On the Security Risks of AutoML

Ren Pang[†] Zhaohan Xi[†] Shouling Ji[‡] Xiapu Luo[⋆] Ting Wang[†]
[†]*Pennsylvania State University, {rbp5354, zxx5113, ting}@psu.edu*
[‡]*Zhejiang University, sji@zju.edu.cn*
[⋆]*Hong Kong Polytechnic University, csxluo@comp.polyu.edu.hk*

Automation is good, so long as you know exactly where to put the machine.

– Eliyahu Goldratt

Abstract

Neural Architecture Search (NAS) represents an emerging machine learning (ML) paradigm that automatically searches for models tailored to given tasks, which greatly simplifies the development of ML systems and propels the trend of ML democratization. Yet, little is known about the potential security risks incurred by NAS, which is concerning given the increasing use of NAS-generated models in critical domains.

This work represents a solid initial step towards bridging the gap. Through an extensive empirical study of 10 popular NAS methods, we show that compared with their manually designed counterparts, NAS-generated models tend to suffer greater vulnerability to various malicious attacks (*e.g.*, adversarial evasion, model poisoning, and functionality stealing). Further, with both empirical and analytical evidence, we provide possible explanations for such phenomena: given the prohibitive search space and training cost, most NAS methods favor models that converge fast at early training stages; this preference results in architectural properties associated with attack vulnerability (*e.g.*, high loss smoothness and low gradient variance). Our findings not only reveal the relationships between model characteristics and attack vulnerability but also suggest the inherent connections underlying different attacks. Finally, we discuss potential remedies to mitigate such drawbacks, including increasing cell depth and suppressing skip connects, which lead to several promising research directions.

1 Introduction

Automated Machine Learning (AutoML) represents a new paradigm of applying ML techniques in real-world settings. For given tasks, AutoML automates the pipeline from raw data to deployable ML models, covering model design [18], optimizer selection [37], and parameter tuning [1]. The use of AutoML greatly simplifies the development of ML systems and propels the trend of ML democratization. Many IT giants have unveiled their AutoML frameworks, such as Microsoft Azure AutoML, Google Cloud AutoML, and IBM Watson AutoAI.

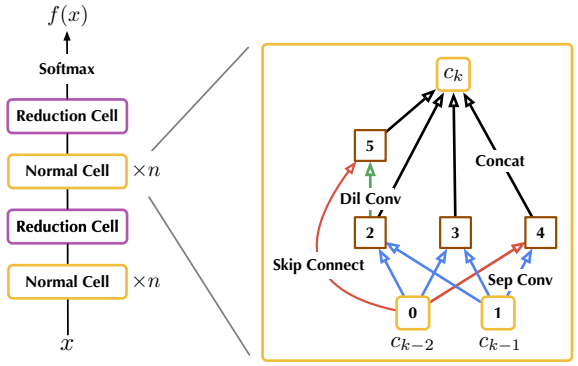

Figure 1: Cell-based neural architecture search.

In this paper, we focus on one primary task of AutoML, Neural Architecture Search (NAS), which aims to find performant deep neural network (DNN) architectures[1] tailored to given tasks. For instance, as illustrated in Figure 1, cell-based NAS constructs a model by repeating the motif of a cell structure following a pre-specified template, wherein a cell is a topological combination of operations (*e.g.*, 3×3 convolution). With respect to the given task, NAS optimizes both the topological structure and the operation assignment. It is shown that in many tasks, NAS finds models that remarkably outperform manually designed ones [11, 35, 39, 46].

Yet, in contrast to the intensive research on improving the capabilities of NAS, its security implications are fairly unexplored. As ML systems are becoming the new targets for malicious attacks [6], the lack of understanding about the potential risks of NAS is highly concerning, given its surging popularity in security-sensitive applications. Specifically,

[1] In the following, when the context is clear, we use the terms of "model" and "architecture" exchangeably.

RQ1 – *Does NAS introduce new weaknesses, compared with the conventional ML practice?*

RQ2 – *If so, what are the possible root causes of such vulnerability?*

RQ3 – *Further, how would ML practitioners mitigate such drawbacks in designing and operating NAS?*

The answers to these key questions are crucial for the use of NAS in security-sensitive domains (*e.g.*, cyber-security, finance, and healthcare).

Our work – This work represents a solid initial step towards answering such questions.

A1 - First, through an extensive empirical study of 10 representative NAS methods, we show that compared with their manually designed counterparts, NAS-generated models tend to suffer greater vulnerability to various malicious manipulations such as adversarial evasion [8, 42], model poisoning [5], backdoor injection [23, 40], functionality stealing [44], and label-only membership inference [13]. The findings suggest that NAS is likely to incur larger attack surfaces, compared with the conventional ML practice.

A2 - Further, with both empirical and analytical evidence, we provide possible explanations for the above observations. Intuitively, due to the prohibitive search space and training cost, NAS tends to prematurely evaluate the quality of candidate models before their convergence. This practice favors models that converge fast at early training stages, resulting in architectural properties that facilitate various attacks (*e.g.*, high loss smoothness and low gradient variance). Our analysis not only reveals the relationships between model characteristics and attack vulnerability but also suggests the inherent connections underlying different attacks.

A3 - Finally, we discuss potential remedies. Besides post-NAS mitigation (*e.g.*, adversarial training [42]), we explore in-NAS strategies that build attack robustness into the NAS process, such as increasing cell depth and suppressing skip connects. We show that while such strategies mitigate the vulnerability to a certain extent, they tend to incur non-trivial costs of search efficiency and model performance. We deem understanding the fundamental trade-off between model performance, attack robustness and search efficiency as an important topic for further investigation.

Contributions – To our best knowledge, this work represents the first study on the potential risks incurred by NAS (and AutoML in general) and reveals its profound security implications. Our contributions are summarized as follows.

– We demonstrate that compared with conventional ML practice, NAS tends to introduce larger attack surfaces with respect to a variety of attacks, which raises severe concerns about the use of NAS in security-sensitive domains.

– We provide possible explanations for such vulnerability, which reveal the relationships between architectural properties (*i.e.*, gradient smoothness and gradient variance) and attack vulnerability. Our analysis also hints at the inherent connections underlying different attacks.

– We discuss possible mitigation to improve the robustness of NAS-generated models under both in-situ and ex-situ settings. This discussion suggests the necessity of improving the current practice of designing and operating NAS, pointing to several research directions.

2 Preliminaries

We first introduce a set of key concepts and assumptions.

2.1 Neural Architecture Search

Deep neural networks (DNNs) represent a class of ML models to learn high-level abstractions of complex data. We assume a predictive setting, in which a DNN f_θ (parameterized by θ) encodes a function $f_\theta : \mathbb{R}^n \to \mathbb{S}^m$, where n and m denote the input dimensionality and the number of classes. Given input x, $f(x)$ is a probability vector (simplex) over m classes.

In this paper, we mainly focus on one primary task of AutoML, neural architecture search (NAS), which searches for performant DNN architectures for given tasks [18]. Formally, let $\mathcal{D}$ be the given dataset, $\ell(\cdot,\cdot)$ be the loss function, $\mathcal{F}$ be the functional space of possible models (*i.e.*, search space), the NAS method A searches for a performant DNN f^* via minimizing the following objective:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(f(x), y) \tag{1}$$

The existing NAS methods can be categorized according to their search spaces and strategies. In the following, we focus on the space of cell-based architectures [39,46,47,58,64], which repeat the motif of a cell structure in a pre-specified arrangement, and the strategy of differentiable NAS [11, 35, 39], which jointly optimizes the architecture and model parameters using gradient descent, due to their state-of-the-art performance and efficiency. Nevertheless, our discussion generalizes to alternative NAS frameworks (details in § 6).

Without loss of generality, we use *DARTS* [39] as a concrete example to illustrate differentiable NAS. At a high level, *DARTS* searches for two cell structures (*i.e.*, normal cell and reduction cell) as the basic building blocks of the final architecture. As shown in Figure 1, a cell is modeled as a directed acyclic graph, in which each node $x^{(i)}$ is a latent representation and each directed edge (i, j) represents an operation $o^{(i,j)}$ applied on $x^{(i)}$ (*e.g.*, skip connect). Each node is computed based on all its predecessors:

$$x^{(j)} = \sum_{i<j} o^{(i,j)}(x^{(i)}) \tag{2}$$

Each cell contains n_{in} input nodes (often $n_{\text{in}} = 2$), n_{out} output nodes (often $n_{\text{out}} = 1$), and n_{mid} intermediate nodes. Each input node takes the output from a preceding cell, the output

node aggregates the latent representations from intermediate nodes, while each intermediate node is connected to m preceding nodes (typically $m = n_{\text{in}}$).

To enable gradient-based optimization of the architecture, *DARTS* applies continuous relaxation on the search space. Letting O be the set of candidate operations, the categorical choice of an operation is reduced to a softmax over O:

$$\bar{o}^{(i,j)}(x) = \sum_{o \in O} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in O} \exp(\alpha_{o'}^{(i,j)})} o(x) \tag{3}$$

where $\alpha_o^{(i,j)}$ represents the trainable weight of operation o. At the end of the search, a discrete architecture is obtained by replacing $\bar{o}^{(i,j)}$ with the most likely operation $\arg\max_o \alpha_o^{(i,j)}$.

The search is thus formulated as a bi-level optimization objective function:

$$\min_{\alpha} \mathcal{L}_{\text{val}}(\theta^*(\alpha), \alpha) \quad \text{s.t.} \quad \theta^*(\alpha) = \arg\min_{\theta} \mathcal{L}_{\text{trn}}(\theta, \alpha) \tag{4}$$

where $\mathcal{L}_{\text{trn}}$ and $\mathcal{L}_{\text{val}}$ are the training and validation losses, and $\alpha = \{\alpha^{(i,j)}\}$ and θ denote the architecture and model parameters, respectively. To handle the prohibitive cost of the nested optimization, single-step gradient descent is applied to avoid solving the inner objective exactly.

2.2 Attack Vulnerability

It is known that DNN models are vulnerable to a variety of attacks at both training and inference phases. Here, we highlight the following major attacks.

Adversarial evasion – At inference time, the adversary generates an adversarial input $(x+\delta)$ by modifying a begin one x with imperceptible perturbation δ, to cause the target model f to misbehave [21]. Formally, in a targeted attack, letting t be the target class desired by the adversary, the attack crafts $(x+\delta)$ by optimizing the following objective:

$$\min_{\delta \in \mathcal{B}_\varepsilon} \ell(f(x+\delta), t) \tag{5}$$

where $\mathcal{B}_\varepsilon$ specifies the set of allowed perturbation(*e.g.*, a ℓ_∞-norm ball of radius ε). Eqn (5) is often solved using projected gradient descent [42] or general-purpose optimizers [8].

