

# ONE BIT MATTERS: UNDERSTANDING ADVERSARIAL EXAMPLES AS THE ABUSE OF REDUNDANCY

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Adversarial examples have somewhat disrupted the enormous success of machine learning (ML) and are causing concern with regards to its trustworthiness: A small perturbation of an input results in an arbitrary failure of an otherwise seemingly well-trained ML system. While studies are being conducted to discover the intrinsic properties of adversarial examples, such as their transferability and universality, there is insufficient theoretic analysis to help understand the phenomenon in a way that can influence the design process of ML experiments. In this paper, we deduce an information-theoretic model which explains adversarial attacks universally as the abuse of feature redundancies in ML algorithms. We prove that feature redundancy is a necessary condition for the existence of adversarial examples. Our model helps to explain the major questions raised in many anecdotal studies on adversarial examples. Our theory is backed up by empirical measurements of the information content of benign and adversarial examples on both image and text datasets. Our measurements show that typical adversarial examples introduce just enough redundancy to overflow the decision making of a machine learner trained on corresponding benign examples. We conclude with actionable recommendations to improve the robustness of machine learners against adversarial examples.

## 1 INTRODUCTION

Deep neural networks (DNNs) have been widely applied to various applications and achieved great successes (Ciregan et al., 2012; Sallab et al., 2017; Hinton et al., 2012). This is mostly due to their versatility: DNNs are able to be trained to fit a target function. Therefore, it raises great concerns given the discovery that DNNs are vulnerable to *adversarial examples*. These are carefully crafted inputs, which are often seemingly normal within the variance of the training data but can fool a well-trained model with high attack success rate (Goodfellow et al., 2015). Adversarial examples can be generated for various types of data, including images, text, audio, and software (Carlini & Wagner, 2018; Ebrahimi et al., 2017), and for different ML models, such as classifiers, segmentation models, object detectors, and reinforcement learning systems (Kos & Song, 2017; Huang et al., 2017). Moreover, adversarial examples are transferable (Tramèr et al., 2017; Liu et al., 2017)—if we generate adversarial perturbation against one model for a given input, the same perturbation will have high probability to be able to attack other models trained on similar data, regardless how different the models are. Last but not the least, adversarial examples cannot only be synthesized in the digital world but also in the physical world (Evtimov et al., 2017; Kurakin et al., 2016), which has caused great real-world security concerns.

Given such subtle, yet universally powerful attacks against ML models, several defensive methods have been proposed. For example, Liu et al. (2018); Feinman et al. (2017) pre-process inputs to eliminate certain perturbations. Other work (Cao & Gong, 2017) suggest to push the adversarial instance into random directions so they hopefully escape a local minimum and fall back to the correct class. The authors are aware of ongoing work to establish metrics to distinguish adversarial examples from benign ones so that one can filter out adversarial examples before they are used by ML models. However, so far, all defense and detection methods have shown to be adaptively attackable. Therefore, intelligent attacks against intelligent defenses become an arms race. Defending against adversarial examples remains an open problem.

In this paper, we propose and validate a theoretical model that can be used to create an actionable understanding of adversarial perturbations. Based upon the model, we give recommendations to modify the design process of ML experiments such that the effect of adversarial attacks is mitigated. We illustrate adversarial examples using an example of a simple perceptron network that learns the Boolean equal operator and then generalize the example into a universal model of classification based on Shannon’s theory of communication. We further explain how adversarial examples fit the thermodynamics of computation. We prove a necessary condition for the existence of adversarial examples. In summary, the contributions of the paper are listed below:

- a model for adversarial examples consistent with related work, physics and information theory;
- a proof that using redundant features is a necessary condition for the vulnerability of ML models to adversarial examples;
- extensive experiments that showcase the relationship between data redundancy and adversarial examples
- actionable recommendations for the ML process to mitigate adversarial attacks.

## 2 RELATED WORK

Given a benign sample  $\mathbf{x}$ , an adversarial example  $\mathbf{x}_{\text{adv}}$  is generated by adding a small perturbation  $\epsilon$  to  $\mathbf{x}$  (i.e.  $\mathbf{x}_{\text{adv}} = \mathbf{x} + \epsilon$ ), so that  $\mathbf{x}_{\text{adv}}$  is misclassified by the targeted classifier  $g$ . Related work has mostly focused on describing the properties of adversarial examples as well as on defense and detection algorithms.

Goodfellow et al. have hypothesized that the existence of adversarial examples is due to the linearity of DNNs (Goodfellow et al., 2015). Later, boundary-based analysis has been derived to show that adversarial examples try to cross the decision boundaries (He et al., 2018). More studies regarding to data manifold have also been leveraged to better understand these perturbations (Ma et al., 2018; Gilmer et al., 2018; Wang et al., 2016). While these works provide hints to obtain a more fundamental understanding, to the best of our knowledge, no study was able to create a model that results in actionable recommendations to improve the robustness of machine learners against adversarial attacks. Prior work do not a measurement process or theoretically show the necessary or sufficient conditions for the existence of adversarial examples.

