Transferable Adversarial Attacks on Vision Transformers with Token Gradient Regularization

Jianping Zhang[1] Yizhan Huang[1] Weibin Wu[2*] Michael R. Lyu[1]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]School of Software Engineering, Sun Yat-sen University
{jpzhang, yzhuang22, lyu}@cse.cuhk.edu.hk, wuwb36@mail.sysu.edu.cn

Abstract

Vision transformers (ViTs) have been successfully deployed in a variety of computer vision tasks, but they are still vulnerable to adversarial samples. Transfer-based attacks use a local model to generate adversarial samples and directly transfer them to attack a target black-box model. The high efficiency of transfer-based attacks makes it a severe security threat to ViT-based applications. Therefore, it is vital to design effective transfer-based attacks to identify the deficiencies of ViTs beforehand in security-sensitive scenarios. Existing efforts generally focus on regularizing the input gradients to stabilize the updated direction of adversarial samples. However, the variance of the back-propagated gradients in intermediate blocks of ViTs may still be large, which may make the generated adversarial samples focus on some model-specific features and get stuck in poor local optima. To overcome the shortcomings of existing approaches, we propose the Token Gradient Regularization (TGR) method. According to the structural characteristics of ViTs, TGR reduces the variance of the back-propagated gradient in each internal block of ViTs in a token-wise manner and utilizes the regularized gradient to generate adversarial samples. Extensive experiments on attacking both ViTs and CNNs confirm the superiority of our approach. Notably, compared to the state-of-the-art transfer-based attacks, our TGR offers a performance improvement of 8.8% on average.

1. Introduction

Transformers have been widely deployed in the natural language processing, achieving state-of-the-art performance. Vision transformer (ViT) [5] first adapts the transformer structure to the computer vision, and manifests excellent performance. Afterward, diverse variants of ViTs have been proposed to further improve its performance [2, 26] and broaden its application to different computer vision tasks [42, 43], which makes ViTs a well-recognized successor for convolutional neural networks (CNNs). Unfortunately, recent studies have shown that ViTs are still vulnerable to adversarial attacks [1, 22], which add human-imperceptible noise to a clean image to mislead deep learning models. It is thus of great importance to understand DNNs [13, 27, 28, 35] and devise effective attacking methods to identify their deficiencies before deploying them in safety-critical applications [16, 17, 37].

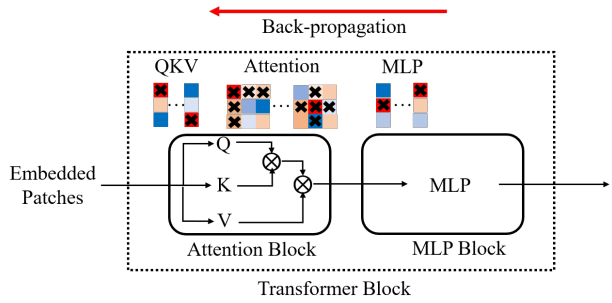

Figure 1. Illustration of our Token Gradient Regularization (TGR) method. The red-colored entry represents the back-propagated gradient with extreme values. The back-propagated gradients corresponding to one token in the internal blocks of ViTs are called the token gradients. Since we regularize the back-propagated gradients in a token-wise manner, we eliminate the token gradients (marked with crosses) where extreme gradients locate during back-propagation to reduce the gradient variance. We then use the regularized gradients to generate adversarial samples.

Adversarial attacks can be generally partitioned into two categories. The first category is the white-box attack, where attackers can obtain the structures and weights of the target models for generating adversarial samples. The second one is the black-box attack, where attackers cannot fetch the information of the target model. Among different black-box attacks, transfer-based methods employ white-box attacks to attack a local source model and directly transfer

*Corresponding author.

the generated adversarial sample to attack the target black-box model. Due to their high efficiency and applicability, transfer-based attacks pose a serious threat to the security of ViT-based applications in practice. Therefore, in this work, we focus on transfer-based attacks on ViTs.

There are generally two branches of transfer-based attacks in the literature [15]. The first one is based on input transformation, which aims to combine the input gradients of multiple transformed images to generate transferable perturbations. Complementary to such methods, the second branch is based on gradient regularization, which modifies the back-propagated gradient to stabilize the update direction of adversarial samples and escape from poor local optima. For example, Variance Tuning Method (VMI) [29] tunes the input gradient to reduce the variance of input gradients. However, the variance of the back-propagated gradients in intermediate blocks of ViTs may still be large, which may make the generated adversarial samples focus on some model-specific features with extreme gradient values. As a result, the generated adversarial samples may still get stuck in poor local optima and possess limited transferability across different models.

To address the weaknesses of existing gradient regularization-based approaches, we propose the Token Gradient Regularization (TGR) method for transferable adversarial attacks on ViTs. According to the architecture of ViTs, TGR reduces the variance of the back-propagated gradient in each internal block of ViTs and utilizes the regularized gradient to generate adversarial samples.