Model poisoning – The adversary aims to modify a target model f's behavior (*e.g.*, overall performance degradation or misclassification of specific inputs) via polluting its training data [5]. For instance, to cause the maximum accuracy drop, letting $\mathcal{D}_{\text{trn}}$ and $\mathcal{D}_{\text{tst}}$ be the training and testing sets and f be the target model, the attack crafts a set of poisoning inputs $\mathcal{D}_{\text{pos}}$ by optimizing the the following objective (note: the adversary may not have access to $\mathcal{D}_{\text{trn}}$, $\mathcal{D}_{\text{tst}}$, or f):

$$\begin{aligned} &\max \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{tst}}} \ell(f_{\theta^*}(x), y) \\ &\text{s.t. } \theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{trn}}\cup\mathcal{D}_{\text{pos}}} \ell(f_\theta(x), y) \end{aligned} \tag{6}$$

Backdoor injection – During training, via perturbing a benign model f, the adversary forges a trojan model f_{θ^*} sensitive to a trigger pattern r^*, which is used in the downstream task by the victim; at inference time, the adversary invokes the malicious function by feeding trigger-embedded input $x + r^*$. Formally, letting $\mathcal{D}_{\text{trn}}$ be the training data and t be the target class desired by the adversary, the attack generates a trojan model parameterized by θ^* and its associated trigger r^* by optimizing the following objective:

$$\min_{r \in \mathcal{R}_q, \theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{trn}}} [\ell(f_\theta(x), y) + \lambda \ell(f_\theta(x+r), t)] \tag{7}$$

where r^* is selected from a feasible set $\mathcal{R}_q$ (*e.g.*, a 3×3 patch with transparency γ), the first term enforces all clean inputs to be correctly classified, the second term ensures all trigger inputs to be misclassified into t, and the hyper-parameter λ balances the two objectives.

Functionality stealing – In functionality stealing [44], the adversary aims to construct a replicate model $\hat{f}$ (parameterized by θ^*) functionally similar to a victim model f via probing f through a black-box query interface. Notably, it is different from model stealing [54] that aims to re-construct f in terms of architectures or parameters. Formally, letting $\mathcal{D}$ be the underlying data distribution, the attack generates the query-prediction set $\mathcal{Q}$ (note: the adversary may not have the labeling of $\mathcal{D}$, has only query access to f, and is typically constrained by the number of queries to be issued), which optimizes the following objective:

$$\begin{aligned} &\min \mathbb{E}_{x\sim\mathcal{D}} \ell(\hat{f}_{\theta^*}(x), f(x)) \\ &\text{s.t. } \theta^* = \arg\min_{\theta} \mathbb{E}_{(x,f(x))\sim\mathcal{Q}} \ell(\hat{f}_\theta(x), f(x)) \end{aligned} \tag{8}$$

Different functionality stealing attacks differ in how $\mathcal{Q}$ is constructed (*e.g.*, random or adaptive construction).

Membership inference – In membership inference [50], given input x and model's prediction $f(x)$, the adversary attempts to predict a binary variable b indicating whether x is included in f's training data: $b \leftarrow \mathcal{A}(x, f)$. The effectiveness of membership inference relies on f's performance gap with respect to the training data $\mathcal{D}_{\text{trn}}$ and testing data $\mathcal{D}_{\text{tst}}$. The adversary may exploit this performance gap by thresholding the confidence score of $f(x)$ if it is available, or estimating other signals (*e.g.*, x's distance to the nearest decision boundary) if only the label of $f(x)$ is provided [13].

3 Measurement

To investigate the security risks incurred by NAS, we empirically compare the vulnerability of NAS-generated and manually designed models to the aforementioned attacks.

	Architecture	CIFAR10	CIFAR100	ImageNet32
Manual Architecture	*BiT* [32]	96.6%	80.6%	72.1%
	DenseNet [28]	96.7%	80.7%	73.6%
	DLA [60]	96.5%	78.0%	70.8%
	ResNet [26]	96.6%	79.9%	67.1%
	ResNext [57]	96.7%	80.4%	67.4%
	VGG [52]	95.1%	73.9%	62.3%
	WideResNet [61]	96.8%	81.0%	73.9%
NAS Architecture	*AmoebaNet* [47]	96.9%	78.4%	74.8%
	DARTS [39]	97.0%	81.7%	76.6%
	DrNAS [11]	96.9%	80.4%	75.6%
	ENAS [46]	96.8%	79.1%	74.0%
	NASNet [64]	97.0%	78.8%	73.0%
	PC-DARTS [59]	96.9%	77.4%	74.7%
	PDARTS [12]	97.1%	81.0%	75.8%
	SGAS [35]	97.2%	81.2%	76.8%
	SNAS [58]	96.9%	79.9%	75.5%
	Random [17]	96.7%	78.6%	72.2%

Table 1. Accuracy of representative NAS-generated and manually designed models on benchmark datasets.

3.1 Experimental Setting

We first introduce the setting of the empirical evaluation. The default parameter setting is deferred to Table 5 in § B.

Datasets – In the evaluation, we primarily use 3 datasets that have been widely used to benchmark NAS performance in recent work [12,35,39,46,58]: CIFAR10 [33] – it consists of 32×32 color images drawn from 10 classes (*e.g.*, 'airplane'); CIFAR100 – it is essentially the CIFAR10 dataset but divided into 100 fine-grained classes; ImageNet32 – it is a subset of the ImageNet dataset [15], downsampled to images of size 32×32 in 60 classes.

NAS methods – We consider 10 representative cell-based NAS methods, which cover a variety of search strategies: (1) *AmoebaNet* [47] applies an evolutionary approach to generate candidate models; (2) *DARTS* [39] is the first differentiable method using gradient descent to optimize both architecture and model parameters; (3) *DrNAS* [11] formulates differentiable NAS as a Dirichlet distribution learning problem; (4) *ENAS* [46] reduces the search cost via parameter sharing among candidate models; (5) *NASNet* [64] searches for cell structures transferable across different tasks by re-designing the search space; (6) *PC-DARTS* [59] improves the memory efficiency by restricting operation selection to a subset of edges; (7) *PDARTS* [12] gradually grows the number of cells to reduce the gap between the model depth at the search and evaluation phases; (8) *SGAS* [35] selects the operations in a greedy, sequential manner; (9) *SNAS* [58] reformulates reinforcement learning-based NAS to make it differentiable; and (10) *Random* [17] randomly samples candidate models from the pre-defined search space.

NAS search space – We define the default search space similar to *DARTS* [39], which consists of 10 operations including: *skip-connect*, 3×3 *max-pool*, 3×3 *avg-pool*, 3×3 *sep-conv*, 5×5 *sep-conv*, 7×7 *sep-conv*, 3×3 *dil-conv*, 5×5 *dil-conv*, $1 \times 7 - 7 \times 1$ *conv*, and *zero*.

Manual models – For comparison, we use 7 representative manually designed models that employ diverse architecture designs: (1) *BiT* [32] uses group normalization and weight standardization to facilitate transfer learning; (2) *DenseNet* [28] connects all the layers via skip connects; (3) *DLA* [60] applies deep aggregation to fuse features across layers; (4) *ResNet* [26] uses residual blocks to facilitate gradient back-propagation; (5) *ResNext* [57] aggregates transformations of the same topology; (6) *VGG* [52] represents the conventional deep convolution structures; and (7) *WideResNet* [61] decrease the depth and increases the width of *ResNet*.

Training – All the models are trained using the following setting: epochs = 600, batch size = 96, optimizer = SGD, gradient clipping threshold = 5.0, initial learning rate = 0.025, and learning rate scheduler = Cosine annealing. The accuracy of all the models on the benchmark datasets is summarized in Table 1. Observe that the NAS models often outperform their manual counterparts.

3.2 Experimental Results

Next, we empirically compare the vulnerability of NAS-generated and manually designed models to various attacks.

Adversarial evasion – We exemplify with the projected gradient descent (PGD) attack [42]. Over each dataset, we apply the attack on a set of 1,000 inputs randomly sampled from the test set and measure the attack success rate as:

$$\text{Attack Success Rate (ASR)} = \frac{\text{\# Successful trials}}{\text{\# Total trials}} \tag{9}$$

A trial is marked as successful if it is classified as its target class within maximum iterations.

Let $f_c(x)$ be the probability that model f assigns to class c with respect to input x. To assess the full spectrum of vulnerability, we consider both "difficult" and "easy" cases for the adversary. Specifically, given input x, we rank the output classes c's according to their probabilities $f_c(x)$ as $c_1, c_2, \ldots, c_m$, where c_1 is x's current classification; the difficult case refers to that the adversary aims to change x's classification to the least likely class c_m, while the easy case refers to that the adversary aims to change x's classification to the second most likely class c_2. Table 5 summarizes the setting of the attack parameters.

Figure 2 illustrates the attack effectiveness against both NAS and manual models. We have the following observations. First, across all the datasets, the NAS models seem more vulnerable to adversarial evasion. For instance, on CIFAR10, the attack attains over 90% and 75% ASR against the NAS models in the most and least likely cases, respectively. Second, compared with the manual models, the ASR of NAS models demonstrates more evident clustered structures, implying their similar vulnerability. Finally, the vul-

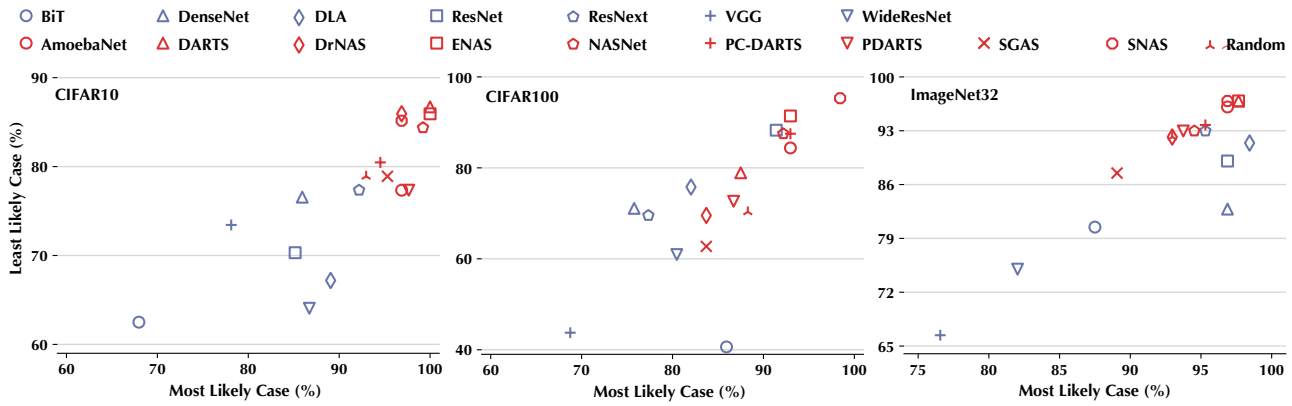

Figure 2: Performance of adversarial evasion (PGD) against NAS and manual models under the least and most likely settings.

nerability of NAS models shows varying patterns on different datasets. For instance, the measures of NAS models show a larger variance on CIFAR100 compared with CIFAR10 and ImageNet32 (especially in the least likely case), which may be explained by that its larger number of classes results in more varying "degree of difficulty" for the attack.

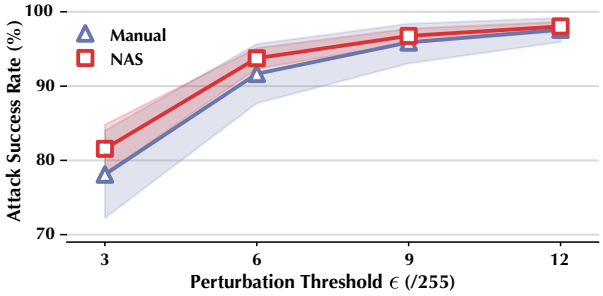

Figure 3: Impact of perturbation threshold (ε) on the vulnerability of different models with respect to PGD on CIFAR10.

We also evaluate the impact of perturbation threshold (ε) on the attack vulnerability. Figure 3 shows the ASR of untargeted PGD as a function of ε against different models on CIFAR10 (with perturbation step $\alpha = \varepsilon/3$). We have the following observations. First, across different settings, the manual models consistently outperform the NAS models in terms of robustness. Second, this vulnerability gap gradually decreases with ε, as the ASR on both NAS and manual models approaches 100%. Third, compared with the manual models, the measures of NAS models show a smaller variance, indicating the commonality of their vulnerability.

Further, by comparing the sets of adversarial examples to which different models are vulnerable, we show the commonality and difference of their vulnerability. We evaluate PGD ($\varepsilon = 4/255$) against different models on CIFAR10 in the least likely case. For each model, we collect the set of adversarial examples successfully generated from 1,000 random samples. Figure 4 plots the distribution of inputs with respect to the number of successfully attacked models.

Overall, PGD generates more successful adversarial examples against the NAS models than the manual models. Moreover, there are more inputs that lead to successful attacks

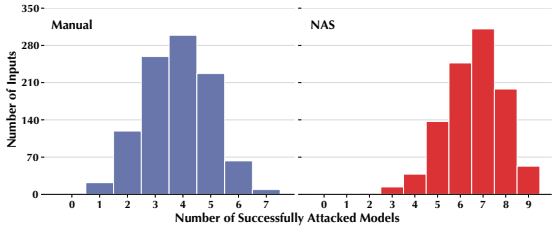

Figure 4: Distribution of inputs with respect to the number of successfully attacked models (PGD with $\varepsilon = 4/255$ on CIFAR10).

against multiple NAS models. For instance, over 300 inputs lead to successful attacks against 7 NAS models; in contrast, the number is less than 10 in the case of manual models. We may thus conclude that the vulnerability of NAS models to adversarial evasion seems fairly similar, pointing to potential associations with common causes.

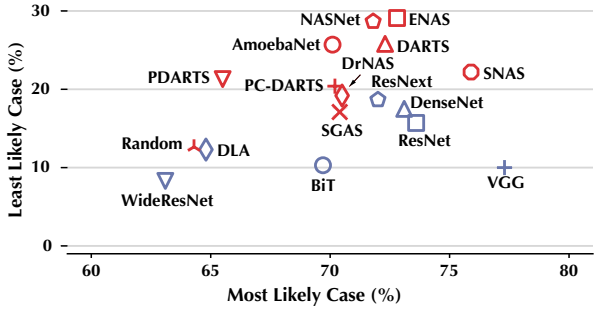

Figure 5: Performance of adversarial evasion (NES) against NAS and manual models under the least and most likely settings.

We also consider alternative adversarial evasion attacks other than PGD. We use natural evolutionary strategies (NES) [29], a black-box attack in which the adversary has only query access to the target model f and generates adversarial examples using a derivative-free optimization approach. Specifically, at each iteration, it generates n_{query} symmetric data points in the vicinity of current input x by sampling from a normal distribution, retrieves their predictions from f, and estimates the gradient $\hat{g}(x)$ as:

$$\hat{g}(x) = \frac{1}{\sigma n_{\text{query}}} \sum_{j=1}^{\lceil n_{\text{query}}/2 \rceil} (f(x+\sigma u_j) - f(x-\sigma u_j))u_j \quad (10)$$

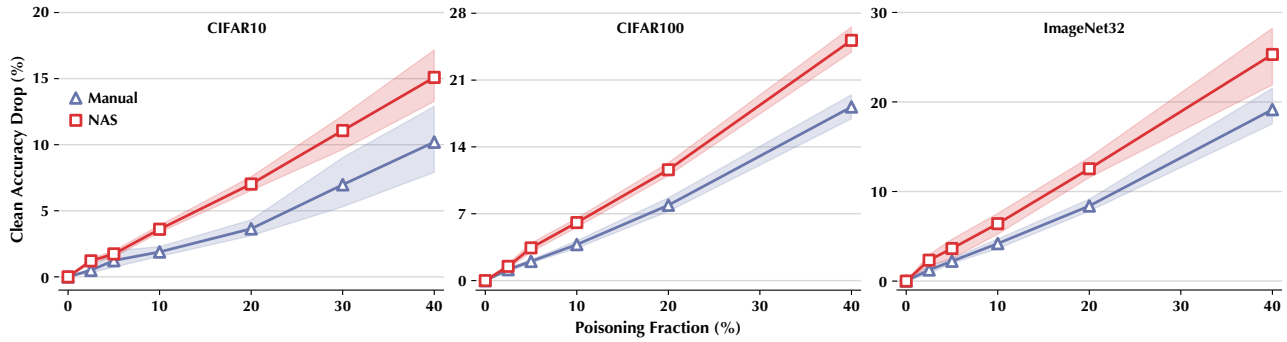

Figure 6: Performance of model poisoning against NAS and manually designed models under varying poisoning fraction p_{pos}.

where each sample u_j is sampled from the standard normal distribution $\mathcal{N}(0, I)$, and σ is the sampling variance.

We evaluate the vulnerability of different models to NES under the same setting of Figure 2 (with $n_{\text{query}} = 400$) on CIFAR10, with results shown in Figure 5. In general, the NAS models show higher vulnerability to NES, especially in the least likely case, indicating that the vulnerability gap between NAS and manual models also generalizes to black-box adversarial evasion attacks.

Model poisoning – In this set of experiments, we evaluate the impact of poisoning attacks on the performance of NAS and manual models. We assume that a fraction p_{pos} of the training data is polluted by randomly changing the class of each input. We measure the performance of various models with respect to varying poisoning fraction p_{pos}, in comparison with the case of clean training data (*i.e.*, $p_{\text{pos}} = 0$). We define the metric of clean accuracy drop:

$$\begin{aligned} &\text{Clean Accuracy Drop (CAD)} \\ &= \text{Acc. of original model} - \text{Acc. of polluted model} \end{aligned} \tag{11}$$

Figure 6 compares the CAD of different models as p_{pos} increases from 0% to 40%. The results are average over the families of NAS and manual models. We have the following observations. First, as expected, larger p_{pos} causes more performance degradation on all the models. Second, with fixed p_{pos}, the NAS models suffer more significant accuracy drop. For instance, on CIFAR100, with p_{pos} fixed as 20%, the CAD of NAS models is 4% higher than the manual models. Further, the CAD gap between NAS and manual models enlarges as p_{pos} increases.

Backdoor injection – Next, we compare the vulnerability of NAS and manual models to neural backdoor attacks [23, 40, 45]. Recall that in backdoor injection, the adversary attempts to forge a trojan model f^* (typically via perturbing a benign model f) that is sensitive to a specific trigger but behaves normally otherwise. We thus measure the attack effectiveness using two metrics: attack success rate (ASR), which is the fraction of trigger-embedded inputs successfully classified by f^* to the target class desired by the adversary; clean accuracy drop (CAD), which is the accuracy difference of f^* and f on clean inputs.

We consider TrojanNN [40], a representative backdoor attack, as the reference attack model. By optimizing both the trigger r and trojan model f^*, TrojanNN enhances other backdoor attacks (*e.g.*, BadNet [23]) that employ fixed triggers. Figure 7 plots the ASR and CAD of all the models, in which the results are average over 1,000 inputs randomly sampled from each testing set. Observe that the attack seems more effective against the NAS models across all the datasets. For instance, on CIFAR10, the attack achieves close to 100% ASR on most NAS models with CAD below 3%. Further, similar to adversarial evasion and model poisoning, the measures of most NAS models (except *Random*) are fairly consistent, indicating their similar vulnerability. Recall that *Random* samples models from the search space; thus, the higher vulnerability of NAS models is likely to be associated with their particular architectural properties.

We further evaluate the impact of the number of target neurons (n_{neuron}) in TrojanNN. Recall that TrojanNN optimizes the trigger with respect to n_{neuron} target neurons. Figure 8 plots the ASR and CAD of TrojanNN against different models under varying setting n_{neuron}. First, across all the settings of n_{neuron}, TrojanNN consistently attains more effective attacks (*i.e.*, higher ASR and lower CAD) on the NAS models than the manual models. Second, as n_{neuron} varies from 1 to 4, the difference of ASR between NAS and manual models decreases, while the difference of CAD tends to increase. This may be explained as follows: optimizing triggers with respect to more target neurons tends to lead to more effective attacks (*i.e.*, higher ASR) but also result in a larger impact on clean inputs (*i.e.*, higher CAD). However, this trade-off is less evident on the NAS models, implying their higher capabilities to fit both poisoning and clean data.

From the experiments above, we may conclude that compared with manual models, NAS models tend to be more vulnerable to backdoor injection attacks, especially under more restricted settings (*e.g.*, fewer target neurons).

Functionality stealing – We now evaluate how various models are subject to functionality stealing, in which each model f as a black box only allowing query access: given input x, f returns its prediction $f(x)$. The adversary attempts to re-construct a functionally similar model f^* based on the query-prediction pairs $\{(x, f(x))\}$.

