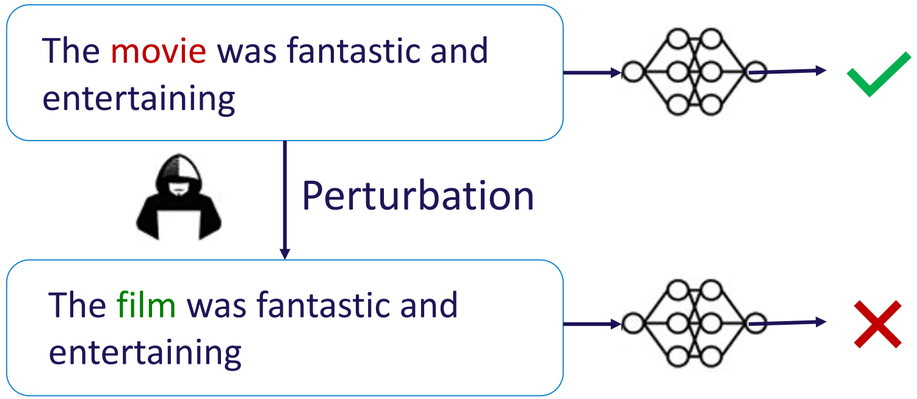
Explaining the role of Intrinsic Dimensionality in Adversarial Training

Enes Altinisik · Safa Messaoud · Husrev Taha Sencar · Hassan Sajjad · Sanjay Chawla

Adversarial Attacks and Training

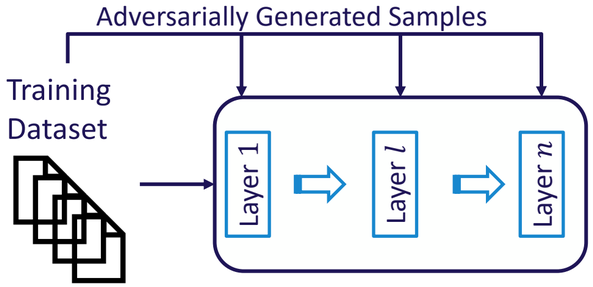
Adversarial Attacks: Finding Vulnerabilities

Small imperceptible perturbations can fool models into incorrect predictions (sentiment)



Adversarial Training (AT): Building Robustness

Training with adversarial examples improves model robustness against attacks

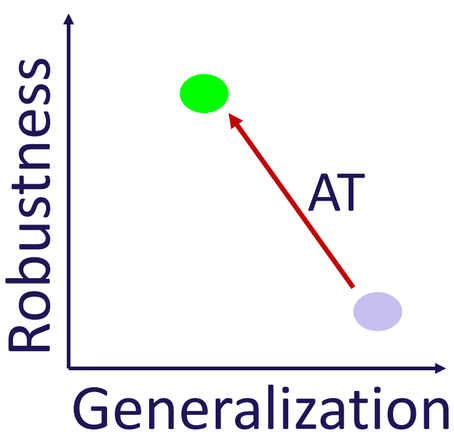


> In adversarial machine learning, we wrote over 9,000 papers in ten years and got nowhere
>
> Nicholas Carlini

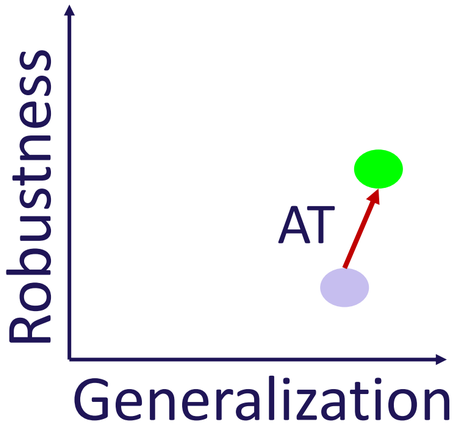
From Theory to Practice: Why Adversarial Training Falls Short in Real Applications

➢ Unclear Trade-Off Between Generalization and Robustness

- In vision models:
  - Significant improvements in robustness
  - Notable drop in generalization performance
- In encoder-based LLMs:
  - Robustness improvements
  - Generalization may also improve



VISION



Encoder LLMs

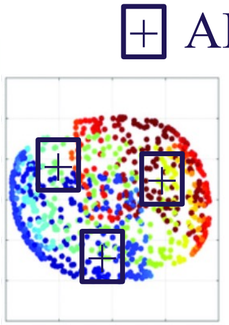
➢ Scalability Challenges

- Significant increase in training time due to cost of adversarial examples (AEs) generation
  - 10-step PGD adversarial training increases training time by 10×

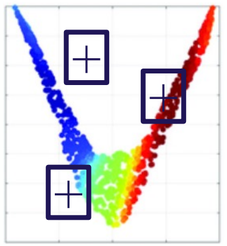
Why Does AT Affect Generalization and Robustness Differently?

