Supplementary Material for Transferable Adversarial Attack for Both Vision Transformers and Convolutional Networks via Momentum Integrated Gradients

Wenshuo Ma[1], Yidong Li[2], Xiaofeng Jia[3], Wei Xu[1]
[1]IIIS, Tsinghua University, [2]Beijing Jiaotong University, [3]Beijing Big Data Centre
mwenshuo@gmail.com, ydli@bjtu.edu.cn, jiaxf@jxj.beijing.gov.cn, weixu@tsinghua.edu.cn

A. Computation Cost of Integrated Gradients

We approximate Integrated Gradients (IG) by averaging gradients at s points from the baseline image to the input image, as detailed in Equation 4 of Section 3. In Figure 1, we provide the *Mean Attack Success Rate (MASR)* and computation time of IG for one image as s varies from 1 to 25, using DeiT-T [12] as the source model. Larger values of s lead to more accurate estimates, but entail increased computation cost. In this section, we further analyze the potential to make our approach more feasible from two perspectives.

Firstly, by grouping these s points into a micro-batch, we conveniently compute their gradients with just a single backpropagation, without the need for s separate backward passes. This significantly reduces the runtime. In this manner, the runtime for $s = 20$ (default) is only **3 times** that of $s = 1$. Secondly, we can further adjust s to strike a balance between accuracy and computation. See the orange curve in Figure 1, reducing s from 20 to 6 only results in a tolerable 1.9% loss in MASR, while still outperforming SOTA method (75.13% vs. 69.56%) and maintaining a similar runtime as $s = 1$. Therefore, when computational resources are limited, it's feasible to adjust s for an easier application without sacrificing much performance.

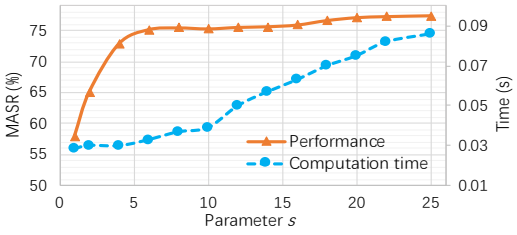


Figure 1. Computation time and MASR of MIG under different s.

B. Additional Experiment Results

We have previously reported the transfer attack performance of our Momentum Integrated Gradients (MIG) method under different settings. In this section, we supplement some additional results from more diverse settings.

B.1. MIG in White-box Attack Settings

We first present additional experiments where we use the same model as both the target and source models, allowing for white-box attacks. The results of these experiments are shown in Table 1, and we draw three main conclusions.

First, attacks with CNNs as the source models show better performance when using CNNs as the target models, with the MASR generally exceeding 80%. Similarly, attacks with ViTs as the source models exhibit better performance when targeting ViTs as the target models, with a generally higher MASR exceeding 75%. Second, MIG achieves excellent white-box attack performance. Specifically, the attack success rates of all eight white-box attacks are above 90%, with six of them achieving a white-box success rate of over 99%. In addition, we find that the transfer attack success rate does not always increase with the number of model parameters. For example, when using DN201 [5] (with 20M parameters) as the source model, the mean attack success rate is higher than that of using BiT [6] (with 26M parameters) as the source model. We speculate that the transferability of adversarial examples may be better when the source and target models are more similar, such as transfers among ViT-S, ViT-B, and ViT-L [3]. Note that although CNeXt [9] and Swin [8] belong to different categories of models (CNN and ViT), we observe that the transfer attack between them is particularly effective, possibly due to the fact that CNeXt is a modernized CNN designed towards the structure of Swin.

B.2. MIG with Ensemble

We present additional results of MIG using diverse model ensembles in Table 2. We mainly focus on ensembles of one CNN model and one ViT model, as prior experiments have shown that this type of ensemble is most effective.

These experiment results confirm our previous conclusions that logit ensemble and integrated gradients (IG) ensemble can both effectively improve the transferability of adversarial examples. By ensembling a CNN and a ViT, we achieve high MASR (above 87%) when using both CNNs and ViTs as target models.

Table 1. Attack success rate (%) and mean attack success rate (MASR, %) of MIG. * indicates the white-box attack.

