Assuring Autonomy

Dr. Ramesh Bharadwaj
Computer Engineer (Code 5546)

ramesh.bharadwaj@nrl.navy.mil

Dr. Ira S Moskowitz
Mathematician (Code 5580)

ira.moskowitz@nrl.navy.mil

Information Technology Division
Naval Research laboratory
4555 Overlook Avenue SW
Washington DC 20375 USA

ABSTRACT

Autonomous systems, including self-driving cars and air vehicles, have caught the imagination of the press and the public. However, broader adoption of such systems in safety-critical applications has been the subject of intense debate and scrutiny. The stunning performance of deep learners compared to extant methods, including pattern matching, statistical methods, and legacy machine learning algorithms, has taken the research world by storm. This has naturally lead the DoD community to ask the question: "How do we harness this technology being unleashed upon the world?" Before we answer this question, however, it is important to note that trust is integral to DoD applications, including autonomous systems, and ensuring reliable system operations is paramount. Therefore, we need strategies that harness deep learning algorithms to provide the DoD and NATO with autonomous systems that are robust, secure, timely and dependable.

1.0 INTRODUCTION

Deep learning has received wide press in the recent past [1]. Although requiring massive computational power to train, these algorithms show great promise by being able to recognize objects with human level precision [2, 3] and translating human speech in real-time [4]. However, in addition to coming to grips with recent advances in deep learning, the DoD and NATO communities must also understand its limitations. For instance, data sparsity and data poisoning attacks may lead to classification and training errors, producing incorrect results which can be both embarrassing and damaging. Two recent examples are Google's image classifier mis-identifying humans as gorillas [5] and Microsoft's chatbot Tay learning to spew racist and misogynistic hate speech minutes after being turned on [6]. More disturbing, the invention of generative adversarial networks (GAN) shows that deep learning algorithms can be deliberately tricked by adversarial examples [7]. A trained neural network can be tricked into grossly misclassifying objects with extremely high confidence, by mere manipulation of input images not discernible to the human eye or even by images that look like noise to the human viewer [8]. The dangers of adversarial attacks can have a profound impact on society -- self-driving vehicles can be hijacked or misdirected with seemingly innocuous signage [9], and system security can be compromised with tampered data.

We must have assurance that Machine Learners cannot be fooled. To obtain 100% assurance may be an impossible task; however, we must raise the bar from the low level that it is at presently. Known as adversarial attacks, a representative paper discussing this issue is Kurakin 2017 [10]. This is also an issue when it comes to the assurance of biometric devices. To make matters even worse, a very recent article in The Register (2018)

[11] has shown how placing random sub-images in an image that we are attempting to classify can throw the Machine Leaner totally off. In addition, our research will focus on ways to improve image classification performed via Artificial Intelligence (AI). We also note that Machine Learners can even make gross errors even without an adversarial attack. Amazon's Rekogntion tool [12], just misclassified 28 members of Congress as criminals using facial recognition (7/2018) without any trickery. Thus, we must make sure that AI Machine Learners do what they are supposed to do, and do not do what they are not supposed to do.

2.0 APPROACH

An approach that shows promise, and something that we plan to work on, is using subtle differential geometric and topological techniques as an add-on to the current machine learning techniques. We hope to develop signatures that must be used in addition to any deep learning neural nets. In addition to image signatures we are developing a theory of image complexity based upon Information Geometry. Our thinking is that the best way to subtly fool a machine learner is in a region of high complexity.

2.1 Misconceptions about ML

A popular misconception about machine learning (ML), which provides an illusion of success, is to test a neural network on its training data. However, what many people do not realize is that the fundamental goal of ML is to *generalize* beyond the examples in the training set [13]. This is because no matter how much training data we provide, it is unlikely that the same data will be encountered during testing. This corresponds to the "no free lunch" theorem of Wolpert and Macready [14], which states that no learner can beat random guessing over all possible functions to be learned. Consider for example learning a Boolean function over 100 variables from a million examples. Of the 2^{100} possible classes to be learned, we have only provided 10^6 examples. There are additionally $2^{100} - 10^6 \approx 1.3 \times 10^{30}$ possible inputs whose classes are yet unknown. Clearly, there is no way to do this that beats flipping a coin. Or is there?

2.1.2 The Manifold Hypothesis

Experts in deep learning algorithms such as John Launchbury (formerly at DARPA and current Chief Scientist at Galois) contend that their phenomenal success is due to what is termed the *manifold hypothesis* [15]. High-dimensional natural data tend to clump and be shaped differently when visualized in lower dimensions. Therefore, our working assumption is that each manifold in a deep neural network represents a unique functional entity, and an understanding how the network classifies its input data can be gained be mapping these manifolds. Unlike extant approaches such as Reluplex [16], which try to reason about the entire network, this insight gives us the ability to map decision boundaries of a feed-forward neural network a layer at a time, thereby mitigating the state explosion problem encountered by extant approaches such as Reluplex.