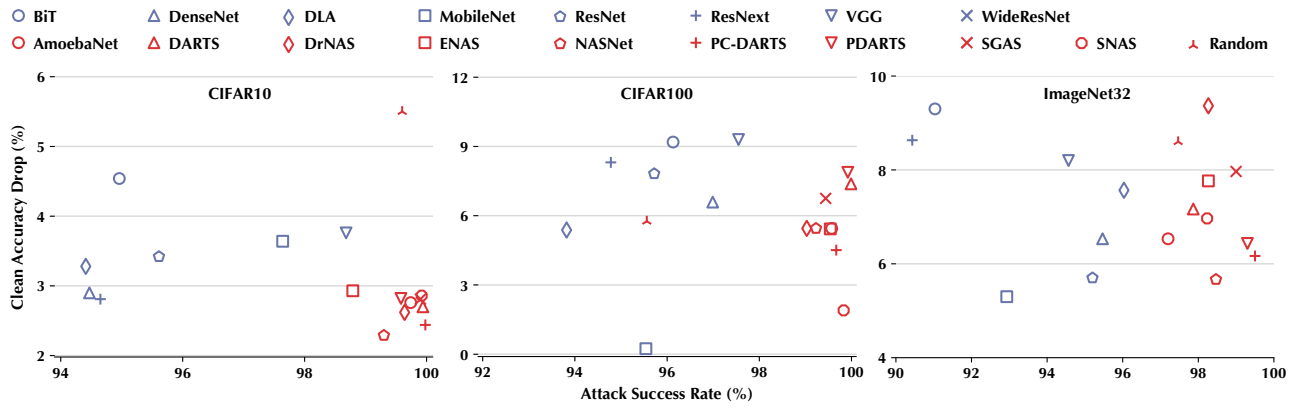

Figure 7: Performance of backdoor injection (TrojanNN) against NAS and manually designed models.

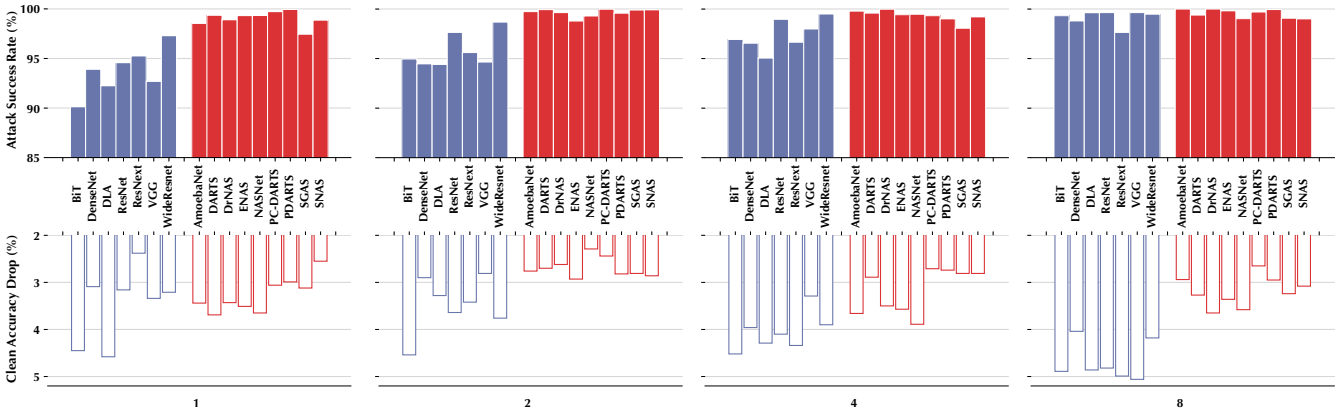

Figure 8: Impact of the number of target neurons (n_{neuron}) on the vulnerability of different models with respect to TrojanNN on CIFAR10.

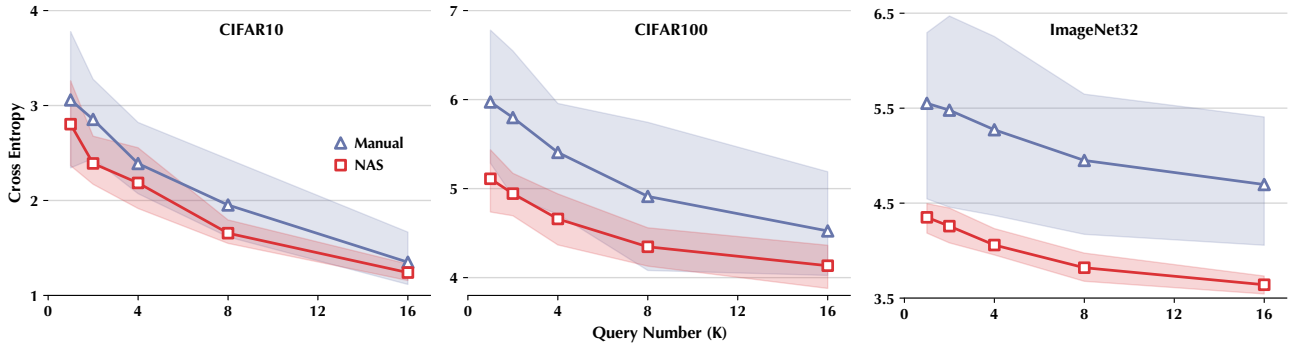

Figure 9: Performance of functionality stealing against NAS and manually designed models under the victim architecture-aware setting.

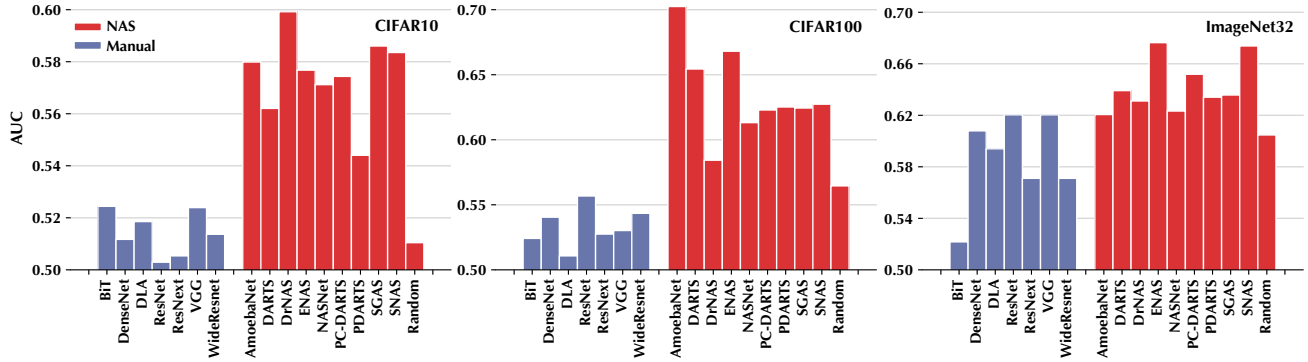

Figure 10: Performance of label-only membership inference attacks against NAS and manually designed models.

We consider two scenarios: (*i*) f and f^* share the same architecture; and (*ii*) the adversary is unaware of f's architecture and instead uses a surrogate architecture in f^*. We apply Knockoff [44], a representative functionality stealing attack that adaptively generates queries to probe f to re-construct f^*. We evaluate the attack using the average cross entropy (ACE) of f's and f^*'s predictions on the testing set, with lower cross entropy indicating more effective stealing.

Victim f \ Replicate f^*		Manual		NAS	
		ResNet	*DenseNet*	*DARTS*	*ENAS*
Manual	*ResNet*	1.286	1.509	1.377	1.455
	DenseNet	1.288	1.245	1.231	1.381
NAS	*DARTS*	1.272	1.115	1.172	1.125
	ENAS	1.259	1.050	1.115	1.151

Table 2. Performance of functionality stealing against NAS and manual models under the victim architecture-agnostic setting.

Figure 9 summarizes the attack effectiveness under the victim architecture-aware setting. Across all the datasets, the attack achieves smaller ACE on the NAS models with much lower variance, in comparison with the manual models. This implies that most NAS models share similar vulnerability to functionality stealing. We further consider the victim architecture-agnostic setting. For each pair of models, we assume one as f and the other as f^*, and measure the attack effectiveness. The results on CIFAR10 (with the query number fixed as 8K) are summarized in Table 2. Observe that with the replicate model f^* fixed, the NAS models as the victim model f result in lower ACE, implying that it tends to be easier to steal the functionality of NAS models, regardless of the architecture of the replicate model.

Membership inference – Recall that in membership inference, the adversary attempts to infer whether the given input x appears in the training set of the target model f. The inference accuracy serves as an indicator of f's privacy leakages. Next, we conduct membership inference attacks on various models to assess their privacy risks.

There are two possible scenarios: (*i*) f's prediction $f(x)$ contains the confidence score $f_c(x)$ of each class c; and (*ii*) $f(x)$ contains only the label $c^* = \arg\max_c f_c(x)$. As (*i*) can be mitigated by removing the confidence scores in $f(x)$ [50], here, we focus on (*ii*). Under the class-only setting, we apply the decision boundary-based attack [13], which determines x's membership (in the training data) by estimating its distance to the nearest decision boundary using label-only adversarial attacks (*e.g.*, HopSkipJump [9]). In each case, we evaluate the attack over 2,000 inputs, half randomly sampled from the training set and the other half from the testing set, and measure the attack effectiveness using the area under the ROC curve (AUC), with the estimated distance as the control of false and true positive rates.

Figure 10 compares the attack performance against different models. Notably, the attack achieves higher AUC scores on the NAS models. For instance, the average scores on the NAS and manual models on CIFAR100 differ by more than 0.05, while the scores on the manual models are close to random guesses (*i.e.*, 0.5). Moreover, most NAS models (except *Random*) show similar vulnerability. Also, note that the manual models seem more vulnerable on ImageNet32, which may be explained as follows: compared with CIFAR10 and CIFAR100, ImageNet32 is a more challenging dataset (see Table 1); the models thus tend to overfit the training set more aggressively, resulting in their higher vulnerability to membership inference.

Remark 1 – *Compared with their manually designed counterparts, NAS-generated models tend to be more vulnerable to various malicious manipulations.*

4 Analysis

The empirical evaluation in § 3 reveals that compared with manually designed models, NAS-generated models tend to be more vulnerable to a variety of attacks. Next, we provide possible explanations for such phenomena.

4.1 Architectural Properties of Trainability

We hypothesize that the greater vulnerability of NAS models stems from their key design choices.

