Explanation-Inspired Transferable Adversarial Attacks with Layer-Wise Increment Decomposition

Shizhe Xue[1], Yiyuan Chen[2], and Wenze Shao[2(✉)]

[1] Bell Honors School, Nanjing University of Posts and Telecommunications, Nanjing, China
[2] School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
shaowenze@njupt.edu.cn

Abstract. Adversarial attacks have gained significant attention in the context of neural network security. In the realm of black-box attacks, feature-level attack methods have substantially enhanced the transferability of adversarial examples. Nevertheless, existing approaches for assessing feature importance exhibit certain weaknesses. In this paper, an explanation method termed Layer-wise Increment Decomposition (LID) for calculating neuron relevance is firstly revisited by further combining it with SoftMax Gradient-LRP (SG-LRP) and Integrated Gradients (IG), making it a more robust and precise tool for guiding adversarial attacks. Building upon this foundation and drawing inspiration from existing intermediate-layer attacks, an alternative transferable attacking loss is proposed by naively adapting the LID-based neuron relevance for intermediate layers. By further incorporating sophisticated numerical schemes for the LID-induced loss, we enhance the transferability of adversarial examples. A series of experiments conducted on both normal and defense models demonstrate that the proposed approach either outperforms or achieves comparable transferability to state-of-the-art methods.

Keywords: Deep Learning · Explainability · Adversarial Attack · Transferability · Intermediate Layer Attack

1 Introduction

Adversarial attacks on deep neural networks (DNNs) have gained significant attention due to their impact on model decision-making process security [1–3]. Adversarial examples are created by introducing imperceptible perturbations into images, which can mislead image classifiers or object detectors into making incorrect decisions. Consequently, adversarial attack is critical for analyzing neural network vulnerabilities. Since Goodfellow et al.'s Fast Gradient Sign Method (FGSM) [1], adversarial attacks have evolved rapidly, including both white-box and black-box attacks. Notably, black-box attacks remain in the early stages, presenting challenges due to the lack of access to internal model parameters. Despite these challenges, transfer-based black-box attacks

D.-S. Huang et al. (Eds.): ICIC 2025, LNCS 15845, pp. 66–77, 2025.
https://doi.org/10.1007/978-981-96-9872-1_6

have seen significant success, with adversarial examples generated by a source model often transferring to other models. This transferability has spurred extensive research into the mechanisms of DNNs, contributing to the development of explainable artificial intelligence (XAI).

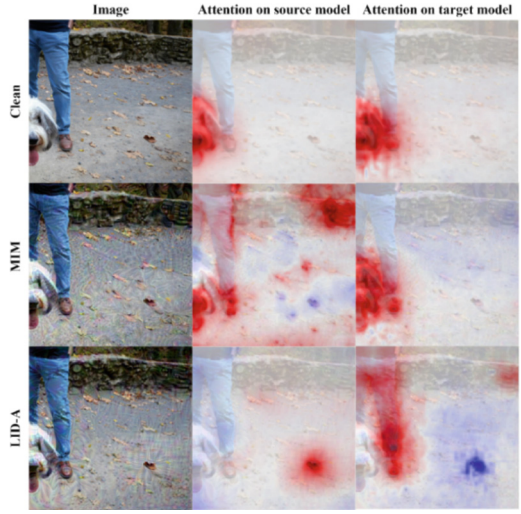

Fig. 1. Attention heatmaps of a clean image (first row) and its adversarial example generated by MIM [3] (second row) and our method (third row) are shown, corresponding to both source and target models. The source model is VGG-16, and the target model is Res-152. The attention heatmaps are obtained by aggregating relevance attribution from multiple convolutional layers, calculated using our explanation approach LID. In the heatmaps, red regions indicate positive relevance, and blue regions represent negative relevance. The results demonstrate that the proposed LID-A method successfully transfers attacks to the target model, while MIM [3] fails, as the attention heatmaps for the clean image and adversarial example are almost identical.

Recent advances in black-box attacks have benefited from XAI, exemplified by the Attack on Attention (AoA) method [4]. AoA utilizes an explanation approach known as SoftMax Gradient-Layer Relevance Propagation (SG-LRP) [5] to generate adversarial examples, back-propagating decision relevance to the input image for pixel-wise attribution. The core idea behind AoA is to disrupt the pixel-wise attribution of the iteratively generated adversarial examples. Similarly, the more recent method, Attacking SGLRP [6] introduces cosine similarity as a loss function alongside SG-LRP. However, SG-LRP faces gradient saturation due to the non-linearity of the SoftMax function, where the gradient vanishes as the output probability approaches one, degrading adversarial attack performance. Thus, a rational and reliable explanation scheme is crucial for explanation-inspired black-box attacks.

We propose LID-A, a novel neuron relevance-guided transferable adversarial attack, which extends our recent Layer-wise Increment Decomposition (LID) [7] explanation approach. The core motivation is that a 'truly' rational explanation method enhances adversarial attacks by generating precise and robust neuron relevance while improving

transferability. The LID method is enhanced by combining it with SG-LRP [5] and Integrated Gradients (IG) [8], improving robustness, precision, and class discriminability. Then, inspired by recent neuron attribution attacks [9], we introduce a transferable attack loss by adapting LID-based relevance to intermediate layers. Sophisticated numerical schemes incorporated into the LID-induced loss further improve transferability. Experiments on both normal and defense models demonstrate the superior or comparable transferability of LID-A to state-of-the-art methods, with visual examples in Fig. 1 illustrating the rationale behind the approach.

The paper is organized as follows. Section 2 presents related works on adversarial attacks and decision explanation. In Sect. 3, we revisit our LID-based explanation method boosted by integration of SG-LRP and Integrated Gradients, then detail our proposed adversarial attack method LID-A. Section 4 presents and analyzes the experimental results. Finally, Sect. 5 concludes the paper.

2 Related Works

2.1 Adversarial Attacks

The purpose of adversarial attack is to seek perturbation within a given upper bound to induce erroneous decisions of neural networks, such that $f(X+\delta) \neq y$, where f is the model, X is the input, δ is the perturbation, and y is the label. The optimization objective for adversarial attack can be naively defined as Eq. (1),

$$\text{argmax}_{\delta} \mathcal{L}(f(X+\delta), y), \text{s.t.} ||\delta||_p < \epsilon \tag{1}$$

where $||\delta||_p$ is often chosen as an infinity norm $p = \infty$, ϵ represents the upper bound of perturbation, and $\mathcal{L}$ denotes the loss function, e.g., cross-entropy in FGSM [1]. As a prominent gradient-based white-box attacking method, FGSM directly employs the gradient sign of the loss function as the perturbation, as shown in Eq. (2). Then, the Basic Iterative Method (BIM) [2] further extends FGSM by iteratively refining the perturbation direction.

$$\delta = \epsilon \cdot sign(\nabla \mathcal{L}(f(X), y)) \tag{2}$$

Transferable black-box attacking approaches aim to generate adversarial examples on a source model to successfully attack any target model. Early methods, like BIM, often show limited transferability due to overfitting to the source model. Later approaches have sought to improve transferability through various strategies. On the aspect of advanced numerical scheme, the Momentum Iterative Method (MIM) [3] enhances transferability by using momentum to guide the update direction, avoiding local minima. Additionally, other methods like DIM [10] improve transferability via data augmentation, while SGM [11] smooths the loss surface with Gaussian filtering. More recent methods focus on intermediate layers. For instance, Feature Disruptive Attack (FDA) [12] emphasizes the feature representation alignment across models, and maximizes the distortion of feature maps in the source model. Similarly, Neural Representation Distortion Method (NRDM) [13] maximizes the feature map distance between adversarial and clean images. Feature

Importance Attack (FIA) [14], as a state-of-the-art (SOTA) approach in this category, locates key features in feature maps to significantly enhance transferability. Indeed, FIA can be also interpreted as an XAI-inspired approach since built on class activation map-style explanation approaches.

