Research Article Summary of Understanding Transferable Adversarial Examples and Black-box Attacks

Sohith Dhavaleswarpu
sohith.dhavaleswarapu@tu-dortmund.de
Technische Universität Dortmund
Dortmund, Nordrhein-Westfalen, Germany

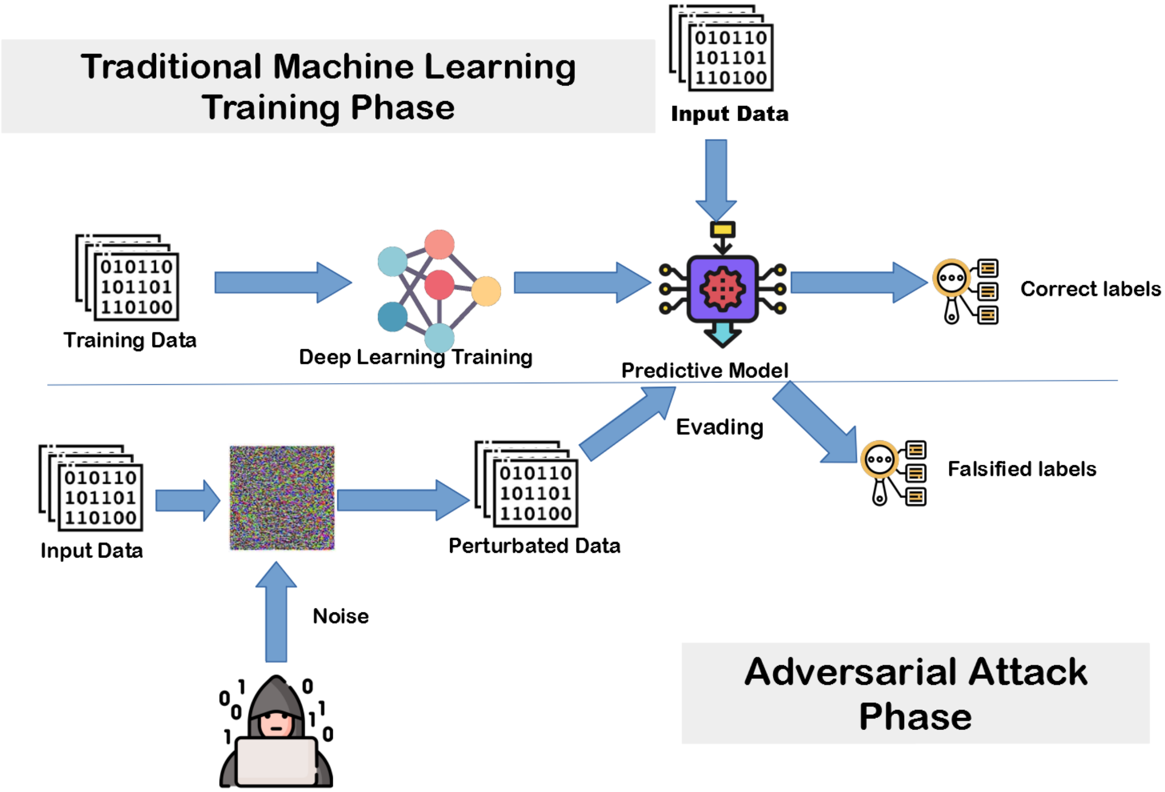


Figure 1: Adversarial attack on classification model based on neural network architecture. [2]

ABSTRACT

Deep neural network architectures have emerged as a powerful tool in the area of computer vision and image analysis. The convolutional neural network was one of the most commonly used machine learning algorithms, particularly for image classification and image processing. Many applications have been developed based on these neural network architectures. These architectures have serious flaws in their systems which results suspicious and exhibit potential dangerous behaviours. One such property is the existence of adversarial examples, which would be transferable among different architectures. Evidences from previous researchers supporting the study of transferability on small-scale datasets. This article summarizes the comprehensive explanations of the transferability over large models and a large-scale dataset inspired by the work made in the article *Delving into transferable adversarial examples and black-box attacks*[10]. This transferability property would be explained further with non-targeted and targeted adversarial examples, and provide reasoning on transferable non-targeted adversarial examples that are easy to incorporate. The targeted adversarial examples generated using existing approaches are almost difficult to transferred with their target labels. Explanation of ensemble-based approach is made to generate transferable adversarial examples and also able to transfer with their target labels. Further, these proposed

Module MD 4: Seminar summer semester 2022 TU Dortmund , June 11,2022,
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

methods are applied to the real-world examples to understand the transferable adversarial examples by performing black box attacks on Clarifai.com which is an existing black-box image classification system application.

