# On the Robustness of Vision Transformers to Adversarial Examples

Kaleel Mahmood
Department of Computer Science and Engineering
University of Connecticut, CT, 06269, USA kaleel.mahmood@uconn.edu

Rigel Mahmood
Department of Computer Science and Engineering
University of Connecticut, CT, 06269, USA

Marten van Dijk
CWI, Amsterdam
The Netherlands

## Abstract

*Recent advances in attention-based networks have shown that Vision Transformers can achieve state-of-the-art or near state-of-the-art results on many image classification tasks. This puts transformers in the unique position of being a promising alternative to traditional convolutional neural networks (CNNs). While CNNs have been carefully studied with respect to adversarial attacks, the same cannot be said of Vision Transformers. In this paper, we study the robustness of Vision Transformers to adversarial examples. Our analyses of transformer security is divided into three parts. First, we test the transformer under standard white-box and black-box attacks. Second, we study the transferability of adversarial examples between CNNs and transformers. We show that adversarial examples do not readily transfer between CNNs and transformers. Based on this finding, we analyze the security of a simple ensemble defense of CNNs and transformers. By creating a new attack, the self-attention blended gradient attack, we show that such an ensemble is not secure under a white-box adversary. However, under a black-box adversary, we show that an ensemble can achieve unprecedented robustness without sacrificing clean accuracy. Our analysis for this work is done using six types of white-box attacks and two types of black-box attacks. Our study encompasses multiple Vision Transformers, Big Transfer Models and CNN architectures trained on CIFAR-10, CIFAR-100 and ImageNet.*

## 1. Introduction

For vision tasks, convolutional neural networks (CNNs) [20] are the de facto architecture [37, 19]. On the other hand, in natural language processing (NLP), attention-based transformers are one of the most commonly used models [35]. Based on the success of transformers in NLP, various works have attempted to apply self-attention (both with and without CNNs) to image processing tasks [4, 36]. In particular, in [12], the training of self-attention transformers is achieved by processing the image in patches. The training in [12] is unique in that the transformer is first trained on the dataset ImageNet-21K (or JFT) before training on a smaller dataset, to achieve near state-of-the-art results on ImageNet, CIFAR-10 and CIFAR-100. These types of transformers are referred to as Vision Transformers (ViT) [12]. It is important to note that the same kind of training regime can be applied to CNNs. In [19], they also propose training on a large dataset (ImageNet-21K or JFT) and fine tuning on a smaller dataset. Using this approach, CNNs are also able to achieve state-of-the-art results on ImageNet, CIFAR-10 and CIFAR-100. CNNs trained in this manner are referred to as Big Transfer Models (BiT-M) [19].

While CNNs are popular for vision tasks, they are not without deficiencies. It has been widely documented that CNNs are vulnerable to adversarial examples [33, 14]. Adversarial examples are benign input images to which small perturbations are added. This perturbation causes the CNN to misclassify the image with high confidence. Broadly speaking, an attacker creates an adversarial example using one of two threat models. Under a white-box adversary [5], the attacker has access to the CNN's parameters (architecture and trained weights). The adversary can directly obtain gradient information from the model to create an adversarial example. The other type of threat is a black-box adversary. In this scenario, the attacker does not know the CNN's parameters or architecture but can repeatedly query the CNN, or build their own synthetic CNN to estimate gradient information and generate adversarial examples.

It has also been shown that adversarial examples generated using CNNs exhibit transferability [28, 21, 29]. Here, transferability refers to the fact that adversarial examples crafted to fool one CNN are often misclassified by other CNNs as well. Overall, CNNs have an expansive body of

literature related to adversarial attacks [6, 11, 10] and defenses [23, 5, 34]. In contrast, Vision Transformers have not been closely studied in the adversarial context. In this work, we investigate how the advent of Vision Transformers advance the field of adversarial machine learning. Here we specifically focus on image based adversarial attacks. Our paper is organized as follows: In Section 2 we first discuss some related NLP work. We then break our analysis of Vision Transformers into several related questions:

*Do Vision Transformers provide any improvement in security over CNNs under a white-box adversary?* We explore this question in Section 3 by attacking Vision Transformers, Big Transfer Models and conventional CNNs (ResNets) with six standard white-box adversarial machine learning attacks. We show that under a white-box adversary, Vision Transformers are just as vulnerable (insecure) as other models. In Section 4, we further delve into white-box attacks and ask: *How transferable are adversarial examples between Vision Transformers and other models?* We perform a transferability study with eight CIFAR-10 and CIFAR-100 models (this includes four Vision Transformers, two Big Transfer Models, and two ResNets). We also study the transferability of ImageNet Vision Transformers using seven models (three Vision Transformers, two Big Transfer Models, and two ResNets). From our experiments we observe an interesting phenomenon. The transferability between Vision Transformers and other non-transformer models is unexpectedly low.

*How can the transferability phenomena be leveraged to provide security?* This is the topic of our final question in Sections 5 and 6. We further break this question down into white-box and black-box analyses. First, we consider a white-box adversary. We develop a new white-box attack called the Self-Attention blended Gradient Attack (SAGA). Using SAGA, we show it is not possible to leverage the transferability phenomena to achieve white-box security. However, achieving black-box security is still possible. To demonstrate this, we consider a black-box attacker that can leverage transfer style [29] and query-based attacks [8]. We show under this threat model, a simple ensemble of Vision Transformers and Big Transfer Models can achieve an unprecedent level of robustness, without sacrificing clean accuracy. Finally, in Section 7, we offer concluding remarks.

