Pre-training also Transfers Non-Robustness

# Introduction

Pre-training has enabled state-of-the-art results on many tasks. In spite of its recognized contribution to generalization, we observed in this study that pre-training also transfers adversarial non-robustness from pre-trained model into fine-tuned model in the downstream tasks. Using image classification as an example, we first conducted experiments on various datasets and network backbones to uncover the adversarial non-robustness in fine-tuned model. Further analysis was conducted on examining the learned knowledge of fine-tuned model and standard model, and revealed that the reason leading to the non-robustness is the non-robust features transferred from pre-trained model. Finally, we analyzed the preference for feature learning of the pre-trained model, explored the factors influencing robustness, and introduced a simple robust pre-training solution.

# Main Result

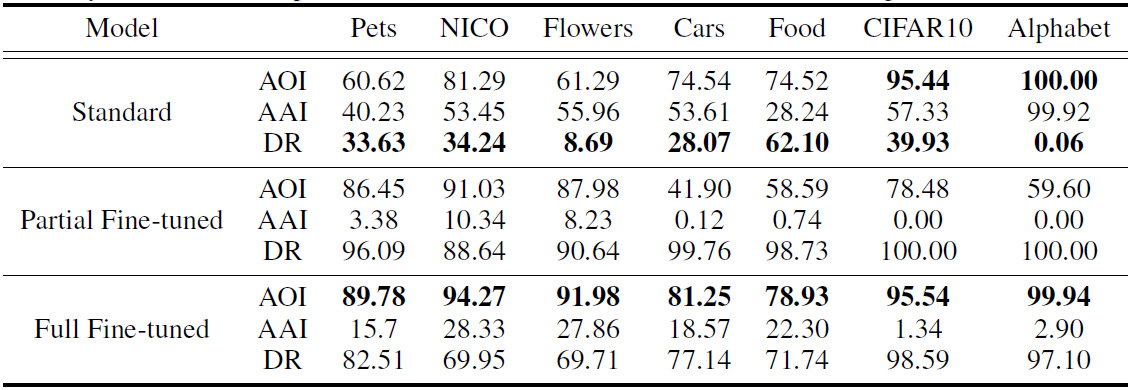


Table.1 Comparison of generalization and robustness between standard model, partial fine-tuned model and full fine-tuned model. For each model, we report accuracy of original inputs (AOI), accuracy of adversarial inputs (AAI), and decline ratio (DR) on 7 different target datasets.

# Paper

https://arxiv.org/abs/2106.10989