- Manifold Hypothesis
  - On manifold AEs → improve generalization
  - Off manifold AEs → improve robustness

- Intrinsic Dimensionality (ID) & On/Off-Manifold Behavior
  - Lower ID → more off manifold AEs
    - Improves robustness, reduces generalization
  - Higher ID → more on manifold AEs
    - Improves generalization, reduces robustness



High ID
On manifold AEs
Better generalization
Lower robustness



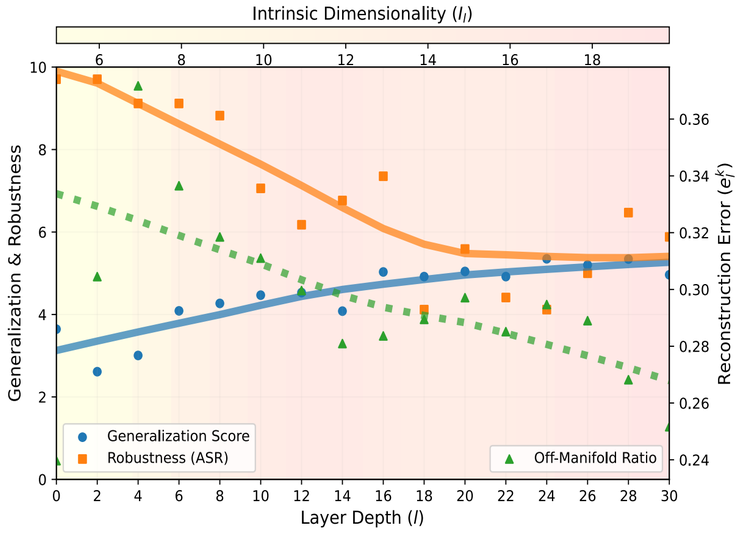
Low ID
Off manifold AEs
Better robustness
Lower generalization

Layer-wise Analysis: How Adversarial Training Affects Model Performance

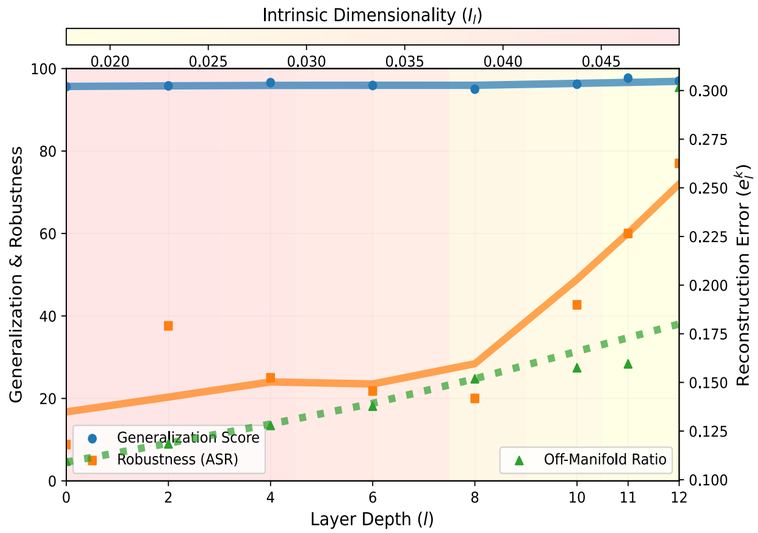
Layer-wise adversarial training analysis reveals:

- Higher ID -> lower off manifold ratio -> better generalization
- Lower ID -> higher off manifold ratio -> better robustness