2.2 Statistical Manifolds

In addition to using standard Gaussian surface geometry and algebraic topological techniques, we will use a natural and intuitive metric on probability distributions, the Fisher-Rao metric [17] from Information Geometry, to form a non-standard Riemannian manifold called the Statistical Manifold. Statistical Manifolds have recently been used in many applications, running the gamut from belief propagation, manifold learning, and neural nets, to Grover's search algorithm in quantum computation. Using a Statistical Manifold is the preferred way to view parameterized distributions as a metric space, so that the concepts of close and far are well defined, intuitive, and tractable. The Fisher-Rao metric, which is the Riemannian metric of a Statistical Manifold, has been shown to be

the natural, parameter-invariant method of computing probabilistic distribution distances. Additionally, the Fisher-Rao metric has recently been applied, albeit in a primitive manner, to image analysis, and its use in stenographical analysis has not been explored at all. We will remedy this situation. The Fisher-Rao metric incorporates, as special/approximate cases, the previous methods of distinguishing distributions, such as mutual information, the Kullback-Liebler divergence, the Jensen-Shannon divergence, the Hellinger distance, the Renyi-Chernoff distance, the Bhattacharyya distance, etc. To discover the relationship between images and shape analysis, and the Fisher-Rao metric induced Statistical Manifold, we will derive new methods to measure the geometric structure of images.

2.2.1 Connecting Differential Geometry and Algebraic Topology

Connections between differential geometry and algebraic topology are numerous and will be utilized in our research. For example, there is the famous Gauss-Bonnet theorem, which relates the surface integral of the sectional curvature to the Euler characteristic (an algebraic concept). The algebraic theory of homology and cohomology can be developed on a differentiable manifold via differential forms. Critical points of the "height" function can determine the homological structure of certain spaces via Morse theory. Ricci curvature lets one obtain bounds on the Betti numbers, which are the rank of the homology groups. There are generalizations of Gauss-Bonnet involving the differential characteristic classes and differential operators that allow the theorem to be extended into the Atiyah-Singer index theorem (which generalizes the relationship between analytic, topological and geometric indices). These and similar techniques must be investigated to fully understand what characterizes an image, and how it can be manipulated.

2.3 Available Variance

A major question we plan to explore is how much variance is available to fool a machine learner, while keeping the curvature and topological signatures essentially unchanged.

We will explore this wiggle room as it applies to machine learning and steganography (hiding information so that its very existence is unknown). There has been minimal research applying the inherent Gaussian Riemannian structure to steganography, and none has been performed using Statistical Manifolds, which have been shown to be valuable in other areas of image analysis. When one is researching steganography, one must utilize the features of the surface and modeling geometries---not just frequency values, or JPEG constraints, for example. We will start with the classical approach of using the standard Gaussian surface normal and principal curvatures to describe a surface, and then add the second geometric structure of the Statistical Manifold structure induced by the Fisher-Rao metric. We will use existing image databases and particular images of Naval interest. Various curvature features will be created to reconstruct the image or certain landmark features. We will investigate the robustness of these geometric features in terms of image integrity and information hiding (steganographic bandwidth). Once basic geometric structures have been identified, we will test their robustness to withstanding manipulation. Concepts that combine differential geometry and algebraic topology will be used. We have discussed some of the relationships between topology and geometry above. However, there has been very little work combining them with respect to analyzing both steganography and/or machine learning.

2.4 Shape Modeling

Shape modeling is an ongoing problem in image analysis and visualization. A recent approach models basic shapes by way of probability distributions for facial recognition. Image morphing and identification is obtained by travelling along geodesics given by the Fisher-Rao metric. The Statistical Manifolds formed from the Fisher-Rao metric can be manipulated via diffeomorphism groups to obtain certain standard forms. This research is in