## 2. Related Work

The transformer has been well studied from an adversarial perspective for NLP applications e.g., [18, 31, 17, 15]. The work in [18] analyzes two popular self-attentive architectures: (a) Transformer for neural machine translation, and (b) BERT for sentiment and entailment classification, and proposes algorithms to generate more natural adversarial examples that preserve the semantics. Theoretical explanations are also provided in [18] to support the claim that self-attentive structures are more robust to small adversarial perturbations in NLP as compared to LSTM based architectures. The work in [31] analyzes the complex relationship between self-attention layers including cross-non-linearity and cross-position, and develops a robustness verification algorithm for Transformers. The authors do not use large-scale pre-trained models such as BERT because they are too challenging to be tightly verified with their approach. The work in [17] studies large pre-trained Transformer models in NLP such as BERT. One of the conjectures drawn by the authors of [17] is that since Transformer models are pre-trained with large amounts of data (e.g., BERT is trained on 3 billion tokens), this may aid robustness. It is also mentioned that perhaps the self-supervised training may also contribute to this robustness. The work in [15] proposes a self-attention attribution method to interpret the information interactions inside a transformer. The authors use BERT as an example to conduct experiments to identify the important attention heads, and extract the most salient dependencies in each layer to construct an attribution tree. This information is used to extract adversarial patterns to implement non-targeted attacks towards BERT.

Thus, as stated above a good body of work has been devoted to the adversarial exploration of the Transformer for NLP applications. To our best knowledge, we are the first to provide an in-depth analysis of the adversarial properties of a Transformer from a vision perspective.

## 3. White-Box Attacks on Vision Transformers

*Do Vision Transformers provide any improvement in security over CNNs under a white-box adversary?* We experimentally analyze Vision Transformers to answer this question. It may seem unorthodox to start with experiments. However, the most expedient way to directly determine the security of the transformer is through attacks and analyses of those attacks. We start with a white-box adversary because it represents the strongest possible adversary.

### 3.1. Adversarial Model, Considered Classifiers and White-Box Attack Selection

**Adversarial Model:** In this section, our adversary has knowledge of the model architecture and trained parameters of the model. We assume the adversary can perturb the original input $x$ to create $x_{adv}$ within a certain amount $\epsilon$ according to $\|x - x_{adv}\|_\infty \leq \epsilon$. For CIFAR-10 and CIFAR-100, the $\epsilon = 0.031$ and for ImageNet $\epsilon = 0.062$, where $x$ is an $n \times m$ color image such that $x \in [0, 1]^{n \times m \times 3}$. The adversary succeeds if they are able to create an input $x_{adv}$ within this bound $\epsilon$ that is misclassified by the classifier (untargeted attack). When we measure security, we do so by taking a set of clean test examples that are correctly identified by the classifier. Using this set of clean examples we generate adversarial examples using one of the six attacks.

We then measure what percent of examples the classifier still correctly identifies. As Vision Transformers are relatively new, we experiment with a wide range of attacks and models. Below, we list the attacks and models we use. We also give our justification for including them in this paper.

**White-Box Attacks:** We run six different types of white-box attacks on our models. We begin with one of the most basic, the Fast Gradient Sign Method (FGSM) [13] as an initial test of robustness. We further build upon this by testing stronger multi-step attacks, the Momentum Iterative Method (MIM) [11], and Projected Gradient Descent (PGD) [24]. We also test the newest iterative attack which uses a variable step size in each iteration, Auto Projected Gradient Descent (APGD) [10]. Aside from the previously mentioned attacks, there are two other possible attack directions. To craft an extremely small, almost imperceptible adversarial noise, the Carlini and Wagner (C & W) attack is often of interest [6].

Lastly, it is possible for some white-box attacks to fail if gradient masking or an obfuscation of the gradient occurs [2]. It is important to note this does not actually mean the classifier is secure, it merely means the gradient for the classifier was not estimated properly. There are attacks designed to overcome gradient masking, such as the Backward Pass Differentiable Approximation (BPDA) [2]. We use BPDA here to ensure gradient masking is not occurring in the self-attention layers, or any other part of the Vision Transformer. Due to the limited space, we cannot give detailed descriptions of each white-box attack here. We urge interested readers to examine the supplemental material where we provide descriptions of each attack.

**Classifier Models:** When considering Vision Transformers, there are several different types of model variants. To begin, the patch size of the transformer needs to be chosen. To test different patch sizes, in our study we include both patch size 32 (ViT-B-32) and patch size 16 (ViT-B-16). The B in the model refers to the model complexity [12]. B models contain 12 layers and L models contain 24 layers. Since model complexity is another factor that can affect security [24], we also test across model complexity (ViT-B-16 and ViT-L-16). It is also possible to use the self-attention layers first and then use a conventional CNN (ResNet) on top. This configuration is denoted as ViT-R50. Experimenting across patch size, model complexity and with the hybrid configuration gives us four Vision Transformer models.

For the Big Transfer Models [19], we vary across model complexity (BiT-M-R50 and BiT-M-R101x3). We do the same for conventional ResNets (ResNet-56 and ResNet-164 [16]). Overall for CIFAR-10 and CIFAR-100, this gives us a total of 8 models to attack: ViT-B-32, ViT-B-16, ViT-L-16, ViT-R50, BiT-M-R50, BiT-M-R101x3, ResNet-56 and ResNet-164. For ImageNet, we run a slight variation of the above set, attacking 7 models: ViT-B-16, ViT-L-16 (image size 224), ViT-L-16 (image size 512), BiT-M-R50, BiT-M-R152x4, ResNet-50 and ResNet-152. For ImageNet, we mainly focus on more complex models (e.g., testing two types of ViT-L-16 instead of ViT-B-32). We do this because the more complex Vision Transformers are better indicative of state-of-the-art performance on ImageNet. We provide full descriptions of the architectures and training parameters for our models in the supplemental material.