Popular NAS methods often evaluate the performance of a candidate model prematurely before its full convergence during the search. For instance, *DARTS* [39] formulate the search as a bi-level optimization problem, in which the inner objective optimizes a given model; to save the computational cost, instead of solving this objective exactly, it approximates the solution using a single training step, which is far from its full convergence. Similar techniques are applied in other popular NAS methods (*e.g.*, [46,47]). As the candidate models are not evaluated on their performance at convergence, NAS tends to favor models with higher "trainability" – those converge faster during early stages – which result in candidate models demonstrating the following key properties:

High loss smoothness – *The loss landscape of NAS models tends to be smooth, while the gradient provides effective guidance for optimization. Therefore, NAS models are amenable to training using simple, first-order optimizers.*

Low gradient variance – *The gradient of NAS models with respect to the given distribution tends to have low variance. Therefore, the stochastic gradient serves as a reliable estimate of the true gradient, making NAS models converge fast.*

Note that the loss smoothness captures the geometry of the loss function in the parameter space (or the input space), while the gradient variance measures the difference between the gradients with respect to different inputs. While related, the former dictates whether a model is easy to train if the

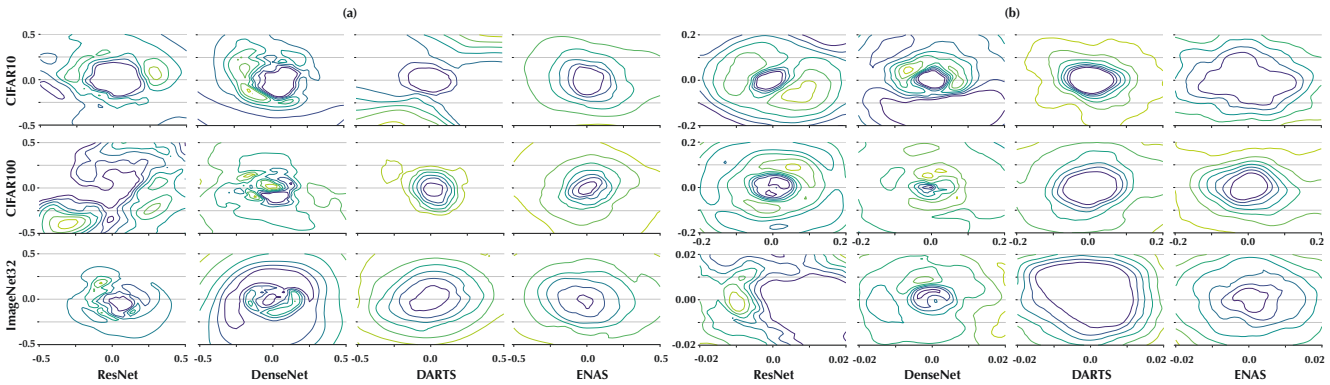

Figure 11: Loss contours of NAS-generated models (*DARTS*, *ENAS*) and manually designed ones (*ResNet*, *DenseNet*) in (a) parameter space and (b) input space.

gradient direction is known and the latter dictates whether it is easy to estimate the gradient direction reliably.

Next, we empirically validate the above hypotheses by comparing the gradient smoothness and variance of NAS-generated and manually designed models.

Loss smoothness – A loss function $\mathcal{L}$ is said to have L-Lipschitz ($L > 0$) continuous gradient with respect to θ if it satisfies $\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\| \leq L\|\theta - \theta'\|$ for any θ, θ'. The constant L controls $\mathcal{L}$'s smoothness. While it is difficult to directly measure L of given model f, we explore its loss contour [22], which quantifies the impact of parameter perturbation on $\mathcal{L}$. Specifically, we measure the loss contour of model f as follows:

$$\Gamma(\alpha, \beta) = \mathcal{L}(\theta^* + \alpha d_1 + \beta d_2) \tag{12}$$

where θ^* denotes the local optimum, d_1 and d_2 are two random, orthogonal directions as the axes, and α and β represent the perturbation steps along d_1 and d_1, respectively. Notably, the loss contour effectively approximates the loss landscape in a two-dimensional space [36].

Figure 11(a) visualizes the loss contours of NAS (*DARTS* and *ENAS*) and manual (*ResNet* and *DenseNet*) models across different datasets. Observe that the NAS models tend to demonstrate a flatter loss landscape. Similar phenomena are observed with respect to other models. This observation may explain why the gradient of NAS models gives more effective guidance for minimizing the loss function, leading to their higher trainability.

Further, for the purpose of the analysis in § 4, we extend the loss smoothness in the parameter space to the input space. We have the following result to show their fundamental connections (proof deferred to § A).

Theorem 1. *If the loss function $\mathcal{L}$ has L-Lipschitz continuous gradient with respect to θ and the weight matrix of each layer of the model is normalized [48], then $\mathcal{L}$ has $L/\sqrt{n}$-Lipschitz continuous gradient with respect to the input, where n is the input dimensionality.*

Empirically, we define f's loss contour with respect to a given input-class pair (x, y) as follows:

$$\Gamma_{(x,y)}(\alpha, \beta) = \ell(f(x + \alpha d_1 + \beta d_2), y) \tag{13}$$

where d_1 and d_2 are two random, orthogonal directions in the input space. Figure 11(b) visualizes the loss contours of NAS and manual models in the vicinity of randomly sampled inputs. It is observed that NAS models also demonstrate higher loss smoothness in the input space, compared with the manual models.

Gradient variance – Meanwhile, the variance of the gradient with respect to inputs sampled from the underlying distribution quantifies the noise level of the gradient estimate used by stochastic training methods (*e.g.*, SGD) [20]. Formally, let g be the stochastic gradient. We define the gradient variance as follows (where the expectation is taken with respect to the given distribution):

$$\mathrm{Var}(g) = \mathbb{E}\left[\|g - \mathbb{E}[g]\|_2^2\right] \tag{14}$$

Assuming g is an unbiased estimate of the true gradient, $\mathrm{Var}(g)$ measures g's expected deviation from the true gradient. Smaller $\mathrm{Var}(g)$ implies lower noise level, thereby more stable updating of the model parameters θ.

In Figure 12, we measure the gradient variance of various models before training (with Kaiming initialization [25]) and after training is complete. It is observed in all the cases that at initialization, the gradient variance of NAS models is more than two orders of magnitude smaller than the manual models and then grows gradually during the training; in comparison, the gradient variance of manual models does not change significantly before and after training. This observation may explain why the stochastic gradient of NAS models gives a reliable estimate of the true gradient, making them converge fast at early training phases.

4.2 Explanations of Attack Vulnerability

We now discuss how the vulnerability of NAS models to various attacks can be attributed to the properties of high loss smoothness and low gradient variance.

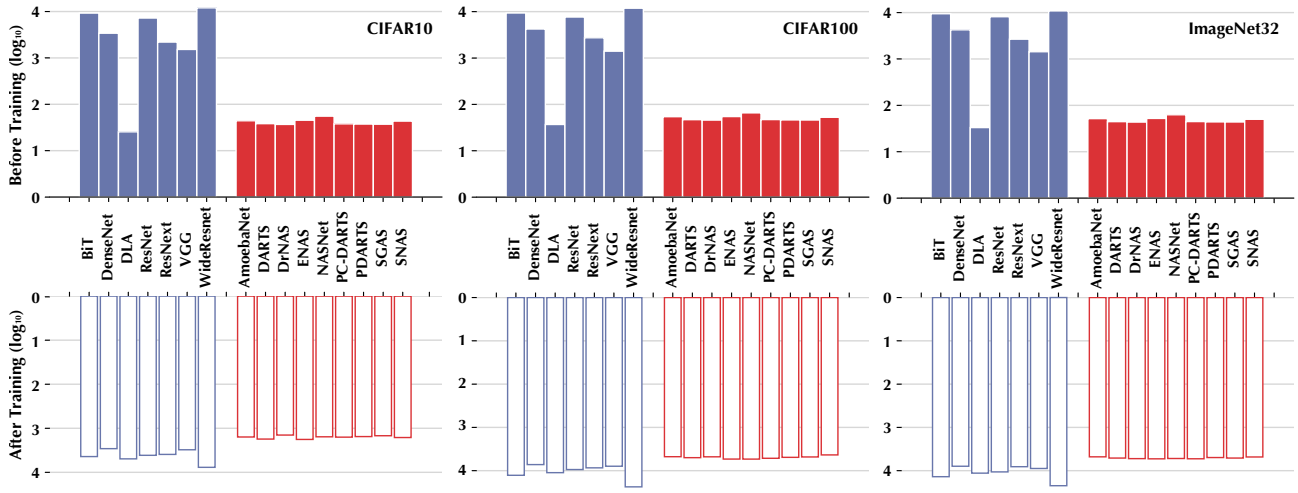

Figure 12: Gradient variance of NAS-generated and manually designed models before and after training.

Adversarial evasion – The vulnerability to adversarial evasion is mainly attributed to the sensitivity of model prediction $f(x)$ to the perturbation of input x. Under the white-box setting, the adversary typically relies on the gradient to craft the adversarial input x^*. For instance, PGD [42] crafts x^* by iteratively updating the input using the following rule:

$$x_{t+1} = \Pi_{x+\mathcal{B}_\varepsilon}\left(x_t + \alpha \operatorname{sgn}\left(\nabla_x \ell(f(x_t), y)\right)\right) \tag{15}$$

where x_t is the perturbed input after the t-th iteration, Π denotes the projection operator, $\mathcal{B}_\varepsilon$ represents the allowed set of perturbation (parameterized by ε), and α is the perturbation step. Apparently, the attack effectiveness relies on whether the gradient $\nabla_x \ell(f(x_t), y)$ is able to provide effective guidance for perturbing x_t.

As shown in § 3.2, compared with the manual models, due to the pursuit of higher trainability, the NAS models often demonstrate a smoother loss landscape wherein the gradient at each point represents effective optimization direction; thus, the NAS models tend to be more vulnerable to gradient-based adversarial evasion. Notably, this finding also corroborates existing studies (*e.g.*, [21]) on the fundamental tension between designing "linear" models that are easier to train and designing "nonlinear" models that are more resistant to adversarial evasion.

The similar phenomena observed in the case of black-box attacks (*e.g.*, NES) may be explained as follows: to perform effective perturbation, black-box attacks often rely on indirect gradient estimation, while the high loss smoothness and low gradient variance of NAS models lead to more accurate and efficient (with fewer queries) gradient estimation.

Model poisoning – The vulnerability to model poisoning can be attributed to the sensitivity of model training to the poisoning data in the training set. Here, we analyze how the property of low gradient variance impacts this sensitivity.

