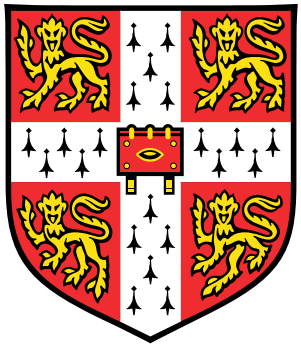
Model Uncertainty for Adversarial Examples using Dropouts



Ambrish Rawat

Department of Engineering

University of Cambridge

This dissertation is submitted for the degree of

Master of Philosophy

Fitzwilliam College

August 2016

Declaration

I, Ambrish Rawat of Fitzwilliam College, being a candidate for the M.Phil in Machine Learning, Speech and Language Technology, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose. (Total word count - 4200)

Ambrish Rawat
August 2016

Acknowledgements

First, I am thankful to my supervisor Zoubin Ghahramani, for all his unique and useful insights. I am specially thankful to Yarin Gal, for all the discussions which greatly helped me comprehend the intricacies of the problem at hand.

I am also thankful to Richard Turner, for bringing in a clarity of thought and helping understand things in a broader perspective.

Most importantly, I am thankful to my Mom and Dad, for their constant inspiration and unconditional support.

Abstract

An image can undergo a visually imperceptible change and yet get confidently misclassified by a trained Neural Network. Puzzled by this counter-intuitive behaviour, a lot of research has been undertaken in search of answers for this inexplicable phenomenon and more importantly, a possibility to impart robustness against adversarial misclassification. This thesis is a first step in the direction of investigating the effect of this adversarial misclassification on Bayesian Neural Networks. With dropouts as a tool for obtaining estimates of model uncertainty, this thesis presents a study of model uncertainty for adversarial images.

Contents

List of Figures

1 | Introduction

Engineering a system for some intuitive tasks involving speech, language, and vision has a fundamental bottleneck - we ourselves do not understand them completely. For instance, our incomplete understanding of speech-recognition or visual-cognition, limits us in writing a look-up table or a deterministic computer program for the task. Deep Learning has proven to be remarkably useful in these settings Bengio and Courville (2016).

However, with their massively parametrised structure, it is difficult to argue about the assumptions these 'deep' models make or inductive biases they exploit (Griffiths et al. (2010), Kohavi and John (1997)). This has become particularly interesting in the light of recent developments, where Deep Learning models have exhibited counter-intuitive behaviour for image recognition Szegedy et al. (2013). Neural networks are vulnerable to adversarial images i.e. images generated by adding imperceptible perturbation(s), which to a human look the same, but which a trained network misclassifies with high confidence. This challenges the inductive biases of a neural network, raising a concern that may be the implicit assumptions in these models are vastly different from the ones humans make for the same task. From a theoretical standpoint, probabilistic formalism allows for a principled approach to reason about uncertainty Cox (1946). So perhaps we should explore this in the ongoing analysis of adversarial misclassification.

Figure 1.1: Adversarial Image. Left: 99% confidence of class being ship. Right: 70% confidence of it being a truck)

The human aspect of adversarial images makes them particularly interesting. We perceive images in an abstract space which is different from the high-dimensional subspaces

they are dealt with by the neural network. However, the inconsistent behaviour of a neural network's confident disassociation for a pair of perceptually similar images, is at the very least, highly disconcerting.

An adversarial example us hypothesised to exist far away in the data space. It is therefore expected that uncertainity estimates for these images can be obtained. This thesis explores the possibility of investigating model uncertainty for adversarial images by observing model output with an added layer of marginalisation - via dropouts at test time.

The thesis has been structured into three chapters which I believe are relevant for motivating the investigation. The first two chapters provide a modest introduction to Bayesian Reasoning, Model Uncertainty and Deep Learning. This is followed by the main contribution of the thesis, where model uncertainty for adversarial images has been studied.

2 | Bayesian Reasoning

In the literature of Machine Learning, the words 'Bayesian' and 'Probabilistic' are often used synonymously as mere placeholders for a common set of underlying principles. With Probability theory as the underlying script, the language of Bayesian reasoning allows for a simultaneous consideration of all forms of uncertainty. In fact, arguably this is one of the most distinguishing aspects of probabilistic approach to modelling. Ghahramani (2015)

Uncertainty has an all-pervasive presence in a Machine Learning model - from noise in the observations, to a model's belief about its parameters, to uncertainty in model predictions, to our belief about the model itself. However, this is more of a boon than a curse, as its explicit and faithful representation in Bayesian probability theory enables us to reason with it in a principled way.

2.1 Theory and Practice

The uncertainty (or equivalently degree of belief) in the variability of x, is quantified mathematically through a function $p(x)$. This function, by virtue of Cox's axioms must uphold the rules of probability theory Cox (1946),Jaynes (2003). [1]

Modelling approach From a Bayesian standpoint, most Machine Learning tasks can be conveniently summarised as follows - a model, accompanied with its explicit and necessary assumptions based on prior knowledge, is capable of making predictions and performing inference in the light of observed data.

This oversimplified picture is in fact quite powerful and accommodates for a wide breadth

[1] It is quite gratifying that all of these probabilistic manipulations are governed by two simple and powerful rules of probability theory, namely, the sum-rule and the product-rule. Ghahramani (2015) Valpola (2000) Ghahramani (2013)

of models. Interpreting the rules of probability for a Machine Learning model highlights some of the salient features of uncertainty representation and manipulation.

