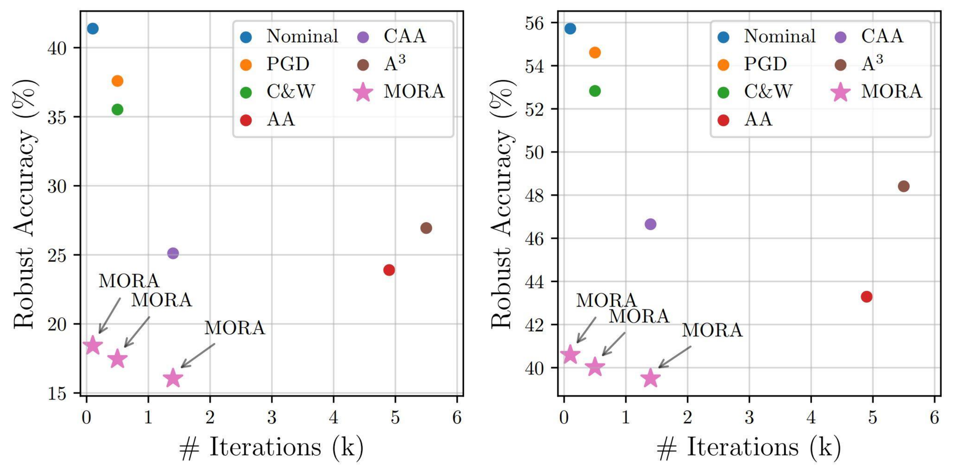
# MORA: Improving Ensemble Robustness Evaluation with Model-Reweighing Attack


Yunrui Yu[1] Xitong Gao[2] Chengzhong Xu[1*]

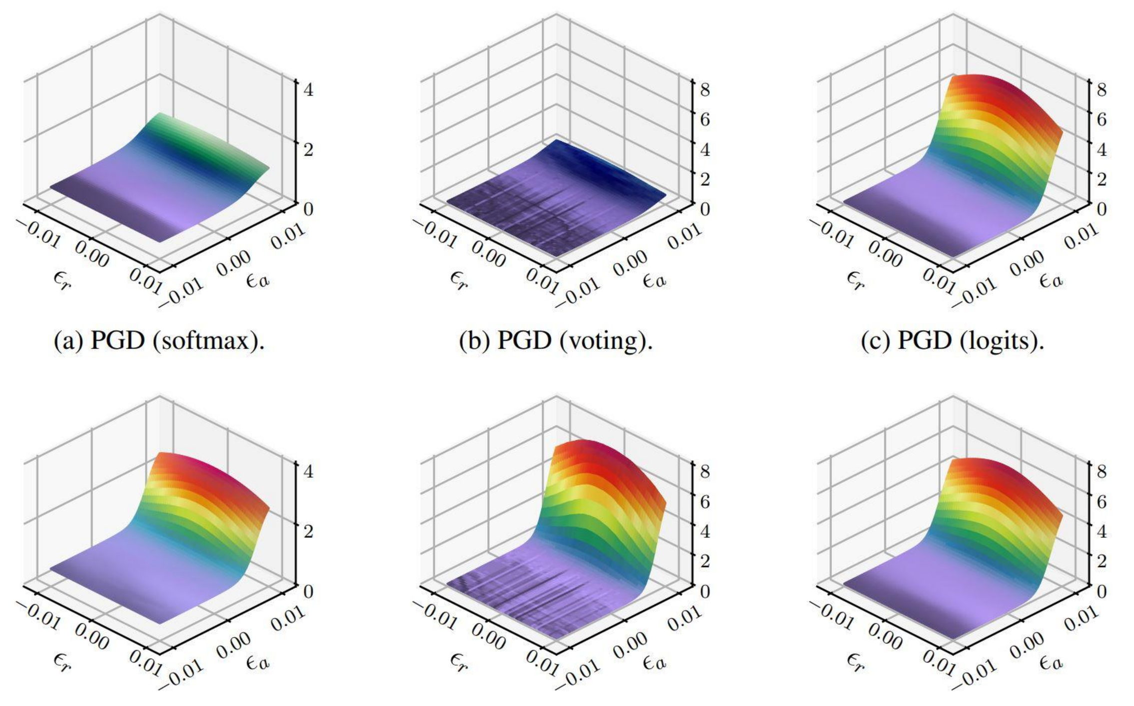
[1]State key lab of IoTSC University of Macau [2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences


## Introduction & Observations

Ensemble defenses offer a promising research direction to improve robustness against such attacks while maintaining a high accuracy on natural inputs. However, recent state-of-the-art (SOTA) adversarial attack strategies cannot reliably evaluate ensemble defenses, sizeably overestimating their robustness.



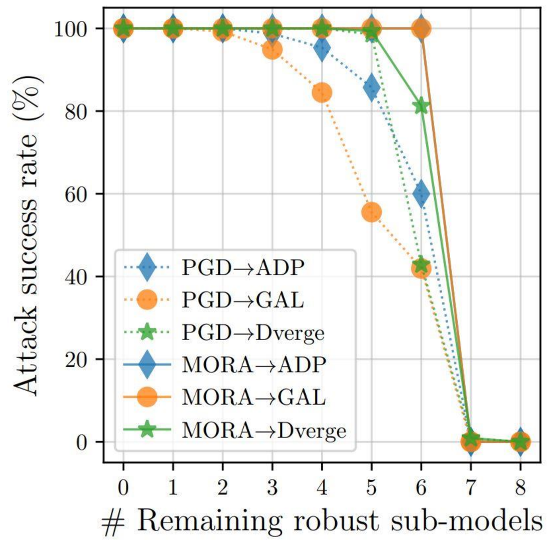

(a) First, these defenses form ensembles that are notably difficult for existing gradient-based method to attack, due to gradient obfuscation



(a) PGD (softmax). (b) PGD (voting). (c) PGD (logits).

(d) MORA (softmax). (e) MORA (voting). (f) MORA (logits).

(b) Second, ensemble defenses diversify sub-model gradients, presenting a challenge to defeat all sub-models simultaneously, simply summing their contributions may counteract the overall attack objective; yet, we observe that ensemble may still be fooled despite most sub-models being correct.




## Proposed LAFEAT Strategy

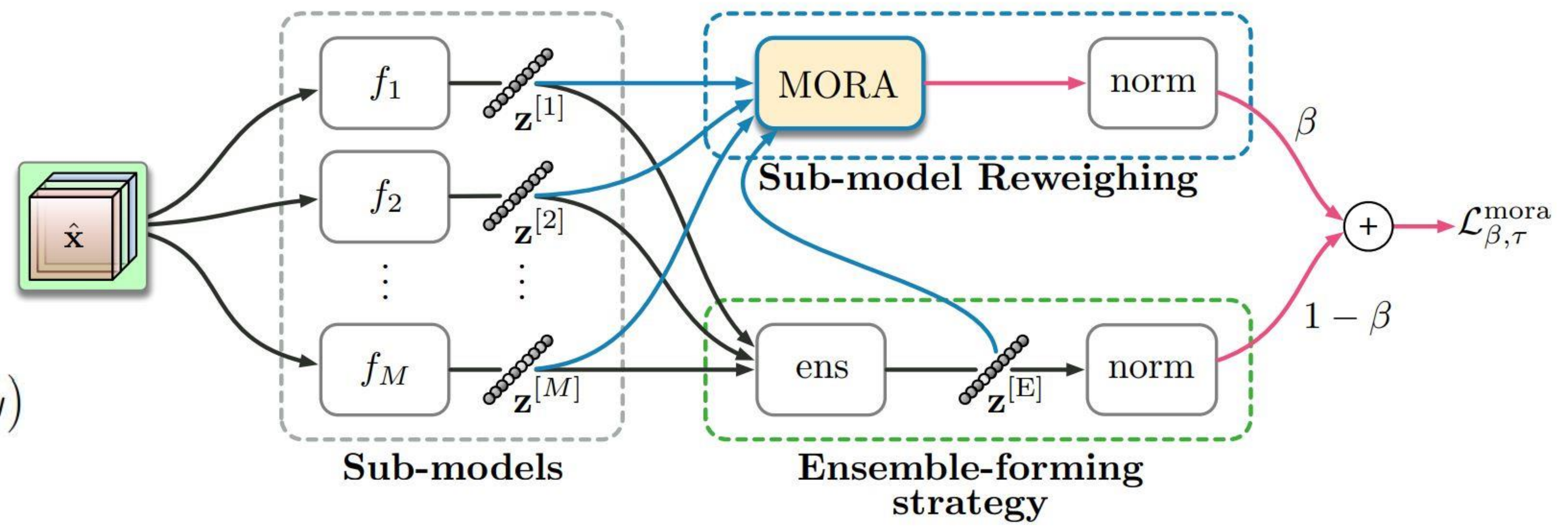
We introduce MORA, a model-reweighing attack to steer adversarial example synthesis by reweighing the importance of sub-model gradients by their respective "Contribution to the change of the ensemble loss value" during attack iterations.