Several approaches have been proposed to generate adversarial examples. For instance, the fast gradient sign method has been proposed to add perturbations along the gradient directions (Goodfellow et al., 2015). Other examples are optimization algorithms that search for the minimal perturbation (Carlini & Wagner, 2017b; Liu et al., 2017). Based on the adversarial goal, attacks can be classified into two categories: targeted and untargeted attacks. In a targeted attack, the adversary’s objective is to modify an input  $\mathbf{x}$  such that the target model  $g$  classifies the perturbed input  $\mathbf{x}_{\text{adv}}$  as a *targeted* class chosen, which differs from its ground truth. In a untargeted attack, the adversary’s objective is to cause the perturbed input  $\mathbf{x}_{\text{adv}}$  to be misclassified in *any class* other than its ground truth. Based on the adversarial capabilities, these attacks can be categorized as white-box and black-box attacks, where an adversary has full knowledge of the classifier and training data in the white-box setting (Szegedy et al., 2014; Goodfellow et al., 2015; Carlini & Wagner, 2017a; Moosavi-Dezfooli et al., 2015; Papernot et al., 2016b; Biggio et al., 2013; Fawzi & Frossard, 2015; Kanbak, 2017; Kurakin et al., 2016), but zero knowledge about them in the black-box setting (Papernot et al., 2016a; Liu et al., 2017; Moosavi-Dezfooli et al., 2016; Mopuri et al., 2017).

Interestingly enough, adversarial examples are not restricted to ML. Intuitively speaking, and consistent with the model that is presented in this paper, acoustic noise masking could be regarded as an adversarial attack on our hearing system. Acoustic masking happens, for example, when a clear sinusoid tone cannot be perceived anymore because a small amount of white noise has been added to the signal (Phillips, 1990). This effect is exploited in MP3 audio compression and privacy applications. Similar examples exist, such as optical illusions in the visual domain (Ponomarenko et al., 2007) and defense mechanisms against sensor-guided attacks (Warm et al., 1997) in the military domain.

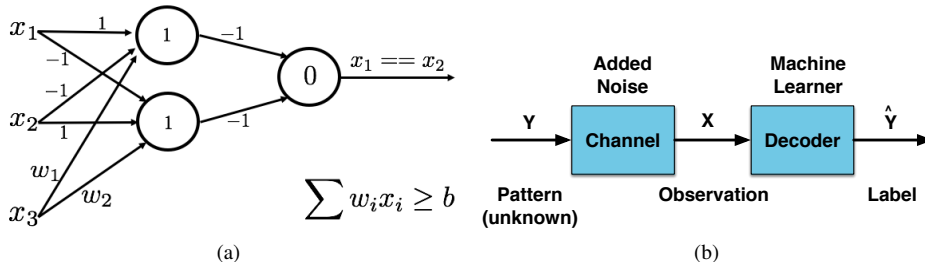


Figure 1: (a) A perceptron network implementing the Boolean equality of  $x_1$  and  $x_2$  with added redundancy  $x_3$  (see Section 3.1). (b) Shannon channel model of machine learning inspired by MacKay (2003). (see Section 3.2).

### 3 A MODEL FOR ADVERSARIAL EXAMPLES

Intuitively speaking, we want to explain the phenomenon shown in Figure 2a, which depicts a plane filled with points that were originally perfectly separable with two 2D linear separations. As the result of perturbing several points by a mere 10% of the original position, the separation of the two classes requires many more than two linear separators. That is, a small amount of noise can overflow the separation capability of a network dramatically. In the following section, we introduce an example model along which we will derive our mathematical understanding, consistent with our experiments in Section 4 and the related work mentioned in Section 2.

#### 3.1 ADVERSARIAL EXAMPLES AND BOOLEAN FUNCTIONS

Consider a perceptron network which implements the Boolean equal function ("NXOR") between the two variables  $x_1$  and  $x_2$ . The input  $x_3$  is redundant in the sense that the result of  $x_1 == x_2$  is not influenced by the value of  $x_3$ .

The first obvious observation is that adding  $x_3$  doubles the input space. Instead of  $2^2 = 4$  possible input pairs, we now have  $2^3 = 8$  possible input triples that the network needs to map to sustain the result  $x_1 == x_2$  for all possible combinations of  $x_1, x_2, x_3$ . The network architecture shown in Figure 1a, for example, theoretically has the capacity to be trained to learn all 8 input triples (Friedland & Krell, 2017). Translating this example into a practical ML scenario, however, this would mean that we have to exhaustively train the entire input space for all possible settings of the noise. This is obviously unfeasible.

We will therefore continue our analysis in a more practical setting. We assume a network like in Figure 1a is correctly trained to model  $x_1 == x_2$  in the absence of a third input. One example configuration is shown. Now, we train weights  $w_1$  and  $w_2$  to try to suppress the redundant input  $x_3$  by going through all possible combinations for  $w_i \in \{-1, 0, 1\}$ . This weight choice is without losing generality as the inputs  $x_i$  are  $\in \{0, 1\}$  (see Rojas (2013)). An adversarial example is defined as a triple  $(x_1, x_2, x_3)$  such that the output of the network is not the result of  $x_1 == x_2$ . Simulating through all configurations exhaustively results in Table 2b. The only case that allows for 100% accuracy, i.e., no adversarial examples, is the setting  $w_1 = w_2 = 0$ , in which case  $x_3$  is suppressed completely. In the other cases, we can roughly say that the more the network pays attention to  $x_3$ , the worse the result (allowing edges). That is, the result is better if one of the  $w_i$  is set to 0 compared to none. Furthermore, the higher the potential, defined as the difference between the maximum and the minimum possible activation value as scaled by the  $w_i$ , the worse the result is. The intuition behind this is that higher potential change leads to higher potential impacts to the overall network.

