Appendix of "Improving Integrated Gradient-based Transferable Adversarial Examples by Refining the Integration Path"

Yuchen Ren[1], Zhengyu Zhao[1*], Chenhao Lin[1], Bo Yang[2] Lu Zhou[3], Zhe Liu[3], Chao Shen[1]

[1]Xi'an Jiaotong University, [2]Information Engineering University, [3]Nanjing University of Aeronautics and Astronautics
{ryc98, zhengyu.zhao, linchenhao, chaoshen}@stu.xjtu.edu.cn, yangbo_hn@163.com, {lu.zhou, zhe.liu}@nuaa.edu.cn

Proof of Proposition 1

This section will prove that the non-monotonic path will degrade the attack performance. To make it more intuitive, we use BIM as the attack and consider the scenario before overfitting occurs during the iterative process

$Proof$. Considering a surrogate model f, an interpolation coefficient α, an adversarial example x_t at the t-th iteration and a baseline b, according to the Definition 1, a nonmonotonic path is charactirized by

$$\text{sign}(\frac{\partial f(x_t + \alpha \cdot (x_t - b))}{\partial x_t}) \neq \text{sign}((x_t - b) \cdot \frac{\partial f(x_t + \alpha \cdot (x_t - b))}{\partial x_t}), \tag{1}$$

where the gradient and the product of the gradient and the path (element-wise multiplication) have inconsistent element symbols.

Assuming a special case when an image with only one pixel, the update direction of the adversarial example at t-th iteration is $\text{sign}((x_t - b) \cdot \frac{\partial f(x_t + \alpha \cdot (x_t - b))}{\partial x_t})$. Since the sign of gradient $\text{sign}(\frac{\partial f(x_t + \alpha \cdot (x_t - b))}{\partial x_t})$ represents the direction of larger loss, the update direction $\text{sign}((x_t - b) \cdot \frac{\partial f(x_t + \alpha \cdot (x_t - b))}{\partial x_t})$, with completely opposite sign, will lead to the lower loss and thus degrade the attack performance.

In the normal case, although not all the sign of elements are opposite, those opposite parts will also prevent the increase of loss, with the extent depending on the number of opposing elements.

Comparison with PAM and NAA

It is important to note that PAM (Zhang et al. 2023) is not an IG-based attack, as the value of its "path" does not contribute to the numerical computation of the gradient during the iterative process. While for NAA (Zhang et al. 2022), the main difference lies in how IG is applied: NAA uses IG to compute feature weights, while our method integrates IG directly into the iterative process. Additionally, NAA requires selecting specific feature layers, with its original implementation focused only on Inception and ResNet architectures. In contrast, our method is model-agnostic and applicable to various CNNs and ViTs.

*Corresponding Author.

We provide a detailed comparison between our method, MuMoDIG, and other concept-related approaches, including PAM and NAA. The results, as shown in Table 1, clearly demonstrate the superiority of our method across different model architectures and defense mechanisms. Specifically, MuMoDIG achieves significantly higher success rates on both CNNs and ViTs compared to PAM and NAA. For instance, MuMoDIG outperforms PAM on CNNs by a margin of 16.0% and on ViTs by 25.5%. Similarly, against adversarial training (AT) and neural representation purifiers (NRP), MuMoDIG consistently demonstrates stronger performance, surpassing PAM by 5.2% and 4.1%, respectively.

These findings highlight the robustness and effectiveness of MuMoDIG, making it a superior choice for crafting transferable adversarial examples in diverse scenarios.

Table 1: Comparison with PAM and NAA. The surrogate model used for all experiments is RN-18, ensuring a fair comparison across methods.

Attack	CNNs	ViTs	AT	NRP
PAM	76.5	36.8	46.7	56.6
NAA	73.2	39.5	46.1	41.4
MuMoDIG	**92.5**	**62.3**	**51.9**	**60.7**

References

Zhang, J.; Huang, J.-t.; Wang, W.; Li, Y.; Wu, W.; Wang, X.; Su, Y.; and Lyu, M. R. 2023. Improving the transferability of adversarial samples by path-augmented method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8173–8182.

Zhang, J.; Wu, W.; Huang, J.-t.; Huang, Y.; Wang, W.; Su, Y.; and Lyu, M. R. 2022. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14993–15002.