST

Select image datasets may be small enough to run locally without the use of GPUs. The data generation process will apply within the bounds of the size of each image during processing.