KEYWORDS

adversarial attacks, black box attacks, neural networks, image classification

ACM Reference Format:
Sohith Dhavaleswarpu. 2022. Research Article Summary of Understanding Transferable Adversarial Examples and Black-box Attacks. In *Proceedings of summarizing the article Liu, Yanpei & Chen, Xinyun & Liu, Chang & Song, Dawn. (2016). Delving into Transferable Adversarial Examples and Black-box Attacks. (Module MD 4: Seminar summer semester 2022 TU Dortmund).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Adversarial examples are the manipulated input data by adding noise to the original input, which can later misclassify an original prediction. Based on the research conclusions about neural network architectures shows that it is really easy and feasible to generate adversarial examples, [4]. They look much more similar to the original input and, in return, they can misclassify deep architecture models. These generated adversarial examples have potential harm and lead to several dangerous consequences. For example, computer vision understanding-based applications like autonomous self-driving and key decision-making image processing applications. Mostly these adversarial instances are made using the understanding of model architectures, and it was still an open question that how can these examples can be created efficiently to find adversarial examples for a black-box model.

Some works from the earlier research have demonstrated that the adversarial examples generated for one model are used to exploit another model and can successfully able be get misclassified output. This property is described to be as transferability. Further, with the help of transferability, black box attacks can be performed with the knowledge of target architecture by developing adversarial examples from a similar neural network model and attacking a black box model [12]. These experiments and research were only conducted on small-scale datasets, which include models such as MNIST [9] and CIFAR-10 [8]. There are very few research proceedings carried in case of large-scale datasets and how the transferability can be applied over models like ImageNet [13] and adversarial examples that can transfer with their target labels.

We study the transferability of different adversarial example generation approaches which are later applied to multiple image classification models which are trained over large-scale databases. There are mainly two types of adversarial examples, which include non-targeted adversarial examples and targeted adversarial examples. The main difference between them is that non-targeted adversarial examples are generated to misclassify a model without the interest of the target label, whereas targeted adversarial examples are generated to manipulate the output of the model to be a particular target label. There are several existing approaches for adversarial examples based on a single model which can successfully able to generate non-targeted adversarial examples that are more likely to be got transferred to another model, and also a few targeted adversarial examples that can transfer with their target labels.

The newly introduced and proposed approach for the creation of transferable adversarial images by ensembling multiple models is evaluated and resulted in findings of better transferability properties compared to the other methods. This new approach performed better when a large proportion of targeted adversarial instances can be transferred with their target labels [10].

Furthermore, during the process of examining the results, some interesting findings were made related to the geometric properties of the models. The gradient directions of different models that are orthogonal to each other and also the decision boundaries of these different models align well with each other, providing the main reason for how adversarial examples can be transferred among the different architectures [10].

Finally, to validate the vulnerability of the existing real-time web application based on deep neural network architecture, a website called Classify.com, an independent commercial company which provide image classification services is exploited. Results from the experiment lead to the successful generation of both non-targeted and targeted adversarial examples from a substitute model and attacks performed without the knowledge of how the application is built or their information on trained dataset [10].

Context and organization

For the detailed explanation of methods are summarized in section 2 and followed by results in section 3 as

- Section 3.1 and Section 3.2: Explains the existing approaches are effective to generate non-targeted transferable adversarial examples only few targeted adversarial examples generated by existing methods can transfer for ImageNet models [13].
- Section 3.3 : Explains the introduced ensemble-based approaches to generate adversarial examples and results in large portion of generation of targeted adversarial examples to transfer among multiple models.
- Section 3.4: The analysis of geometric properties for large models trained over ImageNet [13], and the results explains the several interesting findings like the gradient directions of different models are orthogonal to each other.

Section 4: The attack on real time web application called Clarifai.com is explained and targeted adversarial examples generated for models trained on ImageNet [13] is attacks by setting the Clarifai.com's results label different from ImageNet model.

2 METHODS

2.1 ADVERSARIAL DEEP LEARNING AND TRANSFERABILITY

THE ADVERSARIAL DEEP LEARNING PROBLEM:. let us assume $f_\theta(x)$ is a classifier function where θ is the parameters and the classifier returns the output in the form a label which is a prediction. Lets consider a original image x given as an input to the classifier having a truth label as y. we generate a adversarial counterpart of original image as x^* which is close to original image x.

$$x \rightarrow f_\theta(x) = y \tag{1}$$

In case of non targeted adversarial example, we feed the adversarial image x^* to the classifier and expect a output $f_\theta(x^*) \neq y$.

$$x^* \rightarrow f_\theta(x^*) \neq y \tag{2}$$

In case of targeted adversarial example we feed the adversarial image x^* to classifier and expect a output $f_\theta(x^*) = y^*$ where y^* is the targeted label output and $y^* \neq y$.

$$x^* \rightarrow f_\theta(x^*) = y^* \tag{3}$$

[4].

2.1.1 APPROACHES FOR GENERATING ADVERSARIAL EXAMPLES. This section explains the three classes of approaches for generating adversarial examples namely optimization-based approaches, fast gradient approaches, and fast gradient sign approaches for non-targeted and targeted respectively.

Let us consider image x with original truth label $y = f_\theta(x)$, and non-targeted adversarial example x^* can be modeled by satisfying the following constraints:

$$f_\theta(x^*) \neq y \tag{4}$$

$$d(x, x^*) \leq B \tag{5}$$

where $d()$ is distance function measured by distance between the x and x^* and B is the distortion with the upper bound placed on distance function. Considering the loss generality, the model f is composed of a neural network J_θ(x), which outputs the probability for each label, which implies f also outputs the label with the some probability.

