NEURAL INFORMATION PROCESSING SYSTEMS

Bridging Symmetry and Robustness: On the Role of Equivariance in Enhancing Adversarial Robustness

Dr. Longwei Wang[1], Ifrat Ikhtear Uddin[1], Prof. KC Santosh[1]

Dr. Chaowei Zhang[2], Prof. Xian Qin[3], Dr. Yang Zhou[3]

[1] AI Research Lab, Department of Computer Science, University of South Dakota, Vermillion, SD 57069, USA

[2] School of Information and Artificial Intelligence, Yangzhou University, Yangzhou, China

[3] Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA

Motivation & Problem

- **Adversarial Vulnerability in DNNs**
- Small imperceptible perturbations leads to incorrect predictions
- Standard defense: adversarial training expensive, time consuming and reduce clean accuracy
- **Our Question:**
- Can architectural symmetry priors alone improve robustness?
- **Key Insight:**
- Equivariant architectures enforce consistency under transformations → smoother gradients → robustness

Related Work

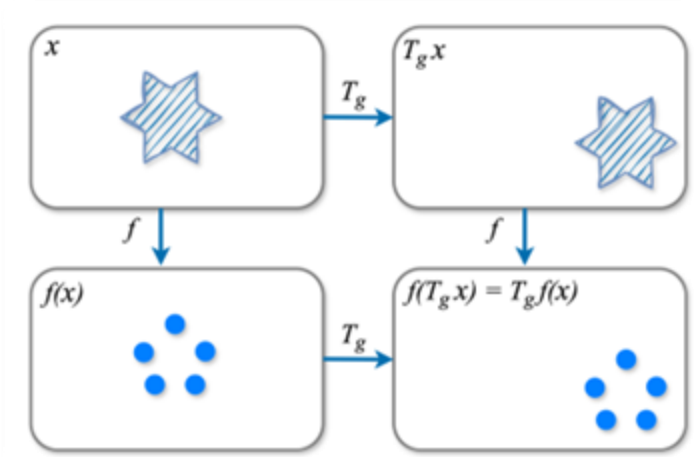
- Equivariant Neural Networks
  - **G-CNNs** (Cohen & Welling, 2016)
    Extended convolutions to group transformations
  - **Scale-Equivariant Networks** (Worrall & Welling, 2019)
    Multi-scale processing without augmentation
- Adversarial Defense
  - **Adversarial Training** (Madry et al., 2018)
    Train on perturbed data,
  - **Certified Defenses:** Randomized smoothing, CLEVER bounds
  - **Limitations:** Computationally expensive, Reduced clean accuracy, Attack-specific defenses

Theoretical framework

Equivariance Definition:

$$f(g \cdot x) = \rho(g)\, f(x), \forall_g \in G$$

Applying a transformation to the input leads to a predictable transformation in the output, thereby promoting stability and consistency in feature representations.



Rotation Equivariance: P4 group (0°, 90°, 180°, 270°)
Scale Equivariance: Multi-resolution processing

Adversarial perturbations often violate known symmetries. By constraining the model to respond consistently along group-induced orbits and suppressing sensitivity to off-orbit perturbations, equivariant architectures provide a natural defense mechanism

Theoretical framework

Mathematical Preliminaries:

Orbit and Quotient Space:

$$[x]_G = \{g \,.\, x \mid \mathrm{g} \in G\}, X/G = \text{set of all distinct orbits}$$

Jacobian and Lipschitz Constant

$$\mathrm{J}_f(x) = \nabla f(x) \in \mathbb{R}^{k \times d}, L(x) = ||J_f(x)||_2$$

Margin Function:

$$g_{c,j}(x) = f_c(x) - f_j(x)$$

Distance between predicted class c and alternative class j

Adversarial Robustness:

A classifier is (ε,p)-robust if

$$f(\mathrm{x} + \delta) = f(\mathrm{x}), \forall \delta \in \mathbb{R}^{\mathrm{d}}, ||\delta||_{\mathrm{p}} \leq \varepsilon$$

Theoretical framework

❑ **Theorem 1: Orbit-Invariant Robustness:**

- Group-equivariant networks preserve gradient norms across symmetric transformations. If $\boldsymbol{\rho}(g)$ and Dg^{-1} are orthogonal (norm, preserving), then:

$$||\nabla g_{c,j}(g \cdot x)||_q = ||\nabla g_{c,j}(x)||_q$$

- **Lipschitz constant invariant across group orbit:**

$$L_q^{(j)} = \sup_{x' \in B_p([x]_{G,r}} ||\nabla g_{c,j}(x')||_q$$

Equivariance preserves gradient norms under symmetric transformations, yielding tighter CLEVER-certified robustness bounds uniformly across orbits and enabling certified defense without adversarial training.