Uncertainty representation: All the the believed uncertainties in the mathematical description of a model are stated via functional definitions - prior $p(\theta|m)$, likelihood $p(D|m)$, noise processes, among other expressions. Through the lens of uncertainty, all these mathematical relationships are identical.

Uncertainty propagation: Intuitively, uncertainty propagation is required at the stage of decision-making (computing utilities), which in general translates to inference or prediction. Mathematically, this involves the use of marginalisation principle (a direct application of sum and product rule).

Inference and Predictive Probability: The task of inference or learning boils down to squeezing prior uncertainties ($p(\theta)$) (often with respect to parameters, latent variables etc.) through the data (D) to posterior uncertainties ($p(\theta|D,m)$) (following Bayes' rule).

$$p(\theta|D,m) = \frac{p(D|\theta,m)p(\theta|m)}{p(D|m)} \qquad p(D|m) = \int p(D|\theta,m)p(\theta|m)\mathrm{d}\theta \tag{2.1}$$

The normalisation constant, $p(D|m)$ is referred to as Marginal Likelihood or Model Evidence. At the level of conditioning on m, Bayesian model selection is naturally implemented as Occam's Razor Jefferys and Berger (1992). These updated uncertainties in the posterior can be sequentially propagated while making predictions about the new data (D^*)

$$p(D^*|D,m) = \int p(D^*|\theta,m)p(\theta|D,m)\mathrm{d}\theta \tag{2.2}$$

The elegance of first principles makes Bayesian reasoning philosophically appealing and intuitively compelling. However, the theory sheds no light on computational or analytical intractabilities. The exact computation of the marginalisation integral is, in most cases, infeasible. Therefore, in practice a range of approximations are used. The inexpensive point estimates (like maximum likelihood and maximum a posteriori), deterministic approximations harnessing structural understanding of the problem (like Variational mean field, Expectation Propagation) and the computationally demanding but often asymptotically exact stochastic methods.

2.2 Model Uncertainity

At first glance it appears that the above mentioned approach is constrained to reason from the point of view of "one" model (or family of models). However, the folklore of Bayesian theory advises you to marginalise all forms of uncertainties, i.e., not just the parametric assumptions but structural assumptions

This formal categorisation of the two uncertainties is often unclear and can only be argued about in context. The supervised learning tasks of regression and classification provide one place for this examination. Given a dataset $D = \{(x_i, y_y)\}_1^N$, with input variables $x \in \mathcal{X}$ and target $y \in \mathcal{Y}$, the goal is to predict for a new point $x^* \in \mathcal{X}$, a point $y^* \in \mathcal{Y}$. Traditionally, for a regression task, $\mathcal{Y}$ is a continuos space while for a classification task, y are nominal labels. A probabilistic classifier models the task as two subproblems - first, a map from the input variables to class conditional probabilities, and second, a decision rule based on the obtained probabilities. While the distinction between uncertainties is clear in a regression framework with continuos, target variables, presence of a discrete space makes the problem intuitively non-obvious.

In the case of regression, as a function family can capture both parameter and structural assumptions. For instance the a GP prior captures the two assumptions in its mean and kernel function. Under these assumptions (and having seen a set of observations D), it models the predictive distribution $p(y^*|x^*, D)$. The present uncertainties being evident in the spread of the predictive distribution.

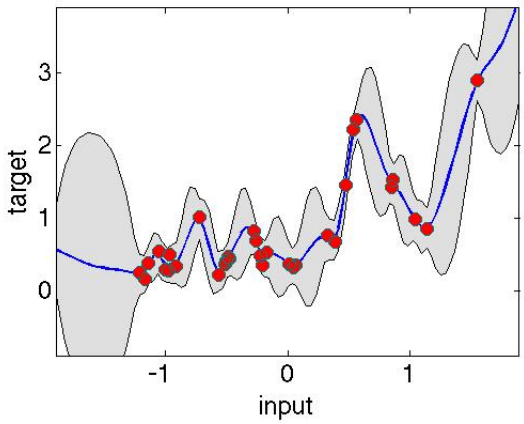


Figure 2.1: A GP based model for regression. Observed points (red), Predicted mean (blue line) and predicted uncertainty (variance of the predictive distribution), at each point in the x-space.

A probabilistic classifier outputs a predictive probability for class labels, which, for a 3-class problem, is a point on the simplex 2.2. Defining a model, as a system which yields one point on this simplex allows us to define model-uncertainty "across models". While defining a model, as a system yielding different points on this simplex for its different parameter settings, allows us to define a "within-model" uncertainty. **Model-uncertainty** in both cases being a representation of structural assumptions. [2].

For the case of classifier, the distinction is fuzzy and boils down to the definition of a model. Nonetheless, a Bayesian outlook coupled with the latter definition of probability would encourage you to marginalise a parameter, if an alteration of which, yields a different point on the probability simplex.

A class-conditional distribution, $p(y = l|x)$ tells us about the uncertainty of 'an' association between x and the class label, l. An alternative interpretation of this underlies in the uncertainty of 'all' associations. This demarcation can also be phrased as *"probability of an association"* (belongingness to a class) or a *"possibility of an associations"* (confidence across different belongingness). At the junction of this desired alignment of subjective-belief and empirical consistency, in the support of reliable predictions, is the idea of calibrated probabilities. While adapting a Bayesian perspective, it is advisable to calibrate a model, if accounting for model-uncertainty.