### 3.2. White-Box Attack Analysis

We report the results of our six white-box attacks for CIFAR-10 and ImageNet in Table 3.1. The robust accuracy (percent of samples correctly identified by the classifier) is reported in Table 3.1 using $1000$ examples for each attack. For this set of attacks, CIFAR-10 and CIFAR-100 follow extremely similar trends. As a result, for brevity, we provide our CIFAR-100 white-box attack results in the supplementary material.

Overall, based on the results in Table 3.1, we can definitively answer the original question posed at the start of the this section. Vision Transformers do not provide any additional security over Big Transfer Models or conventional CNNs. We can clearly see this across all datasets, indicating Vision Transformers have no robustness (i.e. $0\%$) for the C&W and APGD attacks. Likewise, Vision Transformers have less than $6\%$ robustness across all the datasets for the PGD and MIM attacks. While this result may seem expected, it is an important step in understanding the complete security picture of Vision Transformers. Now that we know Vision Transformers are not robust to white-box attacks, we can consider the next important question on transferability.

## 4. Vision Transformers Transferability Study

*How transferable are the adversarial examples created by Vision Transformers?* It was shown in Section 3 that white-box attacks are extremely effective at creating examples that fool Vision Transformers. We further expand on the previous analyses and now examine the *transferability* of adversarial examples misclassified by Vision Transformers. Here, transferability refers to the occurrence of adversarial examples that are misclassified by multiple (i.e., more than one) classifier. The transferability of adversarial examples has been well documented for different CNN architectures. In the literature, the transferability of adversarial examples was first observed in [33]. Consequent studies have shown the transferability of adversarial examples between CNNs on the MNIST dataset in [30] and on the ImageNet dataset in [22]. However, to the best of our knowledge, there have been no large-scale studies on the transferability between CNNs and Visions Transformers at this time. We provide detailed evaluation and analyses on this aspect in this section.

Table 1. White-box attacks on Vision Transformers, Big Transfer Models and ResNets. The attacks are done using the $l_\infty$ norm with $\epsilon = 0.031$ for CIFAR-10 and $\epsilon = 0.062$ for ImageNet. The white-box attack results for CIFAR-100 follow an extremely similar trend to CIFAR-10. Hence for brevity, CIFAR-100 white-box attack results are given in the supplementary material. In this Table the robust accuracy is given for each corresponding attack. The last column "Acc" refers to the clean accuracy of the model.

| | CIFAR-10 | | | | | | |
|---|---|---|---|---|---|---|---|
| | FGSM | PGD | BPDA | MIM | C&W | APGD | Acc |
| ViT-B-32 | 37.9% | 1.8% | 17.6% | 4.4% | 0.0% | 0.0% | 98.6% |
| ViT-B-16 | 39.5% | 0.0% | 20.3% | 0.3% | 0.0% | 0.0% | 98.9% |
| ViT-L-16 | 56.3% | 1.2% | 28.7% | 5.9% | 0.0% | 0.0% | 99.1% |
| ViT-R50 | 40.8% | 0.1% | 13.4% | 0.2% | 0.0% | 0.0% | 98.6% |
| BiT-M-R50x1 | 66.0% | 0.0% | 14.9% | 0.0% | 0.0% | 0.0% | 97.5% |
| BiT-M-R101x3 | 85.2% | 0.0% | 17.1% | 0.0% | 0.0% | 0.0% | 98.7% |
| ResNet-56 | 23.0% | 0.0% | 5.0% | 0.0% | 0.0% | 0.0% | 92.8% |
| ResNet-164 | 29.0% | 0.0% | 5.4% | 0.0% | 0.0% | 0.0% | 93.8% |
| | ImageNet | | | | | | |
| | FGSM | PGD | BPDA | MIM | C&W | APGD | Acc |
| ViT-B-16 | 23.1% | 0.0% | 7.3% | 0.0% | 0.0% | 0.0% | 80.3% |
| ViT-L-16 (224) | 27.9% | 0.0% | 8.4% | 0.0% | 0.0% | 0.0% | 82.0% |
| ViT-L-16 (512) | 29.8% | 0.0% | 8.4% | 0.0% | 0.0% | 0.0% | 85.4% |
| BiT-M-R50x1 | 28.7% | 0.0% | 3.5% | 0.0% | 0.0% | 0.0% | 79.9% |
| BiT-M-R152x4 | 60.9% | 0.0% | 15.2% | 0.0% | 0.0% | 0.0% | 85.3% |
| ResNet-50 | 11.8% | 0.0% | 1.4% | 0.0% | 0.0% | 0.0% | 74.5% |
| ResNet-152 | 18.1% | 0.0% | 2.7% | 0.0% | 0.0% | 0.0% | 77.0% |