Optimization-based approach. let us consider 1_y be the binary encoding of the ground truth label y, and l be a loss function to measure the distance between the prediction and the original truth label, and λ is a constant to balance constraints (4) and (5).

$$\text{argmin}_x^* \lambda \cdot d(x, x^*) - l(1_y, J_\theta(x^*)) \tag{6}$$

Here, loss function l is used to approximate constraint (1), and its value can affect the effectiveness of searching for an adversarial example by minimizing the loss function to local optimum. for the current analysis $l(u; v) = log(1 - u \cdot v)$, is considered [14].

Fast gradient sign (FGS). In this method we need to create the gradients only once to generate the adversarial example. This method is proposed by the existing work [5] which follows the L_∞-norm bound to generate the adversarial example.

$$x^* \leftarrow \text{Clip}(x + \text{Bsgn}(\nabla x l(1_y, J_\theta(x)))) \tag{7}$$

where $clip(x)$ is used to clip each dimension of x to the range of pixel values, i.e., [0; 255].

Fast gradient (FG). This method follows the similar procedure as Fast gradient sign, the main difference is FGS follows in the direction of gradient sign where as FG follows the gradient direction.

$$x^* \leftarrow \text{Clip}(x + \text{B}\frac{(\nabla_x l(1_y, J_\theta(x)))}{||(\nabla_x l(1_y, J_\theta(x)))||})) \tag{8}$$

A targeted adversarial image x^* follows the similar generation approaches as non targeted adversarial image with change in target label from y to y^*.

$$f_\theta(x^*) = y^* \tag{9}$$

[5].

ENSEMBLE-BASED APPROACHES. This method is the newly proposed approach [10] to hypothesize adversarial examples remain adversarial for multiple models so that it can be transferable to multiple models. Let us consider k be the multiple white box models with activation function outputs $J_1, J_2, ...J_k$ repressively. The ensemble approach solves the optimization problem as:

$$\text{argmin}_x^* - log((\sum_{i=1}^{k} \alpha_i J_i(x^*)).1_y) + \lambda d(x, x^*) \tag{10}$$

where $\sum_{i=1}^{k} \alpha_i J_i(x^*)$ be the ensemble model, α_i are the ensemble weights.

Let us consider we have 5 models and this approach is carried as, 4 models are ensemble as white box example and generate the targeted adversarial example and attack the 5th model in black box approach.

2.2 EVALUATION METHODOLOGY

2.2.1 NON-TARGETED ADVERSARIAL EXAMPLES. For a given two models transferability of non targeted examples is measured as percentage of the adversarial examples generated for one model that can be classified correctly for the other. This percentage is considered to be as accuracy indicating lower the accuracy means better the non targeted transferability.

2.2.2 TARGETED ADVERSARIAL EXAMPLES. whereas in case of targeted transferability, Percentage of the adversarial examples generated for one model that are classified as the target label by the other model. We refer to this percentage as matching rate indicating A higher matching rate means better targeted transferability.

Distortion. To calculate the dissimilarities between original and adversarial examples by a measure of distortion. It was calculated as root mean square deviation, (RMSD) as:

$$d(x, x^*) = \sqrt{\sum_i \frac{(x_i^* - x_i)^2}{N}} \tag{11}$$

where x^* and x are the vectors representations of an adversarial image and the original one respectively, N is the dimensionality and X_i is the pixel value ranging from 0 to 255 similarly for x_i^* we measure the

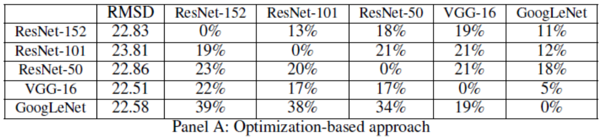
3 RESULTS

The main focus is to examine the properties of transferability of adversarial examples for the both targeted and non targeted adversarial examples. The models considered are trained on ImageNet [13] architecture to perform the transferability. the models are as follows: ResNet-50, ResNet-101, ResNet-152 [6], GoogLeNet [16] and VGG-16 [15]. The transferability is examined by considered the LSVRC 2012 validation set [13], out of 1000 images, 100 images are segregated as test data. The target label is selected manually in case of targeted adversarial attack.

3.1 NON-TARGETED ADVERSARIAL EXAMPLES

In this section, we examine the generation of non targeted adversarial examples using stated approaches in section 2 and discuss about interpretation of results observed.

OPTIMIZATION-BASED APPROACH. For a single model, the Adam optimizer [7] is used to generate adversarial example x^* and original image x to optimize the objective function of optimization based approach (6). For the individual model the learning rate is set to small RMSD which is less than 2 and defined λ. The Adam optimizer finds the adversarial examples with small distortions and can successfully manipulable the target model. Further observations are also recorded by running the optimizer for larger distortions by setting the learning rate as 4 for 100 iterators to generate the adversarial examples.