$$k^{[\mathrm{E}]} = \mathrm{ens}(\mathbf{z}^{[m]})_y - \mathrm{ens}(\mathbf{z}^{[m]})_{\hat{y}} = \mathrm{ens}(\mathbf{z}^{[m]} - \mathbf{z}^{[m]}_{\hat{y}})_y - \mathrm{ens}(\mathbf{z}^{[m]} - \mathbf{z}^{[m]}_{\hat{y}})_{\hat{y}}$$
$$= \mathrm{ens}(k^{[m]}, \cdots)_y - \mathrm{ens}(k^{[m]}, \cdots)_{\hat{y}} \triangleq h_m(k^{[m]}).$$

$$\lambda^{[m]}_{\tau}(\mathbf{z}^{[m]}) = \frac{\partial k^{[\mathrm{E}]}(k^{[m]})}{\partial k^{[m]}} = \frac{\partial}{\partial k^{[m]}}\left(\frac{1}{M}\sum_{m\in[1:M]}\frac{\partial h_m(k^{[m]})}{\partial k^{[m]}}\right) = \frac{1}{M}\frac{\partial h_m(k^{[m]})}{\partial k^{[m]}}$$

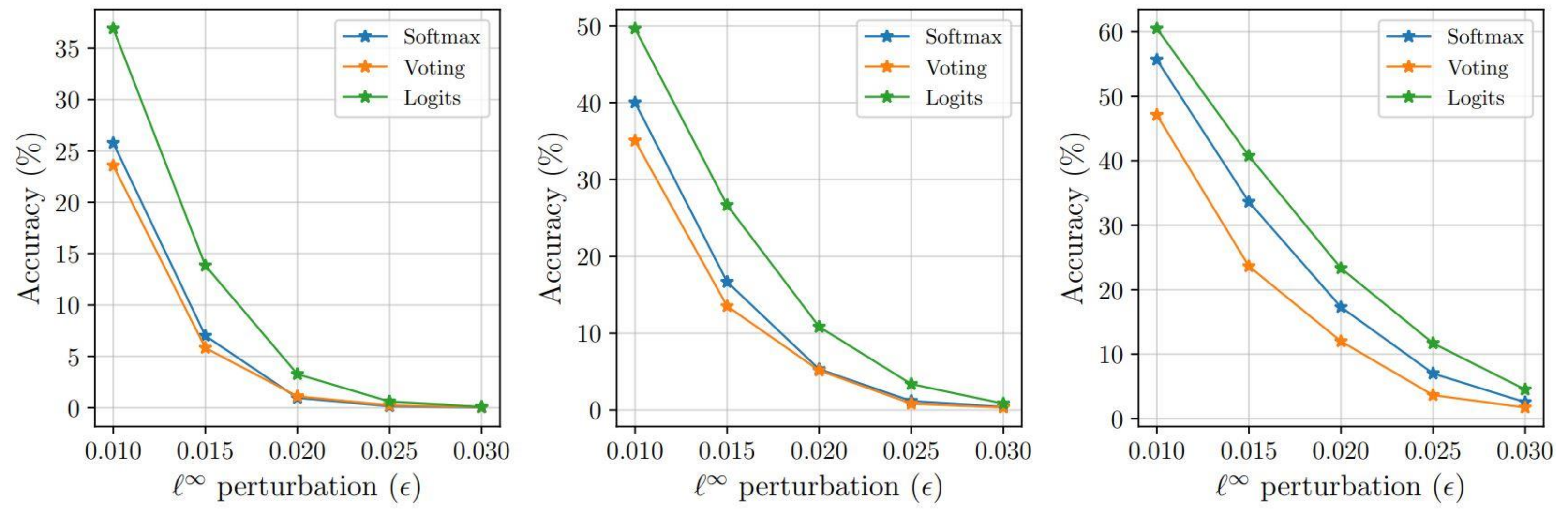
$$\mathcal{L}^{\mathrm{mora}}_{\beta,\tau}(\mathbf{z}^{[1:M]}, \mathbf{z}^{[\mathrm{E}]}, y) \triangleq \mathcal{L}^{\mathrm{sce}}\left(\beta\, \mathrm{norm}\left(\sum_{m\in[1:M]}\lambda^{[m]}_{\tau}(\mathbf{z}^{[m]})\cdot \mathbf{z}^{[m]}\right) + (1-\beta)\,\mathrm{norm}(\mathbf{z}^{[\mathrm{E}]}), y\right)$$

$$\mathrm{norm}(\mathbf{z}) \triangleq 1[\mathbf{z}_y - \mathbf{z}_{\hat{y}} > 0]\cdot \mathbf{z}/\mathrm{detach}(\mathbf{z}_y - \mathbf{z}_{\hat{y}})$$