### 4.1. Measuring Transferability

Formally, we can define non-targeted transferability as follows: We start with a classifier $C_i$ and correctly identified input/label pair $(x, y)$. An attack $A_{C_i}$ is used to generate an adversarial example $x_{adv}$ with respect to classifier $C_i$:

$$x_{adv} = A_{C_i}(x, y) \tag{1}$$

The adversarial example $x_{adv}$ is then said to *transfer* from classifier to $C_i$ to $n-1$ other classifiers if and only if:

$$\forall_{j=1}^{n} \ [\{C_j(x) = y\} \wedge \{C_j(x_{adv}) \neq y\}] \tag{2}$$

Equation 2 states that each classifier $C_j$ must correctly classify $x$ and must misclassify $x_{adv}$. Assuming two classifiers ($n = 2$) and a set of $m$ examples that are correctly classified by both, we can define the transferability from $C_i$ to $C_j$ as follows:

$$t_{i,j} = \frac{1}{m} \sum_{k=1}^{m} \begin{cases} 1 & \text{if } C_j(A_{C_i}(x_k, y_k)) \neq y_k, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

A high transferability between classifiers indicates that they have a shared vulnerability to the same set of adversarial examples. On the other hand, a low transferability may indicate a possible avenue for security. This is due to the fact that the same set of adversarial examples are not misclassified by both classifiers.

### 4.2. Transferability Study Setup

To properly study the transferability between Vision Transformers, Big Transfer Models and conventional CNNs, we use the same 8 models for CIFAR-10 and CIFAR-100 as mentioned in Section 3.1. For ImageNet, we also use the same 7 models listed in Section 3.1. For our transferability study, we consider all possible pairs of classifiers. For each pair of classifiers $(i, j)$, we find a set of $m = 1000$ examples that both classifiers correctly identify. We then measure the transferability between the pair of classifiers using Equation 3. It is important to note that the transferability measurement will be affected by the choice of white-box attack $A_{C_i}$ used to generate the adversarial examples. It has been shown that MIM, PGD and FGSM are good candidates for creating highly transferable examples [25]. As a result, for every pair of classifiers $(i, j)$, we test all three attacks and report the highest transferability result. For these attacks, we use the same $\epsilon$ and $l_\infty$ norm as described in Section 3.1. Additional experimental details are provided in our supplementary material.

In Table 2, we show the transferability results for CIFAR-10, CIFAR-100 and ImageNet. The top row of the table corresponds to the model which was used to generate the adversarial examples, $C_i$ in Equation 3. The first column in the table corresponds to the model which was used to predict the labels of the adversarial examples. The model in the first column is $C_j$ in Equation 3. In the special case when $i = j$, we train an independent copy of model $i$ to generate adversarial examples for CIFAR-10 and CIFAR-100. For ImageNet, due to the high computational cost of model training, we forgo the $i = j$ measurement. It can clearly be seen from the other datasets we study and in the literature [22] that copies of the same model $(i = j)$ already

have high transferability. We also graphically represent the results of Table 2 in Figure 1 for the CIFAR-10 dataset.

### 4.3. Analysis of Transferability Study

From Table 2 and Figure 1, we can see a very interesting phenomenon. The transferability between Vision Transformers and Big Transfer Models is extremely low. For example, consider ViT-L-16 and BiT-M-50x1. Adversarial examples generated using BiT-50x1 are misclassified by ViT-L-16 less than $16\%$ of the time across all datasets ($5.7\%$, $15.5\%$ and $11.8\%$ for CIFAR-10, CIFAR-100 and ImageNet respectively). Likewise, less than half the time BiT-M-50x1 is fooled by adversarial examples generated using ViT-L-16 ($42.5\%$, $47.6\%$ and $34.3\%$ for CIFAR-10, CIFAR-100 and ImageNet).

Broadly speaking, we can consider the ViT models, BiT models and ResNets each as a model genus. In general, the phenomenon of low transferability mostly occurs between model genusus, but not within model genusus. That is to say, adversarial examples generated by one BiT model will likely transfer to a different BiT model, but not to a ViT model or ResNet. Visually, we can see this result represented for CIFAR-10 in Figure 1. The x-axis represents different models used to generate the adversarial examples and the y-axis represents the model used to evaluate those adversarial examples. The z-axis is used to measure the transferability. For clarity, the bars in the plot are color coded. Green, blue and light blue bars represent the transferability measurements between models of different genusus (green is ViT/ResNet transferability, blue is ViT/BiT transferability and light blue is BiT/ResNet transferability). Pink, red, and orange bars represent the transferability between models of the same genus. Pink is the transferability between ViT models, red is the transferability between BiT models and orange is the transferability between ResNet models.

It is important to note while the low transferability phenomenon is a generally observed trend, it is not an absolute rule. For example, the transferability between Big Transfer models (BiT-M-R50x1 and BiT-M-152x4) for ImageNet is also relatively low ($28\%$ and $24.9\%$). However, the most important factor is that the low transferability phenomenon ***does*** happen across multiple datasets and for multiple different pairs of models. The usefulness of these observations may not be apparent immediately. Nevertheless, they have serious security implications which we elaborate on subsequently.