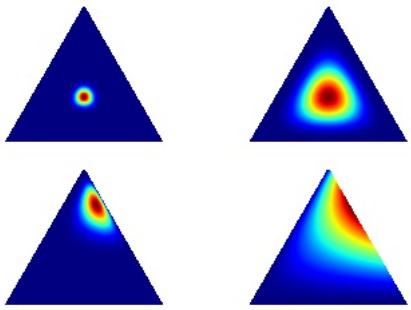


Figure 2.2: A 3-dimensional simplex where class-conditional probabilities from a 3-class classifier reside. The top row is representative of images, where the average confidence is at the centre, but the respective spreads are distributed. Similarly, the images in the bottom row demonstrate spread with confident classification with respect to a class.

[2]It must be emphasised here that from a decision making perspective, the eventual output required is a probability vector. One wouldn't want multiple outputs from a weather-predictor: a model cannot simultaneously predict, both 10% and 20% chance of rain today. Definition of a model as a system yielding one point on the simplex is therefore, more intuitive from a probabilistic-classifier viewpoint.

3 | Deep Learning

3.1 Deep Learning for Classification

Given a dataset $D = \{(x_i, y_y)\}_1^N$ (or $D = \{X, Y\}$), with input variables $x \in \mathcal{X}$ and target variables $y \in \mathcal{Y}$, the goal is to predict for a new point $x^* \in \mathcal{X}$, a point $y^* \in \mathcal{Y}$. The Deep Learning approach, assumes a highly-parameterised non-linear functional relationship $y = f(W, x)$ between the paired values. The goal of prediction is split into two phases - training and testing[1]. During training, a cost function $C(y_1, y_2)$ (for measuring the score or similarity between target vectors) is optimised. The optimisation The adjustable parameters W are then updated to minimise this objective function for the observed data set $min_W \sum_i^N C(f(W, x_i), y_i))$

These learning modes are motivated are thus part of a toolbox approach where a model is optimised (fit) to explain the observations (training). Hence, we often get only point estimates form the output layer. "Optimisation" is outside the vocabulary of Bayesian modelling theory and hence often the assumptions of a Deep Learning model are often unclear. In theory, one can ascribe a Bayesian interpretation to these models, but there are immense computational challenges to arrive at such an interpretation.

For a well trained model, the optimised W, encodes the necessary information from the input image, i.e. required for generalisation during prediction. However a stochasticity of these during test-time is not accounted for in conventional DNN models.

3.2 Model Uncertainty for Classification

As discussed in the previous section, a Bayesian outlook towards classification encourages you to look at uncertainty estimates while taking a decision. Although a Deep Neural

[1] a discrimination absent in a "truly" probabilistic framework

Network can be used as a probabilistic classifier, getting these uncertainty estimates is computationally challenging. Ascribing a Bayesian interpretation to a Neural Network has a long standing history, most of which involve further computation. An inexpensive way to get these uncertainty estimates was proposed in Maeda (2014) and Gal and Ghahramani (2015b). The later also established a grounding for this in Variational Inference.

Model uncertainty is fundamental for decision making. A decision based on one-model is often not encouraged. This philosophy is shared across a variety of DNN-models in different settings. For instance, speech-recognition systems incorporate an engineered model averaging by intricate sequence alignment schemes.

For Deep Neural Network used for classification, the final layer is a soft-max layer. When faced with a large point-estimate for a point away from the data set, softmax yields a high-confidence output. This is the space of extra marginalisation which can be achieved with Dropouts (as demonstrated in Gal and Ghahramani (2015b)). Model uncertainty is critical to deployment of DNN in settings like Reinforcement Learning or self-driving cars as decision of not recognising vs recognising a person on street can be extremely critical.

3.2.1 Dropouts as Variational Approximation

In a supervised learning setting, the predictive distribution for a new point x^*, is given by [2]

$$p(y^*|x^*, X) = \int p(y^*|x^*, W)p(W|X, Y)dW \tag{3.1}$$

As discussed in the previous section, this is often intractable and in most cases because of analytical intractabilities of the posterior, $p(W|X, Y)$. The approximation framework of Variational Inference proposes to use an approximate variational distribution $q(W)$ in place of the exact posterior. It then frames this as an optimisation problem where a similarity measure (Kullback-Liebler (KL) divergence) is minimised between the exact and approximate posteriors. This yields the following approximate predictive

[2] the parameters notation has been changed from theta to W in alignment with the conventional usage the latter in the context of DNN

distribution.

$$q(y^*|x^*, X) = \int p(y^*|x^*, w)q(w)dw \tag{3.2}$$

The minimisation of this KL, is equivalent to maximisation of log evidence lower bound (ELBO), which is the objective function used in practice (L_{VI}).

$$L_{\text{VI}} = \int q(w) \log p(Y|X, w)dw - KL(q(w)||p(w)) \tag{3.3}$$

It was shown that for a specific choice of the prior and a specific choice of the approximate posterior, performing Dropouts during training is equivalent to optimisation of an unbiased estimator of L_{VI}. A strong basis of dropouts in the variational framework has also been suggested in Maeda (2014) and Kingma et al. (2015).

An effective usage of Dropouts for modelling model-uncertainty was observed in Kendall et al. (2015). This, coupled with the theoretical groundings, provides the motivational basis for exploration of model uncertainty for adversarial images.

4 | Adversarial Images

4.1 Framework

The question of robustness is not new to machine learning models. It has been central to their usage in computer-security systems (Asuncion and Newman (2007) Huang et al. (2011)). These scenarios have an intrinsic adversarial setting, like spam filtering, where malicious adversaries are driven to threaten the integrity of security systems. The popular deployment of Machine Learning models in AI-agents has only broadened the scope of this field.