		Target Model								MASR (CNNs)	MASR (ViTs)	MASR
		DN201	BiT	CNeXt	ViT-S	ViT-B	ViT-L	TNT	Swin			
Params (M)		20	26	89	49	87	304	24	88			
Source Model	DN201	**100***	94.18	75.5	77.31	65.06	48.39	73.24	58.79	89.89	64.56	74.06
	BiT	90.06	**99.55***	54.12	64.76	51.96	37.55	56.89	47.34	81.24	51.70	62.78
	CNeXt	83.18	79.47	**91.92***	74.80	70.88	65.41	77.36	78.16	84.86	73.32	77.65
	ViT-S	76.86	71.54	43.47	**99.95***	93.37	76.76	81.48	58.94	63.96	82.10	75.30
	ViT-B	76.50	69.83	47.17	97.17	**99.56***	83.67	78.44	62.22	64.50	84.21	76.82
	ViT-L	84.33	78.42	56.75	96.75	96.17	**99.25***	85.25	72.17	73.17	89.92	83.63
	TNT	81.88	73.39	64.81	85.44	69.98	55.07	**99.95***	64.86	73.36	75.06	74.43
	Swin	72.99	71.39	73.69	78.26	73.95	68.47	75.85	**93.02***	72.69	77.91	75.93

Table 2. Attack success rate (%) and mean attack success rate (MASR, %) of MIG with and without model ensemble. Perturbation, logit, and IG denote using perturbation ensemble, logit ensemble, and integrated gradients ensemble, respectively.

Source Model	Ensemble Strategy	Target Model							MASR (CNNs)	MASR (ViTs)	MASR
		DN201	BiT	CNext	ViT-S	ViT-B	ViT-L	TNT			
VGG19	No Ensemble	95.73	91.92	67.57	60.89	48.84	31.17	64.36	85.07	51.32	65.78
MNAS		94.08	86.65	60.83	70.63	53.56	37.10	71.59	80.52	58.22	67.78
Incep-v4		95.38	90.61	75.15	73.04	62.45	49.40	70.38	87.04	63.82	73.77
DeiT-T		82.28	75.90	51.40	96.98	84.34	79.35	88.62	69.86	87.32	79.84
DeiT-S		84.39	80.63	68.17	97.78	92.52	85.99	96.39	77.73	93.17	86.55
DeiT-B		86.33	82.67	74.73	96.20	91.12	87.53	95.07	81.24	92.48	87.67
VGG19	Perturbation	80.17	74.70	47.74	92.27	76.66	62.75	87.50	67.54	79.80	74.54
+	Logit	95.78	91.87	73.69	96.59	87.05	75.65	94.23	87.11	88.38	87.83
DeiT-T	IG	95.44	91.92	74.02	96.77	87.42	75.39	94.37	**87.13**	**88.49**	**87.91**
Incep-v4	Perturbation	86.33	82.11	70.33	93.89	88.11	82.56	93.89	79.59	89.61	85.32
+	Logit	95.34	91.91	86.40	96.38	92.67	88.85	95.69	**91.22**	93.40	92.46
DeiT-S	IG	95.20	92.00	86.22	96.45	92.63	89.04	95.81	91.14	**93.48**	**92.58**
Incep-v4	Perturbation	81.20	76.31	64.22	87.25	82.40	76.92	85.63	73.91	83.05	79.13
+	Logit	93.63	90.76	84.61	94.83	91.71	87.13	93.97	**89.67**	91.91	**90.95**
DeiT-B	IG	93.52	90.81	84.09	94.53	91.72	87.40	94.08	89.47	**91.93**	90.88
MNAS	Perturbation	83.38	83.67	70.25	93.31	89.88	84.18	90.82	79.10	89.55	85.07
+	Logit	94.68	92.04	84.93	94.98	92.41	89.33	95.20	90.55	92.99	91.95
DeiT-B	IG	94.67	92.22	84.78	95.33	92.67	89.56	95.11	**90.56**	**93.17**	**92.05**

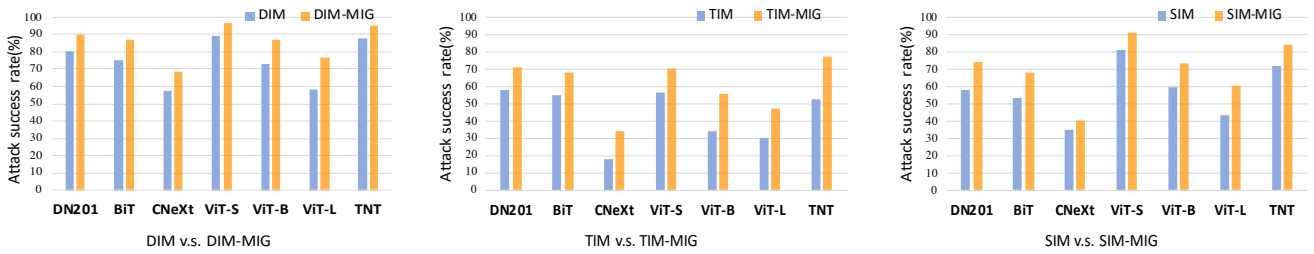
Figure 2 displays some additional MIG results with various input ensemble settings. MIG enhances attack success rates by about 7% $\sim$ 31% for these input ensemble methods when using both ViT (DeiT-T [12]) and CNN (Incep-v4 [11]) as source models. Note that even though MIG improves the performance to some extent, the transfer attack success rate remains relatively low when the target model and source model belong to different categories of models.