Using this simple model, one can see the importance of suppressing noise. Thresholds of neurons taking redundant inputs should be high, or equivalently, weights should be close to 0 (and equal to 0 in the optimal scenario). Now generalizing the example to a large network training images with 'real-valued' weights, it becomes clear that redundant bits of an image should be suppressed by low enough weights otherwise it is easy to generate an exponential explosion of patterns needed to be recognized.

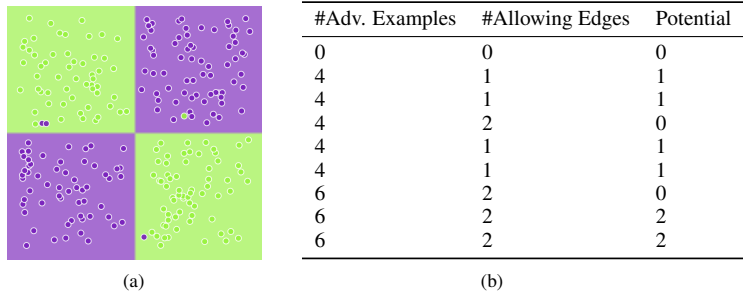


Figure 2: (a) A network trained on XOR cannot separate the two classes anymore perfectly once we add 10% of noise to some of the inputs. (b) Suppression capability of adversarial attacks against the Boolean equality network.

### 3.2 GENERAL MODEL

The generalization of the example from the previous section is shown in Figure 1b. The model shows a machine learner performing the task of matching an unknown noisy pattern to a known pattern (label). For example, a perceptron network implements a function of noisy input data. It quantizes the input to match a known pattern and then outputs the results of the learned function from a known pattern to a known output. Formally, the random variable  $Y$  encodes unknown patterns that are sent over a noisy channel. The observation at the output of the channel is denoted by the random variable  $X$ . For example,  $X$  could represent image pixels. The machine learner then erases all the noise bits in  $X$  to match against trained patterns which are then mapped to known outputs  $\hat{Y}$ , for example, the labels. It is well known from the thermodynamics of computing (Feynman et al., 2000) that setting memory bits and copying them is theoretically energy agnostic. However, resetting bits to zero is not. In other words, we need to spend energy (computation) to reset the noisy bits added by the channel and captured in the observation to get to a distribution of patterns  $\hat{Y}$  that is isomorphic to the original (unknown) distribution of patterns  $Y$ . Connecting back to the NXOR example from the previous section,  $Y$  would be the distribution over the input variables  $x_1$  and  $x_2$ . The noise added is modeled by  $x_3$  and  $\hat{Y}$  is the desired output isomorphic to  $x_1$  and  $x_2$  being equal. Now assuming a fully trained model, this model allows us to explain several phenomena explored in the introduction and Section 2.

First, as illustrated in the previous section, we can view the machine learner as a trained bit eraser. That is, the machine learner has been trained to erase exactly those bits that are irrelevant to the pattern to be matched. This elimination of irrelevance constitutes the generalization capability. For a black box adversarial attack, we therefore just need to add enough irrelevant input to overflow this bit erasure function. As a result, insufficient redundant bits can be absorbed and the remaining bits now create an exponential explosion for the pattern matching functionality. In a whitebox attack, an attacker can guess and check against the bit erasing patterns of the trained machine learner and create a sequence of input bits that specifically overflows the decision making. In both cases, our model predicts that adversarial patterns should be harder to learn as they consist of more bits to erase. This is confirmed in our experiments in Section 4. It is also clear that the theoretical minimum overflow is one bit, which means, small perturbations can have big effects. This will be made rigorous in Section 3.3. It is also well known that, for example, in the image domain one bit of difference is not perceivable by a human eye. Training with noisy examples will most likely make the machine learner more robust as it will learn to reduce redundancies better. However, a specific whitebox attack (with lower entropy than random noise), which constitutes a specific perceptron threshold overflow, will always be possible because training against the entire input space is unfeasible.

Second, with training data available, it is highly likely that a surrogate machine learner will learn to erase the same bits. This means that similar bit overflows will work on both the surrogate and the original ML attack, thus explaining transferability-based attacks.

### 3.3 ABUSE OF REDUNDANCY

In the following we will present a proof based on the model presented in Section 3.2 and the currently accepted definition of adversarial examples (Wang et al., 2016) that shows that feature redundancy is

indeed a necessary condition for adversarial examples. Throughout, we assume that a learning model can be expressed as  $f(\cdot) = g(T(\cdot))$ , where  $T(\cdot)$  represents the feature extraction function and  $g(\cdot)$  is a simple decision making function, e.g., logistic regression, using the extracted features as the input.