More specifically, ViTs crop one image into small patches and treat these patches as a sequence of input tokens to fit the architecture of the transformer. The output tokens of internal blocks in ViTs correspond to the extracted intermediate features. Therefore, we view token representations as basic feature units in ViTs. We then examine the back-propagated gradients of the classification loss with respect to token representations in each internal block of ViTs, which we call token gradient in this work. As illustrated in Figure 1, we directly eliminate the back-propagated gradients with extreme values in a token-wise manner until obtaining the regularized input gradients, which we used to update the adversarial samples. Consequently, we can reduce the variance of the back-propagated gradients in intermediate blocks of ViTs and produce more transferable adversarial perturbations.

We conducted extensive experiments on the ImageNet dataset to validate the effectiveness of our proposed attack method. We examined the transferability of our generated adversarial samples to different ViTs and CNNs. Notably, compared with the state-of-the-art benchmarks, our proposed TGR shows a significant performance improvement of 8.8% on average.

We summarize the contributions of this work as below:

- We propose the Token Gradient Regularization (TGR) method for transferable adversarial attacks on ViTs. According to the architectures of ViTs, TGR regularizes the back-propagated gradient in each internal block of ViTs in a token-wise manner and utilizes the regularized gradient to generate adversarial samples.
- We conducted extensive experiments to validate the effectiveness of our approach. Experimental results confirm that, on average, our approach can outperform the state-of-the-art attacking method with a significant margin of 8.8% on attacking ViT models and 6.2% on attacking CNN models.
- We showed that our method can be combined with other compatible attack algorithms to further enhance the transferability of the generated adversarial samples. Our method can also be extended to use CNNs as the local source models.

2. Related Work

2.1. Adversarial Attacks on CNNs

There are generally two categories of adversarial attacks in the literature [15]: white-box and black-box attacks. White-box attacks get full access to the information of target models, like model structure and weights. In contrast, black-box attacks [31, 41] fail to obtain the specifics of target models. White-box adversarial attacks can directly adopt the gradient information of the target model to craft adversarial samples, like the Fast Gradient Sign Method (FGSM) [7] and Basic Iterative Method (BIM) [14].

However, in reality, the model structure and weights are hidden from the users. Therefore, more research focuses on the adversarial attack under the black-box setting. Among different black-box attack methodologies, transfer-based attacks stand out due to their severe security threat to deep learning-based applications in practice. Therefore, we also focus on transfer-based attacks in this paper. The ability of adversarial samples crafted by a local source model being able to mislead other black-box target models is called transferability. Transfer-based attacks [34,36,40] utilize the transferability of adversarial samples. Specifically, transfer-based attacks usually craft adversarial samples with a local source model using white-box attack methods and input the generated adversarial samples to the target model to cause misclassification.

There are generally two branches of approaches to enhance the transferability of adversarial samples. The first branch is based on input transformation, aiming to combine the gradient of several transformed images during the generation of transferable perturbations. Diverse Input Method (DIM) [38] applies random resizing and padding to the input image with a fixed probability and uses the transformed

image to compute the input gradient for updating adversarial samples. Translation Invariant Method (TIM) [4] employs shifted images for computing the input gradient, which is simplified by convolving the input gradient of the original image with a kernel matrix. Scale Invariant Method (SIM) [15] utilizes the scale-invariant property of CNNs and combines the gradients of the scaled copies of the original image together. However, due to the structural difference between ViTs and CNNs, the input transformation methods tailored for CNNs cannot achieve comparable performance improvement for transferable adversarial attacks on ViTs [32]. Therefore, in this work, we take the special design of ViTs into consideration to improve the transferability of adversarial samples on ViTs.

The second branch is based on gradient regularization [30], which aims to stabilize the update direction of adversarial samples and escape from the poor local optima. Momentum Iterative Method (MIM) [3] incorporates the momentum term into the computation of the input gradient. Skip Gradient Method (SGM) [33] utilizes a decay factor to reduce the back-propagated gradients from the residual module to focus on the transferable low-level information. Variance Tuning Method (VMI) [29] tunes the input gradient with the gradient variance in the neighborhood of the target image to reduce the gradient variance. Different from the existing work, we reduce the gradient variance in each internal block of ViTs based on the architecture of ViTs, which avoids the over-reliance on model-specific features and thus improves adversarial transferability against ViTs.

2.2. Vision Transformer

ViT [5] first adapts the transformer network structure from the natural language processing field to the computer vision field. Specifically, ViT divides the input image into a sequence of small image patches, which are attached with a classification token to constitute the input to the transformer. More advanced versions of ViTs are proposed to improve the accuracy and efficiency of the vanilla ViT. For example, pooling-based vision transformer (PiT) [12] decreases the spatial dimension and increases the channel dimension with pooling to improve model capability. Class-attention in image transformer (CaiT) [26] builds deeper transformers and adds the classification token in the latter layer of the network. The vision-friendly transformer (Visformer) [2] transit a transformer-based model to a convolution-based model. There are also other works trying to improve the performance of the ViT from other points of view [9, 25].