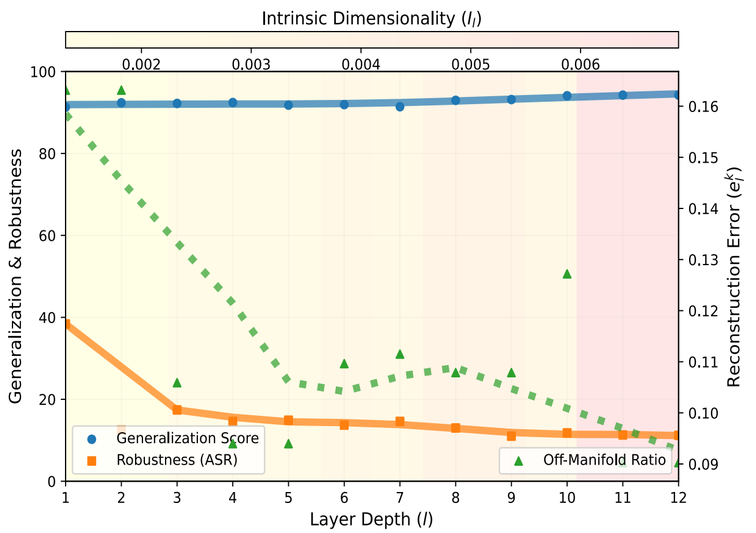
LLaMA-2



BERT

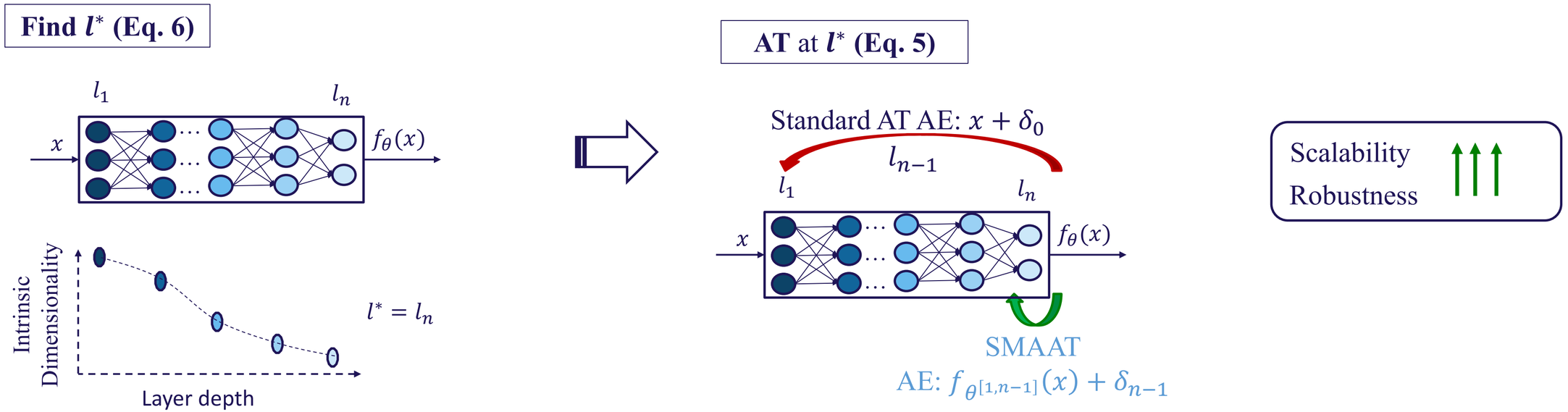


VGG



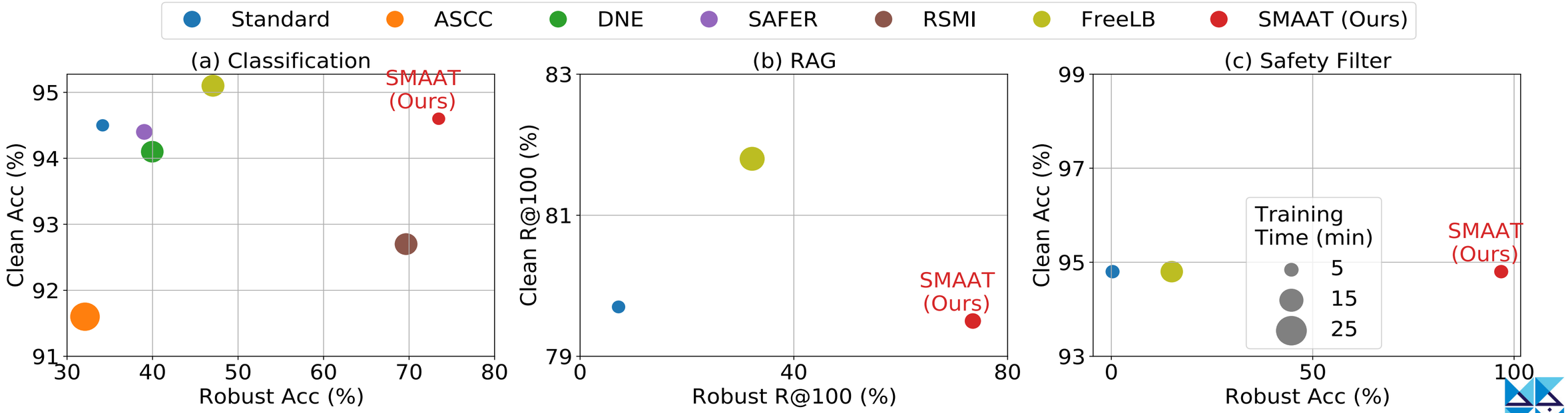
SMAAT: Scalable Manifold Aware Adversarial Training

- Identify the layer with minimum intrinsic dimensionality (ID)
- Apply AT specifically to that layer for optimal robustness and scalability.



SMAAT Evaluation: Generalization and Robustness at No Extra Cost

- SMAAT evaluation across three domains: Text classification, retrieval model in RAG, and LLM safety filtering
- Achieves superior robustness with preserved generalization and no training overhead



THANK YOU

We’d love to talk! Find us at Poster Session 5 on Tuesday.

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY