The Effect of Dimension on Adversarial Learning Phenomena

Computer Science Master's Degree Proposal

Writer: Amit Leibovitz
Advisor: Adi Shraibman
School of Computer Science - The Academic Tel Aviv-Yaffo, Israel
March 2018

Contents

1. Abstract

1.1. Purpose

In recent years we have been experiencing increasing use of products and technologies based on machine learning and artificial intelligence. This phenomenon exists in almost every aspect of our daily lives: web search, online translation, phone's voice personal assistant and fingerprint locking, target advertising and much more. In the coming years, even more advanced technologies that in the past may have been science fiction, including the autonomous vehicle and IOT products, will make a significant change in our lives.

But like the revolution of the personal computer, which become common in every home during the 1980s and the 1990s, the endless race to the upgrade and progress has resulted in security neglect. Machine learning algorithm have always been tested for how much the model was right: accuracy, FP/FN, precision/recall and so on. There has never been a criterion for examining the durability of a model against intentional malformed input. Just imagine the impact of a cyber-attack adding minor changes to the sensors of the autonomous vehicle, causing it to be confused between a stopover sign and a highway sign.

1.2. Problem Statement

In this research we would like to explore a new risen field that began only in 2014 and focuses on methods of deliberate deception of learning models. This field got the name *Adversarial Learning*. We will attempt to examine the effect of the input dimension on the attack surface, focusing on common adversarial attacks in the image processing domain, hoping to have a contribution to the adversarial learning defense methods.

2. Introduction

2.1. Machine Learning

Machine Learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome.

Machine Learning was first mentioned in 1959 by Arthur Samuel. ML has been studied from various perspectives, including artificial intelligent and statistics. Evolved from data mining and pattern recognition, ML explores the study of algorithms that can learn from and make predictions and decisions on a given data.

The ability to 'learn' is defined by progressively improve performance on a specific task.

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks T, as measured by P, improved with experience E." (Tom Mitchell, 1997).

Machine learning is divided into three main different classes. This includes algorithms for supervised, unsupervised and reinforcement learning.

In supervised learning the learning process is guided. For example, in classification problems, the algorithm trains the model with a labeled data, that is, a collection of samples, each of them is assigned with a class label.

In unsupervised learning the range of problems that the algorithm can solve is much smaller and is limited to clustering, dimension reduction and outlier detection.

In the more sophisticated problem of reinforcement learning, the user provides a feedback on the correctness of the algorithm decision.

Above those classes, supervised learning will be the main kind of learning we will focus on. In the basic statistical supervised learning model, the model has a set of *N* labeled points which will be our *training set*. Meaning, for the training process includes N input vectors $\{x^i\}_{i=1}^N, x^i \in \mathbb{R}^n$ and N class labels $\{y^i\}_{i=1}^N, y^i \in \{0,1\}$ for the discrete *classification* problem and $\{y^i\}_{i=1}^N, y^i \in \mathbb{R}$ for the continuous *regression* problem. The training model will give a prediction for every new data point. Testing the model with a new labeled dataset, called the *test* set, can give us a measure of how good our model is. If we define our hypothesis output for new data point *x* as *h(x)* and the real known output as *f(x)*, then the accuracy of our model can be define as

$$(1)\quad L(\Theta) = \sum_{i=1}^{n} l\left(f(x_i), h(x_i, \Theta)\right)$$

where Θ is the learning parameters of the model and *l* is a loss function measuring the error of the prediction with respect to the real value. Some examples for such loss functions can be the square loss function and the logistic loss function. The total L function is referred to as the *Training Loss Function*.

In order to avoid over fitting and keeping the model as simple as possible we also define the regularization function $\Omega\left(h(\Theta)\right)$ which measure the complexity of the model.