2.3. Adversarial Attacks on ViTs

With the wide deployment of ViTs in diverse vision tasks, there is an increasing interest in evaluating the robustness of ViTs to identify their deficiencies. Bhojanapalli et al. [1], and Shao et al. [22] analyzed the robustness of ViTs against white-box attacks. They found that ViTs tend to be more robust than CNNs. Mahmood et al. [19] examined the transferability of adversarial samples crafted by ViTs. They discovered that the transferability from ViTs to CNNs is relatively low due to their structural difference. In this work, to better assess the robustness of ViTs in black-box settings, we attempt to improve the transferability of adversarial samples from ViTs to both different ViT models and CNN models.

Existing transfer-based attacks on ViTs generally follow similar methodologies of those on CNNs. For example, the Pay No Attention (PNA) method adapts the SGM to ViTs. Specifically, PNA skips the gradient of the attention block during back-propagation to improve the transferability of adversarial samples, which accounts for a gradient regularization-based approach [32]. The PatchOut attack strategy randomly samples a subset of patches to compute the gradient in each attack iteration, which acts as an image transformation method for transferable adversarial attacks on ViTs. Different from such methods, our approach proposes to improve transferable adversarial attacks on ViTs from the perspective of gradient variance reduction, which helps to stabilize the update direction of adversarial samples and escape from the poor local optima.

Different from the above line of research, Nasser et al. [20] proposed the Self-Ensemble (SE) method and the Token Refinement module (TR) to improve the transferability of adversarial samples generated from ViTs. SE utilizes the classification token on each layer of ViTs with a shared classification head to perform feature-level attacks. Based on SE, TR further refines the classification token with fine-tuning to improve attack performance. However, the method is time-consuming since it needs to fine-tune the classification token for each source model. Besides, some variants of ViTs do not have or only have a few classification tokens, like Visformer [2] and CaiT [26]. Therefore, the SE and TR method only has limited applicability.

3. Method

We first set up some notations. We denote the benign image as x with the image size of $H \times W \times C$, where H, W, and C represent the height, width, and channel number of the image. ViTs divide the image into a sequence of patches $x_p = \{x_p^1, x_p^1, \cdots, x_p^N\}$, where x_p^i is the i-th patch of the original image. The shape of each patch x_p^i is $P \times P \times C$, and P is the patch size. There are in total $N = \frac{H \cdot W}{P^2}$ patches. The corresponding true label of the image is y. We represent the output of a DNN classifier by $f(x)$. $J(x, y; f)$ stands for the classification loss function of the classifier f, which is usually the cross-entropy loss. Given the target image x, adversarial attacks aim to find an adversarial sample x^{adv}, which can mislead the classifier,

i.e., $f(x^{adv}) \neq f(x)$, while it is human-imperceptible, i.e., satisfying the constraint $\left\|x - x^{adv}\right\|_p < \epsilon$. The $\|\cdot\|_p$ represents the L_p norm, and we focus on the L_∞ norm here to align with previous papers [3].

3.1. Token Gradient Regularization

A large gradient variance [29] triggers an overfitting issue for generating adversarial samples whose update direction is not optimal and easily stuck in the local optimal. Existing efforts focus on regularizing the update gradient on the input and regardless of the gradient variance in intermediate blocks leading to uncontrollable gradient variance in the back-propagation. The motivation behind our method is to regularize the gradient variance in intermediate blocks of ViTs.

In order to reduce the variance of the back-propagated gradients in intermediate blocks of ViTs, we seek to regularize the gradients and consider the structural characteristics of ViTs. Tokens are the fundamental building blocks in the ViTs, and we call the back-propagated gradients corresponding to one token in the internal blocks of ViTs as the *token gradients*. Consequently, regularizing the gradients in the intermediate blocks is to regularize the token gradients. The proposed Token Gradient Regularization (TGR) is to regularize the token gradients in a token-wise manner. We regard that tokens with extreme back-propagated gradient values contribute to the high gradient variance because the extreme back-propagated gradients tend to be model-specific and unstable features [29]. Specifically, if the back-propagated gradient of a token is in the top-k or bottom-k gradient magnitude among all the tokens, then it is called the *extreme tokens*, where k is a hyper-parameter. Since we regularize the back-propagated gradients in the token-wise, we eliminate the extreme token gradients to reduce the gradient variance during the back-propagation.

To design effective attack methods that cater to ViTs structural characteristics, we analyze the workflow of ViTs and select representative components in intermediate blocks to employ TGR. ViTs are composed of transformer blocks and each transformer block has an Attention block and a MLP block. The Attention block deploys the self-attention mechanism to compute the Attention between input tokens by Key and Query and multiply the Attention with Value. Therefore, we consider employing TGR on the QKV component and the Attention component for Attention block. The MLP block utilizes a fully-connected layer to aggregate the channel information for all the tokens. Thus, we select the MLP layer to deploy TGR. In the following sections, we illustrate the detailed implementation of TGR.

3.2. Implementation

As mentioned in the previous section, we aim to regularize three components in the architecture of ViTs: the Attention and QKV component in the Attention block and the MLP component in the MLP block.

Attention Component. The Attention component utilizes the multi-head self-attention mechanism to compute the relationship between tokens. We suppose there are M self-attention operations in one Attention component. Therefore, the Attention component will output an attention map with the size of $N \times N \times M$ for one image, where N is the number of patch tokens. We regard that the self-attention head computes the relationship between tokens independently. Thus, we rank the backward gradient in each output channel of the Attention component independently. We first localize the extreme tokens on the attention map and denote their positions. Then, we eliminate the gradient entries that lie in the same rows and columns of the extreme gradients at a time.

QKV Component. QKV component computes the Query, Key, and Value for the self-attention mechanism. Suppose the QKV component has C channel, so the size of the QKV component is $N \times C$ for one image. We also regard the channels are independent, and we rank the backward gradient in each input channel of the QKV component independently. We define the tokens with the top-k or bottom-k back-propagated gradient magnitude as extreme tokens. Thus, we eliminate the extreme gradient entries at a time.

MLP Block. MLP block aggregates the information of each token along the channels. We denote the MLP block has C channel, so the size of the MLP block is $N \times C$ for one image. Similar to the QKV component, we prioritize each input channel of the MLP component independently. We also eliminate the extreme gradient entries of the token with the top-k or bottom-k back-propagated gradient magnitude.

The illustration of the TGR on each component is shown in Figure 1. Apart from regularizing the extreme tokens, we also introduce a scaling factor s on each component to reduce the overall gradient variance. Equation 1 shows the adversarial attack of step t. In the equation, g' is the regularized backward gradient on the input. $modules$ represents the model structure, and $Grads$ records the backward gradient in the network. $TGR(\cdot)$ is the Token Gradient Regularization method, and details are shown in Algorithm 1.

$$\begin{aligned} g' &= TGR(Grads, modules, k, s) \\ x_{t+1}^{adv} &= x_t^{adv} + \alpha \cdot sgn\{g'\}, \end{aligned} \tag{1}$$

where α is a hyper-parameter to control the step size in update each iteration.

4. Experiments

In this section, we present extensive experiments to evaluate the effectiveness of our proposed method. We first clarify the setup of the experiments. After that, we illustrate the

Algorithm 1 Token Gradient Regularization

Require: network structure $modules$ and gradients $Grads$
Require: scaling factor s and extreme token number k
Ensure: the gradient on the input g'
  for m **in** $modules$ **do**
    if m is MLP or KQV **then**
      $Grads[m] \leftarrow Grads[m] * s$
      $token \leftarrow extreme(Grads[m], k)$ ▷ Extreme
  Tokens on MLP or KQV component
      for $i = 0 \leftarrow 2k - 1$ **do**
        $Grads[m][token[i], :] = 0$
      end for
    else if m is Attention **then**
      $Grads[m] \leftarrow Grads[m] * s$
      $tokens \leftarrow extreme(Grads[m], k)$ ▷ Extreme
  Token Pairs on the Attention Map
      for $i = 0 \leftarrow 2k - 1$ **do**
        $Grads[m][tokens[i, 0], :, :] = 0$
        $Grads[m][:, tokens[i, 1], :] = 0$
      end for
    end if
  end for

attacking results of our method against competitive baseline methods under various experimental settings to show the effectiveness of our proposed attack on both ViT and CNN models. Moreover, we analyze the effect of TGR on the gradient reduction in the internal blocks of ViTs and adapt our approach to CNN attacks. Finally, we perform the ablation study on the selection of the components to employ TGR and the selection of the token number.

4.1. Experiment Setup

We follow the protocol of the baseline method [32] to set up the experiments for a fair comparison to attack image classification models trained on ImageNet [21]. ImageNet is also the most widely utilized benchmark task for transfer-based adversarial attacks [29,32]. Here are the details of the experiment setup.

Dataset. We follow the dataset of the baseline method [32] by randomly sampling 1000 images of different categories from the ILSVRC 2012 validation set [21]. We check that all of the attacking models achieve almost 100% classification success rate in this paper.

Models. We evaluate the transferability of adversarial samples of ViTs under two attacking scenarios. The first one is that the source and target models are both ViT models to validate the transferability across different ViT structures. The other one is that the source model is ViT, but the target models are CNN models to examine the cross-model structure transferability. We choose four representative ViT models as the source models to generate adversarial samples, including ViT-B/16 [5], PiT-B [12], CaiT-S/24 [26], and Visformer-S [2]. In addition to the four source ViT models, the target ViT models contain four more ViTs: DeiT-B [25], TNT-S [9], LeViT-256 [8], and ConViT-B [6]. We keep the four ViT source models to craft adversarial samples under the second experiment setting. We select both undefended (normally trained) models and defended (adversarial training and advanced defense technique) models as the target CNN models. For undefended models, we use four representative target models containing Inception-v3 (Inc-v3) [24], Inception-v4 (Inc-v4) [23], Inception-Resnet-v2 (IncRes-v2) [23] and Resnet-v2-152 (Res-v2) [10, 11]. For defended models, we consider adversarial training and advanced defense models because adversarial training is a simple but effective technique [18,39], and advanced defense models are robust against black-box adversarial attacks. Three adversarial trained models are selected: an ensemble of three adversarial trained Inception-v3 models (Inc-v3$_{ens3}$), an ensemble of four adversarial trained Inception-v3 models (Inc-v3$_{ens4}$), and adversarial trained Inception-Resnet-v2 (IncRes-v2$_{adv}$).

Baseline Methods. We first choose the advanced gradient iterative-based method MIM [3] as our baseline. In addition, VMI [29] is another gradient-based baseline method, which utilizes the gradient variance reduction strategy to regularize the update gradient. Furthermore, we also compare our approach with attacking methods using the structure of the ViTs. We compare our method with SGM [33], which utilizes a decay factor to reduce the gradient from the residual module. In order to show our attacking method outperforming state-of-the-art, we select PNA [32] as our baseline, which is the current state-of-the-art attacking approach for ViTs. In addition, we compose all the attacking methods with input transformation method (PatchOut) [32] for ViT attacks, which is motivated by DIM [38] in CNN attack for ViT models. We denote our method with the PatchOut strategy as TGR-P and the baselines with the PatchOut strategy as MIM-P, VMI-P, SGM-P, and PNA-P, respectively.

Evaluation Metric. In the experiments, the evaluation metric is the attack success rate, the ratio of the adversarial samples which successfully mislead the target model among all the generated adversarial samples.

Parameter. For a fair comparison, we follow the parameter setting in [32] to set the maximum perturbation to $\epsilon = 16$ and the number of iterations to $T = 10$, so the step length $\alpha = \frac{\epsilon}{T} = 1.6$. As for the decay factor μ, we set μ to 1.0 for all the baselines because all the baselines utilize the momentum method as the optimizer. We also keep the hyper-parameter setting of all the baselines to conduct experiments. For the PatchOut strategy, we set the number of sampled patches to be 130 by following PNA [32]. All images are resized to 224×224 to conduct experiments and set the patch size to be 16 for the inputs of ViTs. Therefore,

Model	Attack	ViT-B/16	PiT-B	CaiT-S/24	Visformer-S	DeiT-B	TNT-S	LeViT-256	ConViT-B
ViT-B/16	MIM	100.0	34.5	64.1	36.5	64.3	50.2	33.8	66.0
	VMI	99.6	48.8	74.4	49.5	73.0	64.8	50.3	75.9
	SGM	100.0	36.9	77.1	40.1	77.9	61.6	40.2	78.4
	PNA	100.0	45.2	78.6	47.7	78.6	62.8	47.1	79.5
	TGR	**100.0**	**49.5**	**85.0**	**53.8**	**85.6**	**73.1**	**56.5**	**85.4**
PiT-B	MIM	24.7	100.0	34.7	44.5	33.9	43.0	38.3	37.8
	VMI	38.9	99.7	51.0	56.6	50.1	57.0	52.6	51.7
	SGM	41.8	100.0	57.3	73.9	57.9	72.6	68.1	59.9
	PNA	47.9	100.0	62.6	74.6	62.4	70.6	67.3	61.7
	TGR	**60.3**	**100.0**	**80.2**	**87.3**	**78.0**	**87.1**	**81.6**	**76.5**
CaiT-S/24	MIM	70.9	54.8	99.8	55.1	90.2	76.4	54.8	88.5
	VMI	76.3	63.6	98.8	67.3	88.5	82.3	67.0	88.1
	SGM	86.0	55.8	100.0	68.2	97.7	91.1	74.9	96.7
	PNA	82.4	60.7	99.7	67.7	95.7	86.9	67.1	94.0
	TGR	**88.2**	**66.1**	**100.0**	**75.4**	**98.8**	**92.8**	**74.7**	**97.9**
Visformer-S	MIM	28.1	50.4	41.0	99.9	36.9	51.9	49.4	39.6
	VMI	39.2	60.0	56.6	100.0	54.1	62.8	59.1	54.4
	SGM	18.8	41.8	34.9	100.0	31.2	52.1	52.7	29.5
	PNA	35.4	61.5	54.7	100.0	51.0	66.3	64.5	50.7
	TGR	**41.2**	**70.3**	**62.0**	**100.0**	**59.5**	**74.7**	**74.8**	**56.2**