The model family of Neural Networks has had an unparalleled success in computer vision tasks. This is usually attributed to the expressibility of convolutional-neural networks in a grid-like setting which is conveniently suited for images. Their large inductive bias allows them to make useful abstractions from images for tasks like classification. However, recent developments (Szegedy et al. (2013)) have exposed some bizarre aspects of Neural Networks, which has challenged their generalisability over image spaces.

One of these concerns is the unexplained existence of adversarial images, specifically **images generated by adding imperceptible perturbation(s), which to a human look the same, but which a neural network fails to classify correctly with high confidence**. The implicit human involvement in this definition makes the question of adversarial examples interesting, challenging and fuzzy. Hence, for the sake of clarity I will be utilising the following representation framework (as motivated in Dziugaite et al. (2016)) for discussing adversarial images.

- **Human space**: In this space, images "exist" by definition. Thus every image belongs to this space and consequently so do each of the subspaces defined below. However, for the sake of discussion I will be looking at images which we as humans can observe, interpret and talk about, like the ones printed on a piece of paper or

observed on a screen. [1]

- **Storage space**: This is space of images representation which can be stored in memory. Again, by definition, this includes all tensor representations used in a computer program, or equivalently each of the subspaces defined below. I will, however be using this to refer to images stored as JPEG, PNG and similar bitmap representations.

- **Neural Network feature space**: The space of all network layers which by definition also includes the output layer of class labels.

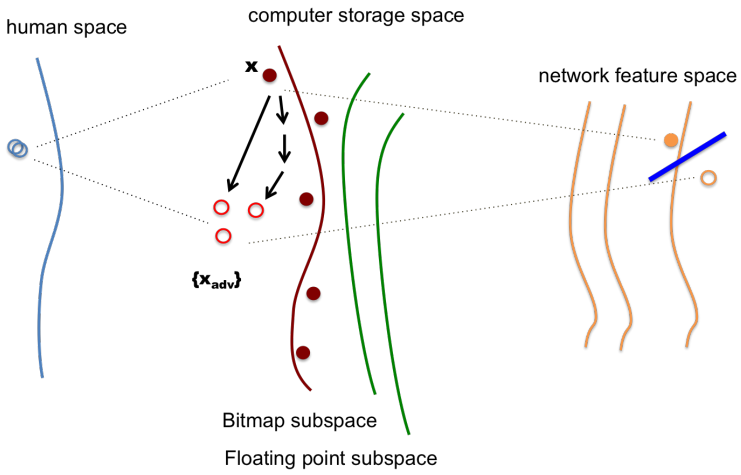


Figure 4.1: A framework for understanding the phenomenon of adversarial examples

Inconsistency: Any two image representations can by pulled to the human-space for a qualitative judgement. The aspect of "imperceptible difference" is loosely defined along this line. As humans, we do not disassociate between a paired image x and x_{adv}, in their representations as bitmaps (in *storage-subspace*) and the class-label (in *feature space*). However, Deep Learning models have been shown to make this dissociation in the space of class-labels, and by definition of adversarial examples, *a strong disassociation*, begging for a close inspection of which assumption(s) neural networks make differently.

[1] Figure 4.1 represents image-similarity for humans as closely-separated circles which is indicative of the imperfect definition of perceptual similarity for humans

4.2 Ongoing Research

Easy generation: An adversarial example can be easily generated by propagating the gradient computed for the Neural Network's cost function to the input space Goodfellow et al. (2014), Szegedy et al. (2013). An iterative version of this generation strategy, with smaller step size can also be adapted to examine the evolution of an input image through the adversarial transformations. I will refer to the first method as `FastGradSign` and the the corresponding iterative version as `SlowGradSign`.

Transferability to other models: An adversarial example generated for one model, is almost always an adversarial example for a different model (Goodfellow et al. (2014)). This indicates that adversarial nature doesn't owe an explanation of model-overfitting which is a common trait of "fit"-ed models. Traditional regularisation techniques like Dropouts and data augmentation techniques have also failed at accommodating adversarial perturbations. The most surprising observation in this regard was the robustness of adversarial images to models trained on a disjoint dataset.

Robustness to bitmap transformations: An invariance to transformation in these subspace has been explored in some recent investigations Dziugaite et al. (2016), Kurakin et al. (2016). Robustness of adversarial images to JPEG compression schemes and transformations via 'physical' world have been reported. These transformations can be argued as modifications of an image in the storage space - JPEG transformation being carried out with-in the space while the latter through an approximation of it. In their experimental setup,Kurakin et al. (2016) print an adversarial image and feed it back to the network through a camera. Based on their observations, they argue of adversarial perturbations to be occurring in an orthogonal space of JPEGs for natural images. Dziugaite et al. (2016) looked at the aspect of varying adversarial perturbation when coupled with JPEG compression. They report of a selective robustness to JPEG transformation.

Different modelling assumptions: Goodfellow et al. (2014) also report of an arguable robustness of their procedure for adversarial image generation across differently trained models, like RBM.

An investigation of CNN layers: Billovits et al. performed extensive observations about the sensitivity of activation to adversarial perturbations. Based on their observations, they hypothesise dropouts to improve robustness against adversarial images.

4.3 A Hypothesis

In theory, `mc`-approximation of neural-networks, performs an additional layer of marginalisation in its process of computing the final probability vector. A robustness against adversarial images can therefore be expected if the required marginalisation is the parameter variation an adversarial misclassification is missing on. A significant presence of this extra-marginalisation can potentially shed light on the neural-network dynamics or at the very least be utilised for an improved decision making for image classification.