For a given dataset $\mathcal{D}$, let $\mathcal{L}(\theta)$ be the loss of a model f_θ parameterized by θ with respect to $\mathcal{D}$:

$$\mathcal{L}(\theta) \triangleq \frac{1}{|\mathcal{D}|}\sum\nolimits_{(x,y)\in\mathcal{D}} \ell(f_\theta(x), y) \tag{16}$$

Further, let θ^* represent f's (local) optimum with respect to $\mathcal{D}$. With θ initialized as θ_0, consider T-step SGD updates with the t-th step update as:

$$\theta_{t+1} = \theta_t - \alpha_t g_t \tag{17}$$

where α_t is the step size and g_t is the gradient estimate. We have the following result describing the convergence property of θ_t $(t = 1, \ldots, T)$.

Theorem 2 ([20]). *Assuming that (i) $\mathcal{L}(\theta)$ is continuous and differentiable, with its gradient bounded by Lipschitz constant L, (ii) the variance of gradient estimate g_t $(t = 1, \ldots, T)$ is bounded by σ^2, and (iii) θ_t is selected as the final parameters with probability proportional to $2\alpha_t - L\alpha_t^2$. Then, the output parameters $\theta_{\bar{t}}$ satisfies:*

$$\mathbb{E}\left[\mathcal{L}(\theta_{\bar{t}}) - \mathcal{L}(\theta^*)\right] \leq \frac{\|\theta_0 - \theta^*\|^2 + \sigma^2 \sum_{t=1}^{T} \alpha_t^2}{\sum_{t=1}^{T}(2\alpha_t - L\alpha_t^2)} \tag{18}$$

where the expectation is defined with respect to the selection of $\bar{t}$ and the gradient variance.

Intuitively, Theorem 2 describes the properties that impact the fitting of model f to the given dataset $\mathcal{D}$. As shown in § 3.2, compared with the manual models, the NAS models tend to have both higher loss smoothness (*i.e.*, smaller L) and lower gradient variance (*i.e.*, smaller σ). Therefore, the NAS models tend to fit $\mathcal{D}$ more easily. Recall that in model poisoning, $\mathcal{D}$ consists of both clean data $\mathcal{D}_{\text{trn}}$ and poisoning data $\mathcal{D}_{\text{pos}}$, fitting to $\mathcal{D}$ more tightly implies more performance drop over the testing data, which may explain the greater vulnerability of NAS models to model poisoning.

Backdoor injection – Recall that in backdoor injection, the adversary forges a trojan model f^* that is sensitive to a trigger pattern r such that any input x, once embedded with r, tends to be misclassified to a target class t: $f^*(x+r) = t$. To train f^*, the adversary typically pollutes the training data $\mathcal{D}_{\text{trn}}$ with trigger-embedded inputs.

Intuitively, this attack essentially exploits the attack vectors of adversarial evasion that perturbs x at inference time and model poisoning that pollutes $\mathcal{D}_{\text{trn}}$ at training time. Therefore, the vulnerability of NAS models to both attack vectors naturally results in their vulnerability to backdoor injection. Due to the space limitations, we omit the detailed analysis here.

Functionality stealing – Recall that in functionality stealing (*e.g.*, Knockoff [44]), the adversary (adaptively) generates queries to probe the victim model f to replicate a functionally similar one f^*. For instance, Knockoff encourages queries that are certain by f, diverse across different classes, and disagreed by f^* and f.

The effectiveness of such attacks depends on f's loss landscape with respect to the underlying distribution; intuitively, the complexity of the loss landscape in the input space implies the hardness of fitting f^* to f based on a limited number of queries. Thus, given their high loss smoothness, the NAS models tend to be more vulnerable to functionality stealing.

Membership inference – It is shown in § 3 that the NAS models seem more vulnerable to membership inference, especially under the label-only setting in which only the prediction labels are accessible. The adversary thus relies on signals such as input x's distance to its nearest decision boundary $\text{dist}(x, f(x))$; intuitively, if x appears in the training set, $\text{dist}(x, f(x))$ is likely to be below a certain threshold. Concretely, the HopSkipJump attack [9] is employed in [13] to estimate $\text{dist}(x, f(x))$ via iteratively querying f to find point x_t on the decision boundary using bin search, walking along the boundary using the estimated gradient at x_t, and finding point x_{t+1} to further reduce the distance to x, which is illustrated in Figure 13.

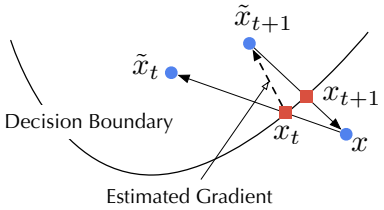

Figure 13: Illustration of the HopSkipJump attack.

The effectiveness of this attack hinges on (*i*) the quality of estimated gradient and (*ii*) the feasibility of descending along the decision boundary. For the NAS models, the gradient estimate tends to be more accurate due to the low gradient variance, while the decision boundary tends to be smoother due to the high loss smoothness, which may explain the greater vulnerability of NAS models to label-only membership inference attacks.

Remark 2 – *The high loss smoothness and low gradient variance of NAS-generated models may account for their greater vulnerability to various attacks.*

4.3 Connections of Various Attacks

It is shown above that the vulnerability of NAS models to various attacks may be explained by their high loss smoothness and low gradient variance, which bears an intriguing implication: different attacks may also be inherently connected via these two factors.

Specifically, most existing attacks involve input or model perturbation. For instance, adversarial evasion, regardless of the white- or black-box setting, iteratively computes (or estimates) the gradient and performs perturbation accordingly; backdoor injection optimizes the trigger and model jointly, requiring to estimate, based on the gradient, how the model responds to the updated trigger.

The effectiveness of such attacks thus highly depends on (*i*) how to estimate the gradient at each iteration and (*ii*) how to use the gradient estimate to guide the input or model perturbation. Interestingly, gradient variance and loss smoothness greatly impact (*i*) and (*ii*), respectively: low gradient variance enables the adversary to accurately estimate the gradient, while high loss smoothness allows the adversary to use such estimate to perform effective perturbation.

Remark 3 – *The effectiveness of various attacks is inherently connected through loss smoothness and gradient variance.*

5 Discussion

In § 3 and § 4, we reveal the relationships between the trainability of NAS-generated models and their vulnerability to various attacks, two key questions remain: (*i*) what are the architectural patterns associated with such vulnerability? and (*iii*) what are the potential strategies to remedy the vulnerability incurred by the current NAS practice? In this section, we explore these two questions and further discuss the limitations of this work.

5.1 Architectural Weaknesses

As shown in § 4, the vulnerability of NAS models is potentially related to their high loss smoothness and low gradient variance, which stem from the preference for models of higher trainability. We now discuss how such preference is reflected in the concrete architectural patterns, which we examine from two aspects, namely, topology selection and operation selection.

Topology selection – Recent studies [51] suggest that in cell-based NAS, the preference for models with faster convergence often results in wide, shallow cell structures. As

Architecture	Cell Depth	Cell Width	# Skip connects
AmoebaNet	4	3c	2
DARTS	3	3c	3
DrNAS	4	2c	1
ENAS	2	5c	2
NASNet	2	5c	1
PC-DARTS	2	4c	1
PDARTS	4	2c	2
SGAS	3	2c	1
SNAS	2	4c	4

Table 3. The cell depth and width, and the number of skip connects of representative NAS-generated models (the width of each intermediate node is assumed to be c).

shown in Figure 1, the cell depth is defined as the number of connections along the longest path from the input nodes to the output node; the width of each intermediate node is defined as the number of channels for convolution operators or the number of features for linear operators, while the cell width is defined as the total width of intermediate nodes connected to the input nodes. Table 3 summarizes the cell depth and width of NAS models used in our evaluation. It is observed that the cell structures of most NAS models are both shallow (with an average depth of 2.8) and wide (with an average width of 3.3c), where the width of each intermediate node is assumed to be c.

It is shown in [51] that under similar settings (*i.e.*, the same number of nodes and connections), wide and shallow cells tend to demonstrate higher trainability. This observation is also corroborated by the recent theoretical studies on the convergence of wide neural networks [34]: neural networks of infinite width tend to evolve as linear models using gradient descent optimization.

Operation selection – The preference for higher trainability also impacts the selection of operations (*e.g.*, 3×3 convolution versus skip connection) on the connections within the cell structure, and typically favors skip connects over other operations.

Recall that differential NAS methods [11, 35, 39] typically apply continuous relaxation on the search space to enable direct gradient-based optimization. The operation on each connection is modeled as a softmax of all possible operations O and discretized by selecting the most likely one $\arg\max_{o\in O}\alpha_o$. It is shown in [55] that in well-optimized models, the weight of skip connection α_{skip} often exceeds other operations, leading to its higher chance of being selected. This preference takes effect in our context, as NAS models tend to converge fast at early training stages. Table 3 summarizes the number of skip connects in each cell of representative NAS models. Observe that most NAS models have more than one skip connection in each cell.

The operation of skip connection is originally designed to enable back-propagation in DNNs [26, 28]. As a side effect, accurate gradient estimation also facilitates attacks that exploit gradient information [56]. Thus, the over-use of skip connects in NAS models also partially accounts for their vulnerability to such attacks.

Remark 4 – *NAS-generated models often feature wide and shallow cell structures as well as overuse of skip connects.*

5.2 Potential Mitigation

We now discuss potential mitigation to remedy the vulnerability incurred by the NAS practice. We consider enhancing the robustness of NAS models under both post-NAS and in-NAS settings. In post-NAS mitigation, we explore using existing defenses against given attacks to enhance NAS models, while with in-NAS mitigation, we explore building attack robustness into the NAS process directly.

Post-NAS mitigation – As a concrete example, we apply adversarial training [41, 49], one representative defense against adversarial evasion, to enhance the robustness of NAS models. Intuitively, adversarial training improves the robustness of given model f by iteratively generating adversarial inputs with respect to its current configuration and updating f to correctly classify such inputs.

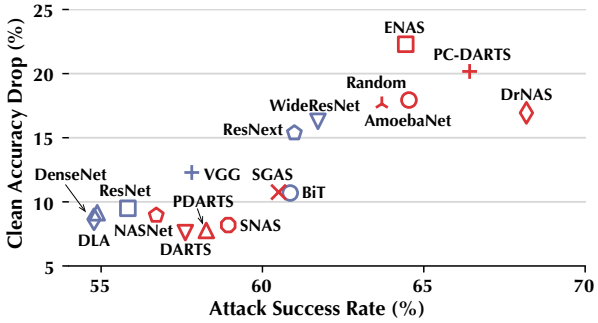

Figure 14: Effectiveness of adversarial training on various models over CIFAR10.

