An Attention-Guided Dynamic Stride Projection Attack Framework

Yiming Chen

Hangzhou Anheng Information Technology Co., Ltd. Zhejiang Hangzhou 310000

Abstract: ***Black-box transfer attacks represent a critical paradigm in adversarial machine learning, whereby adversarial examples crafted against a surrogate (source) model are deployed to deceive unknown target models. This approach serves as a vital tool for conducting security audits and enhancing the robustness of deep neural networks. A primary challenge, however, lies in the tendency of iterative attack methods to overfit the specific characteristics of the source model, thereby diminishing their cross-model transferability. To mitigate this issue, this paper proposes a novel Dynamic Step-length Projection Attack method based on Attention Guidance (DSP-Attack). The core of our method is twofold. First, it introduces a dynamic mechanism that adaptively adjusts the projection step size during the iterative perturbation generation process. This is motivated by the observation that initial perturbations often possess higher transferability; the proposed strategy thus prioritizes these early stages by employing a larger effective step size, while strategically curtailing potentially overfitting and ineffective perturbations in later iterations. Second, the method incorporates an attention guidance mechanism, derived from the source model's gradient-weighted class activation mapping, to focus the perturbation budget on regions that the model deems most salient for its predictions. This ensures that the adversarial modifications are applied to semantically meaningful and model-sensitive areas, thereby increasing the likelihood of the attack transferring to other architectures. Comprehensive experiments on the ImageNet dataset demonstrate the superior efficacy of our approach. The proposed DSP-Attack achieves significant performance improvements in transferability across a diverse set of target models, including ResNet, VGG, and DenseNet architectures, outperforming several state-of-the-art baseline methods. These findings affirm that jointly optimizing the attack trajectory via dynamic step-length control and spatial attention guidance is a potent strategy for crafting highly transferable adversarial examples.***

Keywords: Adversarial Attack, Black-Box Transferability, Dynamic Step-length, Attention Guidance, Model Security, Deep Learning.

1. INTRODUCTION

In recent years, neural networks have achieved remarkable progress in image recognition and natural language processing, yet they are also found to be vulnerable to adversarial attacks. By adding imperceptible perturbations to inputs, adversarial attacks cause models to output incorrect predictions, exposing the security flaws of deep learning and posing potential risks in critical applications such as autonomous driving and face recognition, thereby highlighting the importance of studying adversarial attack and defense mechanisms. Current attack methods, including gradient-based, input-transformation, and feature-interference strategies, show some effectiveness but are often limited in black-box scenarios due to insufficient transferability. Enhancing the cross-model transferability of adversarial examples is essential for understanding model vulnerabilities and building robust systems. Li, Lin, and Zhang (2025) proposed a privacy-preserving framework combining federated learning and differential privacy for personalized advertising, addressing critical data confidentiality concerns [1]. In the domain of urban design, Xu (2025) introduced CivicMorph, a generative modeling approach for public space form development [2]. Concurrently, Tu (2025) presented SmartFITLab, an intelligent platform designed for the execution and validation of 5G field interoperability testing, enhancing network infrastructure robustness [3]. For human resource technology, Xie and Liu (2025) developed EvalNet, a system utilizing sentiment analysis and multimodal data fusion to process recruitment interviews [4]. Zhu (2025) explored language agents with TaskComm, a task-oriented agent aimed at optimizing workflows for small businesses [5]. Further supporting small enterprises, Zhang (2025) investigated reinforcement learning techniques for automated ad campaign optimization in "Learning to Advertise" [6]. Hu (2025) contributed to 3D content creation for small and medium-sized enterprises with "Learning to Animate," focusing on few-shot neural editors [7]. Developer tooling for large language models was advanced by Zhang (2025) through InfraMLForge, enabling rapid LLM development and scalable deployment [8]. In healthcare AI, Ding and Wu (2024) conducted a systematic review on self-supervised learning applications for processing ECG and PPG biomedical signals [9]. Addressing challenges in recommendation systems, Wang (2025) proposed a joint training method for propensity and prediction models using targeted learning, specifically for data missing not at random (MNAR) scenarios [10]. Lin (2025) addressed product management needs in AI systems by introducing a framework for digital experience

observability [11]. Finally, foundational applications were reinforced by Chen (2023), who discussed the utilization of data mining techniques within broader data analysis contexts [12].