**Definition 1** (Adversarial example (Wang et al., 2016)). *Given a ML model  $f(\cdot)$  and a small perturbation  $\delta$ , we call  $x'$  an adversarial example if there exists  $x$ , an example drawn from the benign data distribution, such that  $f(x) \neq f(x')$  and  $\|x - x'\| \leq \delta$ .*

We first observe that  $\forall x, x' \exists \delta$  such that  $\|x - x'\| \leq \delta \implies f(x) = f(x')$  is the generalization assumption of a machine learner. The existence of an adversarial  $x'$  is therefore equivalent to a contradiction of the generalization assumption. This is,  $x'$  could be called a counter example. Practically speaking, a single counter example to the generalization assumption does not make the machine learner useless though. In the following, and as explained in previous sections, we connect the existence of adversarial examples to the information content of features used for making predictions.

**Definition 2** (Feature redundancy). *Let  $X$  and  $Y$  represent the random variables corresponding to features and unknown pattern, respectively. Let  $T_Y^{\min}(X)$  denote the minimal sufficient statistic of  $X$ , i.e.,  $T_Y^{\min}(X) = \arg \min_{T(X): I(X;Y)=I(T(X);Y)} H(T(X))$ , where  $T(X)$  is the sufficient statistic for  $Y$ . The redundancy of using  $T(\cdot)$  as the feature extractor for predicting  $Y$  is defined as*

$$R(T(X); Y) = H(T(X)) - H(T_Y^{\min}(X)) \quad (1)$$

**Theorem 1.** *Suppose that the feature extractor  $T(X)$  is a sufficient statistic for  $Y$  and that there exist adversarial examples for the ML model  $f(\cdot) = g(T(\cdot))$ , where  $g(\cdot)$  is an arbitrary decision making function. Then,  $T(X)$  is not a minimal sufficient statistic.*

We leave the proof to the appendix. The idea of the proof is to explicitly construct a feature extractor with lower entropy than  $T(X)$  using the properties of adversarial examples. Theorem 1 shows that the existence of adversarial examples implies that the feature representation contains redundancy. We would expect that more robust models will generate more succinct features for decision making. We will corroborate this intuition in Section 4.2.

## 4 EXPERIMENTAL RESULTS

In this section, we provide empirical results to justify our theoretical model for adversarial examples. Our experiments aim to answer the following questions. First, are adversarial examples indeed more complex (e.g. they contain more redundant bits with respect to the target that need to be erased by the machine learner)? If so, adversarial examples should require more parameters to memorize in a neural network. Second, is feature redundancy a large enough cause of the vulnerability of DNNs that we can observe it in a real-world experiment? Third, can we exploit the higher complexity of adversarial examples to possibly detect adversarial attacks? Fourth, does quantization of the input indeed not harm classification accuracy?

### 4.1 CAPACITY MEASUREMENTS

Our model implies that adversarial examples generally have higher complexity than benign examples. In order to evaluate this claim practically, we need to show that this complexity increase is in fact an increase of irrelevant bits with regards to the encoding performed in neural networks towards a target function. This can be established by showing that adversarial examples are more difficult to memorize than benign examples. In other words, a larger model capacity is required for training adversarial examples. To quantitatively measure how much extra capacity is needed, we measure the capacity of multi-layer perceptrons (MLP) models with or without non-linear activation function (ReLU) on MNIST. Here we define the model capacity as the minimal number of parameters needed to memorize all the training data. To explore the capacity, we first build an MLP model with one hidden layer (units: 64). This model is efficient enough to achieve high performance and memorize all training data (with ReLU). After that, weights are reduced by randomly setting some of their values to zero and marking them untrainable. The error  $\epsilon$  is set to evaluate the training success (training accuracy is larger than  $1 - \epsilon$ ). We explore the minimal number of parameters and utilize binary search to reduce computation complexity. Finally, we change different  $\epsilon$  and repeat the above steps. As

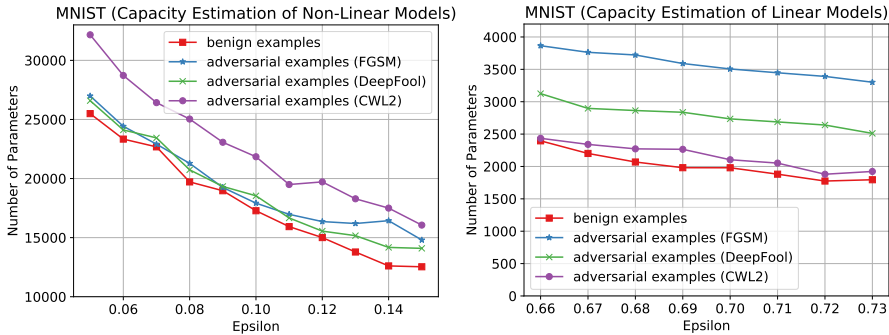


Figure 3: Comparison of estimated model capacity (minimal number of parameters required) as a function of allowed error  $\epsilon$  under different attack algorithms on MNIST. The results are obtained based on the MLPs with and without non-linear activation functions (e.g. ReLU). Note that the non-linear models can generalize the data better, so the  $\epsilon$  is fixed in a much smaller range.