The integration of adversarial attacks with neural network explanation has gained popularity for enhancing adversarial sample transferability. Besides FIA, AoA [4] maximizes the distortion of pixel-level (input layer) neuron relevance using SG-LRP [5], suppressing positively correlated neurons and activating negatively correlated ones. However, AoA works on the pixel layer instead of the more semantically critical intermediate layers, limits its transferability. Additionally, SG-LRP suffers from gradient saturation due to the non-linearity of the SoftMax function, leading to attribution inaccuracies. In contrast, Neuron Attributions Attack (NAA) [9] targets the feature layer by approximating the IG attribution method [8]. In this paper, the adversarial attack is to be taken by constructing an intermediate layer-objected loss using our LID approach.

2.2 Decision Explanation

The goal of decision explanation is to compute the relevance of each neuron to the model's output for a given input. Heatmap methods, such as Class Activation Maps (CAM) [15, 16], visualize neuron relevance in intermediate layers of DNNs. Layer-wise Relevance Propagation (LRP) [17] and related methods [5, 18] utilize the hierarchical structure of DNNs to calculate neuron relevance scores in a top-down manner, identifying each neuron's contribution to the decision. And, variants of LRP arise from different rules designed for different layers, among which, SG-LRP [5] specifically designed to distinguish attribution across classes. This method utilizes the gradient of the SoftMax function as the relevance for the final output layer, as shown in Eq. (3).

$$R(Y_i^L) = \frac{\partial P_c}{\partial Y_i^L} = \begin{cases} (1 - P_c) * P_i & \text{if } i = c \\ -P_c * P_i & \text{else} \end{cases} \tag{3}$$

Here, the number of layers in a model is denoted as $l \in \{1, ..., L\}$, and Y_i^L represents the $i{-}th$ neuron ($i{-}th$ class) of the logits layer; $R(Y_i^L)$ represents the relevance of the corresponding neuron; P_i represents the output of the SoftMax function after the logits layer for each class, i.e., probability. In much difference to SG-LRP, LRP initializes the relevance to the decision function as $R(Y_c^L) = Y_c^L$ itself for class c and zero for other classes by default. However, SG-LRP has set the relevance of non-target classes to $-P_c \cdot P_i$. These negative relevance values help reduce attribution contribution of the corresponding neurons Y_i^L, $i \neq c$, thereby enhancing attribution discriminability between classes.

Given (3), the relevance propagation rule for the fully connected and convolutional layers in SG-LRP is presented in Eq. 4, where W_{ij}^l represents the weight from the $j{-}th$ neuron of layer $l-1$ to the $i{-}th$ neuron of layer l (i.e., Y_j^{l-1} and Y_i^l). The item W^+ stands for making W valued positive, i.e., $max(W, 0)$ [18]. Observed from Eq. (4), the relevance of the $l-1$ layer, denoted as $R(Y_j^{l-1})$, can be computed based on the relevance of the l layer, i.e., $R(Y_i^l)$. As for other layers, alternative rules are applied in SG-LRP,

e.g., winner-take-all [17] is used for the max-pooling layer. In spite that the SoftMax gradient in (3) makes SG-LRP discriminative, the nonlinearity of the SoftMax function would inevitably lead to the problem of gradient saturation. It is not hard to infer that, when the probability of any class approaches one, the gradient tends to zero, causing the corresponding relevance to vanish.

$$R(Y_j^{l-1}) = \sum_i R(Y_i^l) * \frac{W_{ij}^{+l} * Y_j^{l-1}}{\sum_k W_{ik}^{+l} * Y_k^{l-1}} \tag{4}$$

3 Proposed Method

3.1 Neuron Relevance Based on Layer-Wise Increment Decomposition

The various rules in LRP, including SG-LRP, complicate the uniform computation of neuron relevance scores across layers. To streamline explanation-inspired adversarial attacks, this section revisits the more concise and reliable layer-wise increment decomposition [7], formulated within the unified framework of Taylor decomposition.

