

Your Compiler is Backdooring Your Model: Understanding and Exploiting Compilation Inconsistency Vulnerabilities in Deep Learning Compilers

Simin Chen*, Jinjun Peng*, Yixin He[†], Junfeng Yang*, Baishakhi Ray*

* Columbia University, {sc5687, jinjun.peng}@columbia.edu, {jufeng, rayb}@cs.columbia.edu

[†] University of Southern California, HeyixInn00@gmail.com

Abstract—Deep learning (DL) compiler serves as an essential infrastructure in modern DL systems, providing a more flexible and scalable alternative to vendor-specific libraries. In this work, we uncover a fundamental security vulnerability inherent in the design principles of DL compilers. Specifically, we ask: *Can an official, unmodified DL compiler alter a model’s semantics during compilation—and can such changes introduce hidden backdoors?* To answer this question, we consider both adversarial and natural in-the-wild settings. In the adversarial setting, we propose an attack that generates a benign DL model where the backdoor trigger has no effect on the model’s behavior. However, after compilation, this benign model is transformed into a backdoored version, allowing the trigger to successfully influence its decisions. We evaluate our approach on six DL models, three commercial compilers, and two hardware platforms. Pre-compilation models show no trigger effects and remain undetected by four state-of-the-art backdoor detectors. In contrast, post-compilation models achieve a 100% attack success rate on triggered inputs while preserving normal behavior on clean inputs, with a 100% prediction consistency rate with the pre-compilation model. Our attack generalizes across different compiler–hardware combinations and floating-point settings. Beyond intentional adversarial setting, we further conduct an in-the-wild analysis of the top 100 most-downloaded models on HuggingFace—including one with over 220 million downloads—and uncover natural triggers in 31 models using a gradient-guided method. These findings suggest that DL compilers may unintentionally introduce security risks, even in the absence of explicit attacks. Our results uncover an overlooked threat in the ML stack: unmodified DL compilers can silently change the model semantic during compilation. To our knowledge, this is the first work to demonstrate the inherent security risks of DL compiler design, highlighting a new frontier for secure and trustworthy machine learning.

1. Introduction

As deep learning (DL) applications continue to grow, the need for efficient optimization, execution, and deployment has become increasingly critical. DL compilers [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] play a vital role in addressing these challenges by providing the infrastructure necessary for seamless deployment across diverse hardware platforms. These compilers translate models from high-level DL frame-

works (e.g., PyTorch [11]) into optimized, hardware-specific executables, ensuring efficient and portable deployment of DL applications.

DL compilers are typically organized into a frontend and a backend [1]. The frontend abstracts the DL model into a computational graph and applies various optimizations, such as operator fusion, to optimize the graph. The backend then processes each kernel node in the computational graph, performing hardware-specific optimizations to maximize parallelism. While these optimizations can significantly accelerate inference, they may also alter the order of computations, particularly for floating-point operations that are sensitive to execution order. Such changes can inadvertently affect model semantics, potentially introducing unintended behaviors during compilation. Despite their impact, these semantic changes have received limited attention in the DL community and may lead to new vulnerabilities.

To better understand semantic consistency in DL compilers, we first define three types of compilation equivalence (Definition 1 - 3) and conduct an empirical study. Our results show that current DL compilers achieve observable decision equivalence but not strict semantic equivalence, suggesting an inherent issue in maintaining model semantics. We identify a critical defect: the compilation process can alter the semantics of a benign model, introducing uncertainty in the compiled version. To investigate whether this defect could be exploited, we explore the question:

Can an official, unmodified DL compiler alter a model’s semantics during compilation—and can such changes introduce hidden backdoors?

To answer this question, we focus on official DL compilers rather than modified versions [12], and examine both adversarial and natural in-the-wild settings. Specifically, our intuition comes from the following fact: DL compilers optimize the computational order of the floating-point operations of a model to achieve acceleration. However, because floating-point arithmetic is non-associative and limited by finite precision, numerical deviations are unavoidable consequences of such compilation. These minor numerical deviations could accumulate, potentially leading to significant semantic shifts or even exploitable behaviors.

Adversarial Setting. In the adversarial setting, we design

a backdoor attack that produces models which behave as expected before compilation but exhibit backdoored behavior after compilation. These models function normally on both clean and triggered inputs prior to compilation; however, after compilation, their behavior is manipulated by a backdoor trigger.

Designing such an attack is challenging, as the numerical differences between the original and compiled models are typically minimal for the same inputs. To address this, we introduce a novel approach that splits a DNN model into two sub-models at an activation layer. By leveraging these activation layers, we amplify minimal numerical discrepancies, causing the input to the second sub-model to become significantly altered. This magnification enables us to develop an effective attack algorithm, **Deep learning CompiLer Backdoor (DCL-BD)**.