its infancy, and the inability of obtaining closed form geodesics (locally length minimizing curves) is a barrier that we plan to overcome. (We note that diffeomorphism groups in shape analysis have recently been shown to be of interest in the field of quantum gravitation.) Our hypothesis is that regions of high volume (similar to regions of high frequency in discrete Fourier image transforms), with respect to the Fisher-Rao metric, are ripe for steganography and of course affect the distinguishability of information parameters. The Fisher-Rao metric allows one to obtain the volume element for the Statistical Manifold. This result is difficult to obtain in closed form, so we will develop numerical methods, when needed, to determine stego-rich and stego-poor regions of the image in terms of the Fisher-Rao volume element, and in terms of standard Gaussian and topological techniques. Additionally, a detailed analysis of the geodesics for the Fisher-Rao Riemannian metric is a necessity for calculating distance. In general, closed form solutions of these geodesics are difficult to obtain. We will analyze and invent new approximate methods, closed form methods, and associated relationships to well-known partial differential equations from mathematical physics as alternatives. A thorny issue is that the Von-Mises distribution used in shape analysis, when viewed via the Fisher-Rao metric, forms a Statistical Manifold that is not globally diffeomorphic (equivalent) to Euclidean space. The manifold has a non-trivial topological structure that makes geodesic calculations difficult, but this must be pursued.

3.0 EXPERIMENTATION

On the practical side, we are constructing a toolset for robust machine learning centered around data that can be modeled by two-dimensional tensors (2-D tensors), which includes images, video, and handwriting handled by current commercial tools. Datasets within the DoD include numerous other application categories that are modeled by 2-D tensors. These includes spectrograms – visual representations of the spectrum of frequencies of electromagnetic waves, sound, or other signals as they vary with time – and pulse-Doppler videograms of received radar returns. In the cyber domain, as well as in statistics, econometrics, epidemiology, genetics, and related disciplines, causal graphs [18] – also known as path diagrams or causal Bayesian networks – encode assumptions about the data generation process, which can also be modeled as 2-D tensors.

3.1 Testing the Manifold Hypothesis

In order to test the Manifold Hypothesis, we trained a neural network on a simple classification problem modeled by two-dimensional Gaussian processes A and B. The decision problem is to assign each instance of test data to one of the two hypotheses H_A or H_B. The two Gaussian processes A and B have significant overlap; therefore, it is impossible for a classifier to achieve an accuracy of 100%. However, we can set a benchmark for classification accuracy by calculating the optimal decision boundary and the average probability of misclassification of an ideal classifier, i.e., a Gaussian Classifier using Bayesian Decision Theory, using which we determined that the ideal classification boundary is a circle whose radius is approximately equal to twice the variance of Gaussian Process A, and which is centered slightly to the left of the origin. We trained a three-layer feed forward network with two input neurons, four hidden neurons, and two output neurons, using the back-propagation algorithm [19], optimizing the number of hidden neurons, the learning, and momentum constants through a process of trial and error. We determined the average probability of misclassification to be 0.183, which is very close to the theoretical optimum of 0.182. We plotted the decision boundary of the feedforward classifier using test samples uniformly distributed over the two input dimensions and overlaid it with the theoretical optimum boundary of the Bayesian classifier. The results are shown in Figure 1-1.