4.4 Findings

Models and dataset: Adversarial images were generated for CIFAR-10 (Krizhevsky and Hinton (2009)) test set for models trained for the LeNet (LeCun et al. (1998)) architecture (Gal and Ghahramani (2015a)). This includes training with no dropouts (`no-drop`), dropouts used after the inner-product layers (`ip-drop`) and dropouts at every layer (`all-drop`). The Bayesian interpretation of dropouts as suggested in Gal and Ghahramani (2015a), allows for two interpretations for the latter two models, namely a 'std' interpretation - with no dropouts at test time (`ip-std,all-std`) and an 'mc' approximation - with dropouts at test time (`ip-mc,all-mc`).

Adversarial image generation: As per `FastGradSign` method Goodfellow et al. (2014), an adversarial image x_{adv}, can be generated for an image x by adding gradients of the cost function, computed with respect to a true label y_{true}(moving away from the 'true' class). Here, θ represents the pre-trained model parameters ($\{\theta^{no-drop}, \theta^{ip-drop}, \theta^{all-drop}\}$).

$$x_{adv} = x + \epsilon \nabla_x J(x, y_{true}, \theta) \tag{4.1}$$

Note that there are several variants of this method. Some obvious ones being, adding gradients with respect to an adversarial label, y_{adv} (moving towards an adversarial label), or adding gradients iteratively, in the same fashion as backpropagation. The latter being defined as `SlowGradSign`. This method in turn can have invariants as - adding a constant gradient through multiple iterations, or computing new gradients through back-propagation.

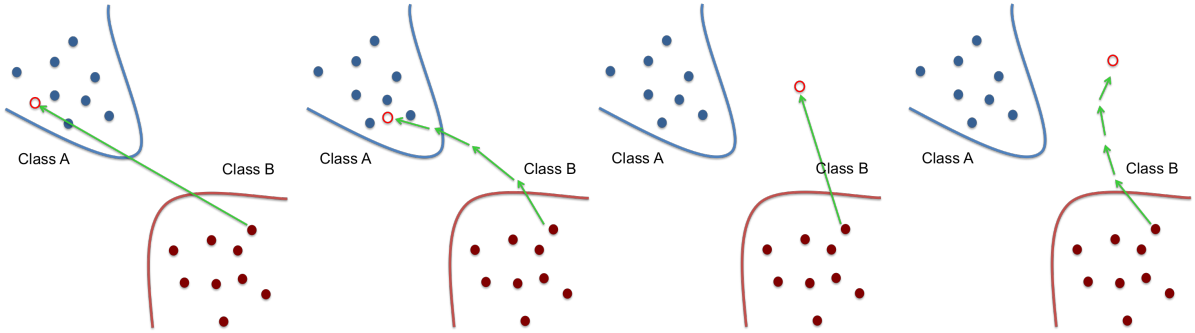


Figure 4.2: 1. FastGradSign - move towards an adversarial label; 2. SlowGradSign - move towards an adversarial label; 3. FastGradSign - move away from true label; 4. SlowGradSign - move away from true label

Experimental Setup: Low recall models are inherently robust to adversarial classification. Hence, for an objective comparison across the three models, 100 images sere chosen from the CIFAR-10 test set which were classified with high confidence by all the three trained-models. Adversarial images for these were constructed using `FastGradSign` **??** ($\epsilon = 0.008$ for $\theta^{\text{no-drop}}$)[2]. The obtained images were observed to be adversarial for each of `fc-drop-std`, `fc-drop-mc`, `all-drop-std`, and `all-drop-mc` Figure 4.4. This is consistent with the observation made by Goodfellow et al. (2014), where regularisation using dropouts was noted to provide no additional robustness against adversarial examples.

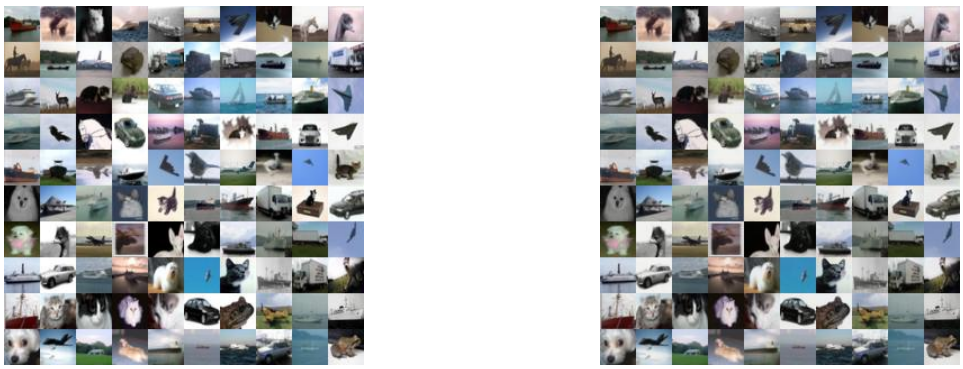


Figure 4.3: A set of 100 adversarial images were generated using `FastGradSign` for `no-drop` with $\epsilon = 0.008$. The `nodrop` model had an average true-class confidence of 75% for images on the left, while it had an average true-class confidence of 10% for the images on the right

.

[2]The images are fed to the network after a pre-processing step involving a combination of global-contrastive-normalisation step and ZCA transform Krizhevsky and Hinton (2009). The gradients computed were back-propagated through these transform

Selective robustness: In addition, I also inspected the misclassification confidence on `mc`-approximation, which were also found to provide no additional layer of global-robustness. Although, for a fixed, ϵ, `mc`-approximated models were observed to be less-confident in misclassification in comparison to their `std`-interpretations. Even if the result after Monte-Carlo approximations demonstrates a lack of robustness, in theory, the model-uncertainty being marginalised might be present. A small but insignificant presence was observed for the adversarial images (generated using `FastGradSign`, 4 samples shown in Figure 4.4).