So, the objective function we wish to minimize in this optimization problem should be

$$(2)\quad Obj(\Theta) = L(\Theta) + \Omega(h(\Theta))$$

The machine learning process illustrated in Figure [1] shows the main stages during a machine learning experiment. The labeled data is divided into *training and test sets*. After Exploratory Data Analysis, which include cleaning, transforming and visualization of the input data, a feature extraction stage gives us the input X. This input will be inserted into a machine learning model giving us a predicted value. The model includes some learned parameters. The actual value and the metric value can be used to define the accuracy of our model.

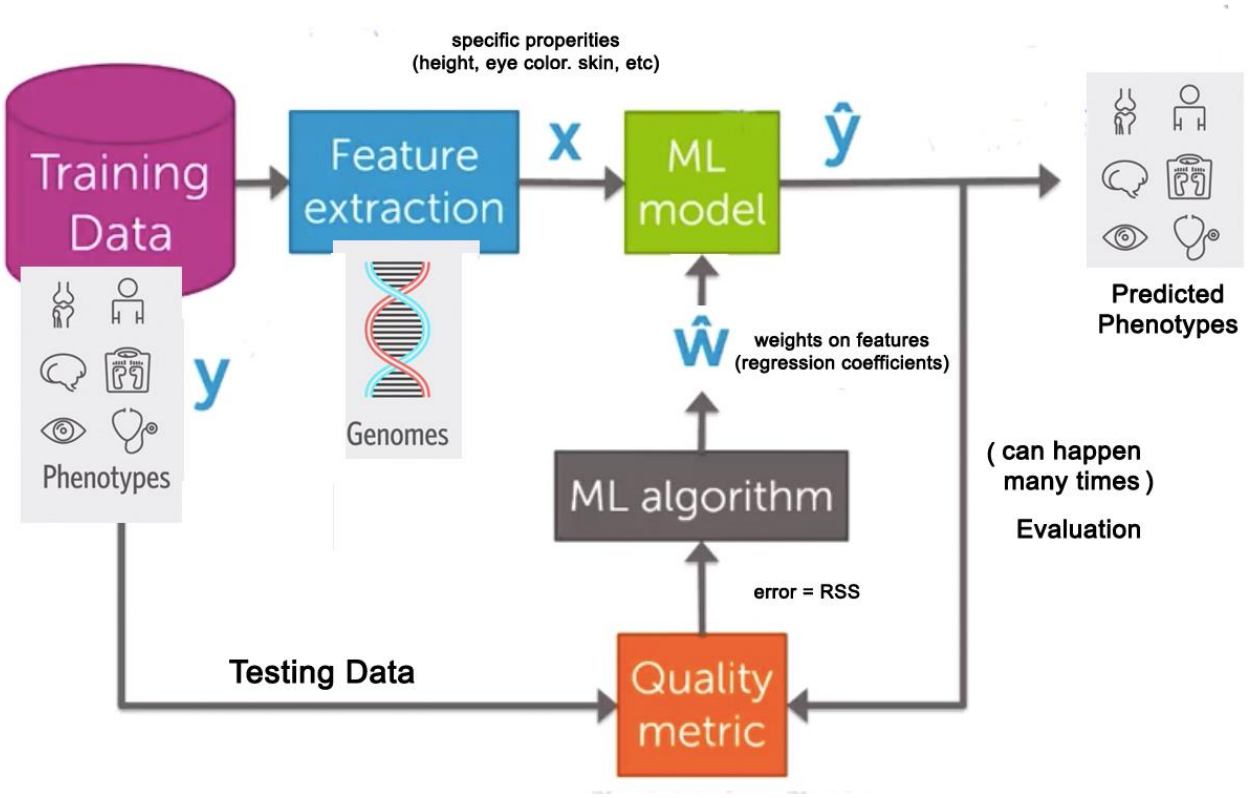


FIGURE 1: THE MACHINE LEARNING PROCESS

2.2. The Curse of Dimensionality

Real life datasets typically come with high dimension, like the number of pixels in an image or the number of different words in a text document. The true dimension is often much lower. For example, if we are looking at 20x20 handwritten digits image the input dimension is $\{0,1\}^{20\times20}$ which is the total possible for input images. If we choose a random image from this domain, most chances it will not be a handwritten digit, but some random black and white image. In fact, of handwritten is only a tiny fraction of events in this large input space. So, why high dimension input space could be a problem?

