Learning Representations for OOD Detection in the Presence of Spurious Correlations

Adrian-Cătălin Luțu, Ștefan-Alexandru Popescu and Elena Burceanu
University of Bucharest, Romania

lutu.adrian.catalin@gmail.com, stefanalex.popescu@gmail.com

1. Problem Introduction

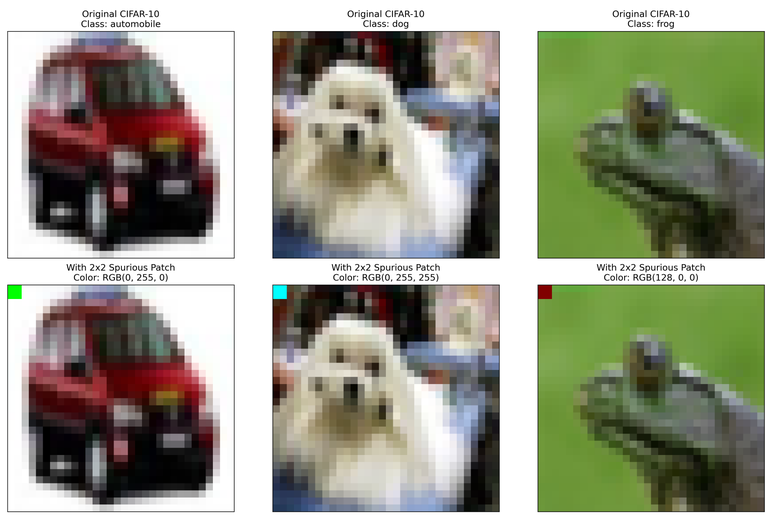
- OOD detection is essential for reliable ML systems, distinguishing between in-distribution (ID) and out-of-distribution (OOD) samples
- Spurious correlations (SCs) in training data can severely compromise OOD detection capability
- Limited investigation exists on impact of SCs on OOD detection tasks

Research Question: How do spurious correlations affect different model architectures' OOD detection performance?

2. Dataset and Methodology

Dataset:

- CIFAR-10 with artificially introduced spurious correlations
- Colored 2×2 patches in top-left corner
- Each class assigned specific color (10 colors total)



Models Tested:

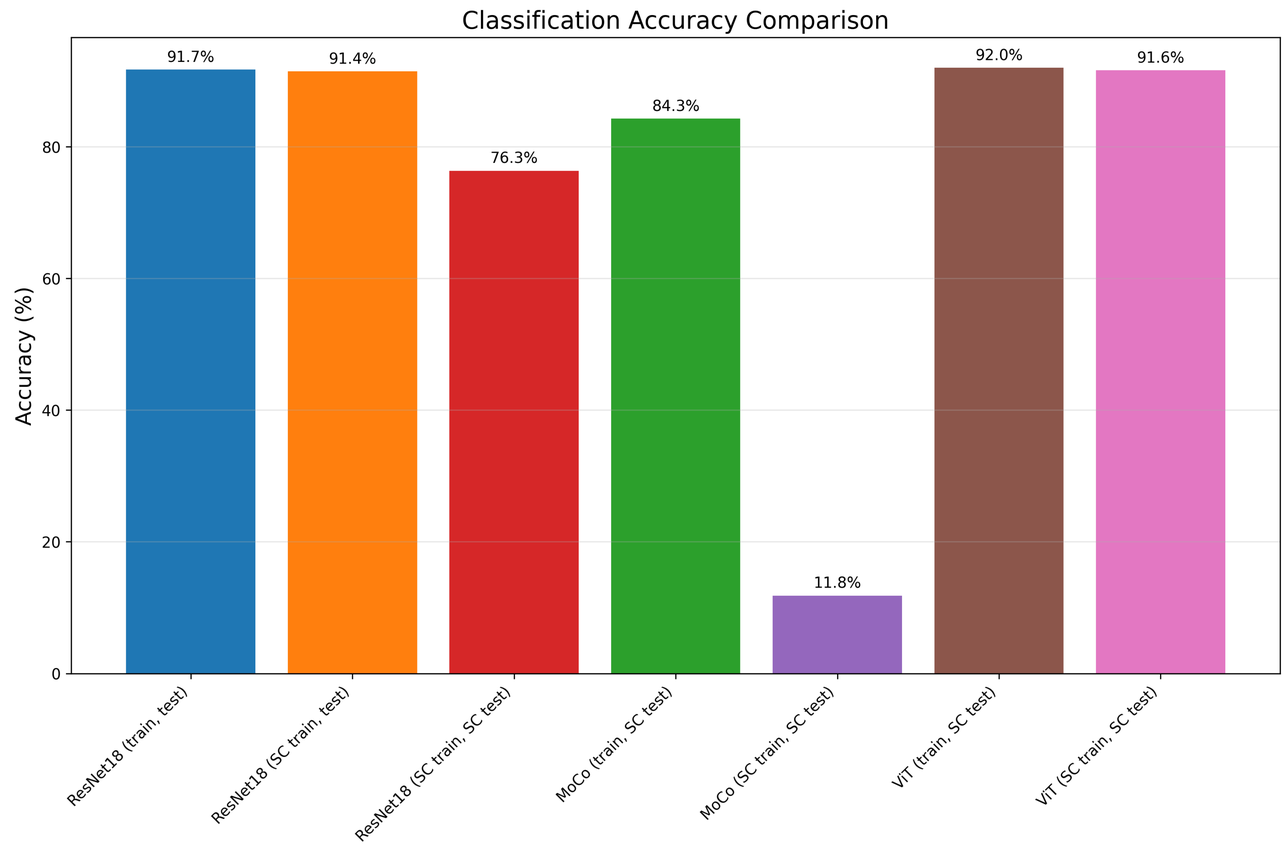
- ResNet18 (trained from scratch)
- MoCo (self-supervised pre-trained)
- Vision Transformer (supervised pre-trained)

Experiments: OpenOOD benchmark framework for comprehensive evaluation.