2. DYNAMIC STEP-SIZE PROJECTION ATTACK METHOD BASED ON ATTENTION GUIDANCE

2.1 Overall model structure

The proposed attention-guided dynamic step-size projection attack comprises two core modules: a dynamic step-size adjustment module and an attention-guided region projection module.

In the dynamic step-size adjustment module, the method adaptively tunes the perturbation step size according to the step-size ratio during the attack and the gradient trend of the loss function. A large step size is used at the beginning to rapidly approach the decision boundary, then gradually reduced to stabilize the optimization. When the gradient direction stabilizes, the step size is moderately enlarged to accelerate convergence; when the gradient fluctuates violently, the step size is shrunk to prevent deviation from the optimal path. This mechanism enhances the flexibility and stability of the attack.

In the attention-guided region projection module, Grad-CAM-generated attention heatmaps highlight the model's focus regions and are used to create region masks. These masks guide the redistribution of cropped redundant perturbations into key areas, improving perturbation utilization and transferability. Compared with traditional methods that directly discard cropped perturbations, this strategy reuses early, general-purpose perturbations, effectively boosting cross-model attack success rates.

Finally, the two modules work in tandem to apply the optimized perturbations to the original image step by step, generating adversarial examples that achieve high attack success rates across multiple target models. The following subsections will detail the implementation of each module.

2.2 Dynamic Step-Size Adjustment Strategy

To fully exploit the more transferable perturbation information in the early attack phase and mitigate overfitting to the source model in the later phase, this paper proposes a dynamic step-size adjustment strategy. Based on the current iteration stage and the gradient trend of the loss function, the strategy adaptively adjusts the perturbation step size: a large step size is used at the beginning to quickly approach the decision boundary, then progressively reduced for fine-grained optimization. Simultaneously, the step size is dynamically tuned according to loss variations, ensuring efficient convergence and enhancing both the aggressiveness and black-box transferability of adversarial examples.

First, the step size is computed using a decay factor $\gamma \in (0,1)$ to control the decay rate; the dynamic step-size formula is defined as:

$$\alpha_t = \epsilon \cdot \frac{1-\gamma}{1-\gamma^T} \cdot \gamma^{t-1} \tag{1}$$

where γ^{t-1} denotes γ raised to the $t-1$ -th power, serving as the exponential decay term that ensures the step size decreases with iteration count, ϵ is the maximum perturbation, and $\frac{1-\gamma}{1-\gamma^T}$ is a normalization coefficient. After normalization, the step size is allocated proportionally to the decay factor each round, ensuring that the total step size over the entire attack equals the maximum perturbation ϵ.

To enhance attack stability and adversarial transferability, this paper adaptively optimizes the step size according to the gradient variation of the loss function and introduces multiple random perturbations into the gradient computation to obtain a more stable gradient estimate and strengthen the attack's generalization ability. Specifically, assuming the current adversarial sample is x_t , when computing the gradient, the paper first applies several Gaussian perturbations $\eta_i \sim \mathcal{N}(0, \sigma^2)$ to the input sample, then calculates the loss-function gradient for each perturbed sample $x_t + \eta_i$, and finally averages the gradients of all perturbed versions to obtain a more stable gradient $\nabla_x \mathcal{L}(x_t + \eta_i, y)$:

$$g_t = \frac{1}{N+1}\sum_{i=0}^{N} \nabla_x \mathcal{L}(x_t + \eta_i, y) \tag{2}$$

where N is the number of Gaussian perturbation samples.

