ResNet-GA: Evolutionary Deep Learning Models for Adversarial Defense (Student Abstract)

Li-Chiao Wang[1, 2], **Chung-Shou Liao**[3], **and Wei Liu** [1]

[1]School of Computer Science, University of Technology Sydney, Australia
[2]Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan
[3]Department of Electrical Engineering, National Taiwan University, Taiwan
li-chiao.wang@student.uts.edu.au, csliao@ntu.edu.tw, wei.liu@uts.edu.au

Abstract

Adversarial attacks remain a major challenge for deep learning models, as they can undermine both performance and reliability in practical applications such as image recognition. Although evolutionary algorithms (EAs) have proven effective in optimizing complex systems, their use for directly enhancing model robustness for adversarial defense has been limited. In this study, we introduce ResNet-GA, a method that applies evolutionary deep learning (EDL) to develop ResNet-like networks specifically designed to resist different forms of adversarial perturbations. The approach evolves network architectures with a genetic algorithm (GA), adapting the Residual Blocks at every stage in ResNet according to the needs of each dataset and attack type. Experimental results show that ResNet-GA strengthens model robustness beyond standard baselines, highlighting the value of iterative evolutionary design for building more dependable deep learning systems under various adversarial conditions.

Introduction

Deep neural networks have achieved strong performance in tasks such as image recognition and pattern detection, yet selecting the right architecture remains crucial for both accuracy and reliability. The structure of a network influences how well it extracts features and generalizes to different datasets, even for the same task. Neural Architecture Search (NAS) provides automated methods for finding effective network structures, reducing reliance on manual trial-and-error and expert knowledge (Sun et al. 2020; Zeng, Li, and Peng 2023; Zeng et al. 2021).

Despite these advances, robustness against adversarial attacks is rarely considered in architecture design. Adversarial attacks add small perturbations to input data that can mislead models while remaining imperceptible to humans. These attacks may be targeted or untargeted and can assume different levels of model knowledge, ranging from white-box to black-box. Prior work mostly focuses on creating attacks (Guo et al. 2020; Chivukula et al. 2020; Yin et al. 2018) or improving defense through specialized training strategies such as customized loss functions, regularization, or gradient-based methods , leaving the design of robust network architectures largely unexplored.

To address this challenge, we propose ResNet-GA, which uses EDL to evolve ResNet-like networks with improved resilience against adversarial perturbations. The approach employs a GA to adapt the Residual Blocks at each stage, optimizing channel sizes and internal structure according to dataset characteristics and attack scenarios. Adversarial examples are incorporated into fitness evaluation to guide evolution toward architectures that maintain both stability and high performance. Experiments show that ResNet-GA consistently improves robustness compared to baseline models and other NAS-optimized networks.

The main contributions of this research are as follows:

- We design block-level optimization that adjusts channel sizes within Residual Blocks to enhance adversarial robustness without significantly increasing the overall parameter count;
- We introduce the integration of evolutionary search with adversarial learning, with encoding block configurations as genes and evolving network structures using crossover, mutation, and selection, while evaluating performance on multiple types of adversarial examples;
- We perform empirical validation across multiple datasets and attack types, demonstrating that the evolved networks achieve higher resilience than standard architectures of similar scale.

The Proposed EDL Models

We propose an innovative framework to automatically discover deep neural network architectures with high adversarial robustness. Focusing on ResNet-like networks, the method treats the number of channels in each Residual Block as a flexible parameter. Candidate architectures are initially generated with random block-level channel configurations and iteratively refined based on their ability to maintain performance under adversarial perturbations. By concentrating variations at the block level, the method improves robustness without increasing overall network depth, reducing computational cost while exploring meaningful architectural diversity.

Genetic Representation of Network Structures

Each candidate network is represented by its Residual Blocks, which encode structural properties such as channel

Dataset	Model	Data used for Searching	Original	FGSM 100	SIA	Wieland	# Params
CIFAR-10	ResNet-32	-	0.819	0.209	0.595	0.717	476.5k
	ResNet-44	-	0.828	0.208	0.622	0.730	670.9k
	ResNet-GA	Original data	0.829	0.225	0.596	0.731	582.6k
	ResNet-GA	FGSM mixed data	**0.830**	**0.229**	**0.638**	**0.747**	601.7k
	ResNet-GA	SIA mixed data	0.820	0.218	0.622	0.735	387.3k
	ResNet-GA	Wieland mixed data	0.825	0.208	0.631	0.745	440.1k
Mini-ImageNet	ResNet-32	-	0.519	0.268	0.197	0.217	472.8k
	ResNet-44	-	0.539	0.296	0.216	0.228	667.2k
	ResNet-GA	Original data	0.534	**0.309**	**0.221**	**0.239**	614.1k
	ResNet-GA	FGSM mixed data	0.539	0.284	0.219	0.227	639.0k
	ResNet-GA	SIA mixed data	0.537	0.306	**0.221**	0.230	680.4k
	ResNet-GA	Wieland mixed data	**0.540**	**0.309**	0.216	0.222	633.6k

Table 1: Test accuracy for different models under various attacks.

sizes and convolutional filter dimensions. Adjustable channel sizes within each block allow the search to explore diverse yet coherent architectures. This block-level representation provides meaningful units for structural variation, enabling targeted modifications that enhance feature extraction and information flow throughout the network.