The Curse of Dimensionality describes a phenomenon in which when the input dimension increases, the volume of the space increases exponentially, making the sparse. If for example 10 data points seems reasonable for 1-dimensional space, in 2-dimensional space we'll need 100 points for the same density of points and 1000 points for the 3-dimensional case. Most machine learning algorithms are statistical by nature, using counting of observations in various regions of some space and distance measures. Those two fail when dimension is increasing.

2.3. Deep Learning

In recent years we have been hearing about more and more domains where the ML technologies based on deep learning had led to breakthrough that has not been seen before. These domains include tasks that not so long ago seemed very difficult, such as voice recognition (Apple's Siri), natural language tasks (translation, auto complete, virtual assistants) and real time computer vision classifiers (the autonomous vehicle for example).

The artificial neural network uses repeated application of a nonlinear function $\sigma(x)$. In more general cases, the sigmoid function will be used.

$$(3)\quad \sigma(x)=\frac{1}{1+e^{-x}}$$

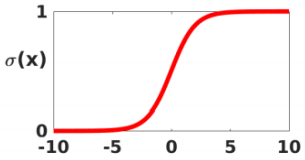
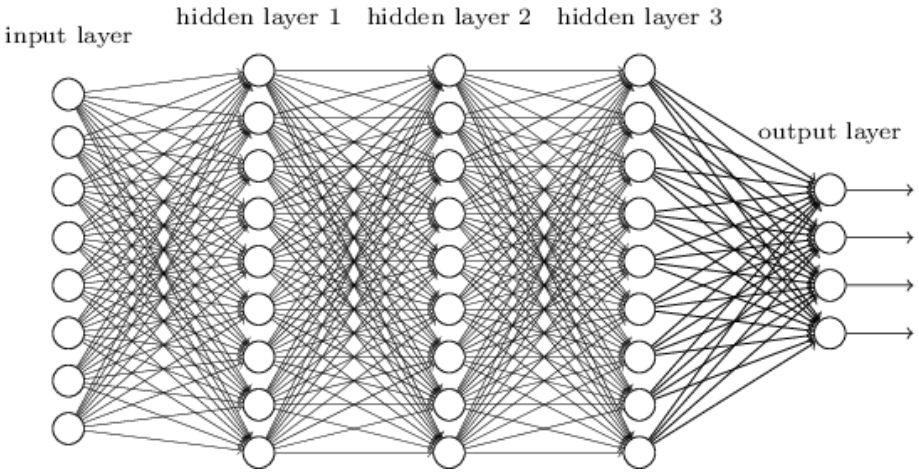


FIGURE2 : THE SIGMOID FUNCTION

The network applies layers of neurons. In each layer, every neuron outputs a signal real number, which is passed to every neuron in the next layer. At the next layer, each neuron forms its own weighted combination of these values, adds its own bias, and applied the sigmoid function. If the real numbers produced by the neurons on one layer are collected into a vector a, then the vector of the outputs from the next layer has the form

$$(4)\quad \sigma(Wa+b)$$

where W is the weights matrix and b is the bias vector. The number of columns in W will be the number of neurons at the previous layer that made the vector a. The number of rows in W and the size of vector b will be the number of neurons at the current layer.



We can now introduce the general form of an artificial neural network. Assume our network has total L layers, where layer 1 is the input layer, L is the output layer, and layer l contains n_l neurons. Notice that our network will map from $\mathbb{R}^{n_l}$ space to $\mathbb{R}^{n_L}$ space. We use $W^{[l]} \in \mathbb{R}^{n_l \times n_{l-1}}$ to note the weights at layer l and $b^{[l]} \in \mathbb{R}^{n_l}$ for the vector of biases in layer l.