To improve attack efficiency, after obtaining a stable gradient direction, the paper dynamically adjusts the step size α. A fixed step size can lead to instability: too large a step may overshoot the optimal direction, while too small a step slows convergence. Therefore, the paper proposes an adaptive step-size strategy based on the gradient change rate, allowing the step size to adjust dynamically with gradient information. Specifically, the step-size adjustment is as follows:

$$\alpha_t' = \frac{\alpha_t}{1+\mu\left(\frac{\|g_t - g_{t-1}\|_2}{E_{t-1}+\epsilon} - 1\right)} \tag{3}$$

where α_t is the step size computed by Equation (3), μ controls the adjustment magnitude, $\|g_t - g_{t-1}\|_2$ represents the gradient change during the current attack, and ϵ is a small constant (1010^{-8}) to prevent division by zero. E_{t-1} is the exponential moving average (EMA) of gradient changes, updated as follows:

$$E_t = \lambda \cdot E_{t-1} + (1-\lambda) \cdot \|g_t - g_{t-1}\|_2 \tag{4}$$

where λ is the smoothing factor that controls the weighting between current and historical data; the smaller λ is, the more sensitive E_t is to recent changes, while a larger λ smooths the trend and relies more on historical data. Its value is computed by:

$$\lambda = \frac{2}{T+1} \tag{5}$$

where T denotes the total number of iterations. The core idea of step-size adjustment is that when the gradient changes sharply, the optimization direction is likely still unstable, so the step size should be reduced to prevent overly rapid adversarial updates that could cause the attack to fail. When the gradient changes little, the attack direction is relatively stable, and the step size can be increased to improve attack efficiency.

To prevent an excessive step size from degrading the attack, this method clips perturbations that exceed the step-size range and retains them for processing in the next stage's attention-guided region projection strategy, projecting them onto the model's most sensitive regions to further enhance the adversarial example's attack strength and transferability. Through this design, the dynamic step-size adjustment strategy enables the perturbation magnitude in each iteration to be adaptively tuned, improving its transferability across different target models. The gradual reduction of the step size effectively avoids overfitting, and while ensuring the adversarial example remains effective, it minimizes unnecessary perturbations as much as possible.

2.3 Attention-Guided Region Projection Strategy

In the first step of this strategy, Grad-CAM is used to generate an attention map for the input sample. Grad-CAM leverages gradient information from convolutional layers to help identify which image regions contribute most to the model's decision. Specifically, Grad-CAM computes a class activation map, performs a weighted summation, and produces a heatmap that visualizes the activation level of different regions. This process is calculated as follows:

$$w_k^c = \frac{1}{z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$
$$L_{Grad-CAM}^c = \mathrm{ReLU}\left(\sum_k w_k^c A^k\right) \tag{6}$$

Here, A_{ij}^k is the k-th feature map of the last convolutional layer in CNN , w_k^c is the gradient-based weight for that feature map, y^c denotes the predicted score for class c , Z is a normalization factor, and ReLU(·) ensures only regions with positive contributions to the target class are retained. The resulting Grad-CAM attention map reveals how much the model attends to each part of the input image, with highlighted areas indicating the most sensitive regions for the model's decision.

Based on the Grad-CAM attention map, this method further extracts the regions that most influence the model's decision—i.e., the key regions. These key regions typically correspond to the parts of the heatmap with the highest activation values, reflecting the model's most sensitive areas. By setting a threshold τ, the most critical parts for the decision are extracted to form a mask matrix M. This mask matrix M guides the projection of clipped perturbations in each attack iteration. The mask matrix M is computed as:

$$M(i,j) = \begin{cases} 1, & \text{GradCAM}(i,j) > \tau \\ 0, & \text{GradCAM}(i,j) \leq \tau \end{cases} \tag{7}$$

This mask matrix indicates whether each pixel in the image belongs to a key region. For every pixel position (i, j), if its corresponding Grad-CAM value exceeds the threshold τ, the position is deemed important, and the corresponding entry in the mask matrix is set to 1; otherwise, the mask is 0.