C. Visualization and Qualitative Analysis

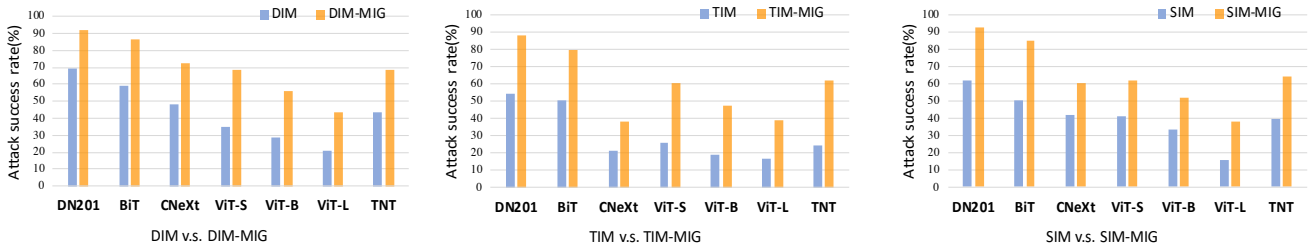
In this section, we give some examples of original images and their corresponding adversarial examples and perturbations generated using previous attacks such as FGSM [4], PGD [10], MI [1], and our MIG method.

We show two groups of original and adversarial images. In each group, the top row contains the clean original images and the remaining rows contain the corresponding adversarial examples and perturbations. All adversarial examples are generated using DeiT-S [12] as the source model. For the first group of images in Figure 3, the images have single-colored backgrounds or objects with relatively uniform colors. As a result, the adversarial examples tend to have more noticeable noise. For the second group of images in Figure 4, the images have more diverse textures and details, then the differences between the original images and the generated adversarial examples are much smaller and harder to perceive. In general, as we use a perturbation budget to control the size of perturbations added to original images, we can ensure that the generated adversarial perturbations remain imperceptible to humans.

Compared to other attacks, perturbations generated using MIG tend to contain more contour or shape information at the location of the main object. This also demonstrates that MIG can selectively attack the semantically relevant regions in images with the guidance of IG.



(a) Using DeiT-T [12] as the source model.



(b) Using Incep-v4 [11] as the source model.

Figure 2. Attack success rate (%) of DIM [13], TIM [2], SIM [7] and their MIG-enhanced versions.

References

[1] 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. 2

[2] 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

[3] 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

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

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[6] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 1

[7] 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. 3

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[9] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1