Theoretical framework

❑ **Theorem 2: Directional Gradient Suppression**

Decompose perturbation:

$$\boldsymbol{\delta} = \boldsymbol{\delta}_G + \boldsymbol{\delta}_\perp$$

- $\boldsymbol{\delta}_G$ **=** along symmetric orbit (rotation/scale)
- $\boldsymbol{\delta}_\perp$ = off−orbit (adversarial direction)

Then:

$$|| \nabla \mathbf{f}(\mathbf{x} + \boldsymbol{\delta}_\perp) - \nabla f(x) ||_2 \gg || \nabla \mathbf{f}(\mathbf{x} + \boldsymbol{\delta}_G) - \nabla f(x) ||_2$$

What This Means:

- Gradients change minimally along symmetric directions
- Gradients change significantly along adversarial directions
- Network naturally suppresses adversarial sensitivity

Key Benefit:

Suppresses gradients along symmetry directions, enhancing robustness and smoothing decision boundaries.

Theoretical framework

❑ Scale Equivariance: Gradient Smoothing

Scale transformations are **not norm-preserving**, but achieve robustness via multi-scale orbit averaging:

$$\bar{\nabla}\,\emptyset_j(x) = \frac{1}{|G_s|}\sum_{s\in G_s}\bar{\nabla}\,\emptyset_j(T_s x)$$

How It Works:

1. Process input at multiple scales simultaneously
2. Average gradients across scale transformations
3. Smoothing effect acts as regularization

Which results in Multi-scale averaging smooths gradients, reducing vulnerability to sharp local changes.

Theoretical framework

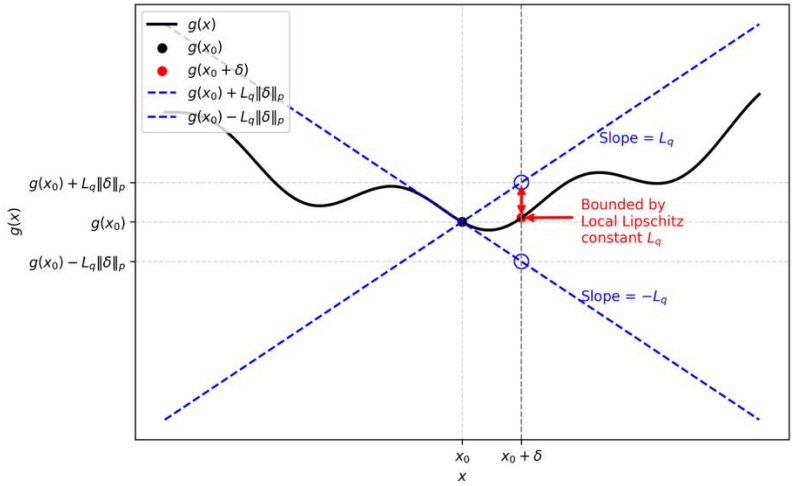
- CLEVER Certified Robustness Framework

CLEVER Bound:

$$\boldsymbol{\varepsilon}_{\min}^{(\boldsymbol{p})}(\boldsymbol{x}) = \min_{\boldsymbol{j}\neq\boldsymbol{c}} \frac{g_{c,j}(x)}{\mathrm{L}_q^{(j)}}, \text{ where}$$

$$g_{c,j}(x) = f_c(x) - f_{\mathrm{j}}(x)\,, L_q^{(j)} = \sup_{\mathrm{x}'\in \mathrm{B_p}([\mathrm{x}]_{\mathrm{G,r}}} ||\nabla \mathrm{g_{c,j}}(\mathrm{x}')||_{\mathrm{q}},$$