## 5. White-Box Security and Transferability

*How can the transferability phenomena be leveraged to provide security?* From Section 4, we know that the transferability of adversarial examples between different model genusus is generally low. Therefore, we propose testing an ensemble of different models as a defense. To further clarify

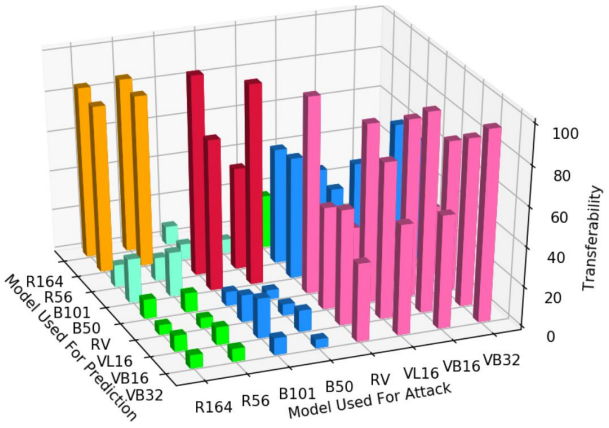



Figure 1. Visual representation of Table 2 for CIFAR-10. The x-axis corresponds to the model used to generate the adversarial examples. The y-axis corresponds to the model used to evaluate the adversarial examples. The z-axis measures transferability between the two models. The bars are color coded based on the two models. Pink, red, and orange bars represent the transferability between models of the same genus. Green, blue and light blue bars represent the transferability measurements between models of different genusus.

the original question, we break it down into two parts: Can an ensemble defense provide security against a white-box adversary, and can an ensemble provide security against a black-box adversary? In this section we answer the white-box question by proposing a novel attack that simultaneously breaks both Transformers and CNNs. In Section 6, we investigate the black-box question.

We first define our base case ensemble defense.

**Ensemble Models:** In this paper, we have already examined multiple Vision Transformers, Big Transfer Models and ResNets. The simplest ensemble would be to choose two types of classifiers from this group. Therefore, as a base case we use the most complex BiT model and ViT model. For CIFAR-10 and CIFAR-100 datasets, the ensemble is comprised of ViT-L-16 and BiT-M-101x3. For ImageNet, this ensemble is made up of ViT-L-16 (image size 512) and BiT-M-152x4. Here, we do not consider ResNets as they have significantly less clean accuracy and we do not want to pay such a security cost. In the supplementary material, we do provide some ResNet ensemble experiments for the sake of completeness.

**Ensemble Output:** In our ensemble defense, there are several possible ways to combine the output of the models. Here we consider three ways commonly found in the literature, majority voting [27], absolute consensus [26] and random selection [32]. Majority voting is a weak method of evaluating adversarial examples because not every classifier must be fooled, resulting in diminishing returns as the number of classifiers increases. The alternative to majority voting is absolute consensus [26]. In this setup, if ev-

Table 2. Transferability results for CIFAR-10, CIFAR-100 and ImageNet. The first column in each table represents the model used to generate the adversarial examples, $C_i$. The top row in each table represents the model used to evaluate the adversarial examples, $C_j$. Each entry is the maximum transferability computed using $C_i$ and $C_j$ over three different attacks, FGSM, PGD and MIM using Equation 3.

CIFAR-10

| | ViT-B-32 | ViT-B-16 | ViT-L-16 | R50-ViT | BiT-50x1 | BiT-101x3 | ResNet-56 | ResNet-164 |
|---|---|---|---|---|---|---|---|---|
| ViT-B-32 | 95.8% | 84.1% | 75.5% | 34.9% | 60.8% | 62.0% | 18.6% | 19.9% |
| ViT-B-16 | 57.1% | 99.6% | 88.9% | 22.6% | 43.4% | 45.0% | 13.9% | 14.0% |
| ViT-L-16 | 55.6% | 78.4% | 89.6% | 30.3% | 42.5% | 44.7% | 13.0% | 14.8% |
| R50-ViT | 39.6% | 58.1% | 51.5% | 98.3% | 61.0% | 58.0% | 26.7% | 29.0% |
| BiT-50x1 | 4.5% | 10.9% | 5.7% | 4.7% | 100.0% | 51.4% | 7.0% | 9.0% |
| BiT-101x3 | 8.6% | 20.3% | 13.7% | 7.2% | 75.9% | 100.0% | 7.8% | 9.3% |
| ResNet-56 | 6.6% | 9.0% | 5.3% | 9.7% | 22.5% | 11.8% | 85.9% | 87.2% |
| ResNet-164 | 6.8% | 8.1% | 5.0% | 9.7% | 22.3% | 11.2% | 83.6% | 85.7% |

CIFAR-100

| | ViT-B-32 | ViT-B-16 | ViT-L-16 | R50-ViT | BiT-50x1 | BiT-101x3 | ResNet-56 | ResNet-164 |
|---|---|---|---|---|---|---|---|---|
| ViT-B-32 | 96.2% | 88.5% | 83.6% | 52.2% | 60.5% | 61.1% | 14.9% | 14.0% |
| ViT-B-16 | 71.3% | 99.3% | 93.2% | 38.6% | 44.5% | 47.9% | 9.0% | 7.5% |
| ViT-L-16 | 67.8% | 88.3% | 94.2% | 48.1% | 47.6% | 50.0% | 9.9% | 9.5% |
| R50-ViT | 51.6% | 65.0% | 62.3% | 98.9% | 64.1% | 61.2% | 11.0% | 9.9% |
| BiT-50x1 | 17.7% | 25.0% | 15.5% | 18.2% | 100.0% | 56.5% | 4.9% | 5.2% |
| BiT-101x3 | 24.9% | 39.0% | 26.3% | 23.5% | 74.0% | 99.0% | 5.7% | 3.2% |
| ResNet-56 | 20.1% | 22.2% | 15.3% | 22.7% | 31.4% | 21.9% | 70.8% | 68.9% |
| ResNet-164 | 22.1% | 24.5% | 15.5% | 24.2% | 35.9% | 26.5% | 74.5% | 79.2% |