Table 1. The attack success rates (%) against eight models by various transfer-based attacks. The best results are marked in bold.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{adv}$
ViT-B/16	MIM	31.7	28.6	26.1	29.4	22.3	19.8	16.5
	VMI	43.1	41.6	37.9	42.6	31.4	30.6	25.0
	SGM	31.5	27.7	23.8	28.2	20.8	18.0	14.3
	PNA	42.7	37.5	35.3	39.5	29.0	27.3	22.6
	TGR	**47.5**	**42.3**	**37.6**	**43.3**	**31.5**	**30.8**	**25.6**
PiT-B	MIM	36.3	34.8	27.4	29.6	19.0	18.3	14.1
	VMI	47.3	45.4	40.7	43.4	35.9	34.4	29.7
	SGM	50.6	45.4	38.4	41.9	25.6	20.8	16.7
	PNA	59.3	56.3	49.8	53.0	33.3	32.0	25.5
	TGR	**72.1**	**69.8**	**65.1**	**64.8**	**43.6**	**41.5**	**32.8**
CaiT-S/24	MIM	48.4	42.9	39.5	43.8	30.8	27.6	23.3
	VMI	58.5	50.9	48.2	52.0	38.1	36.1	30.1
	SGM	53.5	45.9	40.2	45.9	30.8	28.5	21.0
	PNA	57.2	51.8	47.7	51.6	38.4	36.2	30.1
	TGR	**60.3**	**52.9**	**49.3**	**53.4**	**39.6**	**37.0**	**31.8**
Visformer-S	MIM	44.5	42.5	36.6	39.6	24.4	20.5	16.6
	VMI	54.6	53.2	48.5	52.2	33.0	32.0	22.2
	SGM	43.2	41.1	29.6	35.7	16.1	13.0	8.2
	PNA	55.9	54.6	46.0	51.7	29.3	26.2	21.1
	TGR	**65.9**	**66.8**	**55.3**	**60.9**	**36.0**	**32.5**	**23.3**

Table 2. The attack success rates (%) against seven models by various transfer-based attacks. The best results are marked in bold.

the total number of tokens is $16 \times 16 = 196$ *without* the classification token or $16 \times 16 + 1 = 197$ *with* the classification token respectively. We set the scaling factor for the Attention component and MLP block to be 0.25 and the scaling factor for the QKV component to be 0.75. In addition, we set $k = 1$, which means we only consider the tokens with the maximum or the minimum back-propagated gradient magnitude as extreme tokens.

4.2. Transferability

In this section, we analyze the performance of our approach against the undefended ViTs, undefended CNNs,

Model	Attack	ViTs	CNNs	CNNs-adv
ViT-B/16	MIM-P	61.3	31.3	21.7
	VMI-P	69.1	42.8	30.9
	SGM-P	64.8	29.2	18.9
	PNA-P	70.8	42.6	29.9
	TGR-P	**76.0**	**46.7**	**33.3**
PiT-B	MIM-P	47.3	32.5	17.5
	VMI-P	59.5	46.2	35.8
	SGM-P	70.0	45.6	21.3
	PNA-P	73.1	57.8	32.7
	TGR-P	**82.3**	**68.9**	**41.3**
CaiT-S/24	MIM-P	70.3	44.0	29.3
	VMI-P	76.8	57.8	38.4
	SGM-P	85.1	49.2	29.3
	PNA-P	81.6	56.6	39.3
	TGR-P	**88.8**	**60.5**	**40.5**
Visformer-S	MIM-P	54.9	45.7	23.4
	VMI-P	64.8	56.6	32.6
	SGM-P	51.6	44.3	15.0
	PNA-P	68.8	61.8	32.3
	TGR-P	**70.4**	**64.3**	**33.5**

Table 3. The average attack success rates (%) against ViTs, CNNs, and adversarially trained CNNs by various transfer-based attacks with PatchOut strategy. The best results are marked in bold.

Methods	Deep	Middle	Shallow	Average
MIM	7.5	37.6	70.9	38.7
VMI	4.0	19.1	34.0	19.1
TGR	**1.7**	**5.1**	**6.6**	**4.4**

Table 4. The average gradient variance of ViT-B/16 by different attacking methods. The best results are marked in bold.

Model	Attack	Inc-v3	IncRes-v2	Inc-v3$_{ens4}$	IncRes-v2$_{adv}$
Res-v2	MIM	59.1	50.8	18.5	11.7
	VMI	66.3	58.6	**33.4**	**21.5**
	TGR	**75.5**	**67.6**	29.7	17.2

Table 5. The attack success rates (%) against four CNN models by various transfer-based attacks on Res-v2. The best results are marked in bold.

and adversarially trained CNN models respectively. Specifically, we attack a given source model and directly test the other different models by crafted adversarial samples, which is the black-box setting. We also test the adversarial samples on the source model itself in a white-box setting.