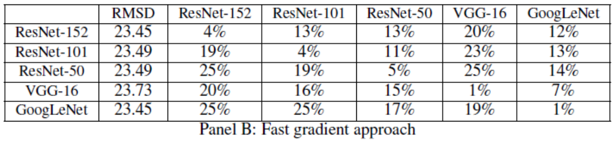
	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	22.83	0%	13%	18%	19%	11%
ResNet-101	23.81	19%	0%	21%	21%	12%
ResNet-50	22.86	23%	20%	0%	21%	18%
VGG-16	22.51	22%	17%	17%	0%	5%
GoogLeNet	22.58	39%	38%	34%	19%	0%

Panel A: Optimization-based approach

Figure 2: Transferability of non-targeted adversarial images generated between pairs of models. The first column indicates the average RMSD of all adversarial images generated for the model in the corresponding row. The cell (i; j) indicates the accuracy of the adversarial images generated for model i (row) evaluated over model j (column) [10].

From the figure 2 panel A refers to the results obtained by adversarial examples generated using optimization based approach on one network and evaluated on another network. The diagonal values from the results indicates that, all adversarial images generated for one model can mislead the same model. The rest of the values interpreted as the large proportion of non-targeted adversarial images generated for one model using the optimization-based approach can transfer to another. we also interpret that the three ResNet models having similar architectures which differ only in the hyper parameters.The adversarial examples generated against a ResNet model do not particularly transfer to another ResNet model better than other non-ResNet models. For instance, the adversarial examples generated for VGG-16 have lower accuracy on ResNet-50 than those generated for ResNet-152 or ResNet-101.

FAST GRADIENT-BASED APPROACHES. For Fast gradient based approach we generate the adversarial examples as discussed in the equation (7). The generated examples lie in the 1-D subspace and it is easily able to approximate the minimal distortion in this subspace of transferable adversarial examples between two models. We know that the hyper parameter distortion B and the RMSD of the generated adversarial images are highly correlated and obtain the distortion B to generate adversarial images with a given RMSD.



	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.45	4%	13%	13%	20%	12%
ResNet-101	23.49	19%	4%	11%	23%	13%
ResNet-50	23.49	25%	19%	5%	25%	14%
VGG-16	23.73	20%	16%	15%	1%	7%
GoogLeNet	23.45	25%	25%	17%	19%	1%

Panel B: Fast gradient approach

Figure 3: Transferability of non-targeted adversarial images generated between pairs of models. The first column indicates the average RMSD of all adversarial images generated for the model in the corresponding row. The cell (i; j) indicates the accuracy of the adversarial images generated for model i (row) evaluated over model j (column) [10].

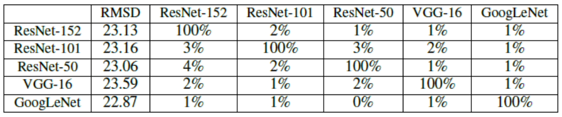
From the figure 3 panel B refers to the results obtained by adversarial examples generated using fast gradient based approach on one network and evaluated on another network. we observe the average RMSD from the first column is almost similar to the generation using optimization approach. The diagonal values from the results are all positve indicates that all adversarial images generated for one model cannot fully mislead the output. the values of non-diagonal cells are accuracy values of adversarial images generated for one model but evaluated on another are having good accuracy and in comparable to the counterparts of optimization-based approach have a similar or less accuracy. The results shows that non-targeted adversarial examples generated by FG exhibit transferability as well.

Similar experiments are also considered by using Fast gradient sign app approach but the transferability results observed are not performed well when compared to Optimization based and fast gradient based approach.

3.2 TARGETED ADVERSARIAL EXAMPLES

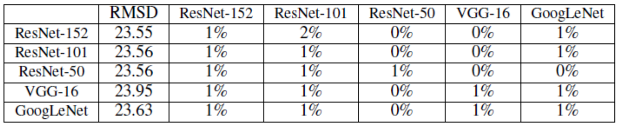
In this section, we examine the generation of targeted adversarial examples using stated approaches in section 2 and discuss about interpretation of results observed.

OPTIMIZATION-BASED APPROACH. From the figure 4 gives the results of targeted adversarial generated using optimization-based approach and interprets the results of transferability of targeted adversarial images. we can observe that the diagonal values of the table stated that, the prediction of targeted adversarial images can match the target labels when evaluated on the same model. In contrast, if we observe the non diagonal elements, the targeted adversarial images can be rarely predicted as the target labels by a different model. We can conclude from the observations that the target labels do not transfer among the other models. In case we increase the distortion as well the results still do not see any improvements on making target label transfer.



	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.13	100%	2%	1%	1%	1%
ResNet-101	23.16	3%	100%	3%	2%	1%
ResNet-50	23.06	4%	2%	100%	1%	1%
VGG-16	23.59	2%	1%	2%	100%	1%
GoogLeNet	22.87	1%	1%	0%	1%	100%

Figure 4: The matching rate of targeted adversarial images generated using the optimization-based approach. First column indicates the average RMSD of the generated adversarial images. Cell (i; j) indicates that matching rate of the targeted adversarial images generated for model i (row) when evaluated on model j (column) [10].