ImageNet

| | ViT-B-16 | ViT-L-16 | ViT-L-16 (512) | BiT-50x1 | BiT-152x4 | ResNet-50 | ResNet-152 |
|---|---|---|---|---|---|---|---|
| ViT-B-16 | + | 89.1% | 39.6% | 40.8% | 27.4% | 44.0% | 40.1% |
| ViT-L-16 | 90.9% | + | 64.5% | 40.0% | 26.9% | 43.7% | 40.8% |
| ViT-L-16 (512) | 28.0% | 43.4% | + | 34.3% | 26.3% | 28.4% | 23.2% |
| BiT-50x1 | 9.8% | 8.4% | 11.8% | + | 24.9% | 24.7% | 18.7% |
| BiT-152x4 | 8.2% | 7.6% | 13.5% | 28.0% | + | 15.1% | 12.0% |
| ResNet-50 | 23.8% | 18.8% | 24.7% | 55.3% | 24.4% | + | 86.7% |
| ResNet-152 | 25.9% | 22.1% | 26.6% | 54.1% | 26.8% | 89.4% | + |

ery classifier does not agree on the same class label then the sample is marked as adversarial. Absolute consensus removes the diminishing returns disadvantage of majority voting, though at the cost of clean accuracy. In absolute consensus, it is typical that many clean samples are marked as adversarial [26]. Due to this, we use random selection in all our ensemble defenses for the remainder of the paper. In random selection, a single model is selected randomly and used to evaluate the input at run time.

### 5.1. The Self-Attention Gradient Attack

**Attack Motivation:** A naïve approach would be to assume that an ensemble defense would provide security against a white-box adversary if only the low transferability results in Section 3 and Section 4 were taken into account. Consider the following analysis: Let us focus on the ImageNet models ViT-L-16 (image size 512) and BiT-M-152x4. From Section 4, we know a white-box MIM attack has a $100\%$ attack success rate ($0\%$ robust accuracy) on ViT-L-16 (see Table 2). Now let us introduce an additional model, BiT-M-152x4 into the ensemble with ViT-L-16. From Section 4 Table 2, we know the adversarial examples generated from ViT-L-16 will be misclassified by BiT-M-152x4 only $26.3\%$ of the time. If we make an ensemble of ViT-L-16 and BiT-M-152x4 with random selection, this means the attack success rate on average would drop to $63.15\%$. It seems as if we went from $0\%$ robust accuracy using only ViT-L-16 to $36.85\%$ robust accuracy just by using an ensemble with random selection. However, this is not the case as the adversarial examples we are using only come from attacking one model. We demonstrate the flaws in this type of analysis by proposing a new attack which generates adversarial examples that are simultaneously misclassifed by both Vision Transformers and CNNs. We call this new attack, the Self-Attention Gradient Attack (SAGA).

**Mathematical Description:** To derive SAGA, we assume the same white-box adversary we detailed in Section 3. Such an adversary has knowledge of the models and trained parameters in an ensemble defense. Instead of focusing completely on optimizing over one of the models, SAGA focuses on breaking multiple models at once. Assume we are given an ensemble with a set of Vision Transformers $V$ and a set of CNNs $K$. The goal of the attacker is to craft an adversarial example $x_{adv}$ from $x$ within perturba-

tion bounds $\epsilon$ that is misclassifed by all members $v \in V$ and $k \in K$. We can iteratively compute the adversarial example as follows:

$$x_{adv}^{(i+1)} = x_{adv}^{(i)} + \epsilon_s * sign(G_{blend}(x_{adv}^{(i)})) \quad (4)$$

where $x_{adv}^{(1)} = x$ and $\epsilon_s$ is the step size for the attack. Further, we define $G_{blend}(x_{adv}^{(i)})$ as follows:

$$G_{blend}(x_{adv}^{(i)}) = \sum_{k \in K} \alpha_k \frac{\partial L_k}{\partial x_{adv}^{(i)}} + \sum_{v \in V} \alpha_v \phi_v \odot \frac{\partial L_v}{\partial x_{adv}^{(i)}} \quad (5)$$

In Equation 5, the first summation is for the models in set $K$ which are CNNs. $\partial L_k / \partial x_{adv}^{(i)}$ is the partial derivative of the loss function of the $k^{th}$ CNN with respect to the adversarial input $x_{adv}^{(i)}$. Each model $k$ has an associated weighting factor $\alpha_k$. In a more refined approach, $\alpha_k$ could be optimized over as well, but here we simply leave $\alpha_k$ as a hyperparameter in the attack. Note that PGD [24] without randomized start is a special case of our attack when $V = \emptyset$, $K$ has exactly one element and $a_1 = 1$. However, when attacking an ensemble, $V \neq \emptyset$ and hence we have a second term.

In Equation 5, the second term[1] $\alpha_v \phi_v \odot \partial L_v / \partial x_{adv}^{(i)}$ is used to craft adversarial examples that are misclassified by the Vision Transformers in the ensemble. Here $\partial L_v / \partial x_{adv}^{(i)}$ is the loss function of the transformer with respect to the adversarial input. Likewise, $\alpha_v$ is a weighting factor selected by the attacker to balance the emphasis on different models. We also bring in one additional term which is specific to Vision Transformers, $\phi_v$. The term $\phi_v$ is the self-attention map associated with the $v^{th}$ transformer in the ensemble.