Without loss of generality, given a reference input $\hat{X}$ (e.g., zero) to a target input X, Taylor decomposition can be applied to the increment of the decision function f, i.e., $\Delta f = f(X) - f(\hat{X})$, leading to a power series of the input increment, represented as $\Delta X = X - \hat{X}$. Inheriting the layer-wise idea of LRP, the relationship between inner neurons from two consecutive layers could be described by the Taylor expansion, too. As such, Eq. (5) presents a Taylor expansion of neuron increment $\Delta Y_i^l = Y_i^l - \hat{Y}_i^l$ for each layer $l \in \{1, ..., L\}$, where $D_{ij}^l = \partial Y_i^l / \partial Y_j^{l-1}$ stands for the partial derivative of Y_i^l with respect to Y_j^{l-1}, and ϵ is the higher-order term. Following LRP and assuming ϵ can be neglected, Eq. (5) shows that upper layer neuron relevance is proportionally distributed to each lower layer neurons. In other words, each lower layer neuron absorbs the relevance from corresponding upper layer neurons, as provided in Eq. (6). Comparing this with Eq. (4) yields a concise method for calculating relevance across layers based on Taylor decomposition.

$$\Delta Y_i^l = \sum_j D_{ij}^l \cdot \Delta Y_j^{l-1} + \epsilon \tag{5}$$

$$R(Y_j^{l-1}) = \sum_i R(Y_i^l) * \frac{D_{ij}^l * \Delta Y_j^{l-1}}{\Delta Y_i^l} \tag{6}$$

Here, following LRP [17], as for a class label c, the top layer neuron's relevance is represented as $R(Y^L) = \Delta Y^L * e_c$, where e_c denotes the one-hot unit vector. After straightforward calculation, the neuron relevance for each layer l can be written as Eq. (7), where $\cdot$ represents the dot product, and $D^L \cdot ... \cdot D^{l+1}$ can be denoted as $G^{l+1} = \partial Y^L / \partial Y^l$ according to the chain rule, and $e_c \cdot G^{l+1} = \partial Y_c^L / \partial Y^l$.

$$R(Y^l) = e_c \cdot D^L \cdot ... \cdot D^{l+1} * \Delta Y^l = \frac{\partial Y_c^L}{\partial Y^l} * \Delta Y^l \tag{7}$$

The relevance attribution approach in Eq. (7) is named Layer-wise Increment Decomposition (LID). LID reveals a connection between relevance, increment, and gradient. In fact, the relevance $R(Y^l)$ in LID is equivalent to the first-order Taylor expansion of the decision increment Δf.

3.2 Improved Strategy

Inspired by SG-LRP [5], this section explores using the SoftMax function to enhance class discriminability, and applying IG [8] to address gradient saturation, thereby enhancing robustness.

$$\Delta Y_i^l = \sum_j \int_{path^l} D_{ij}^l * \mathrm{d}Y_j^{l-1} \approx \sum_j \overline{D}_{ij}^l * \Delta Y_j^{l-1} \tag{8}$$

According to the IG [8], the vector integral of the gradient along any path equals to the function increment. Multiple integration points along the path will alleviate the issue of gradient saturation. For the $l-th$ layer, the IG is represented as Eq. (8), where $path^l$ denotes any integration path, defaulting to a straight line from the reference point to the target point. The average gradient $\overline{D}_{ij}^l = 1/m * \sum_{k=1}^m D_{ij}^l(\xi_k^{l-1})$, where $\xi_k^{l-1} = \hat{Y}_j^{l-1} + k/m * \Delta Y_j^{l-1}$, is adopted in practice for the numerical calculation of the integration. Especially, the SoftMax function decomposes to $\Delta P_c = \sum_i \overline{P}_{c'} * \Delta Y^L$, where $\overline{P}_{c'} = \partial P_c / \partial Y^L$.

This paper applies IG to improve LID. Firstly, the top-layer relevance is set to the integrated gradients of SoftMax (SIG), that is $R(Y^L) = \overline{P}_{c'} * \Delta Y^L$, further enhancing robustness towards class discriminability. To be noted, SIG differs from SG-LRP in that, with the integration sampling count of 1, its average gradient reduces to the gradient, but SIG includes an additional increment term. Moreover, due to the uniformity of relevance calculation, IG can be applied across layers to mitigate gradient saturation. Finally, as guided by Eq. (7), the robust relevance is shown in Eq. (9), where the approximate average gradient $AG = \overline{P}_{c'} \cdot \overline{G}^{l+1} \approx \overline{P}_{c'} \cdot \overline{D}^L \cdot \ldots \cdot \overline{D}^{l+1}$ is obtained via the chain rule.

$$R(Y^l) = \overline{P}_{c'} \cdot \overline{D}^L \cdot \ldots \cdot \overline{D}^{l+1} * \Delta Y^l \approx AG * \Delta Y^l \tag{9}$$

3.3 Transferable Attack Guided by Neuron Relevance Explanation

With Eq. (9), this paper proposes a novel transferable attacking approach guided by neuron relevance, with the loss function defined as Eq. (10). This method is partly inspired by the recent neuron attribution attack approaches [9] targeting the intermediate layers. In addition, a method termed Intermediate-level Perturbation Decay (ILPD) [19] also reports that, as iterations progress, the feature gradients of adversarial examples gradually overfit to the source model, leading to a rapid decline in transferability. Those methods all reveal that the clean examples' features exhibit stronger transferability. Hence, the average gradient $AG(X)$ of the original example is adopted in (10) to guide optimization of the adversarial example.

$$\mathcal{L}(X_{adv}) = AG(X) \cdot \Delta Y^l(X_{adv}) \tag{10}$$

Algorithm 1 Layer-wise Increment Decomposition-based Attack (LID-A)

Require: Clean sample X, Ground-truth label c, Model f, Intermediate layer l, Integration steps m, Max perturbation ϵ, Number of iterations T, Momentum decay μ, Iteration step α, Global search factor S, Pre-attack iterations T_p, Step amplifier β

Output: Adversarial example X_{adv}

1: Initialize $X_{adv} = X, \hat{X} = 0, \overline{P}_{c'} = 0, AG = 0, \beta = \|sign(g)\| / \|g\|, S = 2.0$

2: Gaussian noise $n \sim N(0, \sigma^2)$

3: **for** k from 1 to m **do**

4: $\overline{P}_{c'} = \overline{P}_{c'} + \frac{1}{m} * P_{c'}(\hat{Y}^L + \frac{k}{m} * (Y^L - \hat{Y}^L) + n)$

5: **end for**

6: **for** k from 1 to m **do**

7: $AG = AG + \frac{1}{m} * \overline{P}_{c'} \cdot G^{l+1}(\hat{Y}^l + \frac{k}{m} * (Y^l - \hat{Y}^l) + n)$

8: **end for**

9: **for** t from 1 to $T + T_p$ **do**

10: **if** $t == T_p$: $X_{adv} = X$

11: $\mathcal{L}(X_{adv}) = AG(X) \cdot \Delta Y^l(X_{adv})$

12: $g = \mu * g + \nabla \mathcal{L}(X_{adv})$

13: **if** $t \le T_p$: $X_{adv} = X_{adv} + S * \alpha * \beta * g$

14: **else**: $X_{adv} = X_{adv} + \alpha * \beta * g$

15: $X_{adv} = clip_{\epsilon}(X_{adv})$

16: **end for**

17: **return** X_{adv}

The method, Layer-wise Increment Decomposition-based Attack (LID-A), offers strong interpretability guided by LID-based explanations. Note that, a key aspect is the gradient-based optimization technique that enhances transferability. Inspired by the trade-off between accuracy and efficiency in MIM [3], we employ a momentum-based optimization scheme to minimize LID-induced attack loss. Specifically, a pre-attack strategy is initially employed to facilitate the convergence of momentum direction, given that momentum is an accumulated variable over time-series iterations. The stability of momentum is closely linked to the number of accumulated iterations. By evaluating the cosine similarity of momentum across different iteration rounds, we observe.

that the cosine similarity is relatively low in the initial iterations; however, as the iterations progress, the cosine similarity of the momentum gradually converges, further substantiating the temporal characteristics of the momentum. This issue is also

Table 1. The results of attacking normal models.