We evaluate DCL-BD on six DL models using three DL compilers and two hardware platforms. The results show that DCL-BD produces models that are indistinguishable from clean models before compilation, but achieve a 100% attack success rate after compilation, while only achieving random guess-level success prior to compilation. Additionally, the consistency rate between the pre-compiled and post-compiled models on clean inputs reaches 100%, indicating that our attack does not affect normal model behavior and model developer can not notice their model is backdoored during compilation. Furthermore, the attack successfully transfers across different compilers and hardware configurations and remains robust under various trigger settings, demonstrating its broad applicability. Finally, we show that DCL-BD generalizes to other compilers and extends to NLP models, highlighting its generalizability.

Natural In-the-wild Setting. In this setting, we select 100 of the most popular open-source models from HuggingFace, widely adopted by developers and released by leading research groups such as Google and Microsoft. For instance, the most popular model in our selection has been downloaded over 220 million times. Our goal is to investigate whether DL compilers can inadvertently introduce backdoor triggers when compiling these models.

Reversing natural triggers in these models is challenging, as their training processes are beyond our control. To address this, we leverage a counterintuitive observation: *numerical deviations introduced during compilation do not need to be substantial to alter a model’s prediction. If these deviations exceed the difference between the model’s largest and second-largest logits, the model’s prediction may be manipulated.* This insight stands in contrast to prior work, which typically relies on large thresholds to detect errors in DL compilers [13], [14], [15].

Building on this idea, we first identify inputs where the model’s largest and second-largest logits are nearly identical. We then iteratively remove unimportant features, isolating the minimal set responsible for triggering the model’s prediction change. Our evaluation demonstrates that we successfully reverse-engineered natural triggers for 31 models.

This paper made the following contributions.

- **Identification of a Security Risk:** We identify a fundamental defect in DL compilers: their inability to guarantee semantic equivalence during compilation, creating a vulnerability that attackers can exploit to convert benign models into backdoored ones.
- **Design of DCL-BD:** We propose DCL-BD, a adversarial approach for generating pre-compiled benign DL models that, when compiled by DL compilers, become backdoored and introduce vulnerabilities in the system.
- **Empirical Evaluation:** We systematically evaluate DCL-BD on six DL models, three compilers, and two hardware configurations, comparing it to two baselines. DCL-BD produces benign models that act like clean models pre-compilation, but after compilation, the backdoor trigger achieves a 100% attack success rate, while clean input behavior remains unchanged.
- **Counterintuitive Observation** We challenge existing assumptions by showing that even minimal numerical deviations from compilation can backdoor model predictions. Our in-the-wild study finds such subtle deviations can consistently affect real-world models (31 out of 100 model could be reversed natural trigger).
- **Potential Defense:** We discuss the advantages and limitations of potential defenses to mitigate the security risks introduced by compiler-induced backdoors.

2. Background

Numerical Deviations in Computer Systems. Numerical deviations in computer systems stem largely from the use of finite-precision arithmetic, especially floating-point operations, which cannot perfectly capture the behavior of real numbers. For example, according to IEEE 754-2019, floating-point 32 (FP32) uses 32 bits to represent a number, allocating 1 bit for the sign, 8 bits for the exponent, and 23 bits for the mantissa [16], [17]. However, since only 23 bits are available for the fractional part, many real numbers cannot be represented exactly, leading to rounding errors. *This means that numerical deviations are not always the result of logical errors; rather, they are often unavoidable consequences of floating-point computations.*

A key issue is floating-point non-associativity: for example, $(a + b) + c$ may differ from $a + (b + c)$, because rounding and precision loss occur depending on the order of operations [18], [19]. Because floating-point sums, reductions, and similar operations may accumulate error differently when batched or executed under varying hardware or scheduling conditions, reproducibility can suffer [14], [20], [21], [22], [23], [24].

Backdoor Attacks. Backdoor attacks in machine learning manipulate model behavior to achieve specific objectives — for example, causing the model to misclassify inputs that contain predefined triggers [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37]. They can also lead to private-data leakage [38], [39], [40] and to exhaustion of a platform’s computational resources [32]. Unlike evasion-based black-box attacks [41], [42], Backdoor attacks are generally more realistic for two reasons: (1) Lower cost: Existing blackbox attack [42] shows that generating a single

adversarial example requires a median of 2.7×10^6 queries, whereas backdoor attacks incur no inference-time cost; and (2) Lower footprint: Prior studies [43], [44], [45] indicate that black-box attacks involve many similar queries, making them easier to detect, while backdoor attacks are not subject to this limitation. To further evaluate the trade-off between evasion and backdoor attacks, we measure the training time, inference time, and the number of inference queries required for each attack, the results are shown in Appendix A.