Evolutionary Search with Adversarial Training

Candidate architectures are evaluated based on their performance under adversarial examples. Each generation produces new candidates by recombining high-performing structural components and applying stochastic adjustments to block-level parameters. Within each stage of the ResNet backbone, all residual blocks share the same number of channels to maintain smooth residual connections and consistent computational cost:

$$C_l^{(b)} = C_l, \quad \forall b \in \{1, \ldots, B_l\}, \tag{1}$$

where $C_l^{(b)}$ denotes the number of channels in block b of stage l, and C_l is the shared channel size.

Networks are selected according to their performance under a chosen adversarial attack. To guide the evolutionary search, pre-generated adversarial examples are incorporated during training of each candidate network, and the training objective combines losses on clean and adversarial samples:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{CE}(\hat{y}_i, y_i) + \frac{\lambda}{M}\sum_{j=1}^{M} \mathcal{L}_{CE}(\hat{y}_j^{adv}, y_j), \tag{2}$$

where $\hat{y}_i$ and $\hat{y}_j^{adv}$ are the predicted logits for clean and adversarial inputs, y_i and y_j are the corresponding labels, N and M are the numbers of samples, $\mathcal{L}_{CE}$ is the cross-entropy loss, and λ balances the adversarial contribution.

Integrating adversarial evaluation directly into the search ensures that candidate architectures maintain stability across perturbations, guiding the evolutionary process toward networks that preserve both accuracy and robustness while respecting computational constraints.

Experiments

We conducted ResNet-GA experiments using ResNet-32 as the backbone, maintaining 15 residual blocks. The evolutionary search employs two-point crossover at stage boundaries and a stage-wise mutation that modifies all residual channels within a stage. To evaluate robustness against adversarial attacks, we tested three attack types: gradient-based FGSM (Goodfellow, Shlens, and Szegedy 2015), score-based SIA (Wang, Zhang, and Zhang 2023), and transfer-based Wieland (Brendel et al. 2019), aiming to identify the optimal architecture for each scenario.

Table 1 presents experiment results, including ResNet-32 and ResNet-44 as baselines. Results show that ResNet-GA consistently outperforms ResNet-32 and ResNet-44. These results demonstrate the superiority of our structure search method, as ResNet-44 has more layers and parameters than our models in most cases. This illustrates that the evolutionary search effectively identifies architectures that balance adversarial robustness and efficiency.

Conclusion and Future Work

In this research we introduce ResNet-GA, which uses evolutionary deep learning to evolve ResNet-like networks, adapting Residual Blocks to improve robustness against diverse adversarial attacks without increasing network depth. Experiments show it outperforms baseline models, confirming that block-level adaptation strengthens defense. In future, we plan to expand encoding strategies, explore larger search spaces, and improve computational efficiency for broader real-world applications.

References

Brendel, W.; Rauber, J.; Kümmerer, M.; Ustyuzhaninov, I.; and Bethge, M. 2019. Accurate, reliable and fast robustness evaluation. In *Advances in Neural Information Processing Systems*.

Chivukula, A. S.; Yang, X.; Liu, W.; Zhu, T.; and Zhou, W. 2020. Game theoretical adversarial deep learning with variational adversaries. *IEEE Transactions on Knowledge and Data Engineering*, 33(11): 3568–3581.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

Guo, M.; Yang, Y.; Xu, R.; Liu, Z.; and Lin, D. 2020. When nas meets robustness: In search of robust architectures against adversarial attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Sun, Y.; Xue, B.; Zhang, M.; Yen, G. G.; and Lv, J. 2020. Automatically designing CNN architectures using the genetic algorithm for image classification. *IEEE transactions on cybernetics*, 50(9): 3840–3854.

Wang, X.; Zhang, Z.; and Zhang, J. 2023. Structure Invariant Transformation for better Adversarial Transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4607–4619.

Yin, Z.; Wang, F.; Liu, W.; and Chawla, S. 2018. Sparse feature attacks in adversarial learning. *IEEE Transactions on Knowledge and Data Engineering*, 30(6): 1164–1177.

Zeng, N.; Li, H.; and Peng, Y. 2023. A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease. *Neural Computing and Applications*, 35(16): 11599–11610.

Zeng, N.; Li, H.; Wang, Z.; Liu, W.; Liu, S.; Alsaadi, F. E.; and Liu, X. 2021. Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip. *Neurocomputing*, 425: 173–180.