Source	Method	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-152	VGG-16	VGG-19
Inc-v3	MIM	99.8%	43.4%	41.8%	40.8%	32.7%	43.3%	39.5%
	FDA	81.9%	43.4%	36.2%	43.2%	30.0%	34.3%	30.5%
	FIA	96.0%	73.5%	71.4%	63.7%	60.9%	63.1%	63.0%
	NAA	98.7%	86.6%	84.8%	76.3%	70.2%	75.3%	73.7%
	LID-A	**98.8%**	**87.4%**	**87.1%**	**76.0%**	**72.1%**	**76.6%**	**76.9%**
Res-152	MIM	67.9%	58.8%	55.8%	94.7%	100.0%	80.9%	80.0%
	FDA	73.3%	62.1%	58.2%	88.5%	95.5%	79.5%	78.4%
	FIA	84.5%	79.2%	77.6%	95.6%	99.5%	87.9%	86.5%
	NAA	84.4%	80.6%	79.5%	94.8%	98.8%	87.4%	86.5%
	LID-A	**91.3%**	**86.5%**	**87.0%**	**97.6%**	**99.2%**	**92.4%**	**93.1%**
IncRes-v2	MIM	62.1%	53.4%	98.6%	47.2%	40.0%	47.7%	45.3%
	FDA	69.2%	67.5%	78.2%	61.5%	51.6%	49.5%	44.8%
	FIA	60.1%	54.9%	70.0%	53.6%	48.7%	53.5%	52.2%
	NAA	84.2%	80.4%	94.4%	73.4%	70.4%	73.5%	73.9%
	LID-A	**85.8%**	**81.3%**	**94.9%**	**73.7%**	**71.3%**	**74.6%**	**74.0%**
VGG-16	MIM	88.2%	88.3%	82.2%	93.5%	89.1%	99.9%	99.7%
	FDA	81.5%	82.8%	68.7%	81.9%	78.3%	95.1%	95.4%
	FIA	95.9%	95.8%	92.9%	95.8%	94.5%	99.9%	99.5%
	NAA	95.3%	93.1%	90.6%	95.6%	93.8%	98.8%	99.0%
	LID-A	**97.1%**	**97.0%**	**95.7%**	**98.0%**	**97.1%**	**99.9%**	**99.6%**

highlighted by the Global Momentum Initialization (GI) [20], which introduces a pre-attack to accumulate momentum. While, GI relies entirely on gradients to optimize the momentum direction, incorporating feature-level information derived from LID into the optimization process is expected to enable faster convergence, more accurate direction, and improved attack performance. On the other hand, since MIM utilizes a rough sign method to optimize the gradient, this results in a considerable deviation between the actual momentum direction and the gradient direction [21]. To mitigate this discrepancy, we incorporate a step amplifier alongside the gradient to further reduce the deviation between the momentum update direction and the gradient direction. To mitigate overfitting, Gaussian noise is introduced into the samples during optimization, utilizing the Smooth-Grad (SGM) [11] technique. The details are outlined in Algorithm 1, where the reference point is chosen as a black image and the integration path follows a straight line between the reference and endpoint.

4 Experimental Results

4.1 Experimental Setup

Table 2. The results of attacking defense models.

Source	Method	Adv-Inc-v3	Adv-IncRes-v2	Ens3-Inc-v3	Ens4-Inc-v3	Ens-IncRes-v2
Inc-v3	MIM	23.1%	17.9%	16.8%	16.5%	7.8%
	FDA	19.7%	12.6%	9.2%	12.3%	5.2%
	FIA	48.5%	49.7%	39.1%	39.3%	24.5%
	NAA	61.6%	60.6%	48.8%	50.9%	**29.7%**
	LID-A	**67.7%**	**65.3%**	**52.2%**	**51.4%**	29.1%
Res-152	MIM	53.2%	47.5%	48.6%	50.9%	36.4%
	FDA	61.9%	42.2%	50.3%	50.2%	36.8%
	FIA	76.8%	71.1%	72.4%	72.2%	62.3%
	NAA	76.6%	70.3%	74.5%	72.9%	62.1%
	LID-A	**78.7%**	**79.0%**	**75.3%**	**74.2%**	**65.7%**
IncRes-v2	MIM	25.6%	28.7%	21.1%	21.3%	12.2%
	FDA	34.2%	29.6%	16.9%	15.2%	8.1%
	FIA	38.0%	41.3%	36.5%	35.0%	27.8%
	NAA	66.2%	67.9%	58.2%	53.5%	46.5%
	LID-A	**71.1%**	**71.6%**	**62.6%**	**59.2%**	**50.2%**
VGG-16	MIM	82.8%	73.7%	80.2%	77.1%	66.2%
	FDA	72.5%	60.7%	66.8%	68.3%	57.8%
	FIA	91.7%	87.3%	90.5%	89.3%	84.7%
	NAA	91.7%	88.9%	91.0%	90.8%	82.6%
	LID-A	**92.5%**	**91.2%**	**93.0%**	**92.2%**	**86.8%**