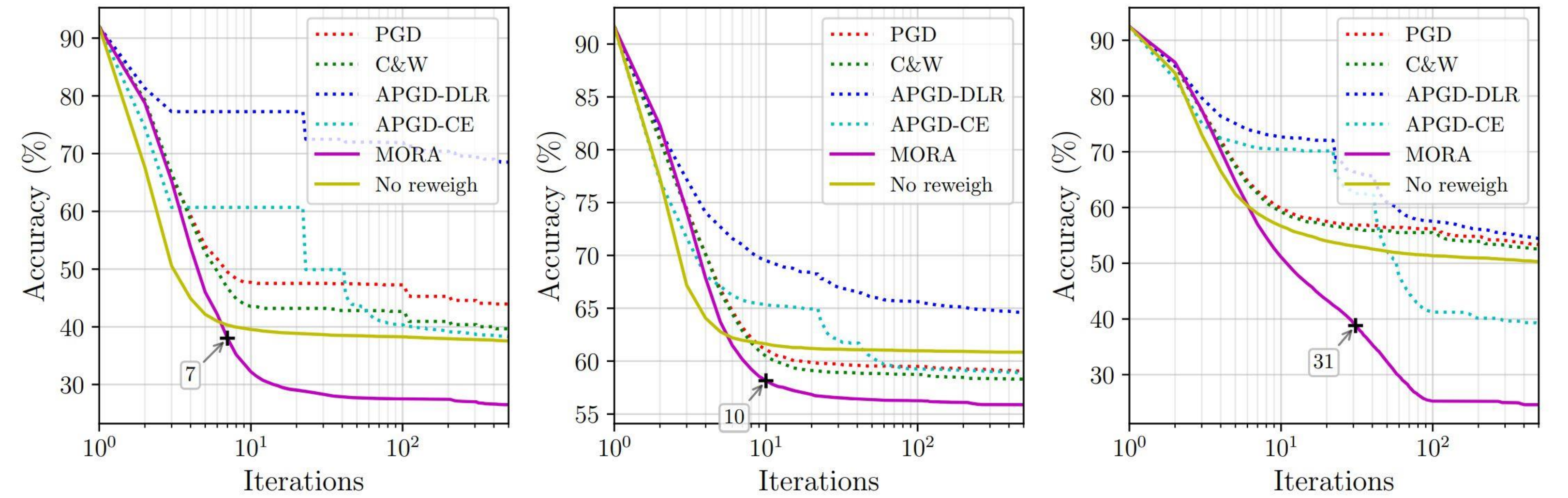



## Experiments & Visualization Results

| Defense Complexity | # | Clean 1 | Nominal — | PGD 500 | CW 500 | MORA 500 | $A^3$ 12k | AA 4.9k | CAA 1.8k | $\mathrm{MORA}^{\mathrm{mt}}$ 1.4k | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Softmax** | | | | | | | | | | | |
| ADP | 3 | 92.88 | 29.12 | 5.98 | 7.72 | 0.59 | 2.12 | 0.98 | 3.34 | **0.34** | 28.78 |
| | 5 | 93.34 | 25.14 | 7.10 | 8.70 | 0.97 | 3.62 | 2.18 | 4.25 | **0.67** | 24.47 |
| | 8 | 93.48 | 20.20 | 9.22 | 9.59 | 1.70 | 4.84 | 3.94 | 6.04 | **1.32** | 18.88 |
| Dverge | 3 | 91.99 | 47.42 | 44.49 | 40.17 | 25.77 | 33.36 | 30.58 | 32.98 | **25.26** | 22.16 |
| | 5 | 92.38 | 55.72 | 54.61 | 52.83 | 40.02 | 48.41 | 43.29 | 46.65 | **39.50** | 16.22 |
| | 8 | 91.65 | 59.63 | 59.13 | 58.25 | 55.68 | 57.29 | 56.71 | 56.89 | **55.57** | 4.06 |
| GAL | 3 | 89.41 | 19.48 | 8.13 | 11.57 | 0.67 | 0.70 | 0.85 | 1.00 | **0.51** | 18.97 |
| | 5 | 90.93 | 41.38 | 37.59 | 35.52 | 17.45 | 26.94 | 23.90 | 25.11 | **16.05** | 25.33 |
| | 8 | 92.45 | 56.31 | 53.39 | 52.56 | 28.71 | 36.51 | 37.46 | 35.30 | **27.44** | 28.87 |
| TRS† | 3 | 70.02 | 19.71 | 14.01 | 10.87 | 8.11 | 8.72 | 8.46 | 9.75 | **7.60** | 12.11 |
| | 5 | 69.00 | 23.17 | 15.91 | 15.28 | 12.67 | 13.22 | 13.20 | 13.78 | **12.47** | 10.70 |
| | 8 | 73.01 | 23.64 | 18.02 | 17.59 | 15.90 | 16.22 | 16.51 | 16.73 | **15.64** | 8.00 |
| **Voting** | | | | | | | | | | | |
| ADP | 3 | 91.84 | 41.62† | 9.32 | 11.84 | 0.64 | 3.06 | 6.13 | 8.29 | **0.29** | 41.33 |
