It is critical that the deep learning models of AI-driven industrial systems are robust and quick to adapt to changes in the environment. This is typically achieved through adversarial training, but acquiring new data can be costly, and the performance of the technique heavily depends on its quality. In this work, we try to balance the need for quality data and the problem of its cost by selecting the most `important' of it for adversarial training. We propose a task-based Robust ACTive (RACT) learning method that selects training data near to the boundary by applying adversarial attacks. Our method improves 68.67% speed of model training on CIFAR-10. Comparing with other data selection methods, our method has a 10\% higher accuracy with 10% training data selected and a 7% higher robustness with 4% training data selected. Moreover, we show our method can significantly improve the efficiency by at least 25% on adversarial training with the same performance. Finally, we evaluate our method on a physical robotic arm system with object detection, generating adversarial patches as our attack, and adopting our method as the defense. The results indicate that RACT can defend against 60% of untargeted attacks and 20% of targeted attacks, suggesting that it is applicable for protecting real-world systems against physical attacks.

The deep learning models of AI-enabled industrial systems must be robust and quick to adapt to changes in the environment, which is typically achieved through adversarial training. However, acquiring new data can be costly, and the technique’s performance heavily depends on its quality. In this work, we try to balance the need for quality data and the problem of its cost by selecting the most ‘important’ of it for adversarial training. In particular, we propose a task-based robust fast (RAST) learning method that selects training data near the boundary by considering adversarial samples. Our method improves 68.67% speed of model training on CIFAR-10. Compared with other data selection methods, our method has a 10% higher accuracy with 10% of training data selected and a 7% higher robustness with 4% of training data selected. Moreover, we show our method can significantly improve the efficiency by at least 25% on adversarial training with the same performance. Finally, we evaluate our method on a physical robotic arm system with object detection, generating adversarial patches as our attack, and adopting our method as the defense. The results indicate that RAST can defend against 60% of untargeted attacks and 20% of targeted attacks, suggesting that it is applicable for protecting real-world systems against physical adversarial attacks.