This paper conducts transfer attack experiments using 1000 images from the ImageNet [22] dataset, testing both normal models and adversarially trained defense models. Normal models include Inception-V3(Inc-v3), Inception-V4(Inc-v4), Inception-ResNet-V2(IncRes-v2), ResNet-V1-50(Res-50), ResNet-V1-152(Res-152), VGG-16, and VGG-19. The defense models are Adv-Inc-v3, Adv-IncRes-v2, Ens3-Inc-v3, Ens4-Inc-v3 and Ens-IncRes-v2. Benchmark methods compared with our method include MIM [3], FDA [12], FIA [14], NAA [9].

The maximum perturbation is limited to $\epsilon = 16$, the iteration number $T = 10$, and the iteration step $\alpha = 1.6$. Following the settings in FIA [14], for all feature-level attacks, the chosen feature layers are Mixed_5b for Inc-v3, Conv3_3 for VGG-16, Conv_4a for IncRes-v2, and Block2_Unit8 for Res-152. For LID-A, the integration steps $m = 10$,

momentum decay factor $\mu = 1$, the noise magnitude $\sigma = 1.0 * \epsilon$ and the Pre-attack iterations $T_p = 5$.

4.2 Transferable Attacks

In the transfer attack experiments, adversarial examples are generated from source models Inc-v3, Res-152, IncRes-v2, and VGG-16, targeting normal models (Table 1) and defense models (Table 2). Table 1 shows the attack results on normal models, with the leftmost column representing the source models and the first row representing the target models. LID-A slightly trails MIM in white-box attack success rate but outperforms most other methods in black-box attacks. The average attack success rates of LID-A across different source models are 82.1, 92.4, 79.4, and 97.8, respectively, showing improvements over FIA by 11.9, 5.2, 23.2, and 1.4, and over NAA by 1.3, 5.0, 0.8, and 2.6, respectively.

The results of attacking defense models are presented in Table 2. LID-A outperforms all other methods, with average success rates of 53.1, 74.6, 62.9, and 91.1 on different source models, respectively. Compared to FIA, these rates exhibit improvements of 12.9, 3.6, 27.2, and 2.4, and compared to NAA, the improvements are 2.8, 3.3, 4.5, and 2.1, respectively. This demonstrates the robust attack capability of LID-A against defense models. The results on both normal and defense models demonstrate LID-A's high transferability, emphasizing the accuracy and robustness of the LID explanation.

4.3 Ablation Study

This section conducts experiments to analyze the impact of various improvements to LID-A in Algorithm 1. The source model is VGG-16 only, and the following degenerated losses are tested: (1) LID-A w/o SIG: removes the SIG top-layer relevance, by $R(Y^L) = \Delta Y^L * e_c$ instead; (2) LID-A w/o IG: eliminates IG in intermediate layer, using Eq. (7) for computation; (3) LID-A w/o Noise: excludes Gaussian smoothing; (4) LID-A w/o GI: removes GI, by adopting FGSM as in BIM and MIM.

The experimental results are presented in Table 3. The introduction of IG in the intermediate layer and GI significantly enhances the transferability of adversarial samples. Introducing Gaussian noise also contributes to improved transferability to some

Table 3. The results for ablation study.