| | 5 | 93.13 | 40.29† | 12.42 | 12.05 | 1.17 | 6.03 | 10.13 | 0.67 | **0.62** | 39.67 |
| | 8 | 93.28 | 30.10† | 12.53 | 10.50 | 3.16 | 6.11 | 9.21 | 1.69 | **1.65** | 28.45 |
| Dverge | 3 | 91.72 | 39.05† | 31.48 | 28.00 | 23.57 | 24.95 | 24.98 | 27.65 | **22.91** | 16.14 |
| | 5 | 92.18 | 49.36† | 44.28 | 42.28 | 35.06 | 39.15 | 39.20 | 40.85 | **34.46** | 14.90 |
| | 8 | 91.58 | 56.85† | 53.72 | 52.35 | 47.12 | 50.58 | 50.04 | 51.15 | **46.10** | 10.75 |
| GAL | 3 | 89.09 | 21.48† | 5.85 | 7.64 | 0.87 | 0.71 | 0.56 | 0.78 | **0.35** | 21.13 |
| | 5 | 90.77 | 37.32† | 29.33 | 27.62 | 12.96 | 18.55 | 20.82 | 22.17 | **12.25** | 25.07 |
| | 8 | 92.37 | 55.39† | 49.56 | 48.02 | 21.66 | 30.35 | 31.39 | 30.93 | **20.16** | 35.23 |
| TRS† | 3 | 68.95 | 13.79 | 10.19 | 8.71 | 5.73 | 11.89 | 6.69 | 8.08 | **5.44** | 8.35 |
| | 5 | 68.31 | 15.36 | 12.71 | 11.88 | 8.82 | 10.08 | 10.30 | 11.21 | **8.38** | 6.98 |
| | 8 | 72.05 | 17.00 | 14.57 | 13.48 | 11.39 | 11.99 | 11.85 | 12.80 | **10.69** | 6.31 |
| **Logits** | | | | | | | | | | | |
| ADP | 3 | 92.86 | 3.44† | 0.87 | 2.05 | 0.48 | 0.25 | 0.22 | 0.31 | **0.21** | 3.23 |
| | 5 | 93.48 | 4.57† | 1.97 | 4.24 | 1.12 | 1.00 | 0.97 | 1.09 | **0.89** | 3.68 |
| | 8 | 93.38 | 5.39† | 3.57 | 4.77 | 2.13 | 2.20 | 2.05 | 2.11 | **1.93** | 3.46 |
| Dverge | 3 | 92.19 | 38.31† | 37.99 | 38.60 | 36.89 | 36.94 | 36.96 | 37.07 | **36.84** | 1.47 |
| | 5 | 92.28 | 50.77† | 50.57 | 51.28 | 49.65 | 49.72 | 49.66 | 49.75 | **49.59** | 1.18 |