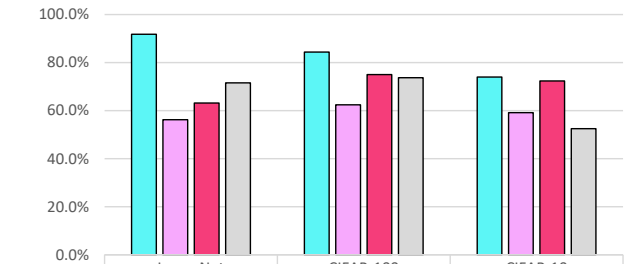
The self-attention $\phi_v$ is computed using attention rollout [1] and is defined as:

$$\phi_v = \left( \prod_{l=1}^{n_l} \left[ \sum_{i=1}^{n_h} (0.5 W_{l,i}^{(att)} + 0.5 I) \right] \right) \odot x. \quad (6)$$

where $n_h$ is the number of attention heads per layer, $n_l$ is the number of attention layers, $W_{l,i}^{(att)}$ is the attention weight matrix in each attention head, $I$ is the identity matrix and $x$ is the input image. This technique takes into account the attention flow from each layer of the transformer to the next layer, including the effect of skip connections. The attention values from the different attention heads within the same layer are averaged, and the attention values are recursively multiplied between different layers.

**Experimental Results:** We demonstrate the SAGA results by attacking a simple ensemble of Vision Transformers and Big Transfer Models for CIFAR-10, CIFAR-100 and ImageNet. We use 1000 clean correctly identified examples with the same attack parameters as described in Section 3.

[1] $\odot$ is the element wise Hadamard product; $x$ in (5) and (6) is an image matrix and the partial derivative w.r.t $x$ in (5) is represented as a matrix.




| | ImageNet | CIFAR-100 | CIFAR-10 |
|---|---|---|---|
| SAGA | 91.8% | 84.4% | 74.0% |
| Single PGD | 56.2% | 62.4% | 59.2% |
| Single MIM | 63.2% | 75.0% | 72.4% |
| Basic | 71.6% | 73.7% | 52.5% |

Figure 2. Attack success rate of the Self-Attention Gradient Attack (SAGA), the Single MIM attack and Basic attack on an ensemble containing one ViT-L-16 model and one BiT-M-R101x3 model (or BiT-M-R152x4 for ImageNet). For full descriptions of each attack see Section 5.1.

For CIFAR-10 and CIFAR-100, we use Bit-M-R101x3 and ViT-L-16. For ImageNet, we use Bit-M-R152x4 and ViT-L-16. We also test three other simple attacks which are denoted in Figure 2 as Basic, single PGD and Single MIM. The basic attack is a combination of the model gradients without weighted coefficients and self-attention included. The single MIM/PGD attack is the best transfer attack on the ensemble as reported from Table 2.

The main contribution of this attack is to demonstrate that Vision Transformer/Big Transfer type of ensembles are not secure under a white-box adversary. This is precisely what is shown in Figure 2. SAGA has an attack success rate of $74.0\%$, $84.4\%$ and $91.8\%$ on the ensemble for CIFAR-10, CIFAR-100 and ImageNet, respectively. In Figure 2, we also show SAGA outperforms the other white-box multi-model attacks across all datasets. For brevity, many details are omitted here such as the hyperparameter selection for SAGA and attacks on Transformer/ResNet ensembles. We provide this information fully in the supplementary material.

## 6. Black-Box Security and Transferability

In this section, we consider the transferability phenomena and its security implications under a black-box adversarial model. We once again use an ensemble of classifiers with random selection as described in Section 5. From Section 5.1, we know that such an ensemble is not secure against white-box adversaries. Using attacks like SAGA, an adversary can blend the gradients of different models and the self-attention of Transformers. This results in a high percentage of adversarial examples that are misclassified by all the classifiers. However, this type of attack relies heavily on the white-box capabilities of the adversary. Without knowledge of the models in the ensemble and their trained

parameters, this type of attack would not work. This brings up a new possibility. Can transferability (through an ensemble) provide security when individual model gradients are not available to the attacker?

### 6.1. Black-Box Attack Parameters and Adversarial Model

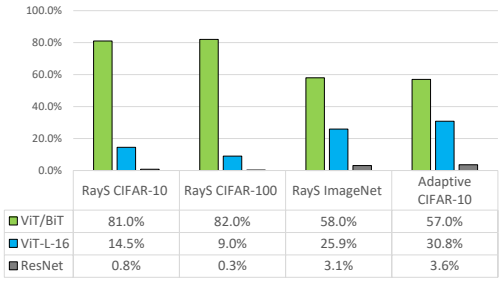
**Adversarial Model:** In this section we consider two of the main types of black-box adversaries, query-based [3] and transfer-based adversaries [29]. For the query based adversary, we test one of the most recent attacks, the RayS attack [8]. In this attack, the adversary generates an adversarial example by repeatedly querying the defense and adjusting the noise accordingly. For the transfer attack, we implement the Adaptive Black-Box Attack [26]. This attack is a stronger version of the Papernot attack originally proposed in [29]. Here the attacker has access to a percentage of the original training data, query access to the defense and the ability to train a synthetic model to generate adversarial examples. In this attack, the adversary queries the defense to obtain labels for the training data. It then uses the data labeled by the defense to train an independent classifier (synthetic model). An attack is then performed on the trained synthetic model. The resulting adversarial examples are then tested on the defense.