We first craft adversarial samples on ViTs and transfer the adversarial samples to other ViT models. As shown in Table 1, our approach achieves nearly 100% white-box attacking success rate. In addition, our method outperforms all the other baselines with a large margin of 8.8% attacking accuracy, demonstrating the high transferability of adversarial samples generated by our approach. Although VMI and SGM are attacking methods for CNN models, VMI directly regularizes the gradient on the input, and SGM utilizes skip connections, which are largely used in ViTs. VMI and SGM can achieve good attacking performance on ViTs, but they are still inferior to our approach. In addition, VMI also utilizes the idea of gradient variance reduction to generate adversarial samples, but our approach has a higher transferability, which reveals that regularizing the gradient variance in intermediate blocks of the model is effective.

Then, we study the performance of our proposed attacking method against undefended CNN models and adversarially trained CNN models to validate the cross-structure transferability. We transfer the generated adversarial samples from the source ViT models to the target CNN models. The attacking performance is summarized in Table 2. The transferability of all the attacking methods drops significantly on target CNN models because of the different architectures of ViTs and CNNs. Our approach consistently outperforms all the baselines with a margin of 6.2% on average, which demonstrates the superiority of cross-structure transferability of our proposed attacking method. The difference in the transferability between the undefended CNN models and adversarially trained CNN models is small, which demonstrates the adversarial samples crafted by the ViTs can figure out similar defects inside undefended or adversarially trained CNN models. Furthermore, our approach achieves 39.3% attacking accuracy on adversarial trained CNNs by attacking the PiT-B model on average, which shows a serious threat to the CNN defense methods. Therefore, new defense methods are required to defend the transferable adversarial samples crafted by ViT models.

Furthermore, we compose all the attacking methods with the input transformation method in ViT: the PatchOut to further improve transferability as shown in Tabel 3. Our approach, combined with the PatchOut strategy, also outperforms all the baseline methods by a considerable margin of 4.9% on average under the black-box setting, which further demonstrates the superiority of our method.

4.3. Analysis

In addition to evaluating the attacking performance compared with baselines, we aim to understand why our proposed method can achieve good performance. Moreover, we also craft adversarial samples on CNNs based on our proposed TRG method on the CNN layer to show the effectiveness of regularizing the gradient variance in intermediate blocks of the network.

We first compute the gradient variance of each block in the ViT-B/16 during the generation of adversarial samples to show the benefit of regularizing the gradient of intermediate blocks in the network. We randomly sample 100 images and compute the average gradient variance of the network. The whole network is divided into three parts – shallow level (Block 1 - Block 4), middle level (Block 5 - Block

Attention	QKV	MLP	ViTs	CNNs	CNNs-adv
-	-	-	56.2	29.0	19.5
✓	-	-	67.4	38.1	25.4
-	✓	-	64.1	33.7	23.1
-	-	✓	57.3	30.0	19.9
✓	✓	-	69.7	40.0	27.3
✓	-	✓	69.3	39.4	26.6
-	✓	✓	66.0	35.5	23.7
✓	✓	✓	73.6	42.7	29.3

Table 6. The average attack success rates (%) against ViTs, CNNs, and adversarially trained CNNs by various setting of components.

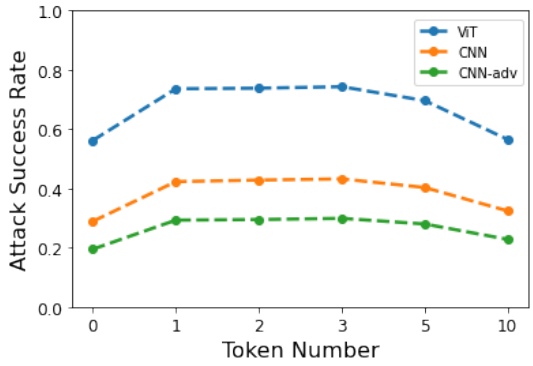

Figure 2. The attack success rates of TGR with different number of extreme tokens.

8), and deep level (Block 8 - Block 12), and we compute the average gradient of each level of the network during the adversarial sample generation. As shown in Table 4, the deep level has the least average gradient variance, and the shallow level has the largest average gradient variance because the gradient variance of the blocks increases during the back-propagation process. Furthermore, our proposed TGR achieves the least average gradient variance in all the levels compared with the baselines. Although VMI can reduce the gradient variance of each component during iteration, the gradient variance in the network is still too large. Therefore, regularizing the gradient variance in intermediate blocks by our proposed TGR can mitigate the increment of the gradient variance during the back-propagation.

In addition, we also adapt TGR to CNN attacks to show the general effectiveness of our approach. We apply TGR on the gradient of intermediate feature maps in the CNN during back-propagation. We generate adversarial samples by attacking Res-v2 and test the transferability on four CNN models, as shown in Table 5. Our proposed TGR outperforms MIM by more than 12.5% and achieves similar performance with the baseline VMI. The experiment result shows the effectiveness of our approach on attacking CNN models. Although our TGR is not designed for CNN attacks, our approach can reduce the gradient variance in intermediate blocks of the CNNs contributing to competitive performance.