| | 8 | 91.73 | 61.06† | 60.95 | 61.51 | 60.52 | 60.59 | 60.52 | 60.55 | **60.49** | 0.57 |
| GAL | 3 | 89.50 | 15.47† | 10.01 | 10.53 | 0.52 | **0.02** | **0.02** | 0.08 | 0.03 | 15.44 |
| | 5 | 90.93 | 36.36† | 33.97 | 35.14 | 22.24 | 33.43 | 20.24 | 21.66 | **19.40** | 16.96 |
| | 8 | 92.54 | 56.08† | 53.67 | 54.69 | 31.52 | 40.90 | 30.89 | 31.17 | **30.66** | 25.42 |
| TRS† | 3 | 69.72 | 13.31 | 13.06 | 13.80 | 12.11 | 12.13 | 12.16 | 12.21 | **12.07** | 1.24 |
| | 5 | 68.90 | 16.89 | 16.65 | 17.34 | 15.88 | 15.86 | 15.90 | 15.95 | **15.82** | 1.07 |
| | 8 | 72.24 | 19.40 | 19.20 | 19.67 | 18.20 | 18.18 | 18.27 | 18.34 | **18.17** | 1.23 |




(a) Diverge (3 sub-models). (b) Diverge (5 sub-models). (c) Diverge (8 sub-models).




(a) Dverge 3. (b) Dverge 8. (c) GAL 8.

## Contributions & Conclusions

- ✓ This paper presents the first extensive study on the robustness of ensemble defenses under multiple ensemble-forming strategies.
- ✓ By reweighing the importance weights of sub-models to steer adversarial example synthesis, we show that gradient-based attacks on ensemble defenses can often be orders of magnitude faster, while enjoying a higher success rate.
- ✓ Empirical results on a wide variety of different ensemble defenses show that MORA outperforms competing attacks in both performance and convergence rate.
- ✓ Misleading a minority of sub-models is sufficient to fool the ensemble.
- ✓ Summing by logits is the simplest yet most robust way to form ensembles.