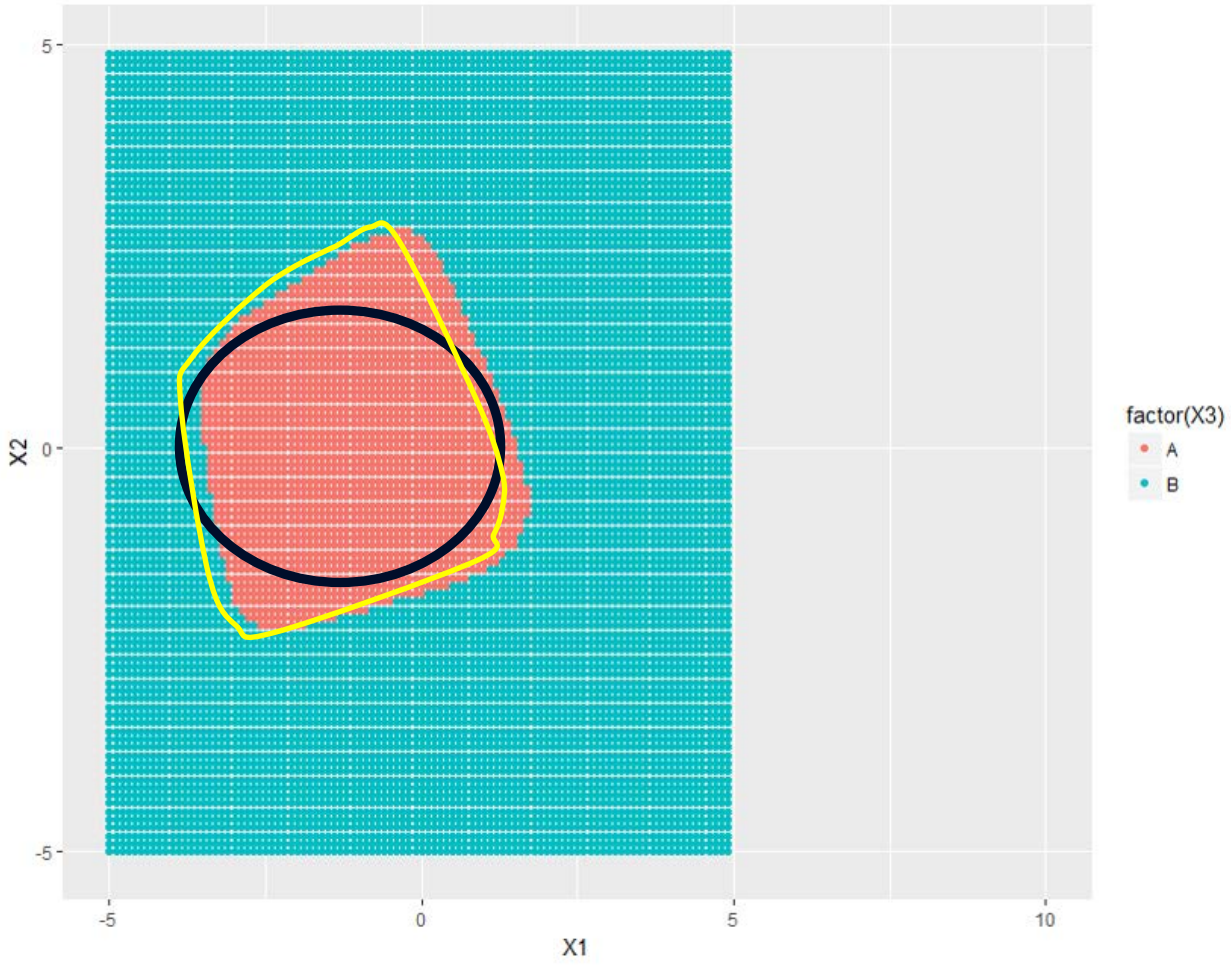

Figure 1-1: Decision Boundary of a Feedforward Classifier Overlaid on Optimum Bayesian Boundary.Legend: [ ] Optimum Bayesian Boundary [ ] Learned Boundary

Notice however that this approach soon becomes intractable for inputs of high dimensions -- the biggest problem in machine learning is the "curse of dimensionality" [20] which posits that generalizing correctly becomes exponentially harder with increasing input dimension. We were able to visually inspect and confirm our Smoothness Assumption, i.e., that each point in the input layer is enclosed by a classification manifold. We also identified the minimum norm required to perturb a test example into a misclassification region, i.e., into a region of no overlap. Robustness of our training algorithms can be improved by regularization, such as dropouts, which has been shown to approach the performance of the ideal Bayesian classifier. We plan to defend against adversarial attacks using a novel approach we call Stochastic Pruning.

4.0 FUTURE WORK

For autonomous systems that employ deep learning, additional measures have to be instituted to ensure trust. The following summarizes our approach:

- Develop a universal attack framework to provide a comprehensive set of adversarial manipulations using measures of variance pioneered by Carlini and Wagner [21].
- Include these adversarial examples during training to harden the neural network against data poisoning and data sparsity attacks.
- Develop metrics against which to measure robustness of trained networks. We shall evaluate and build upon extant approaches such as adaptive regularization to ensure robustness against all adversarial attacks, and semidefinite relaxation methods which provide certificates guaranteeing that, for a given network and test inputs, no attack can force classification errors to exceed a given value.
- Extant approaches, such as Reluplex of Stanford researchers [16], reason about a trained feed-forward network in its entirety and therefore do not scale. Reluplex handles around 300 neurons, whereas the CNN used in Nvidia's autonomous car [22] has around a quarter million. We propose instead a mathematical framework to propagate verification conditions layer-by-layer, ensuring scalability.
- Create techniques analogous to model checking [23] and Satisfiability Modulo Theories (SMT) [24] solving for mapping the manifolds of each layer of a feed-forward network by exhaustive search near regions of misclassification due to adversarial attacks. Investigate decision procedures for real closed fields and propose methods for cylindrical algebraic decomposition. Plain Simplex may be sufficient for polyhedral regions.
- Explore reduction methods such as counterexample guided abstraction refinement (CEGAR) [25] to ensure finiteness of the search.
- Construct tools to provide certificates of guarantee (at a certain confidence level) that are implied by the absence of adversarial examples.