Figure 14 compares the effectiveness of adversarial training on various models over CIFAR10. For each model, we measure its accuracy (in terms of accuracy drop from before adversarial training) and robustness (in terms of the success rate of the untargeted PGD attack). Observe that a few NAS models (*e.g.*, *DARTS*) show accuracy and robustness comparable with manual models, while the other NAS models (*e.g.*, *DrNAS*) underperform in terms of both accuracy and robustness, which may be explained by their diverse architectural patterns associated with adversarial training (*e.g.*, dense connections, number of convolution operations, and cell sizes) [24]. This disparity also implies that adversarial training may not be a universal solution for improving the robustness of all the NAS models.

In-NAS mitigation – We further investigate how to build attack robustness into the NAS process directly. Motivated by the analysis in § 5.1, we explore two potential strategies.

(*i*) Increasing cell depth – As the vulnerability of NAS

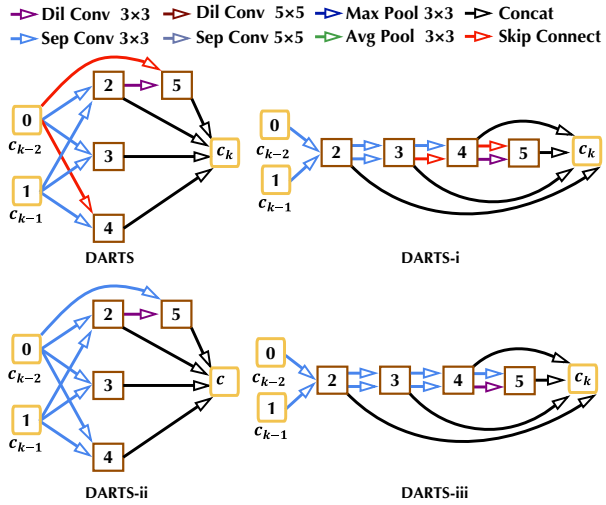

Figure 15: Illustration of cell structures of *DARTS*, *DARTS-i*, *DARTS-ii*, and *DARTS-iii*.

models tends to be associated with their wide and shallow cell structures, we explore increasing their cell depth. To this end, we may re-wire existing NAS models or modify the performance measure of candidate models. For the latter case, we may increase the number of training epochs before evaluation. For instance, *DARTS*, without fully optimizing model parameters θ with respect to architecture parameters α, uses a single-step gradient descent ($n_{\text{step}} = 1$) to approximate the solution [39]. We improve the approximation by increasing the number of training steps (*e.g.*, $n_{\text{step}} = 5$) at the cost of additional search time.

(*ii*) Suppressing skip connects – As the vulnerability of NAS models is also associated with skip connects, we explore purposely reducing their overuse. To this end, we may replace the skip connects in existing NAS models with other operations (*e.g.*, convolution) or modify their likelihood of being selected in the search process. For the latter case, at each iteration, we may multiply the weight of skip connection α_{skip} by a coefficient $\gamma \in (0, 1)$ in Eqn (3).

We evaluate the effectiveness of such strategies within the *DARTS* framework. Let *DARTS-i*, *DARTS-ii*, and *DARTS-iii* be the variants of *DARTS* after applying the strategies of (*i*), (*ii*), and (*i*) and (*ii*) combined. Figure 15 compares their cell structures. Notably, *DARTS-i* features a cell structure deeper than *DARTS* (5 versus 2), while *DARTS-ii* and *DARTS-iii* substitute the skip connects in *DARTS* and *DARTS-i* with 3×3 convolution, respectively.

Architecture	Evasion		Backdoor		Membership
	ASR (M)	ASR (L)	ASR	CAD	AUC
DARTS	100.0%	86.7%	99.9%	2.7%	0.562
DARTS-i	88.3%	72.7%	90.4%	4.6%	0.534
DARTS-ii	93.0%	75.0%	98.8%	3.0%	0.531
DARTS-iii	82.0%	65.6%	84.2%	4.6%	0.527

Table 4. Vulnerability of *DARTS* and its variants to adversarial evasion (M - most likely case, L - least likely case), backdoor injection, and membership inference on CIFAR10.

Table 4 compares their vulnerability to adversarial evasion, backdoor injection, and membership inference on CIFAR10. The experimental setting is identical to that in § 3. Observe that both strategies may improve the robustness of NAS models against these attacks. For instance, combining both strategies in *DARTS-iii* reduces the AUC score of membership inference from 0.562 to 0.527. Similar phenomena are observed in the case of model extraction attacks. As shown in Figure 16, increasing the cell depth significantly augments the robustness against model extraction, while suppressing skip connects further improves it marginally.

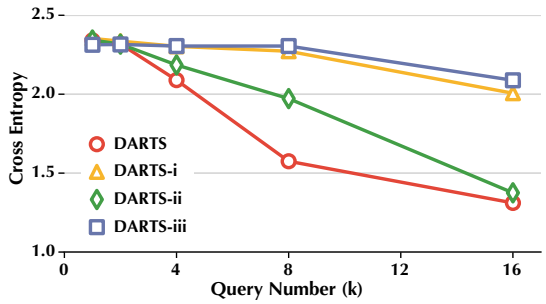

Figure 16: Vulnerability of *DARTS* and its variants to model extraction on CIFAR10.

Yet, such strategies seem to have a negative impact on the robustness against model poisoning. As shown in Figure 17, both strategies, especially increasing the cell depth, tends to exacerbate the attack vulnerability. This may be explained by that while more difficult to fit the poisoning data, it is also more difficult to fit deeper structures to the clean data, which results in a significant accuracy drop. This may also explain why the backdoor injection attack has higher CAD on *DARTS-i* and *DARTS-iii* as shown in Table 4. The observation also implies a potential trade-off between the robustness against different attacks in designing NAS models.

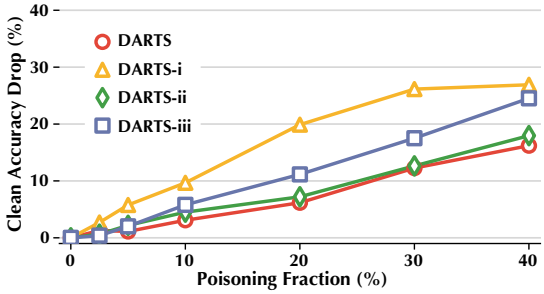

Figure 17: Vulnerability of *DARTS* and its variants to model poisoning on CIFAR10.

Remark 5 – *Simply increasing cell depth and/or suppressing skip connects may only partially mitigate the vulnerability of NAS-generated models.*

5.3 Limitations

Next, we discuss the limitations of this work.

Alternative NAS frameworks – In this work, we mainly consider the cell-based search space adopted by recent NAS methods [11, 14, 39, 46, 62], while other methods have considered the global search space (*e.g.*, chain-of-layer structures) [3, 7]. Further, while we focus on the differentiable search strategy, there are other strategies including random search [31], Bayesian optimization [4], and reinforcement learning [3, 63, 64]. We consider exploring the vulnerability of models generated by alternative NAS frameworks as our ongoing research.

Other trainability metrics – In this work, we only consider loss smoothness and gradient variance as two key factors impacting the trainability (and vulnerability) of NAS models. There are other trainability metrics (*e.g.*, condition number of neural tangent kernel [10]) that are potentially indicative of attack vulnerability as well.

Robustness, accuracy, and search efficiency – It is revealed that the greater vulnerability incurred by NAS is possibly associated with the preference for models that converge fast at early training phases (*i.e.*, higher trainability). It is however unclear whether this observation implies fundamental conflicts between the factors of robustness, accuracy, and search efficiency; if so, is it possible to find an optimal balance between them? We consider answering these questions critical for designing and operating NAS in practical settings.

6 Related Work

Next, we survey the literature relevant to this work.

Neural architecture search – The existing NAS methods can be categorized along three dimensions: search space, search strategy, and performance measure.

The search space defines the possible set of candidate models. Early NAS methods focus on the chain-of-layer structure [3], consisting of a sequence of layers. Motivated by that hand-crafted models often consist of repeated motifs, recent methods propose to search for such cell structures, including the connection topology and the corresponding operation on each connection [39, 46, 47, 58, 64].

The search strategy defines how to efficiently explore the pre-defined search space. Early NAS methods rely on either random search [31] or Bayesian optimization [4], which are often limited in terms of search efficiency and model complexity. More recent work mainly uses the approaches of reinforcement learning (RL) [3] or neural evolution [39, 47]. Empirically, neural evolution- and RL-based methods tend to perform comparably well [47].

The performance measure evaluates the candidate models and guides the search process. Recently, one-shot NAS has emerged as a popular performance measure. It considers all candidate models as different sub-graphs of a super-net (*i.e.*, the one-shot model) and shares weights between candidate models [39, 46, 58]. The differentiable NAS methods considered in this paper belong to this category. Different one-shot methods differ in how the one-shot model is trained. For instance, *DARTS* [39] optimizes the one-shot model with continuous relaxation of the search space.

ML Security – With their wide use in security-sensitive domains, ML models are becoming the new targets for malicious manipulations [6]. A variety of attack vectors have been exploited: adversarial evasion crafts adversarial inputs to force the target model to misbehave [8, 21]; model poisoning modifies the target model's behavior (*e.g.*, performance drop) via polluting its training data [30]; backdoor injection creates a trojan model such that any input embedded with a specific trigger is likely to be misclassified by the model [23, 40]; functionality stealing constructs a replicate model functionally similar to a victim model [27, 44]; membership inference breaches data privacy via inferring whether a given input is included in the model's training data based on the model's prediction [50].

In response, another line of work strives to improve the resilience of ML models against such attacks. For instance, against adversarial evasion, existing defenses explore new training strategies (*e.g.*, adversarial training) [42, 53] and detection mechanisms [19, 43]. Yet, such defenses often fail when facing even stronger attacks [2, 38], resulting in a constant arms race between the attackers and defenders.

Despite the intensive research on NAS and ML security in parallel, the robustness of NAS-generated models to malicious manipulations is fairly under-explored [24]. Concurrent to this work, it is shown in [16] that NAS models tend to be more vulnerable to adversarial evasion, while our work differs in considering a variety of attacks beyond adversarial evasion, providing possible explanations for such vulnerability, and investigating potential mitigation.