[10] 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. 2

[11] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 2, 3

[12] 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. 1, 2, 3, 4

[13] 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. 3

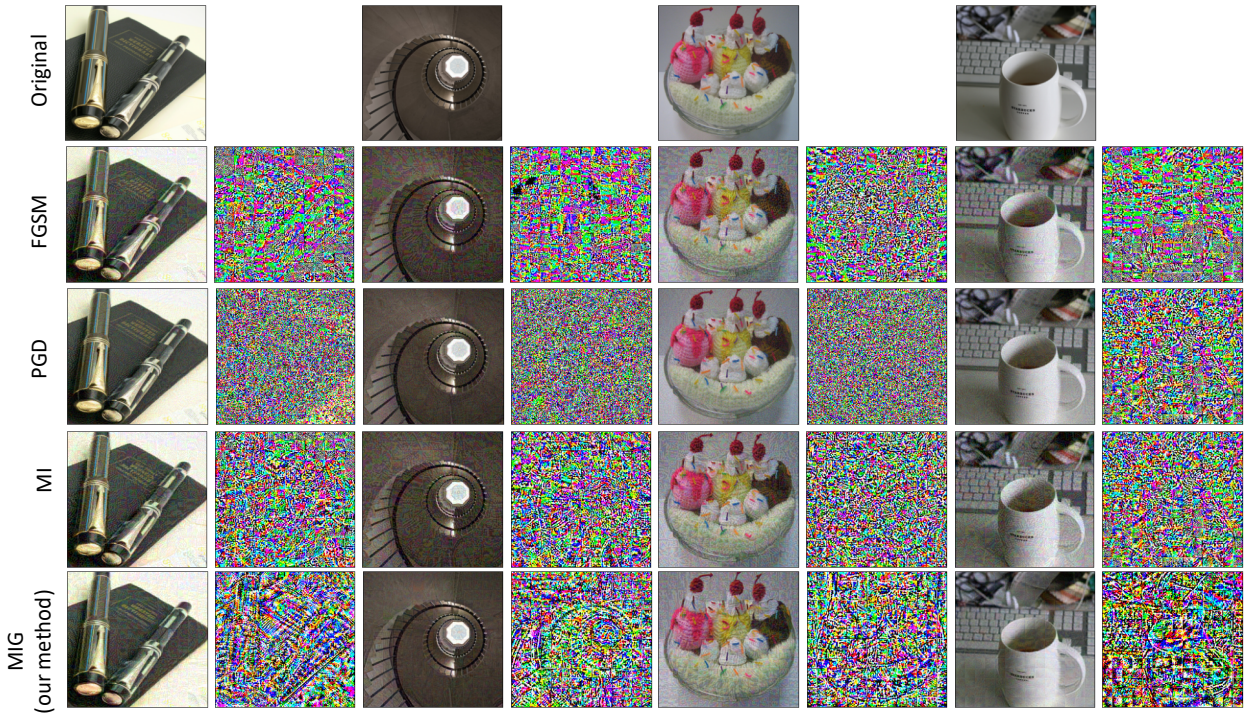


Figure 3. Examples of images with monotone background or object colors and their corresponding adversarial examples and perturbations generated via different attacks, using DeiT-S [12] as the source model. We set the perturbation budget $\epsilon = 16/255$.

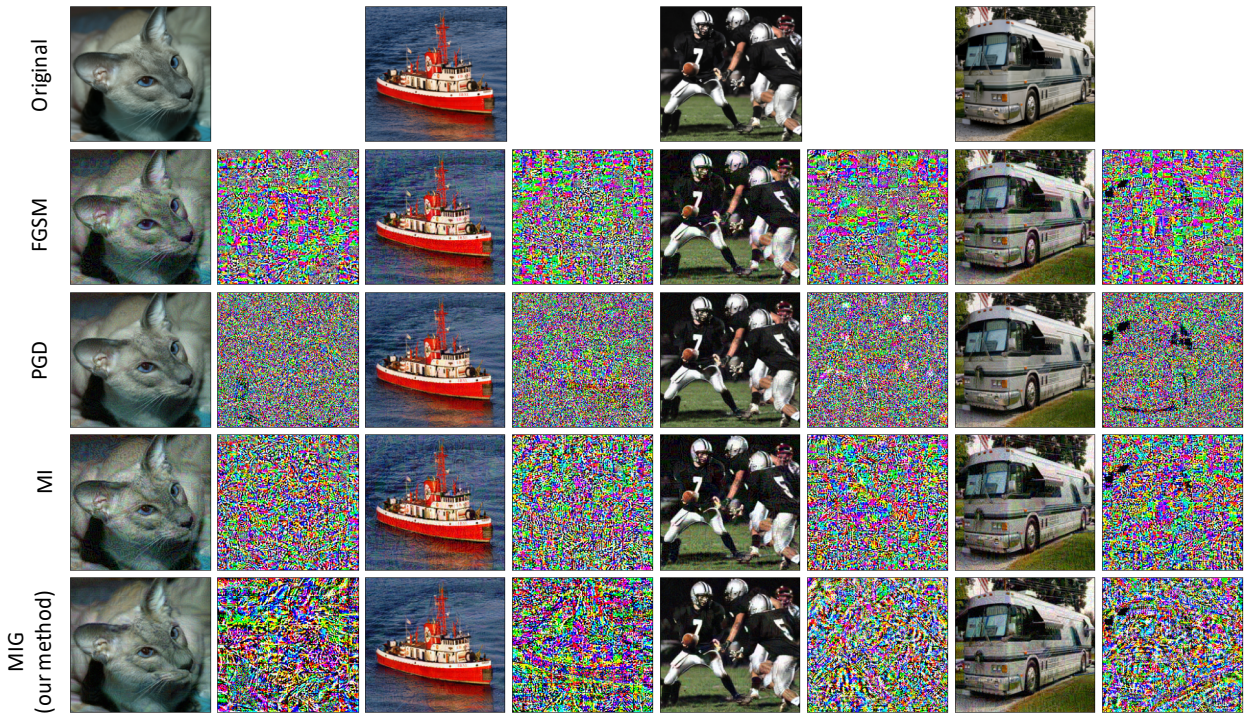


Figure 4. Examples of images with more details and textures and their corresponding adversarial examples and perturbations generated via different attacks, using DeiT-S [12] as the source model. We set the perturbation budget $\epsilon = 16/255$.