$$And\ \frac{1}{p} + \frac{1}{q} = 1$$



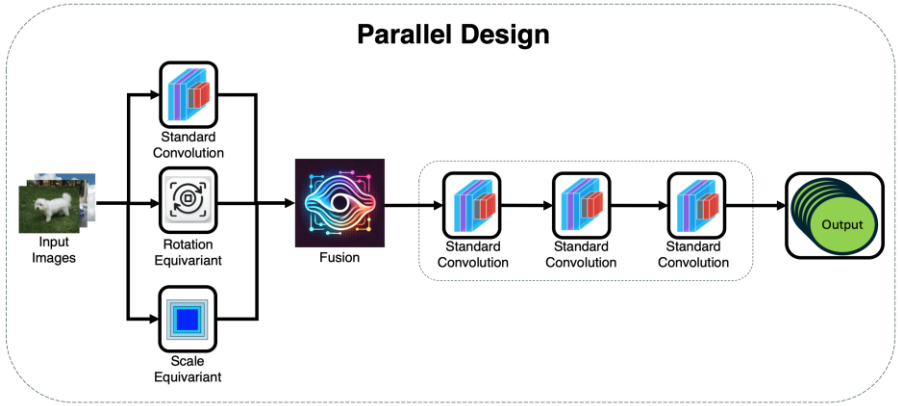
Standard CNN:
$L_q^{(j)}$ varies across transformations
Must estimate worst-case Lipschitz constant
⇒ Loose certified bounds

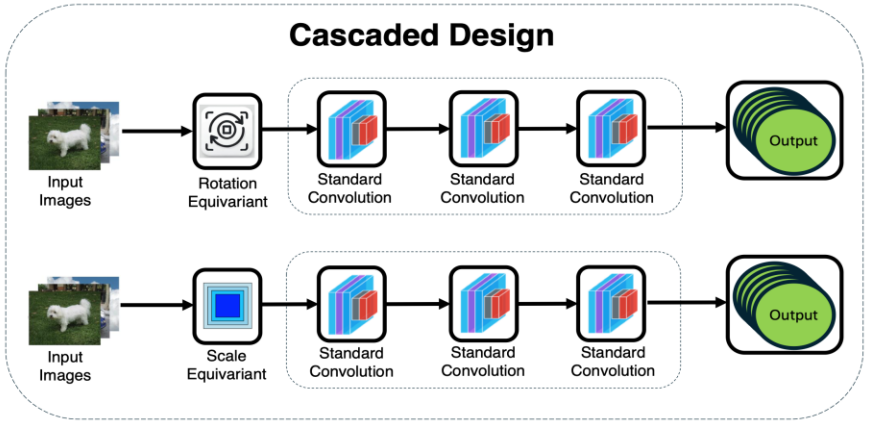
Equivariant CNN:
$L_q^{(j)}$ **invariant** across group orbit $[x]_G$:

$$||\nabla \mathrm{g_{c,j}}(\mathrm{g}\,.\,\mathrm{x})\,||_{\mathrm{q}} = ||\nabla \mathrm{g_{c,j}}(\mathrm{x})||_{\mathrm{q}}$$

Single Lipschitz constant shared across the orbit
⇒ **Tighter, more stable certified bounds**

Architectures





Parallel Design:

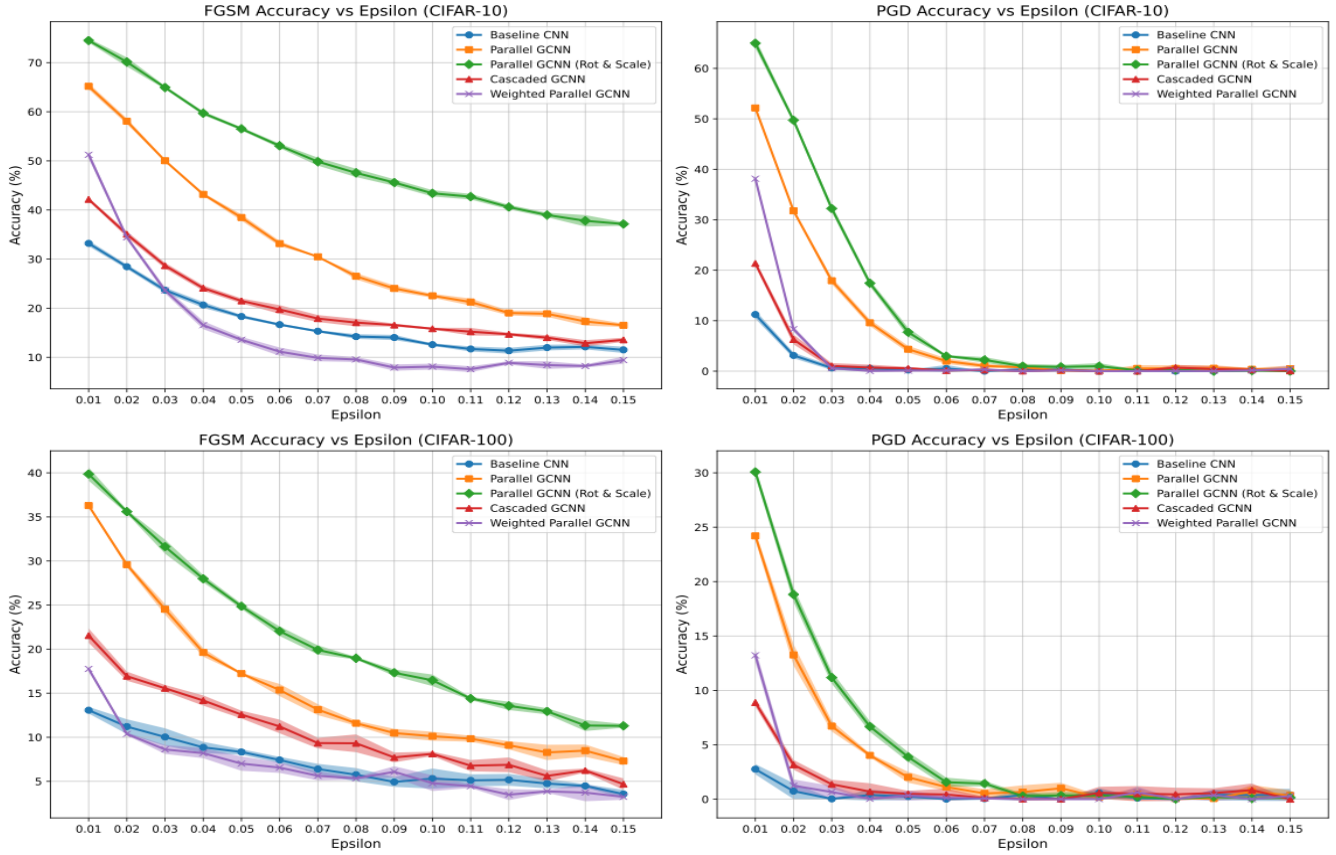
Input processed through three independent branches. Three parallel branches process standard, rotation (P4), and scale-equivariant features independently before fusion via concatenation. This design preserves complementary feature spaces and achieves superior adversarial robustness.

Cascade Design:

Sequential application of equivariant transformations. Rotation-equivariant layer processes input first, followed by scale-equivariant processing, then standard convolutions.

Adversarial Robustness Evaluation

We evaluated our architectures on CIFAR-10 and CIFAR-100 under FGSM and PGD attacks at varying perturbation levels. The Parallel GCNN with combined rotation and scale branches consistently achieved the highest robustness across all settings, outperforming baseline CNNs and alternative fusion strategies.



Performance Highlight

10-Layer Parallel GCNN (Rotation + Scale) Robust without adversarial training (ε=0.01-0.15)

Dataset	Model	ε	FGSM	PGD
CIFAR 10	Baseline CNN	0.01	33.20%	11.00%
	Parallel GCNN		**75.08%**	**64.82%**
CIFAR100	Baseline CNN	0.01	13.60%	2.50%
	Parallel GCNN		**39.97%**	**29.62%**

Fully Equivariant Models

To validate our theory, we evaluate architectures where ALL layers are rotation-equivariant (P4).

Depth	ε=0.01		ε=0.05		ε=0.10	
	FGSM	**PDG**	**FGSM**	**PGD**	**FGSM**	**PGD**
4-layer Baseline	29.3%	3.03%	18.47%	0.01%	15.35%	0.00%
4-layer Equivariant	**65.7%**	**52.2%**	**47.1%**	**15.9%**	**38.0%**	**7.0%**
10-layer baseline	33.2%	11.0%	18.60%	0.14%	12.93%	0.00%
10 Layer Equivariant	**73.0%**	**65.0%**	**60.2%**	**37.8%**	**44.9%**	**12.5%**

Depth-dependent improvement (7-15% gain) confirms orbit-invariant gradient regularization compounds beneficially when symmetry enforced end-to-end.

Conclusion

Key Takeaways:

- ❑ Theory: Equivariance yields tighter CLEVER certified bounds via orbit-invariant gradients
- ❑ Architecture: Parallel design with rotation-scale branches achieves best performance
- ❑ Empirical: Competitive with adversarial training at fraction of computational cost
- ❑ Scaling: Full end-to-end equivariance essential for maintaining robustness in deep networks