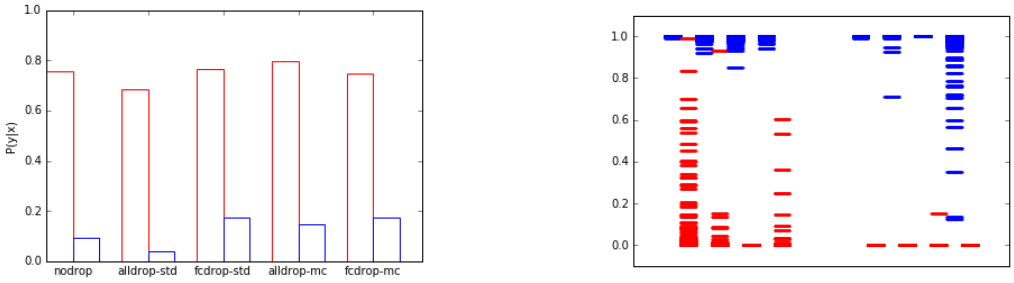


Figure 4.4: The image on the left compares average classification to the correct labels for the original image (red) and adversarial image (blue). The image on the right is the scatter plot for 4 image-pairs (original-red, adversarial-blue), with red-scattered lines being samples corresponding to true labels and blue being samples corresponding to eventual adversarial class). The left set of columns are the ip-drop-mc and the right set of columns are for all-drop-mc. The second figure is only demonstrative and doesn't lead to any tangible comclucions

The `SlowGradSign`, method allows for a finer-but-qualitiative inspection of the manifold where adversarial images are known to exist. The observed curves (Figure 4.5) are consistent with the global observation where images were misclassified by all the models. Since there is some extra information present in terms of Monte-Carlo samples in the `mc`-approximation method, this can be utilised as a debugging or a diagnostic tool [3]

[3]It must be noted that these are probability vectors in themselves. An interpretation of distribution over probability vectors might be misguided. By performing the Monte Carlo estimate the neural network model merely performs an additional marginalisation before computing the eventual probability vector.

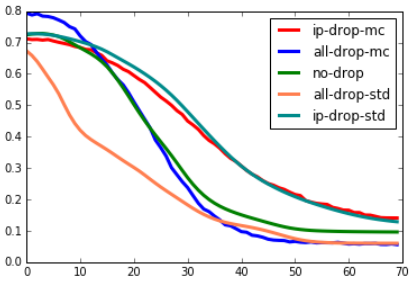


Figure 4.5: Evolution of class conditional probability for the true label (averaged across 100 images). The adversarial images were obtained using SlowGradSign with negative gradients added with respect to an adversarial label (epsilon=1e-3). The x-axis represents gradients-steps.

Statistical dispersion (moving-away from true): For a qualitative analysis, a variance of class-conditional probability for the true-label was observed for the same evolution of adversarial misclassification. The hump suggests that the `all-drop-mc` model exhibits some statistical-dispersion in these probability vectors. A better grasp at the statistical dispersion can be observed in variational-ratio as it measures dispersion across all labels (Figure 4.6 (right)). The `alldrop` model is observed to be more confused in the process of picking an adversarial class (when the `SlowGradSign` method is used to move away from the true label). As the `all-drop` model becomes less confident on the original label with slow addition of perturbations, it goes though a regime of confusion before picking an adversarial label.

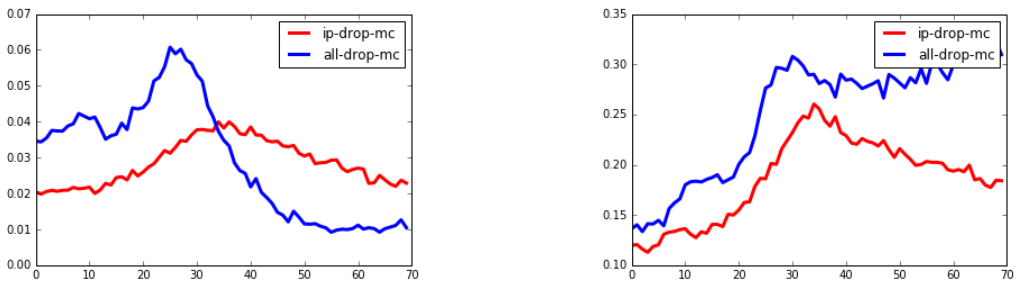


Figure 4.6: Statistical dispersion in probability vectors (**moving away from the true class**). The figure on the left quantifies the dispersion using variance in the class-conditional probabilities of the true label. The figure on the right quantifies the dispersion in the probability vectors as the variation ratio of class labels, where the labels were computed as argmax of the probability vectors. Both these quantities have been averaged across the 100 images. The x-axis represents gradients-steps.