We will assign an input data point x to first layer as:

$$(5)\quad a^{[1]} = x \in \mathbb{R}^{n_1}$$

The recursive flow of the network will be as follow

$$(6)\quad a^{[l]} = \sigma\left(W^{[l]}a^{[l-1]} + b^{[l]}\right) \in \mathbb{R}^{n_l}, \quad for\ l = 2,3,\ldots,L$$

Where $y = a^{[L]}$ is the output of the model.

2.4. Adversarial Learning

Deep Learning technologies brought great changes in our daily lives, no question about it. A new field that has grown in parallel in recent years has shown that alongside the success of deep learning in solving complex problems, there is a very surprising weakness for deliberate deceptions. These attacks include small changes in the data, which will usually will not be noticed by a human, which can cause the model to make errors with high confidence.

Szegedy et al. [7] were the first to show this weakness in their research from 2014. In this article they demonstrate how small perturbations to image can cause the model to change its prediction.

Since then, several methods were published for adversarial attacks on deep learning in Computer Vision. A notable work is that of Moosavi-Dezfooli et al. [8] who showed the existence of 'universal perturbations' that can cause a classifier to mistake on any image.

Szegedy et al. [7] defined a small perturbation to the images, such that the perturbed images could lead to misclassification. Let $I_c \in \mathbb{R}^m$ be a vectorized clean image. We wish to compute an additive perturbation $\rho \in \mathbb{R}^m$ that would change the image very slightly but change the model's prediction. They defined the following problem:

$$(7)\quad \min_{\rho} \|\rho\|_2 \quad s.t. C\left(I_c + \rho\right) = l;\ I_c + \rho \in [0,1]^m$$

where *'l'* is the label of the image and should be different from its original label, and C is the model classifier function.

The results of their research, and those that followed, surprised the deep neural network and the computer vision communities and continue to challenge them.

2.5. Related Works

This section will give a list of studies from recent years that explore relations between increasing dimension and the effect of adversarial learning. Those works will be researches and reviewed during the research process.

All researches are related to the effect of dimension on adversarial attack, not just in the computer vision domain. Examining other fields will allow us to try adding abilities that were already tested and has some success in our domain.

Study	Date	Link
High Dimensional Spaces, Deep Learning and Adversarial Learning	Jan, 2018	https://arxiv.org/pdf/1801.00634.pdf
The Vulnerability of Learning to Adversarial. Perturbation Increases with Intrinsic Dimensionality	Jun, 2018	http://www.nii.ac.jp/TechReports/public_html/16-005E.pdf
Adversarial Vulnerability of Neural Networks Increases With Input Dimension	Feb, 2018	https://arxiv.org/pdf/1802.01421.pdf
Dimensionality reduction as a defense against evasion attacks on machine learning classifiers	2016	http://www.princeton.edu/~abhagoji/DCAPS_fall2016.pdf
Enhancing Robustness of Machine Learning Systems via Data Transformations	Nov, 2017	https://arxiv.org/pdf/1704.02654.pdf

3. Solution Description

3.1. Aims and Objective

In this study we would like to investigate the effect of the input dimension on the adversarial attack surface for deep learning-based system in an image classification task.

We will examine different methods for reducing dimension, both in image processing and other fields, while checking the effect on:

1. The classification accuracy for clean images.
2. The robustness to adversarial attacks, and perhaps also other non-malicious changes.
3. The damage to the input quality.

We would like to come up with recommendations for methods to protect against adversarial attacks while minimizing the loss to the image quality.

3.2. Stages

To better understand the domain, this research project will first begin with learning and reviewing existing common methods for adversarial attacks in the field of computer vision. This will be done by an in-depth literature review of articles and surveys since 2014.

In the second stage we must achieve those abilities, that will be used in the experimental phase later. First, we will determine the common datasets we wish to work on, according to what is common on this domain (for example MNIST). We will then wish to obtain those offensive abilities, which means that for the common methods that will be chosen we will look for open-source frameworks or even decide to implement them if necessary.