5.0 CONCLUSIONS

Preventing machine learners from being fooled is of extreme importance as the DoD relies more and more on Artificial Intelligence. The accuracy and robustness of biometric devices is also of importance as the DoD attempts to secure its assets and provide internal security. Furthermore, the ongoing issue of image steganography is of continued importance to the DoD as both an offensive technique and as a tool for detection. Invention of new theories on manifold learning and implementation of developer-friendly tools, methods, and guidelines, will assure safety and trust in autonomous unmanned systems running custom DoD software. By following a process for certifying autonomous systems prescribed by the outcome of this research, DoD organizations can gain authority to operation (ATO) expeditiously, at low cost, and with limited expertise in the mathematics underlying our formal methods and tools.

REFERENCES

[1] MarketWatch. "Deep Learning Market Projected to Touch a Valuation of $261,113.0 Mn by 2027-end," https://www.marketwatch.com/press-release/deep-learning-market-to-projected-to-touch-a-valuation-of-us-2611130-mn-by-2027-end-2018-09-07 (ALL URLs validated as of 9-September-2018)

[2] Roman Steinberg, uKit AI (VentureBeat). "6 areas where artificial neural networks outperform humans," https://venturebeat.com/2017/12/08/6-areas-where-artificial-neural-networks-outperform-humans/

[3] Andrej Karpathy, (blog). "What I learned from competing against a ConvNet on ImageNet," http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

[4] Jeremy Hsu. "Starkey's AI Transforms Hearing Aids Into Smart Wearables," IEEE Spectrum, Aug 2018.

[5] Tom Simonite. "When It Comes to Gorillas, Google Photos Remains Blind," WIRED Magazine, Jan 2018.

[6] The Guardian. "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter," 24 Mar 2016. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

[7] Nguyen A, Yosinski J, Clune J. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.

[8] Ian J Goodfellow, Jonathan Shiens, and Christian Szagedy. "Explaining and Harnessing Adversarial Examples," in Proc. International Conference on Learning Representations 2015 (ICLR 2015).

[9] Simson Garfinkel. "Hackers Are the Real Obstacle for Self-Driving Vehicles," MIT Technology Review, August 2017.

[10] Alexey Kurakin, Ian J Goodfellow, Samy Bengio. "Adversarial Examples in the Physical World," Proc Workshop Track International Conference on Learning Representations 2017 (ICLR 2016).

[11] Katyanna Quach. "AI image recognition systems can be tricked by copying and pasting random objects," The Register, 28 August 2018.

[12] Congressional Black Caucus. Letter to Jeffrey Bezos, Chairman, President and CEO of Amazon.com, Inc., At: https://cbc.house.gov/uploadedfiles/final_cbc_amazon_facial_recognition_letter.pdf (24 May 2018.)

[13] Pedro Domingos. "A Few Useful Things to Know about Machine Learning," Communications of the ACM, 55(10) October 2012.

[14] David H Wolpert and William G Macready. "No Free Lunch Theorems for Optimization," in IEEE Transactions on Evolutionary Computation, 1(1), April 1997.

[15] Lawrence Cayton. "Algorithms for manifold learning," Technical Report, UCSD, 12:1-17, 2005.

[16] Guy Katz, Clark Barrett, David L Dill, Kyle Julian and Mykel J Kochenderfer. " Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," Lecture Notes in Computer Science, Volume 10426.

[17] Shun-ichi Amari. “Differential-Geometrical Methods in Statistics.” Lecture Notes in Statistics, Number 28, Springer-Verlag, 1985.

[18] Judea Pearl, “Causal inference in statistics: An overview,” Statistics Surveys, Vol. 3, pp 96—146, 2009.

[19] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. “Learning representations by back-propagation errors,” Nature 323, 533-536 (09 October 1986).

[20] R E Bellman, “Adaptive Control Processes,” Princeton University Press, Princeton NJ, 1961.

[21] Nicholas Carlini and David Wagner. “Towards Evaluating the Robustness of Neural Networks,” IEEE Symposium on Security and Privacy, 2017.

[22] Nvidia Inc., (Scalable AI Platform for Autonomous Driving). “World’s First Functionally Safe AI Self-Driving Platform,” https://www.nvidia.com/en-us/self-driving-cars/drive-platform/

[23] Edmund M Clarke, Orna Grumberg and Doron Peled. Model Checking, MIT Press, January 2001.

[24] Armin Biere et al. Handbook of Satisfiability, ISO Press, 2008.

[25] Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, Helmut Veith. “Counterexample-Guided Abstraction Refinement,” in Proc. International Conference on Computer Aided Verification (CAV 2000). Lecture Notes in Computer Science Volume 1855.