	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.55	1%	2%	0%	0%	1%
ResNet-101	23.56	1%	1%	0%	0%	1%
ResNet-50	23.56	1%	1%	1%	0%	0%
VGG-16	23.95	1%	1%	0%	1%	1%
GoogLeNet	23.63	1%	1%	0%	1%	1%

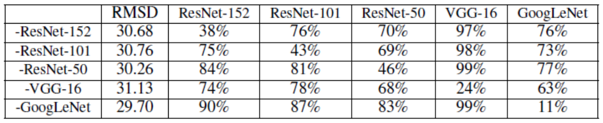
Figure 5: The adversarial images are generated using the targeted FG. The first column indicates the average RMSD of all adversarial images generated for the model in the corresponding row. The first column indicates the average RMSD of the generated adversarial images. Cell (i; j) indicates that top-1 matching rate of the targeted adversarial images generated for model i (row) when evaluated on model j (column). Higher value indicates more successful transferable target labels.

FAST GRADIENT-BASED APPROACHES. From the figure 5 We tend to examine the targeted adversarial images generated by fast gradient-based approaches, and we observe that the matching rates of across the combination pair of models are very low which indicated that the target labels do not transfer as well. results tend to close the conclusion as most targeted adversarial images cannot mislead the model for which the adversarial images are generated and to predict the target labels.

From the method of calculations of fast gradient it is to the fact that the fast gradient-based approaches only search for attacks in a 1-D subspace. In this particular subspace, the total possible predictions may contain a small subset of all labels, which usually does not contain the target label. We see the matching rate of any of the 5 models is 0% which lead to the conclusion of the attacker cannot generate successful targeted adversarial examples and also targeted transferability.

3.3 ENSEMBLE-BASED APPROACHES

The proposed ensemble-based model is discussed in section 2.1.1 of methods. The targeted adversarial examples are generated using this novel approach idea and by the adam optimizer. Equal ensemble weights are considered across all five models used for earlier explanations in the ensemble. For this analysis, the learning rate is set to 8 for each model and in each iteration, adam update is applied for each model and later aggregated into one input image. This resulted in the generation of targeted adversarial images whose target labels can transfer among the other models.

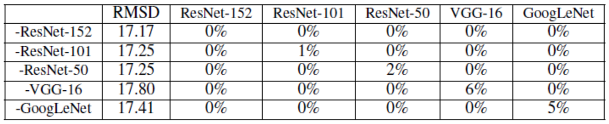


	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	30.68	38%	76%	70%	97%	76%
-ResNet-101	30.76	75%	43%	69%	98%	73%
-ResNet-50	30.26	84%	81%	46%	99%	77%
-VGG-16	31.13	74%	78%	68%	24%	63%
-GoogLeNet	29.70	90%	87%	83%	99%	11%

Figure 6: The matching rate of targeted adversarial images generated using the optimization-based approach. The first column indicates the average RMSD of the generated adversarial images. Cell (i; j) indicates that percentage of the targeted adversarial images generated for the ensemble of the four models except model i (row) is predicted as the target label by model j (column). In each row, the minus sign "-" indicates that the model of the row is not used when generating th attacks [10].

From the figure 6 we observe the results of the matching rate of targeted adversarial images generated by ensembling and by using the optimization-based approach. From the diagonal values, we can interpret that the transferability to ResNet models is better than to VGG-16 or GoogLeNet when adversarial examples are generated against all models except the target model. For the non-diagonal values not all targeted adversarial images can be misclassified to the target labels by the models used in the ensemble which indicates looking for an adversarial example for the ensemble model, there is no direct supervision to mislead any individual model in the ensemble to predict the target label.

A similar approach is also applied for non-targeted adversarial examples from the results in figure 7 clearly states that the generated adversarial images have almost perfect transferability by looking at their matching rate.

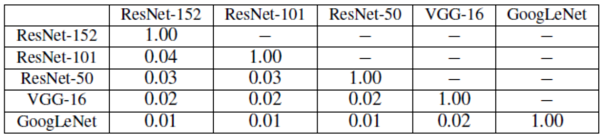


	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	17.17	0%	0%	0%	0%	0%
-ResNet-101	17.25	0%	1%	0%	0%	0%
-ResNet-50	17.25	0%	0%	2%	0%	0%
-VGG-16	17.80	0%	0%	0%	6%	0%
-GoogLeNet	17.41	0%	0%	0%	0%	5%

Figure 7: Accuracy of non-targeted adversarial images generated using the optimization-based approach. The first column indicates the average RMSD of the generated adversarial images. Cell (i; j) corresponds to the accuracy of the attack generated using four models except model i (row) when evaluated over model j (column). In each row, the minus sign "-" indicates that the model of the row is not used when generating the attacks [10].

3.4 GEOMETRIC PROPERTIES OF DIFFERENT MODELS

The study was conducted during the analysis to check whether the adversarial directions of different models align with each other by calculating the cosine value of the angle between gradient directions of different models. The results from figure 8 concluded to the gradient directions of different models are almost orthogonal to each other by observing that all non-diagonal values are close to 0 for most images. For detailed results, can be referred to in 9.