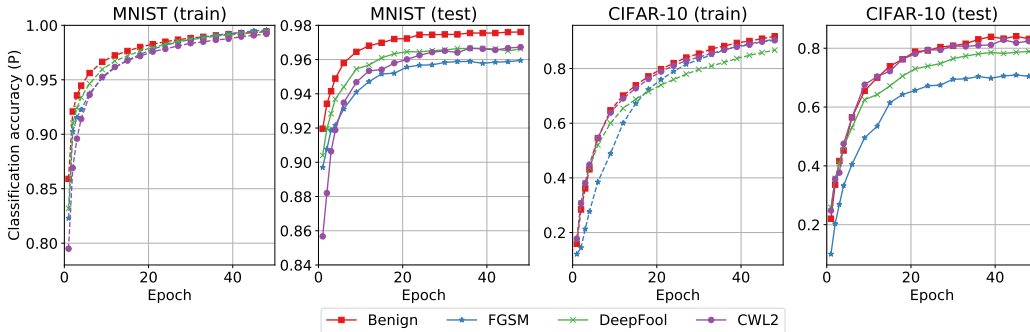


Figure 4: Classification accuracy of different attack models as a function of training epoch. Adversarial examples are generated with different attack methods on both MNIST and CIFAR-10.

illustrated in Figure 3, the benign examples always require fewer number of weights to memorize on different datasets with various attack methods. It is shown that adversarial examples indeed require larger capacity. From the training/testing process given in Figure 4, we can draw the same conclusion. The benign examples are always fitted and predicted more efficiently than adversarial examples given the same model. That is to say, adversarial examples have more complexity and therefore require higher model capacity.

#### 4.2 INPUT COMPLEXITY ESTIMATION

We now investigate if there are possible ways to exploit the higher complexity of adversarial examples to possibly detect adversarial attacks. That is to say, we need a machine-learning independent measure of entropy to evaluate how much benign and adversarial examples differ. For images, we utilized Maximum Likelihood (MLE), Minimax (JVHW) (Jiao et al., 2014) and compression estimators for

Table 1: Comparison of complexity for benign and adversarial examples on MNIST and CIFAR-10.

Dataset	Examples	H (MLE)	H (JVHW)	Original Size	Compressed Size
MNIST	<b>Benign</b>	<b>1.741</b>	<b>1.887</b>	<b>988.89 B</b>	<b>431.40 B</b>
	FGSM (2015)	2.488	2.601	1690.36 B	503.54 B
	DeepFool (2015)	4.844	5.088	1654.99 B	510.41 B
	CW ( $L_2$ ) (2017a)	4.094	4.301	1159.01 B	437.27 B
CIFAR-10	<b>Benign</b>	<b>9.595</b>	<b>7.104</b>	<b>1845.98 B</b>	<b>741.36 B</b>
	FGSM (2015)	9.937	7.710	2717.01 B	872.40 B
	DeepFool (2015)	9.675	7.147	1880.41 B	743.02 B
	CW ( $L_2$ ) (2017a)	9.621	7.113	1850.54 B	741.56 B

Table 2: Comparison of complexity for benign and adversarial examples on IMDB and Reuters2.

Dataset	Examples	Mean Bits	H (BW)	H (bW)	Compressed Size
IMDB	<b>Benign</b>	<b>4.556</b>	<b>0.569</b>	<b>0.99775</b>	<b>2.235 B</b>
	FGSM (2015)	4.671	0.584	0.99926	3.027 B
	FGVM (2015)	4.701	0.588	0.99944	3.481 B
	DeepFool (2015)	4.632	0.580	0.99953	3.156 B
Reuters2	<b>Benign</b>	<b>4.946</b>	<b>0.618</b>	<b>0.99457</b>	<b>1.934 B</b>
	FGSM (2015)	5.032	0.629	0.99712	3.181 B
	FGVM (2015)	5.035	0.629	0.99754	3.237 B
	DeepFool (2015)	5.202	0.650	0.99545	3.301 B

three kind of adversarial examples (FGSM (Goodfellow et al., 2015), DeepFool (Moosavi-Dezfooli et al., 2015), CW ( $L_2$ ) (Carlini & Wagner, 2017a)) on both MNIST and CIFAR-10 dataset. These are four metrics for entropy measurement, all of which indicate higher unpredictability with the value increasing. For compression estimation, prior work (Friedland et al., 2018) has found that an optimal quantification ratio exists for DNNs and appropriate perceptual compression is not harmful. Therefore, we consider such information as redundancy and set the quality scale to 20 following their strategy. We also reproduce the experiments in our settings and obtain the same results shown in Figure 5.

As shown in Table 1, the benign images have smallest complexity in all of the four metrics, which suggests less entropy (lower complexity) and therefore higher predictability. Similarly, we also design four metrics for text entropy estimation including mean bits per character, byte-wise entropy (BW), bit-wise entropy (bW) and compression size. More specifically, BW and bW are calculated based on the histogram of the bytes or the bits per word. It is worthwhile to note that all of the metrics are measured on adversarial-benign altered pairs, because adversarial algorithms only modify specific words of the texts. In our evaluations, FGSM, FGVM (Miyato et al., 2015) and DeepFool attacks are implemented. From Table 2, we can draw the conclusion that adversarial texts introduce more redundant bits with regards to the target function which results in higher complexity and therefore higher entropy. A reduction of adversarial attacks via entropy measurement is therefore potentially possible for both data types.