Microscopic analysis for 1 image: The fishing expedition was furthered towards an image-specific analysis. An image of a 'ship' was selected for this purpose. As discussed previously, the flexibility of choosing the y label, allows for two different explorations of the manifold - away from the true-label (with positive gradients) and towards an adversarial label (with negative gradients). The observations in 4.7 were consistent with the global phenomenon Figure 4.5. The only interesting behaviour being exhibited by `all-drop` model for the away-from-ship case. While it moved away from the ship label, (4.7), as expected, it moved towards 'dog' instead of 'bird', which was the class picked by `no-drop` and `ip-drop`. This is not a generalisable quality as it was observed for a specific image and the stochastic-optimisation algorithms are highly sensitive to choice of step size. The other interesting observation was in the **structural similarity of `ip-drop` and `no-drop`** reflected in the case when the optimisation is done towards an adversarial class (Figure 4.7 first column, right set; the spread of the scatter plots is thin and the red-line (ip-drop) converges in the same fashion as the green (no-drop))). These are consistent with the observations of variational ratio as explained below.

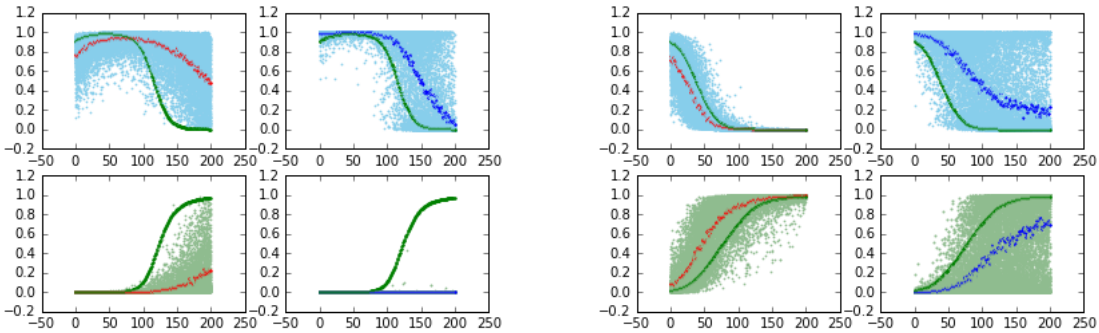


Figure 4.7: Right set of images : away from'ship' (SlowGradSign with positive epsilon). Left set of images : towards 'truck' (SlowGradSign with negative epsilon). Top row class-condition probabilities for 'ship'. Bottom row - left set: class 'bird', right set: class 'truck'. Plot lines - 'green' : nodrop, 'red' : ip-drop-mc, 'blue': all-drop-mc. The blue and green scatter points are samples with dropouts at test time. The x-axis represents gradients-steps.

Statistical dispersion (moving-towards an adversrial) The contrast of moving-away and moving towards was inspected for all the 100-images. Variational Ratio and variance were used as diagnostic tools. The results indicate of a structural similarity between `no-drop` and `ip-drop`. As the images were generated for gradients computed with θ^{nodrop}, a smoother transitioning suggests of the structural similarity. The `ipdrop`

model 'agrees-with' `nodrop`. The `alldrop` model undergoes an intermediary phase of disagreement, but eventually agrees with `no-drop` model. These observations are inconclusive in a generic setting and are speculated to reflect on the general behaviour of dropouts with respect to ReLU units.

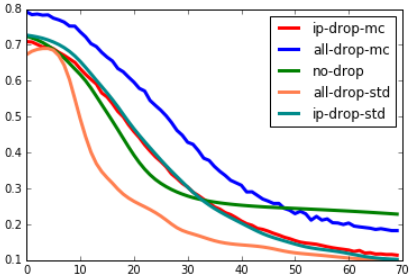


Figure 4.8: (moving-towards an adversarial label). Evolution of class conditional probability for the true label (averaged across 100 images). The adversarial images were obtained using SlowGradSign with positive gradients added with respect to true label (epsilon=1e-3). The x-axis represents gradient steps

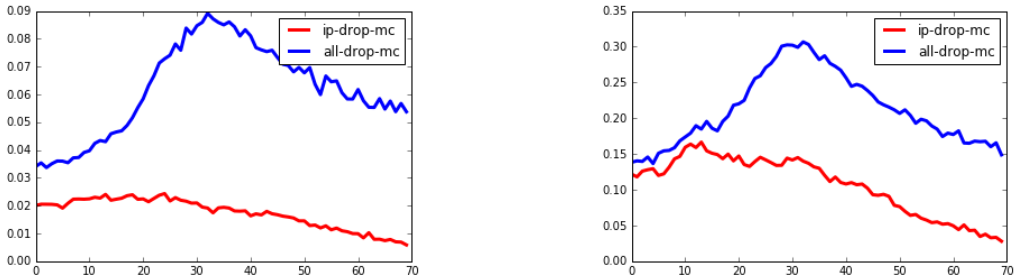


Figure 4.9: Statistical dispersion in probability vectors **(moving-towards an adversarial class)**. The figure on the left quantifies the dispersion using variance in the class-conditional probabilities of the true label. The figure on the right quantifies the dispersion in the probability vectors as the variation ratio of class labels, where the labels were computed as argmax of the probability vectors. Both these quantities have been averaged across the 100 images. The x-axis represents gradient steps

Noise images: A sensitivity to ReLU units is also evident in the adversarial behaviour of noisy-images. [4]. An interesting property lies in high interpretability even in the

[4]The noisy image was sampled from sampled in the space storage-space of CIFAR-10 data set as downloaded from the source. Each pixel of the image was sampled from a Normal distribution with mean and variance set to empirical mean and variance of the CIFAR10 test set. Before feeding it into the network, the image underwent the mandated preprocessing of gcn and ZCN.

initial stage Figure 4.10 (as has been observed in Nguyen et al. (2015)). This image was observed to jump to its adversarial label, with steeper transitions and no uncertainty regions (as opposed to the smoother banded transition observed for a natural image). This is a characteristic of ReLU units which assign high significance to numbers which exist away from the data set. A softmax activation is sensitive to such inputs from ReLU and squashes these to either 0 or 1, which is what is evident in the scatter-points in the Figure 4.10 (right).