	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	1.00	–	–	–	–
ResNet-101	0.04	1.00	–	–	–
ResNet-50	0.03	0.03	1.00	–	–
VGG-16	0.02	0.02	0.02	1.00	–
GoogLeNet	0.01	0.01	0.01	0.02	1.00

Figure 8: Average cosine value of the angle between gradient directions of two models. Notice that the dot-product of two normalized vectors is the cosine value of the angle between them, for each image, we compute the dot-product of normalized gradient directions with respect to model i (row) and model j (column), and the value in cell (i; j) is the average over dot-product values of all images. Notice that this table is symmetric [10].

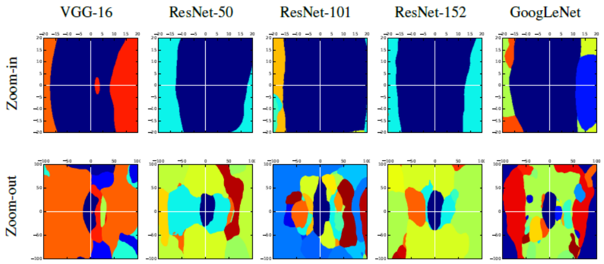


Figure 9: Decision regions of different models. We pick the same two directions for all plots: one is the gradient direction of VGG-16 (x-axis), and the other is a random orthogonal direction (y-axis). Each point in the span plane shows the predicted label of the image generated by adding a noise to the original image (e.g., the origin corresponds to the predicted label of the original image) [10].

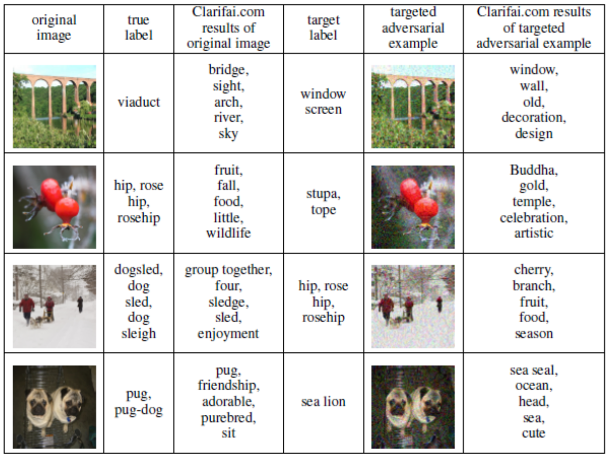
4 BLACK BOX ATTACKS

A black box attack is a penetration testing methodology where the attacker has no idea about the model architecture and dataset which is used to train the model. For this analysis to test the proposed models, a real-world application called Clarifai.com which is a commercial company providing state-of-the-art image classification services performed a black box attack. In this procedure, 100 original images are passed to Clarifai.com and the returned labels are correct based on a subjective measure. The results indicate that labels returned from Clarifai.com are also different from categories in ILSVRC 2012.

In the current analysis, 400 adversarial images are passed on to the network out of which 200 of them are targeted adversarial examples, and the remaining 200 are non-targeted ones. As for the 200 targeted adversarial images, 100 of them are generated using the optimization-based approach based on VGG-16 and the rest 100 are generated using the optimization-based approach based on an ensemble of all models except ResNet-152. Similar stratification is made for non-targeted adversarial examples as well.

From the results of figure 10, for non-targeted adversarial examples, the results exhibited that for those generated using VGG-16 and those generated using the ensemble, most of them can transfer to Clarifai.com. For a large proportion of considered targeted adversarial examples are misclassified by Clarifai.com. Out of which 57% of the targeted adversarial examples generated using VGG-16 and 76% of the ones generated using the ensemble approach are classified as the incorrect label which is different from the ground truth.

Similarly, In the case of targeted adversarial examples, 18% of those generated using the ensemble model and 2% of VGG-16 can predict the labels close to the expected target label.



original image	true label	Clarifai.com results of original image	target label	targeted adversarial example	Clarifai.com results of targeted adversarial example
	viaduct	bridge, sight, arch, river, sky	window screen		window, wall, old, decoration, design
	hip, rose hip, rosehip	fruit, fall, food, little, wildlife	stupa, tope		Buddha, gold, temple, celebration, artistic
	dogsled, dog sled, dog sleigh	group together, four, sledge, sled, enjoyment	hip, rose hip, rosehip		cherry, branch, fruit, food, season
	pug, pug-dog	pug, friendship, adorable, purebred, sit	sea lion		sea seal, ocean, head, sea, cute

Figure 10: Original images and adversarial images evaluated over Clarifai.com. For labels returned from Clarifai.com, we sort the labels firstly by rareness: how many times a label appears in the Clarifai.com results for all adversarial images and original images, and secondly by confidence. Only top 5 labels are provided [10].