**Attack Parameters:** For all black-box attacks, we use the same basic set of constraints as described in Section 3.1. The noise the adversary can generate is bounded by the $l_\infty$ norm with $\epsilon = 0.031$ for CIFAR-10/CIFAR-100 and $\epsilon = 0.062$ for ImageNet. For the RayS attack, we give the adversary a budget of $10,000$ queries per sample. For the Adaptive attack, we give the adversary $100\%$ of the training data. For the synthetic model in this attack, we used ViT-B-32 pre-trained on ImageNet-21K. We also experimented with CNN based synthetic models, however these did not perform as well on our ensemble defense. It should also be noted the $100\%$ strength attack requires a huge amount of computation. Due to this we only show the results for CIFAR-10 for the Adaptive attack. For RayS, we show results for all three datasets.

### 6.2. Black-Box Attack Analysis

In Figure 3, we show the results graphically for the RayS and Adaptive attack. We consider three different model configurations. We test an ensemble of one Vision Transformer (ViT-L-16) and one Big Transfer Model (BiT-M-101x3 for CIFAR-10/CIFAR-100 and BiT-M-152x4 for ImageNet). We also test a single ViT-L model and a single CNN (ResNet-56 for CIFAR-10/CIFAR-100 and ResNet-50 for ImageNet). While slightly redundant, we do test other ensemble configurations (and individual Big Transfer Models) in the supplementary material for those interested.

The robust accuracy (percent of adversarial samples correctly identified by the defense) are shown in Figure 3 for each attack. Here we observe the most significant result of our paper: a simple ensemble including a Vision Transformer and Big Transfer model drastically improves security. For RayS, we observe an increase of $66.5\%$, $73\%$ and $32.1\%$ in robust accuracy for CIFAR-10, CIFAR-100 and ImageNet respectively. For the CIFAR-10 Adaptive attack, even when the adversary has $100\%$ of the training data, query access and a synthetic model pre-trained on the same dataset as the defense (ImageNet-21K), we can still achieve a robust accuracy of $57\%$. For the Adaptive attack that represents an improvement of $26.2\%$ over a single model.




| | RayS CIFAR-10 | RayS CIFAR-100 | RayS ImageNet | Adaptive CIFAR-10 |
|---|---|---|---|---|
| ViT/BiT | 81.0% | 82.0% | 58.0% | 57.0% |
| ViT-L-16 | 14.5% | 9.0% | 25.9% | 30.8% |
| ResNet | 0.8% | 0.3% | 3.1% | 3.6% |

Figure 3. Robust accuracy (higher is better) of different model configurations under black-box attacks. Here ViT/BiT is an ensemble containing a Vision Transformer (ViT-L-16) and a Big Transfer Model (BiT-M-101x3 for CIFAR-10/CIFAR-100 and Bit-M-R152x4 for ImageNet.

We also stress that this improvement does not come at the cost of clean accuracy. The average clean accuracy of the ensemble is $98.2\%$, $92.83\%$ and $85.37\%$ for CIFAR-10, CIFAR-100 and ImageNet respectively. By leveraging the low transferability phenomena we previously studied, we are able to create a defense that achieves near state-of-the-art performance on clean data and gives significant black-box robustness.

## 7. Conclusion

The introduction of Vision Transformers represents new opportunities for the field of adversarial machine learning. By analyzing these new models, we are the first to uncover several intriguing properties. We demonstrated that the transferability between different model genusus are in general, remarkably low. We then showed this phenomena does not yield white-box security by developing a new white-box attack, the Self-Attention Gradient Attack (SAGA). Finally, we showed that under a black-box adversary, the transferability phenomena can be used to achieve robustness, all while maintaining near state-of-the-art clean accuracy on CIFAR-10, CIFAR-100 and ImageNet. Through our comprehensive experiments and analyses, we show how Vision Transformers advance security in the field of adversarial machine learning.

## References


[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pages 274–283, 2018.

[3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, 2018.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *Lecture Notes in Computer Science*, page 213–229, 2020.

[5] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019.

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[7] Nicholas Carlini and David A. Wagner. Towards Evaluating the Robustness of Neural Networks. *CoRR*, 2016.

[8] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1739–1747, 2020.

[9] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020.

[10] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020.

[11] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 9185–9193, 2018.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[13] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[15] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. *arXiv preprint arXiv:2004.11207*, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[17] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, July 2020.

[18] Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529, 2019.

[19] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *Lecture Notes in Computer Science*, page 491–507, 2020.

[20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, Dec. 1989.

[21] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[22] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *CoRR*, abs/1611.02770, 2016.

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learn- ing Representations (ICLR)*, 2018.

[25] Kaleel Mahmood, Deniz Gurevin, Marten van Dijk, and Phuong Ha Nguyen. Beware the black-box: on the robustness of recent defenses to adversarial examples, 2020.

[26] Kaleel Mahmood, Phuong Ha Nguyen, Lam M. Nguyen, Thanh Nguyen, and Marten van Dijk. Buzz: Buffer zones for defending adversarial examples in image classification, 2020.

[27] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving Adversarial Robustness via Promoting Ensemble Diversity. In *ICML*, pages 4970–4979, 2019.

[28] 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*, 2016.

[29] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *2017 ACM Asia Conference on Computer and Communications*

*Security, ASIA CCS 2017*, pages 506–519. Association for Computing Machinery, Inc, 2017.

[30] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016.

[31] Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *ICLR*, 2020.

[32] Siwakorn Srisakaokul, Zexuan Zhong, Yuhao Zhang, Wei Yang, and Tao Xie. Muldef: Multi-model-based defense against adversarial examples for neural networks. *arXiv preprint arXiv:1809.00065*, 2018.

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[34] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33, 2020.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[36] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020.

[37] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020.

[38] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning*, pages 11278–11287. PMLR, 2020.