#### 4.3 OUTPUT COMPLEXITY ESTIMATION

Inspired by Theorem 1, we investigate the relation between feature redundancy and the robustness of a ML model. We expect that more robust models would employ more succinct features for prediction. To validate this, we design models with different levels of robustness and measure the entropy of extracted features (i.e., the input to the final activation layer for decisions). In experiments, we choose All-CNNs network for it has no fully-connected layer and the last convolutional layer is directly followed by the global average pooling and softmax activation layer, which is convenient for the estimation of entropy. In other words, we can estimate the entropy of the feature maps extracted by the last convolutional layer using perceptual compression and MLE/JVHW estimators. Specifically, we train different models on benign examples and control the ratios of adversarial examples in adversarial re-training period to obtain models with different robustness. In general, the larger ratio of adversarial examples we re-train, the more robust models we will obtain. The robustness in experiments is measured by the test accuracy on adversarial examples. Then we obtain the feature maps on adversarial examples generated by these models and compress them to  $q = 20$ , following Friedland et al. (2018). Finally, we measure the compressed entropy of using MLE and JVHW estimator like Section 4.2. As illustrated in Figure 6, the estimated entropy (blue dots) decreases as the classification accuracy (red dots) increases for all the three adversarial attacks (FGSM, DeepFool, CW) and the two datasets (MNIST, CIFAR), which means that the redundancy of last-layer feature maps is lower when the models become more robust.

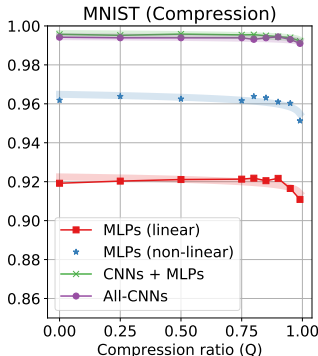


Figure 5: Classification accuracy as a function of the JPEG compression quality  $q$ . The shadow curve represents the properly scaled version of the theoretical curve in Friedland et al. (2018).

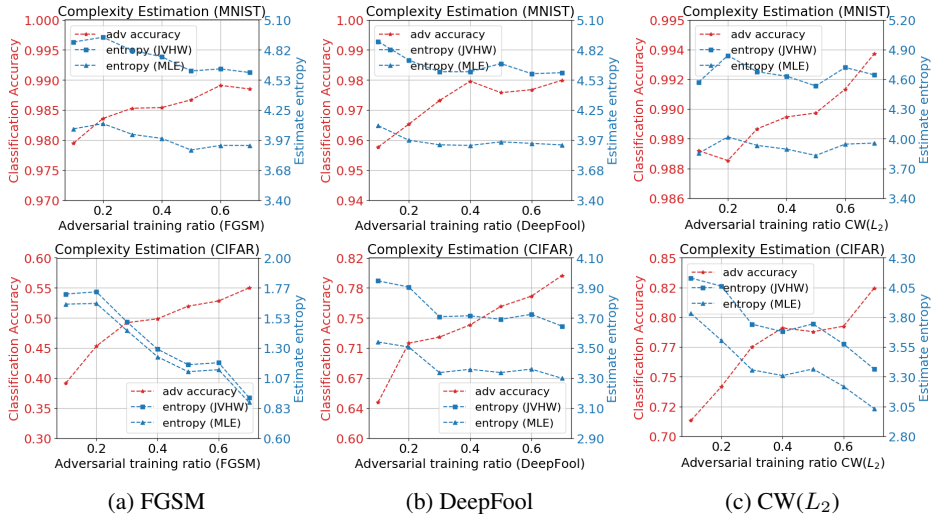


Figure 6: Estimated entropy of the output layer under All-CNNs architecture with different robustness. Three adversarial attacks (FGSM, DeepFool, CW) are tested on MNIST and CIFAR classification dataset. Note that only the adversarial examples that successfully fool the networks are calculated.

Surprisingly, adding adversarial examples into the training set serves as an implicit regularizer for feature redundancy.

## 5 CONCLUSION AND RECOMMENDATIONS

Our theoretical and empirical results presented in this paper consistently show that adversarial examples are enabled by irrelevant input that the networks was not trained to suppress. In fact, a single bit of redundancy can be exploited to cause the ML models to make arbitrary mistakes. Moreover, redundancy exploited against one model can also affect the decision of another model trained on the same data as that other model learned to only cope with the same amount of redundancy (transferability-based attack). Unfortunately, unlike the academic example in Section 3.1, we almost never know how many variables we actually need. For image classification, for example, the current assumption is that each pixel serves as input and it is well known that this is feeding the network redundant information e.g., nobody would assume that the upper-most left-most pixel contributes to an object recognition result when the object is usually centered in the image.