4.4. Ablation Study

In this section, we do ablation studies on the influence of two factors on transferability in our proposed TGR: 1) components in ViTs. We want to figure out the contribution of each component to the transferability. 2) The extreme token number k.

Components. We craft adversarial samples by utilizing TGR on different choices of components and observe the transferability. We choose ViT-B/16 as the source model and observe the attacking performance on ViTs, undefended CNNs, and adversarially trained CNNs. As shown in Table 6, the Attention component contributes the most transferability. We believe the Attention component computes the relationships between token pairs, which exert a large influence on the output. The ablation study also validates the effectiveness of TGR on different components in ViTs.

Token Number. We measure the transferability of adversarial samples generated from the ViT-B/16 model by altering the extreme token number k in the attacking algorithm TRG. We observe from Figure 2 that when the token number increases from 0 to 1, the transferability boosts. Then the transferability drops after $k = 3$. We regard the performance improvement is due to the regularized tokens, which reduces the gradient variance inside the network. However, the original gradient will be changed largely by regularizing more tokens, contributing to the observed transferability drop. Therefore, in order to balance the performance and the efficiency, we choose $k = 1$.

5. Conclusion

In this paper, We first analyze the reasons for the low transferability of the gradient regularization-based methods. Although they regularize the gradient variance on the input, the variance in intermediate blocks of the network is still large, and thus models are stuck in local optima. To address the weakness of existing works, we propose the Token Gradient Regularization (TGR) method for transferable attacks. According to the architecture of ViTs, TGR reduces the variance of the back-propagated gradient in each internal block of ViTs and utilizes the regularized gradient to generate adversarial samples. Extensive experiments on attacking both ViTs and CNNs confirm the superiority of our approach.

Acknowledgment

The work described in this paper was supported by the National Natural Science Foundation of China (Grant No. 62206318) and the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 14206921 of the General Research Fund).

References

[1] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10231–10241, 2021. 1, 3

[2] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 589–598, 2021. 1, 3, 5

[3] 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*, pages 9185–9193, 2018. 3, 4, 5

[4] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3, 5

[6] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 5

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[8] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 5

[9] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 3, 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] 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. 5

[12] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 3, 5

[13] Jen-tse Huang, Jianping Zhang, Wenxuan Wang, Pinjia He, Yuxin Su, and Michael R Lyu. Aeon: a method for automatic evaluation of nlp test cases. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 202–214, 2022. 1

[14] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 2

[15] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019. 2, 3

[16] Zihan Liu, Yun Luo, Lirong Wu, Zicheng Liu, and Stan Z Li. Towards reasonable budget allocation in untargeted graph structure attacks via gradient debias. In *Advances in Neural Information Processing Systems*, 2022. 1

[17] Zihan Liu, Yun Luo, Zelin Zang, and Stan Z Li. Surrogate representation learning with isometric mapping for gray-box graph adversarial attacks. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 591–598, 2022. 1

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 5

[19] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021. 3

[20] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, 2021. 3

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[22] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. 1, 3

[23] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 5

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3, 5

[26] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 1, 3, 5

[27] Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael R. Lyu. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2022. 1

[28] Wenxuan Wang and Zhaopeng Tu. Rethinking the value of transformer components. In *International Conference on Computational Linguistics*, 2020. 1

[29] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021. 2, 3, 4, 5

[30] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting Adversarial Transferability through Enhanced Momentum. In *British Machine Vision Conference*, 2021. 3

[31] Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle attack: A query-efficient decision-based adversarial attack. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 156–174. Springer, 2022. 2

[32] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2668–2676, 2022. 3, 5

[33] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*, 2020. 3, 5

[34] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1161–1170, 2020. 2

[35] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2020. 1

[36] Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9024–9033, 2021. 2

[37] Weibin Wu, Hui Xu, Sanqiang Zhong, Michael R Lyu, and Irwin King. Deep validation: Toward detecting real-world corner cases for deep neural networks. In *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 125–137. IEEE, 2019. 1

[38] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 2, 5

[39] Zhuoer Xu, Guanghui Zhu, Changhua Meng, Zhenzhe Ying, Weiqiang Wang, GU Ming, Yihua Huang, et al. A2: Efficient automated attacker for boosting adversarial training. In *Advances in Neural Information Processing Systems*. 5

[40] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14993–15002, 2022. 2

[41] Zeliang Zhang, Peihan Liu, Xiaosen Wang, and Chenliang Xu. Improving adversarial transferability with scheduled step size and dual example. *arXiv preprint arXiv:2301.12968*, 2023. 2

[42] Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu. Vit-yolo: Transformer-based yolo for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2799–2808, 2021. 1

[43] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 1