Scan for Paper

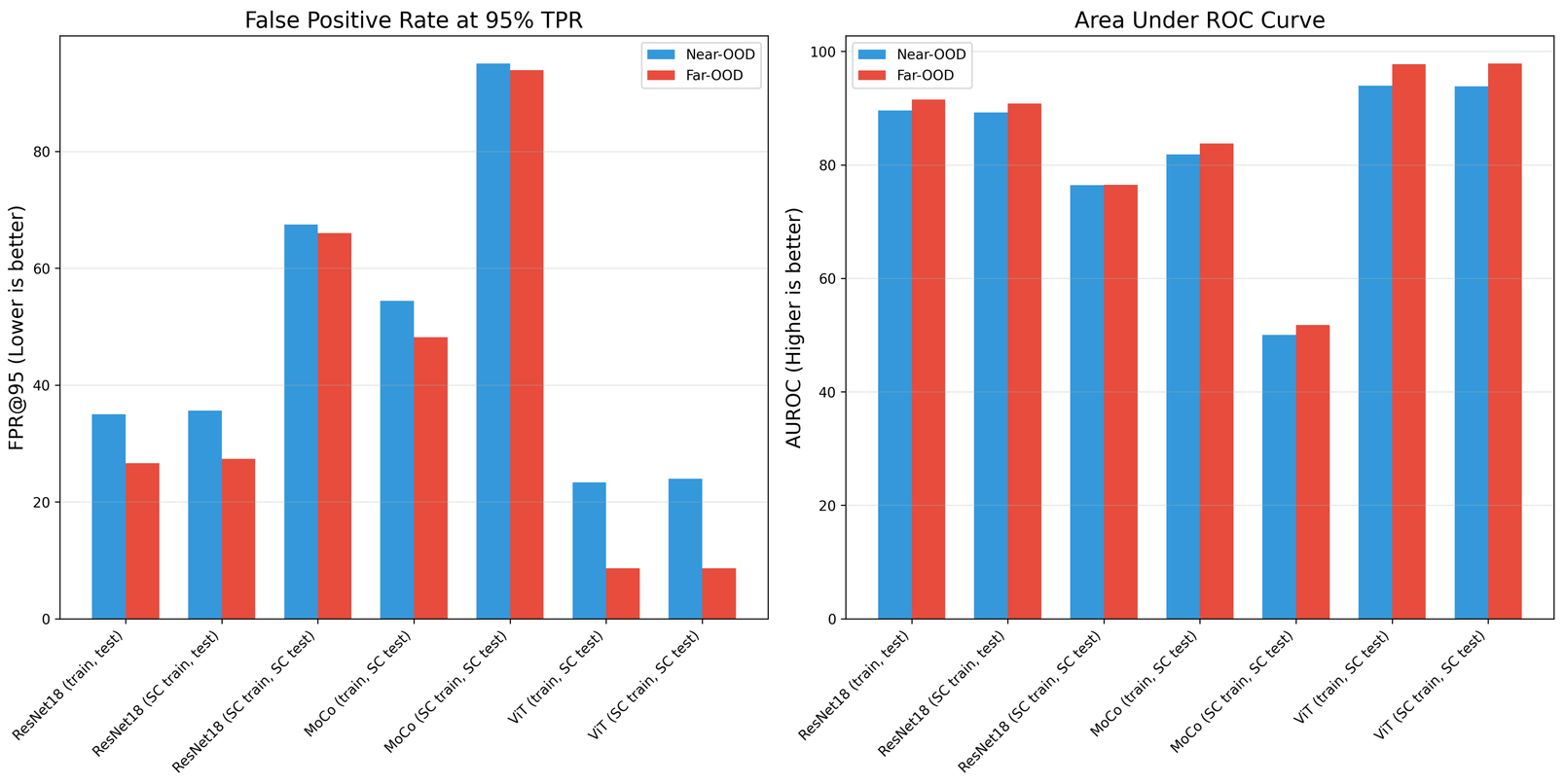
3. Key Findings

1. Architecture-Specific Robustness



- **Vision Transformers**: Remarkably robust (91.98% → 91.61%), minimal accuracy drop
- **MoCo**: Severely affected (84.28% → 11.83%), 72.45% drop
- **ResNet18**: Moderate impact (91.73% → 76.34%), 15.39% drop

4. OOD Detection Performance



- ViT consistently outperforms other architectures across all OOD metrics
- MoCo shows dramatic OOD detection degradation when tested on SC data
- Supervised training (ViT) demonstrates superior robustness compared to self-supervised (MoCo)

5. Insights and Conclusions

Main Insights:

- Training paradigm and architecture together determine robustness to spurious correlations
- Self-supervised learning (MoCo) amplifies rather than mitigates SC effects
- Transformer-based architectures exhibit superior resilience to SCs

Implications: The combination of training paradigm and architecture selection is crucial for OOD detection in presence of spurious correlations. **Future Work:** Investigate additional architectures and training methods to understand robustness mechanisms.