Nevertheless, the highest priority actionable recommendation has to be to reduce redundancies. Before deep learning, manually-crafted features reduced redundancies assumed by humans before the data entered the ML system. This practice has been abandoned with the introduction of deep learning, explaining the temporal correlation with the discovery of adversarial examples. Short of going back to manual feature extraction, automatic techniques can be used to reduce redundancy. Obviously, adaptive techniques, like auto encoders, will be susceptible to their own adversarial attacks. However, consistent with our experiments in Section 4.2, and dependent on the input domain, we recommend to use lossy compression. Similar results using quantization have been reported for MP3 and audio compression (Friedland et al., 2018) as well as molecular dynamics (Liu, 2018). In general, we recommend a training procedure where input data is increasingly quantized while training accuracy is measured. The point where the highest quantization is achieved at limited loss in accuracy, is the point where most of the noise and least of the content is lost. This should be the point with least redundancies and therefore the operation point least susceptible to adversarial attacks. In terms of detecting adversarial examples, we showed in Section 4 that estimating the complexity of the input using surrogate methods, such as different compression techniques, can serve as a prefilter to detect adversarial attacks. We will dedicate future work to this topic. Ultimately, however, the only way to practically guarantee adversarial attacks cannot happen is to present every possible input to the machine learner and train to 100% accuracy, which contradicts the idea of generalization in ML itself. There is no free lunch.



## REFERENCES

- Battista Biggio, Iginio Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 387–402. Springer, 2013.
- Xiaoyu Cao and Neil Zhenqiang Gong. Mitigating evasion attacks to deep neural networks via region-based classification. *arXiv preprint arXiv:1709.05583*, 2017.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy, 2017*, 2017a.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 39–57. IEEE, 2017b.
- Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv preprint arXiv:1801.01944*, 2018.
- Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*, pp. 3642–3649. IEEE, 2012.
- Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for nlp. *arXiv preprint arXiv:1712.06751*, 2017.
- Ivan Evtimov, Kevin Eykholt, Earlene Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2017.
- Alhussein Fawzi and Pascal Frossard. Manitest: Are classifiers really invariant? *arXiv preprint arXiv:1507.06535*, 2015.
- Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.
- Richard Phillips Feynman, Anthony J Hey, and Robin W Allen. *Feynman lectures on computation*. Perseus Books, 2000.
- Gerald Friedland and Mario Krell. A capacity scaling law for artificial neural networks. *arXiv preprint arXiv:1708.06019*, 2017.
- Gerald Friedland, Jingkang Wang, Ruoxi Jia, and Bo Li. The helmholtz method: Using perceptual compression to reduce machine learning complexity. *arXiv preprint arXiv:1807.10569*, 2018.
- Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- Warren He, Bo Li, and Dawn Song. Decision boundary analysis of adversarial examples. In *ICLR-International Conference on Learning Representations*, 2018.
- Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.
- Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.
- Jiantao Jiao, Kartik Venkat, and Tsachy Weissman. Order-optimal estimation of functionals of discrete distributions. *CoRR*, abs/1406.6956, 2014. URL <http://arxiv.org/abs/1406.6956>.

- Can Kanbak. Measuring robustness of classifiers to geometric transformations. Technical report, 2017.
- Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*, 2017.
- Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
- Shuai Liu. *Compression and Analysis of Molecular Dynamics Data*. University of California, Berkeley, 2018.
- Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.
- Zihao Liu, Qi Liu, Tao Liu, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. *arXiv preprint arXiv:1803.05787*, 2018.
- Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Michael E Houle, Grant Schoenebeck, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.
- David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *stat*, 1050:25, 2015.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *arXiv preprint arXiv:1511.04599*, 2015.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *arXiv preprint arXiv:1610.08401*, 2016.
- Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. *arXiv preprint arXiv:1707.05572*, 2017.
- Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016a.
- Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387. IEEE, 2016b.
- Dennis P Phillips. Neural representation of sound amplitude in the auditory cortex: effects of noise masking. *Behavioural brain research*, 37(3):197–214, 1990.
- Nikolay Ponomarenko, Flavia Silvestri, Karen Egiazarian, Marco Carli, Jaakko Astola, and Vladimir Lukin. On between-coefficient contrast masking of dct basis functions. In *Proceedings of the third international workshop on video processing and quality metrics*, volume 4, 2007.
- Raúl Rojas. *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.
- Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.
- Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

Beilun Wang, Ji Gao, and Yanjun Qi. A theoretical framework for robustness of (deep) classifiers against adversarial examples. *arXiv preprint arXiv:1612.00334*, 2016.

Berndt Warm, Detlev Wittmer, and Matthias Noll. Apparatus for defending against an attacking missile, February 4 1997. US Patent 5,600,434.

## A PROOF OF THEOREM 1

*Proof.* Let  $\mathcal{X}$  be the set of admissible data points and  $\mathcal{X}'$  denote the set of adversarial examples. We prove this theorem by constructing a sufficient statistic  $T'(X)$  that has lower entropy than  $T(X)$ . Consider

$$T'(x') = \begin{cases} T(x') & \text{if } x' \in \mathcal{X} \setminus \mathcal{X}' \\ T(x) & \text{if } x' \in \mathcal{X}' \end{cases} \quad (2)$$

where  $x$  is an arbitrary benign example in the data space. Then, for all  $x' \in \mathcal{X}'$ ,  $g(T(x)) \neq g(T(x'))$ . It follows that  $T(x) \neq T(x')$ ,  $\forall x' \in \mathcal{X}'$ . On the other hand,  $T(x) = T'(x)$  by construction.