In each attack iteration, this paper employs a mask matrix M to guide the perturbation application, so the perturbation is influenced not only by the loss function but also dynamically adjusted according to attention weights. During the dynamic projection phase, the excess perturbation clipped in the previous round is combined with the current perturbation to intensify the attack on key regions. For mask entries equal to 1, gradient projection fuses historical and current perturbations to apply strong perturbations; in non-critical regions (value 0) only small perturbations are added, thereby improving overall attack effectiveness and transferability. Specifically, the method reuses the excess perturbation δ_{excess} clipped in the previous iteration, dynamically projecting it onto the most sensitive regions and updating it together with the current perturbation. The formula is as follows:

$$\delta_t(x) = \text{Clip}_\epsilon\left(\alpha'_t \cdot \text{sign}(g_t) + \Pi_M\left(\beta \cdot \delta_{excess}(x)\right)\right) \tag{8}$$

Here, $\delta_t(x)$ is the perturbation at round t, α'_t is the current step size computed via Equation (3), g_t is the gradient information calculated by Equation (2), M is the mask matrix marking key regions in the attention map, β is a dynamically adjusted factor controlling the influence of the previous round's clipped perturbation, and $\delta_{excess}(x)$ is the excess perturbation clipped in the previous iteration. The excess perturbation $\delta_{excess}(x)$ is adjusted according to β and the mask matrix M, ensuring it is applied only to regions most sensitive to model decisions and has no effect on non-critical regions. Π_M denotes projecting the excess perturbation into the constraint mask matrix M. Cli p_ϵ clips the perturbation so each pixel value stays within the maximum perturbation bound, while temporarily storing the clipped excess for use in the next iteration.

Table 1: Attack success rates (%) against normally trained models

Model	Attack methods	Inc − v3	Inc − v4	IncRes −v2	Res-1 52	Res- 50	Res- 101	AVG
Inc-v 3	MI	100.0 *	51.1	46.9	39. 3	46.6	41.6	54.3
	S^2I	100.0 *	64.8	59.6	48.9	57.0	52.4	63.8
	VMI	100.0 *	74.5	70.7	63. 3	68.0	62.5	73.2
	Admix	100.0 *	78.6	73.2	68.0	74.3	69.2	77.2
	PI	100.0 *	52.1	34.7	38.5	44. 2	40.9	51.7
	SI-NI	100.0 *	76.3	75.1	67.6	73.0	69.8	77.0
	GE-AdvGAN	100.0*	90.9	75.1	88.1	74.0	69.9	83. 0
	Ours	100.0 *	81.4	79.6	75.1	74.6	70.2	80. 2
Inc-v 4	MI	61.6	100.0 *	45. 3	42.4	45.2	42.8	56.2
	S^2I	71.9	100.0 *	55.6	49.4	55.6	48.8	63.6
	VMI	83.0	100.0*	76.1	68.8	71.4	68. 2	77.9
	Admix	88.4	100.0 *	82.9	77.8	79.6	75.9	84.1
	PI	52.6	100.0 *	30.7	36.8	40.7	37.0	49.6
	SI-NI	85.5	99. 9*	79.1	72.8	75.2	73.0	80.9
	GE-AdvGAN	88.4	100.0 *	69.1	81.4	81.4	80. 3	83.4
	Ours	90.1	100.0 *	85.6	82.1	82.5	81.1	86.9
IncRe s-v2	MI	61.4	53.9	99.3*	45.4	50.2	45.0	59.2
	S^2I	75.9	66.8	98.3*	55.5	61.9	57.9	69.4
	VMI	80.7	76.4	99.3*	65.7	69. 3	68. 2	76.6
	Admix	90.9	88.8	99.5*	83.4	84.5	84. 2	88. 6

	PI	53.8	48.1	98.6*	38.1	40.7	39.1	53.1
	SI-NI	87.8	83. 1	99. 9*	75.7	78.5	77.1	83.7
	GE-AdvGAN	87.4	83.4	98.9*	80. 3	79.2	79.0	84.7
	Ours	91.9	90.0	99. 9*	86.6	85.4	85.9	89.9

Superscript "*" indicates white-box attacks.

3. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the performance of the proposed method against normally trained models, Inc-v3, Inc-v4, and IncRes-v2 are selected as white-box models to generate adversarial examples. These adversarial examples are then used to attack the same models and two additional black-box models: Res-50 and Res-101. As shown in Table 3-1, compared with seven mainstream adversarial attack methods, the proposed method significantly improves black-box transferability without reducing the attack success rate on white-box models. For example, when Inc-v3 is used as the white-box model to generate adversarial examples and attack other black-box models, the proposed method raises the average attack success rate from 51.7% to 80.2% compared with the classic PI-FGSM algorithm. Compared with the advanced methods Admix and VMI-FGSM, the average attack success rate is increased by 3% and 7% , respectively. Moreover, compared with the recent method GE-AdvGAN, the proposed method achieves higher average attack success rates when Inc-v4 and IncRes-v2 are used as white-box models.

4. CONCLUSION

With the widespread adoption of deep neural networks, their vulnerability to adversarial examples has drawn significant attention. Existing attack methods often struggle to balance success rate and transferability, limiting black-box attack effectiveness. To address this, the paper proposes an attention-guided dynamic step-size projected attack method from the perspective of gradient optimization. By employing a dynamic step-size strategy, the method accelerates early perturbation convergence to the decision boundary and reduces later overfitting, while leveraging attention mechanisms to project highly transferable perturbations onto key regions. Experimental results demonstrate significant advantages in improving transferability, attack success rate, and stealth.

REFERENCES

[1] Li, X., Lin, Y., & Zhang, Y. (2025). A Privacy-Preserving Framework for Advertising Personalization Incorporating Federated Learning and Differential Privacy. arXiv preprint arXiv:2507.12098.
[2] Xu, Haoran. "CivicMorph: Generative Modeling for Public Space Form Development." (2025).
[3] Tu, Tongwei. "SmartFITLab: Intelligent Execution and Validation Platform for 5G Field Interoperability Testing." (2025).
[4] Xie, Minhui, and Boyan Liu. "EvalNet: Sentiment Analysis and Multimodal Data Fusion for Recruitment Interview Processing." (2025).
[5] Zhu, Bingxin. "TaskComm: Task-Oriented Language Agent for Efficient Small Businesses Workflows." (2025).
[6] Zhang, Yuhan. "Learning to Advertise: Reinforcement Learning for Automated Ad Campaign Optimization for Small Businesses." (2025).
[7] Hu, Xiao. "Learning to Animate: Few-Shot Neural Editors for 3D SMEs." (2025).
[8] Zhang, Yuhan. "InfraMLForge: Developer Tooling for Rapid LLM Development and Scalable Deployment." (2025).
[9] Ding, C.; Wu, C. Self-Supervised Learning for Biomedical Signal Processing: A Systematic Review on ECG and PPG Signals. medRxiv 2024.
[10] Wang, Hao. "Joint Training of Propensity Model and Prediction Model via Targeted Learning for Recommendation on Data Missing Not at Random." AAAI 2025 Workshop on Artificial Intelligence with Causal Techniques. 2025.
[11] Lin, Tingting. "Digital Experience Observability in AI-Enhanced Systems: A Framework for Product Managers." ResearchGate, Mar (2025).
[12] Chen, Rensi. "The application of data mining in data analysis." International Conference on Mathematics, Modeling, and Computer Science (MMCS2022). Vol. 12625. SPIE, 2023.