Method	Inc-v4	IncRes-v2	Res-152	Ens4-Inc-v3	Ens-IncRes-v2
LID-A	97.0%	97.5%	97.1%	90.2%	79.8%
w/o SIG	96.7% (−0.3)	94.5% (−3.0)	96.0% (−1.1)	89.5% (−0.7)	79.5% (−0.3)
w/o IG	94.5% (−2.5)	93.1% (−4.4)	94.5% (−2.6)	86.4% (−3.8)	72.9% (−6.9)
w/o Noise	95.0% (−2.0)	93.5% (−4.0)	94.5% (−2.6)	87.3% (−2.9)	76.9% (−2.9)
w/o GI	95.7% (−1.3)	92.5% (−5.0)	95.4% (−1.7)	85.3% (−4.9)	70.8% (−9.0)

extent. Meanwhile, the introduction of SIG only brings marginal improvement, possibly because most samples do not exhibit category interference, eliminating the need for enhanced class distinctiveness.

5 Conclusion

This paper proposes an explanation-inspired adversarial attack method LID-A. Firstly, the paper revisits a Layer-wise Increment Decomposition (LID) method to calculate neural relevance, improving class discriminability and addressing gradient saturation. LID establishes a connection between relevance and gradient concepts. Based on this, the paper utilizes feature level neural relevance of clean samples to guide adversarial attacks and combines sophisticated gradient optimization methods to further enhance transferability. This method is evaluated on various models and compared against existing methods, demonstrating its comparable or superior performance. With significant progress being made, there is still room for further research in relevance-guided transfer attacks. Future work in this field may involve a deeper understanding of relevance and exploration of more feature-based attack methods.

Acknowledgments. The present work is supported in part by the NSF of China (No. 92470126).

References

1. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
2. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security, pp. 99–112 (2018)
3. Dong, Y., Liao, F., Pang, T., et al.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193 (2018)
4. Chen, S., He, Z., Sun, C., et al.: Universal adversarial attack on attention and the resulting dataset damagenet. IEEE Trans. Pattern Anal. Mach. Intell. **44**(4), 2188–2197 (2020)
5. Iwana, B.K., Kuroki, R., Uchida, S.: Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 4176–4185. IEEE (2019)
6. Chen, Z., Dai, R., Liu, Z., et al.: An interpretive adversarial attack method: Attacking softmax gradient layer-wise relevance propagation based on cosine similarity constraint and TS-invariant. Neural Process. Lett. **55**(4), 4623–4639 (2023)
7. Chen, Y., Li, J., Shao, W., et al.: Layer-wise increment decomposition-based neuron relevance explanation for deep networks. Acta Automatica Sinica **50**(10), 20f49−2062 (2024)
8. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328 (2017)
9. Zhang, J., Wu, W., Huang, J.t., et al.: Improving adversarial transferability via neuron attribution-based attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14993–15002 (2022)
10. Xie, C., Zhang, Z., Zhou, Y., et al.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2730–2739 (2019)

11. Wu, L., Zhu, Z.: Towards understanding and improving the transferability of adversarial examples in deep neural networks. In: Asian Conference on Machine Learning, pp. 837–850. PMLR (2020)
12. Ganeshan, A., BS, V., Babu, R.V.: FDA: feature disruptive attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8069–8079 (2019)
13. Naseer, M., Khan, S.H., Rahman, S., et al.: Task-generalizable adversarial attack based on perceptual metric. arXiv preprint arXiv:1811.09020 (2018)
14. Wang, Z., Guo, H., Zhang, Z., et al.: Feature importance-aware transferable adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7639–7648 (2021)
15. Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
16. Zhou, B., Khosla, A., Lapedriza, A., et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
17. Bach, S., Binder, A., Montavon, G., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
18. Montavon, G., Lapuschkin, S., Binder, A., et al.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recogn. **65**, 211–222 (2017)
19. Li, Q., Guo, Y., Zuo, W., et al.: Improving adversarial transferability by intermediate-level perturbation decay. arXiv preprint arXiv:2304.13410 (2023)
20. Wang, J., Chen, Z., Jiang, K., et al.: Boosting the transferability of adversarial attacks with global momentum initialization. Expert Syst. Appl. **255**, 124757 (2024)
21. Cheng, Y., Song, J., Zhu, X., et al.: Fast gradient non-sign methods. arXiv preprint arXiv:2110.12734 (2021)
22. Russakovsky, O., Deng, J., Su, H., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**, 211–252 (2015)