After reviewing and implanting the offensive side, we will now want to move to the defensive part. At this stage we will review existing defense methods for adversarial attacks.

The forth part will try to look for existing literature on the relationship between dimension and adversarial attacks.

The next step will prepare all that is needed to conduct the experiment to test the dimensional effect on the adversarial attack surface. We will determine at this stage the methods for dimensionality reduction (such as changing resolution or the number of colors, compression, PCA etc.), achieve (or implement) those abilities, and determine the quality metric.

In the six stages we will conduct the experiments and process the results to get conclusions. We will try to explain the results in a mathematical manner.

A bonus stage if will be possible - try to figure mathematical barriers and rules of thumb for each method.

3.3. Benefits and Contributions

There is nothing to add to the tremendous benefits of deep learning-based system. These systems have become and will be in the future even more integral part of our daily lives. Alongside the improvement of existing machines and the race for solving more problems, it sometimes seems that the security has been neglected. The autonomous vehicle will be a fact in a few years, so just imagine the impact of an attack causing it to misclassify a stop sign with a highway sign.

In this research we aim to bring our small contribute to a fast-evolving field of security awareness and methods in the machine learning based daily systems. We hope this study can be a good base for upcoming researches in the field.

3.4. Innovations

Because the subject is brand new and is at the heart of current research, all studies are from recent years and a lot of work is yet to be done. As can be seen from the related works section, all articles are from the last two years, some of them are from the last two months. So, this field very new, but has a lot of potential.

3.5. Success Criteria

To define this project as successful at the end of the process, we would like to come up with some clear conclusions regarding the impact of various dimension reduction methods on the defense against adversarial attacks. We would like also to make recommendations for defense methods that will increase security while minimizing the damage on the input and on its classification.

3.6. Timeline

Stage	Details	Due Date
Proposal submission		30.3.18
Reviewing Adversarial Attack Methods	• Review and study deeply for existing adversarial attack methods.	1.4.18-15.4.18 (2 weeks)
Implementing Attack methods	• Determine common datasets/ • Obtain offensive ability (open source / implementation)/	15.4.18 – 1.5.18 (2 weeks)
Reviewing Adversarial Defense Methods	• Review and study of existing defense methods for adversarial attack (not necessary in the field of computer vision).	1.5.18 – 15.5.18 (2 weeks)
Reviewing of Dimension Effect on Adversarial Learning	• Review for existing studies on the relation between dimension and adversarial learning.	15.5.18 – 1.6.18 (2 weeks)
Experiment Preparation	• Determine and implement methods for dimensionality reduction. • Determine quality measures	1.6.18 – 15.6.18 (2 weeks)
Experimenting	• Conduct the experiment	15.6.18-8.7.18 (3 weeks)
Result Evaluation	• Collect the results and get to conclusions	8.7.18 – 1.8.18 (3 weeks)
Final Project Writing	• Paper writing	1.8.18 - 1.9.18 (1 months)
-- Buffer --	--	1.9.18-10.10.18 (1 month)
Submission		10.10.18

4. Bibliographic

[1] Machine Learning / Stanford University, Coursera, Andrew Ng

[2] Mitchell, T.M. (1997). Machine Learning. McGraw-Hill.

[3] C.F. Higham and D.J. Higham (2018). Deep Learning: An Introduction for applied Mathematicians

[4] T.Chen (2014). Introduction to Boosted Trres.

[5] S.Shalev-Shwartz and S.Ben-David (2014). Understanding Machine Learning: From Theory to Algorithm, 33-35.

[6] N. Akhtar and A.Mian (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey.

[7] C. Szegeday, W.Zaremba, I.Sutskever, J.Bruna, D.Erhan, I.Goodfellow, R.Fergus (2014). Intriguing Properties of Neural Networks.

[8] S.M. Moosavi-Dezfooli, A. Fawzi, O.Fawzi and P.Frossard (2017). Universal Adversarial Perturbations.

[9] M.Verelysen and D.Francois (2005). The Curse of Dimensionality in Data Mining and Time Series Pre