5 CONCLUSION

The transferability of both non-targeted and targeted adversarial instances was produced using various methodologies over big models and a large size dataset. The results demonstrate that even for large models and a sizable dataset, the transferability for non-targeted adversarial cases is possible. On the other hand, it is challenging to generate targeted adversarial cases with transferable target labels using current techniques. The new ensemble-based technique shows how effectively it can transferable targeted adversarial examples. In contrast to existing models, the novel techniques perform better at producing non-targeted transferable adversarial examples. Clarifai.com, a black-box image classification system, can be successfully attacked using both non-targeted and targeted adversarial examples produced using novel methodologies. In addition, some geometrical characteristics to supports the transferable adversarial instances.

6 DISCUSSION

Neural networks are emerged as key machine leaning algorithm in the area of image classification models. These neural network architecture poses a serious security flaw in the form of adversarial examples. The adversarial examples are the maniputed input by adding an pertubation or noise to original images. These manipulated images are feed into classification model to misclassify the output. These adversarial examples could hinder the deep neural network architecture and results in suspicious and potential harmful behaviors.

The traditional approach to generating these adversarial examples involves an optimization-based approach and fast gradient approaches. Using these approaches one can create adversarial examples based on the target label to misclassify the output. The adversarial examples are described as "non-targeted adversarial examples" which can misclassify the network regardless of misclassified label it predicts and "targeted adversarial examples" which can misclassify a network with a specific target label.

The adversarial examples created for one model can also be fed to another model and could successfully misclassify the prediction. This property is described as "transferability". In the current discussion, we mainly focus on the transferability property of adversarial examples among both the targeted and non-targeted manners.

From the results of experiments, one could successfully transfer the non-targeted adversarial examples when adversarial examples were generated using existing approaches. In contrast, targeted adversarial examples could not able to transfer among other models when adversarial examples are generated by existing approaches. A new approach called the ensemble-based approach is proposed by the combination of multiple models. Later using the ensemble-based approach one could generate targeted adversarial examples to attack another model.

Furthermore, we discuss the results on how these adversarial examples were generated upon a white box model and later transferred to attack over image classification model called Clarifai.com which is a black box system.

6.1 Core Concept(s)

The researchers are explaining the core concepts about the serious flaw in the neural networks in the form of adversarial examples [4]. Additionally, the main focus was on the transferability of targeted and non-targeted adversarial examples. This was the first work to conduct a transferability over large models and large datasets [10]. They have a newly proposed method called the ensemble-based approach in the generation of adversarial examples which can transfer among the other networks and misclassify the prediction with the target label. During experiments, the geometric properties of different models were also analyzed. they have evidence to support that the gradient directions of different models in the evaluation are almost orthogonal to each other [10].

6.2 Scope of Research

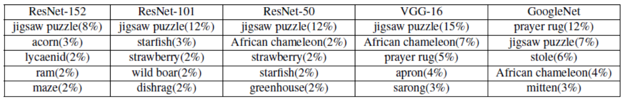
The results observed from the article [10] support the evidence, transferability of targeted adversarial examples and can successfully attack the state of art classification models with their black box access. In the current work, there were only five models whose model architecture and dataset they trained are known to generate the adversarial examples. Later they tend to attack a black box model which is a substitute for used white box models. This scope can be further applied by considering some different combinations of model architectures and generating adversarial examples. Further, implement to attack a similar architecture and observe the results.

From the results of the black box attack on Clarifai.com, the adversarial examples generated by the ensemble-based approach, out of all the predictions only 18% of the targeted adversarial examples can misclassify and predict close to the target label. The results indicate a very low percentage of the adversarial examples generated performing with a good misclassification rate and most of the time they cannot misclassify to the target label. This scope can furthermore be studied about the reasons for not achieving a higher percentage and required an in-depth analysis of how the ensemble-based approach can be more exploitable to the vast range of different neural network architectures.

In the current work, the adversarial examples were generated using existing approaches which include optimization-based and fast gradient approaches. This scope could be replaced by considering the alternative method like Projected Gradient Descent (PGD) [11] and Carlini and Wagner (C&W) attack [1]. These methods would generate a much more robust adversarial examples and could explore the transferability of these adversarial instances over the black box accessible models.

6.3 Implications of Findings

In the current analysis, the researchers could successfully achieve the transferability of non-targeted and targeted adversarial examples based on the evidence upon the geometric properties of different models. There are supporting works which also discuss the geometric models for the Robustness of classifiers from adversarial to random noise [3] and [5]. The main difference between them is large models trained over a large dataset with 1000 labels are additionally discovered whose geometric properties are never examined before. This made new observations to better understand the models and their adversarial examples.



ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogleNet
jigsaw puzzle(8%)	jigsaw puzzle(12%)	jigsaw puzzle(12%)	jigsaw puzzle(15%)	prayer rug(12%)
acorn(3%)	starfish(3%)	African chameleon(2%)	African chameleon(7%)	jigsaw puzzle(7%)
lycaenid(2%)	strawberry(2%)	strawberry(2%)	prayer rug(5%)	stole(6%)
ram(2%)	wild boar(2%)	starfish(2%)	apron(4%)	African chameleon(4%)
maze(2%)	dishrag(2%)	greenhouse(2%)	sarong(3%)	mitten(3%)

Figure 11: When using non-targeted optimization-based approach for VGG-16 model to generate adversarial images, column i indicates the top 5 common incorrect labels predicted by model i. The value in the parentheses is the percentage of the predicted label..