Let the probability density of  $T(X)$  be denoted by  $p(t)$ , where  $t \in T(\mathcal{X})$ , and the probability density of  $T'(X)$  be denoted by  $q(t)$  where  $t \in T(\mathcal{X} \setminus \mathcal{X}')$ . Then,  $q(t) = p(t) + w(t)$  for  $t \in T(\mathcal{X} \setminus \mathcal{X}')$ , where  $w(t)$  corresponds to the part of benign example probability that is formed by enforcing an originally adversarial example's feature to be equal to the feature of an arbitrary benign example according to (2). Furthermore,  $\sum_{t \in T(\mathcal{X} \setminus \mathcal{X}')} w(t) = \sum_{t \in T(\mathcal{X}')} p(t)$ . We now compare the entropy of  $T(X)$  and  $T'(X)$ :

$$H(T(X)) - H(T'(X)) \quad (3)$$

$$= - \sum_{t \in T(\mathcal{X}')} p(t) \log p(t) - \sum_{t \in T(\mathcal{X} \setminus \mathcal{X}')} p(t) \log p(t) + \sum_{t \in T(\mathcal{X} \setminus \mathcal{X}')} (p(t) + w(t)) \log(p(t) + w(t)) \quad (4)$$

$$= - \underbrace{\sum_{t \in T(\mathcal{X}')} p(t) \log p(t)}_{U_1} + \underbrace{\sum_{t \in T(\mathcal{X} \setminus \mathcal{X}')} (p(t) + w(t)) \log(p(t) + w(t)) - p(t) \log p(t)}_{U_2} \quad (5)$$

It is evident that  $U_1 \geq 0$ . Note that for any  $p(t)$ , there always exists a configuration of  $w(t)$  such that  $U_2 \geq 0$ . For instance, let  $t^* = \arg \max_{t \in T(\mathcal{X} \setminus \mathcal{X}')} p(t)$ . Then, we can let  $w(t^*) = \sum_{t \in T(\mathcal{X}')} p(t)$  and  $w(t) = 0$  for  $t \neq t^*$ . With this configuration of  $w(t)$ ,

$$U_2 = (p(t^*) + w(t^*)) \log((p(t^*) + w(t^*))) - p(t^*) \log p(t^*) \quad (6)$$

Due to the fact that  $x \log x$  is a monotonically increasing function,  $U_2 \geq 0$ .

To sum up, both  $U_1$  and  $U_2$  are non-negative; as a result,

$$H(T(X)) > H(T'(X)) \quad (7)$$

Thus, we constructed a sufficient statistic  $T'(\cdot)$  that achieves lower entropy than  $T(\cdot)$ , which, in turn, indicates that  $T(X)$  is not a minimal sufficient statistic.  $\square$

## B SUPPLEMENTARY EXPERIMENTAL RESULTS

Apart from the adversarial examples, we also observed the same phenomenon for random noise that redundancy will lead to the failure of DNNs. We tested datasets with different signal-to-noise ratios (SNR), generated by adding Gaussian noise to the real pixels. The SNR is obtained by controlling the variance of the Gaussian distribution. Finally, we derived the testing accuracy on hand-crafted noisy testing data. As shown in Figure 7a and 7b, a small amount of random Gaussian noise will add complexity to examples and cause the DNNs to fail. For instance, noisy input sets with one tenth the signal strength of the benign examples result in only 34.3% test accuracy for DenseNet on CIFAR-10. This indeed indicates, and is consistent with related work, that small amounts of noise can practically fool ML models in general.

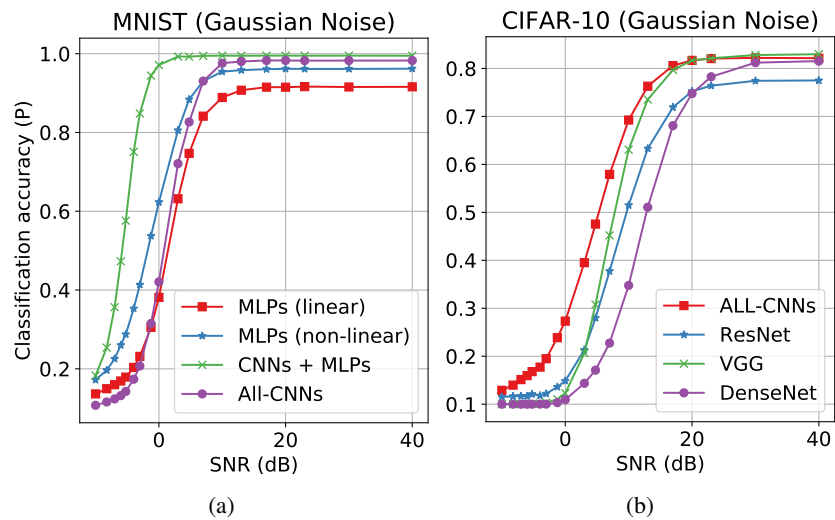


Figure 7: (a, b) Classification accuracy of benign and adversarial examples as a function of signal-to-noise ratio (SNR). (c) Classification accuracy as a function of the JPEG compression quality  $q$ . The shadow curve represents the properly scaled version of the theoretical curve in Friedland et al. (2018).