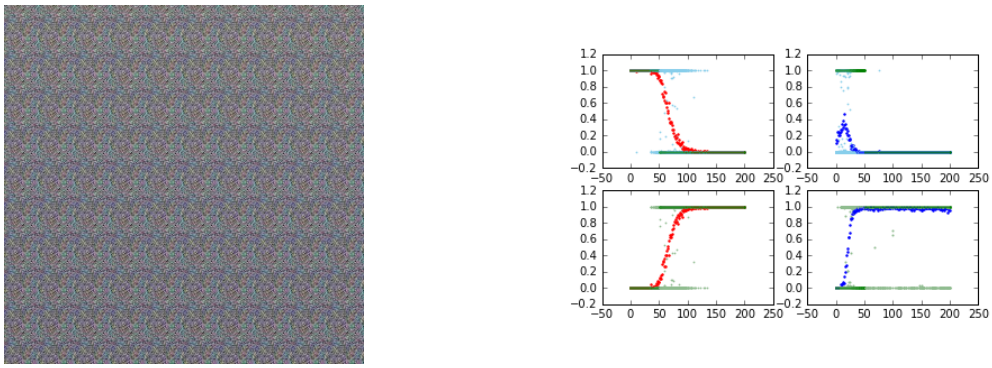


Figure 4.10: Evolution of a noisy-image (moving-towards an adversarial class). The figure on the left shows the evolution of noisy image across 100 gradient steps of Slow-GradSign. The image was observed to belong to the class 'frog'. The gradients however were set to guide it towards 'tree'.

5 | Conclusion

In this thesis, model uncertainty for adversarial images was explored. The recent demonstrations of existence of adversarial images have raised concerns about the reliability of Neural Networks. Their confident misclassification of adversarial images is an invitation for investigation of uncertainty estimates for these adversarial images. Bayesian Neural Networks are a natural starting point for this investigation. However, most Bayesian interpretations are computationally expensive. With its root in variational inference, estimates based on dropouts have been proposed to capture model uncertainty. Benefiting form this inexpensive proposal, model uncertainty for adversarial images was explored.

Neural Networks with dropout-approximation at test time were not found to be robust to adversarial images. An additional layer of marginalisation by incorporating dropouts at test time was not known to ascribe the desired robustness to a neural network. A finer resolution with small gradient increments was indicative of a period of confusion that a neural network undergoes before assigning an adversarial label with high-confidence. Thus, a selective robustness to small adversarial perturbations can be claimed for dropout-approximated neural networks. As the experiments were conducted on a relatively small data set, the speculations may not generalise in multi-dimensional classifications like ImageNet and CIFAR100.

The bizarre behaviour of neural networks, when faced with noisy (humanly) uninterpretable images is an open ground and a potential ground of answers. A detailed investigation of human-disassociation and neural-network 4.1 disassociation across a more refined space hierarchy might shed light on the currently unexplained phenomenon of adversarial images.

Robustness against adversarial images is the eventual desired property. Many proposals have been suggested in this direction. An interesting investigation would be the use a

non-parametric approximation on the final layer The classical Gaussian Process models are known to push the uncertainty to regions to their prior when faced with a point away from a data space. With an appropriate kernel choice, this can be hypothesised to account for the desired marginalisation of adversarial perturbation. This has a similar flavour to the proposal made in litertature, where they propose to push the a data distribution estimate as a prior before the softmax.

It is also contended that the phenomenon of adversarial images may not owe its explanation in dissimilarity for natural images but in the strong interpretability and correspondingly, strong association and disassociation for noisy images. This also calls for an investigation of data-augmentation schemes directed towards disassociation rather the, more conventional, association.

References

Arthur Asuncion and David Newman. Uci machine learning repository, 2007. (Page 10)

Ian Goodfellow Yoshua Bengio and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2016. URL http://www.deeplearningbook.org. (Page 1)

Chris Billovits, Mihail Eric, and Nipun Agarwala. Hitting depth: Investigating robustness to adversarial examples in deep convolutional neural networks. (Page 12)

Richard T Cox. Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13, 1946. (Pages 1 and 3)

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. (Pages 10 and 12)

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015a. (Page 13)

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *arXiv preprint arXiv:1506.02142*, 2015b. (Page 8)

Zoubin Ghahramani. Bayesian non-parametrics and the probabilistic approach to modelling. *Phil. Trans. R. Soc. A*, 371(1984):20110553, 2013. (Page 3)

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015. (Page 3)

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. (Pages 12, 13, and 15)

Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010. (Page 1)

Ian Hacking. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, 2006.

Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58. ACM, 2011. (Page 10)

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003. (Page 3)

William H Jefferys and James O Berger. Ockham's razor and bayesian analysis. *American Scientist*, 80(1):64–72, 1992. (Page 4)

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. (Page 9)

Diederik P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *arXiv preprint arXiv:1506.02557*, 2015. (Page 9)

Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997. (Page 1)

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. (Pages 13 and 15)

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. (Page 12)

Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits, 1998. (Page 13)

Shin-ichi Maeda. A bayesian encourages dropout. *arXiv preprint arXiv:1412.7003*, 2014. (Pages 8 and 9)

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 427–436. IEEE, 2015. (Page 19)

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. (Pages 1, 10, and 12)

Harri Valpola. *Bayesian ensemble learning for nonlinear factor analysis.* Finnish Academies of Technology, 2000. (Page 3)