The results from figure 11 we observe that not all non targeted adversarial examples could be transferable which are generated by optimization based approach. There are some instances providing evidences for transferable non-targeted adversarial images are classified as the same wrong labels. The model used for generating adversarial examples is VGG-16 which in return consists of 999 categories. This indicates increase in number of categorical labels in the model leads to wrong predictions of target labels with same name and hence more vulnerable to adversarial examples.

The current work made on some intersecting findings by proposing concepts of constructing a substitute model to attack a black-box aces-sable target model. To train the substitute model, a technique was inspired which can synthesizes a training set and label them by searching the target model for labels [12]. Using this approach, one can perform black-box attacks on machine learning model which are hosted by by Amazon, Google, and portals.

6.4 Limitations

The article provides the evidence of supporting the main hypothesis to explore the possibilities of vulnerability in deep neural network architecture. The exploiting of these neural network architecture is performed based on the vulnerability of the models and tend to misclassify the prediction. The researcher intention is generating a adversarial examples and transferring of these instance across the models. In the final conclusions they are successful in generating some concrete evidences that proves models have a serious flaw in their architectures. No further discussion was made about how can these vulnerabilities could be avoided by adversarial examples. The method of generating these adversarial examples could prove a scope of applying patches to vulnerable models and can also be avoid further mode of attacks on these models.

some research techniques like Adversarial training with perturbation or noise shows that models can withstand for adversarial attacks. Adversarial training with perturbation or noise helps the model training with adv adversarial examples generated using the existing approach in the training phase of model and reduce the classification errors [18].

Ensemble Adversarial Training Attacks and Defenses is the one of the contracting methodology where the proposed novel approach in the article. This method mainly focus on adversarial training converges to a degenerate global minimum, wherein small curvature artifacts near the data points can unclear a linear approximation of the loss. Also, This technique protect augments of training data with perturbations transferred from other models [17].

6.5 Summary

Adversarial examples are manipulated inputs generated to fool machine learning models. Current existing machine learning algorithms especially neural network architectures can severely effected and the prediction of these models are misclassified. The current article explains the possible vulnerability of these models by generating the adversarial examples and transferring the generated adversarial examples across the models.

From the analysis, The non-targeted adversarial examples generated using the optimization approach and fast gradient approach can successfully be transferred to another model and misclassify their prediction. In the case of targeted adversarial examples generated using optimization and fast gradient, approaches are almost impossible to transfer among the models and misclassify them. Using the Ensemble-based approach up to some extent we can able to misclassify the prediction with the target label. For instance, for the VGG model up to 18% of its predictions are got misclassified as per mentioned target label.

These vulnerabilities in the models can posses serious consequences when deployed in crucial sectors like Medical, defense, and social aspects. The explained adversarial attacks need to be tackled with proper security updated for existing models and train the new model with proposed methods like Adversarial training with perturbation or noise or Ensemble Adversarial Training Attacks. Furthermore, research is to be conducted in the field of neural networks to identify possible vulnerabilities like adversarial examples and methods to not get exploited in near future.

ACKNOWLEDGEMENTS

This work is a part of MD:4 module of the course Uncertainty quantification Seminar. I would like to thank Prof. Dr. Emmanuel Müller, Chair of Data Science and Data Engineering, Department of Computer Science, Mr. Simon Klüttermann and Ms. Carina Newen for organizing the course. Special regards to Mr. Benedikt Böing for supervising and mentoring the course work.

REFERENCES

[1] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP)* (2017), 39–57.

[2] Ferhat Ozgur Catak. 2020. Adversarial Machine Learning Mitigation: Adversarial Learning. Retrieved July 23, 2021 from https://towardsdatascience.com/adversarial-machine-learning-mitigation-adversarial-learning-9ae04133c137

[3] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2016. Robustness of classifiers: from adversarial to random noise. In *NIPS*.

[4] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *arXiv 1412.6572* (12 2014).

[5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *arXiv 1412.6572* (12 2014).

[6] Asifullah Khan and Noorul Wahab. 2016. Deep Residual Learning.

[7] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2015).

[8] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[10] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Xiaodong Song. 2017. Delving into Transferable Adversarial Examples and Black-box Attacks. *ArXiv* abs/1611.02770 (2017).

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJzIBfZAb

[12] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. (05 2016).

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[14] Sean Saito and Sujoy Roy. 2018. Effects of Loss Functions And Target Representations on Adversarial Robustness.

[15] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556

[16] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. (12 2013).

[17] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. Ensemble Adversarial Training: Attacks and Defenses. https://doi.org/10.48550/ARXIV.1705.07204

[18] Pulei Xiong, Scott Buffett, Shahrear Iqbal, Philippe Lamontagne, Mohammad Mamun, and Heather Molyneaux. 2022. Towards a Robust and Trustworthy Machine Learning System Development: An Engineering Perspective. *J. Inf. Secur. Appl.* 65, C (mar 2022), 20 pages. https://doi.org/10.1016/j.jisa.2022.103121