7 Conclusion

This work represents a systematic study on the security risks incurred by AutoML. From both empirical and analytical perspectives, we demonstrate that NAS-generated models tend to suffer greater vulnerability to various malicious manipulations, compared with their manually designed counterparts, which implies the existence of fundamental drawbacks in the design of existing NAS methods. We identify high loss smoothness and low gradient variance, stemming from the preference of NAS for models with higher trainability, as possible causes for such phenomena. Our findings raise concerns about the current practice of NAS in security-sensitive domains. Further, we discuss potential remedies to mitigate such limitations, which sheds light on designing and operating NAS in a more robust and principled manner.

Acknowledgments

We thank our shepherd Carmela Troncoso and anonymous reviewers for valuable feedback. This work is supported by the National Science Foundation under Grant No. 1951729, 1953813, and 1953893. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation. S. Ji is partly supported by NSFC under No. 61772466, 62102360, and U1836202, the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under No. LR19F020003, and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform). X. Luo is partly supported by Research Grants Council of the Hong Kong Special Administrative Region, China (No. PolyU15224121)

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to Learn by Gradient Descent by Gradient Descent. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *Proceedings of IEEE Conference on Machine Learning (ICML)*, 2018.

[3] Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing Neural Network Architectures using Reinforcement Learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[4] J. Bergstra, D. Yamins, and D. D. Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of IEEE Conference on Machine Learning (ICML)*, 2013.

[5] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning Attacks against Support Vector Machines. In *Proceedings of IEEE Conference on Machine Learning (ICML)*, 2012.

[6] Battista Biggio and Fabio Roli. Wild Patterns: Ten Years after The Rise of Adversarial Machine Learning. *Pattern Recognition*, 84:317–331, 2018.

[7] Han Cai, Jiacheng Yang, Weinan Zhang, Song Han, and Yong Yu. Path-Level Network Transformation for Efficient Architecture Search. In *Proceedings of IEEE Conference on Machine Learning (ICML)*, 2018.

[8] Nicholas Carlini and David A. Wagner. Towards Evaluating the Robustness of Neural Networks. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2017.

[9] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2020.

[10] Wuyang Chen, Xinyu Gong, and Zhangyang Wang. Neural Architecture Search on ImageNet in Four {GPU} Hours: A Theoretically Inspired Perspective. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[11] Xiangning Chen, Ruochen Wang, Minhao Cheng, Xiaocheng Tang, and Cho-Jui Hsieh. DrNAS: Dirichlet Neural Architecture Search. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[12] Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive Differentiable Architecture Search: Bridging the Depth Gap between Search and Evaluation. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.

[13] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. *ArXiv e-prints*, 2020.

[14] Xiangxiang Chu, Xiaoxing Wang, Bo Zhang, Shun Lu, Xiaolin Wei, and Junchi Yan. DARTS-: Robustly Stepping out of Performance Collapse Without Indicators. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[15] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-scale Hierarchical Image Database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[16] Chaitanya Devaguptapu, Devansh Agarwal, Gaurav Mittal, Pulkit Gopalani, and Vineeth N Balasubramanian. On Adversarial Robustness: A Neural Architecture Search perspective. In *RobustML Workshop of International Conference on Learning Representations*, 2021.

[17] Xuanyi Dong and Yi Yang. NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[18] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural Architecture Search: A Survey. *Journal of Machine Learning Research*, (20):1–21, 2019.

[19] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev. AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2018.

[20] Saeed Ghadimi and Guanghui Lan. Stochastic First- and Zeroth-order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[21] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[22] Ian J. Goodfellow, Oriol Vinyals, and Andrew M. Saxe. Qualitatively Characterizing Neural Network Optimization Problems. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[23] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *ArXiv e-prints*, 2017.

[24] Minghao Guo, Yuzhe Yang, Rui Xu, Ziwei Liu, and Dahua Lin. When NAS Meets Robustness: In Search of Robust Architectures against Adversarial Attacks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] Sanghyun Hong, Michael Davinroy, Yiğitcan Kaya, Dana Dachman-Soled, and Tudor Dumitraş. How to 0wn NAS in Your Spare Time. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[28] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. In *Proceedings of IEEE Conference on Machine Learning (ICML)*, 2018.

[30] Yujie Ji, Xinyang Zhang, Shouling Ji, Xiapu Luo, and Ting Wang. Model-Reuse Attacks on Deep Learning Systems. In *Proceedings of ACM SAC Conference on Computer and Communications (CCS)*, 2018.

[31] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An Empirical Exploration of Recurrent Network Architectures. In *Proceedings of IEEE Conference on Machine Learning (ICML)*, 2015.

[32] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big Transfer (BiT): General Visual Representation Learning. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[33] Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. *Technical report, University of Toronto*, 2009.

[34] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models under Gradient Descent. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[35] Guohao Li, Guocheng Qian, Itzel C. Delgadillo, Matthias Müller, Ali Thabet, and Bernard Ghanem. SGAS: Sequential Greedy Architecture Search. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[36] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the Loss Landscape of Neural Nets. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[37] Ke Li and Jitendra Malik. Learning to Optimize. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[38] X. Ling, S. Ji, J. Zou, J. Wang, C. Wu, B. Li, and T. Wang. DEEPSEC: A Uniform Platform for Security Analysis of Deep Learning Model. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2019.

[39] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable Architecture Search. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[40] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning Attack on Neural Networks. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2018.

[41] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[43] Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.

[44] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff Nets: Stealing Functionality of Black-Box Models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] Ren Pang, Hua Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex Liu, and Ting Wang. A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models. In *Proceedings of ACM SAC Conference on Computer and Communications (CCS)*, 2020.

[46] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient Neural Architecture Search via Parameter Sharing. In *Proceedings of IEEE Conference on Machine Learning (ICML)*, 2018.

[47] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized Evolution for Image Classifier Architecture Search. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[48] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[49] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial Training for Free! In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[50] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks against Machine Learning Models. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2017.

[51] Yao Shu, Wei Wang, and Shaofeng Cai. Understanding Architectures Learnt by Cell-based Neural Architecture Search. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[52] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2014.

[53] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[54] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *Proceedings of USENIX Security Symposium (SEC)*, 2016.

[55] Ruochen Wang, Minhao Cheng, Xiangning Chen, Xiaocheng Tang, and Cho-Jui Hsieh. Rethinking Architecture Selection in Differentiable NAS. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[56] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[57] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[58] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. SNAS: Stochastic Neural Architecture Search. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[59] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong. PC-DARTS: Partial Channel Connections for Memory-Efficient Architecture Search. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[60] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep Layer Aggregation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[61] Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *Proceedings of British Machine Vision Conference (BMVC)*, 2016.

[62] Arber Zela, Thomas Elsken, Tonmoy Saikia, Yassine Marrakchi, Thomas Brox, and Frank Hutter. Understanding and Robustifying Differentiable Architecture Search. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[63] Zhao Zhong, Junjie Yan, Wei Wu, Jing Shao, and Cheng-Lin Liu. Practical Block-wise Neural Network Architecture Generation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[64] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning Transferable Architectures for Scalable Image Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

A Proofs

Proof. (Theorem 1) Without loss of generality, we assume the loss function $\mathcal{L}$ is computed over a single input-label pair:

$$\mathcal{L}(x;\theta) = \ell(f(x;\theta),y) \tag{19}$$

We split the model f into its first layer and the remaining layers (with parameters $\bar{\theta}$). Typically, the first layer (without the non-linear activation) can be modeled as a linear function $Ax+b$ (*e.g.*, fully-connected or convolutional layer). We thus rewrite $\mathcal{L}$ as a composite function: $\mathcal{L}(x;\theta) = \bar{\mathcal{L}}(Ax+b;\bar{\theta})$, where $\bar{\mathcal{L}}$ (parameterized by $\bar{\theta}$) is $\mathcal{L}$ excluding f's first layer.

According to the assumption that $\mathcal{L}(x;\theta)$ has L-Lipschitz continuous gradient with respect to θ, with $\bar{\theta}$ and A fixed,

$$\|\nabla_b \bar{\mathcal{L}}(Ax+b)\| = \|\nabla \bar{\mathcal{L}}\| \leq L \tag{20}$$

Thus, the gradient of $\mathcal{L}$ with respect to x is also bounded:

$$\|\nabla_x \bar{\mathcal{L}}(Ax+b)\| = \|A^\intercal \nabla \bar{\mathcal{L}}\| \leq \|A^\intercal\| \|\nabla \bar{\mathcal{L}}\| \leq L\|A^\intercal\| \tag{21}$$

With weight normalization, $\forall i \sum_j A_{ij} = 0$, $\sum_j A_{ij}^2 = 1$. Applying the Chebyshev's inequality, we bound $\|A^\intercal\|_1$ as:

$$\|A^\intercal\|_1 = \|A\|_\infty = \max_{1\leq i\leq n} \sum_j |A_{ij}| \leq \max_{1\leq i\leq n} \sqrt{\frac{\sum_j A_{ij}^2}{n}} = \frac{1}{\sqrt{n}} \tag{22}$$

where n is the number of rows of A (*i.e.*, the input dimensionality). Putting everything together,

$$\|\nabla_x \mathcal{L}(x;\theta)\|_1 \leq \frac{L}{\sqrt{n}} \tag{23}$$

Therefore, $\mathcal{L}(x;\theta)$ has $\frac{L}{\sqrt{n}}$-Lipschitz continuous gradient with respect to input x.

□

B Parameter Setting

Table 5 summarizes the default parameter setting in § 3.

Type	Parameter	Setting
Training	Optimizer	SGD
	Initial learning rate	0.025
	LR scheduler	Cosine annealing
	Gradient clipping threshold	5.0
	Training epochs	600
	Batch size	96
Adversarial Evasion	Perturbation threshold	$\epsilon = 8/255$
	Learning rate	$\alpha = 2/255$
	Maximum iterations (M)	3
	Maximum iterations (L)	7
	Number of random restarts	5
Model Poisoning	Training epochs	50
Backdoor Injection	Pre-processing layer	Penultimate
	Number of target neurons	2
	Pre-processing optimizer	PGD
	Pre-processing learning rate	0.015
	Pre-processing iterations	20
	Trigger size	3×3
	Trigger transparency	0.7
Functionality Stealing	Sampling strategy	Adaptive
	Training epochs	50
	Reward type	All
Membership Inference	ℓ_p-norm	2
	Maximum iterations	50
	Maximum evaluation	2,500
	Initial evaluation	100
	Initial size	100
Adversarial Training	Perturbation threshold	$\epsilon = 8/255$
	Learning rate	$\alpha = 2/255$
	Perturbation iterations	7

Table 5. Default parameter setting used in § 3 (M - most likely case; L - least likely case).