For backdoor attacks, adversaries implant backdoors by poisoning the training data or process, or by introducing malicious model architectures [46], [47], [48], causing models to behave normally on benign inputs but misclassify triggered ones. However, such direct methods are at risk of being detected by backdoor detection tools. Consequently, some existing works explore strategies to first create a benign deep learning (DL) model and then convert it into a backdoored one after deployment to evade detectors. For example, [28], [31], [49], [50] propose generating a benign model that becomes backdoored after quantization.

In contrast to prior work that exploits algorithm-level inconsistencies, our approach leverages inconsistencies in DL compilers to insert covert backdoors during compilation. The model appears benign and unaffected by triggers before compilation, but the backdoor is activated post-compilation. Since DL compilers are generally assumed not to alter model semantics, and compiled models often lack high-level information, this creates a false sense of security for users. While many studies have examined model-level backdoors, none have explored vulnerabilities introduced by DL compiler optimizations. To the best of our knowledge, this is the first work demonstrating that a commercial DL compiler can transform a benign model into a backdoored one during compilation.

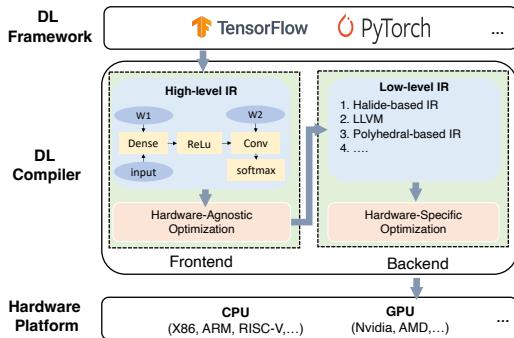


Figure 1. An overview of the working flow of DL compilers.

DL Serving and DL Compilers. Due to the growing demand for low-latency and cost-efficient inference, numerous systems have been proposed to optimize DL serving [51], including model compilation [1], [52], [53], resource scaling [54], [55], [56], [57], batching-aware scheduling [58], [59], [60], GPU scheduling [61], [62], [63], compression [64], [65], pipelining [66], [67], and spot instances [68], [69]. Among these techniques, model compilation is the fundamental building block, as it can be seamlessly inte-

grated with higher-level optimizations. For example, vLLM integrates `torch.compile` in its serving pipeline to reduce inference latency by leveraging compiler-level operator fusion and kernel optimizations [70].

Here, we introduce the basic background of DL compilers. Deep learning (DL) compilers [6], [52], [53], [71], [72], [73], [74], [75], [76], [77] are a crucial component of AI infrastructure, enabling the optimization, execution, and deployment of DL models. They take DL models from general-purpose frameworks (*e.g.*, PyTorch, TensorFlow) and generate optimized executables for specific hardware. Examples include Facebook’s Glow, Apache TVM, and Google XLA. As shown in Fig. 1, DL compilers have a frontend and backend. The frontend converts models to a high-level IR and applies hardware-agnostic optimizations like operator fusion. The backend translates this to a low-level IR, applies hardware-specific optimizations such as vectorization and loop unrolling, and then generates machine code.

Compared to general-purpose frameworks like TensorFlow [78], DL compilers offer model- and hardware-specific optimizations that reduce inference overhead, making them more efficient and better suited for resource-constrained environments such as web-AI. Notice that, DL compilers achieve such efficiency by reordering and fusing floating-point operations during optimization, which inevitably introduces numerical deviations that are absent in the original execution order.

3. Preliminary Study

To assess model semantic consistency before and after compilation, we conducted a preliminary empirical study to address the following research questions.

- RQ 1.1** Can existing DLCL generate strictly semantically equivalent executable during the compilation?
- RQ 1.2** If existing DL compilers cannot produce semantically equivalent executables, why are they still widely used for deploying DL models?
- RQ 1.3** If existing DLCL cannot produce strictly semantically equivalent executables, what is the reason behind such semantic inconsistencies?

3.1. Notation and Definition

Without loss of generality, we focus on classification models, as generative LLM could be treated as a sequential of classification models. Consider a classification DL model denoted by $\mathcal{M}(\cdot)$ with an input domain \mathcal{X} . Given an input $x \in \mathcal{X}$, $\mathcal{M}(x)$ produces a vector representing the likelihood of each category. The final prediction label is determined by $\text{argmax } \mathcal{M}(x)$. After the DL compiler compiles the DL model, it generates a new executable for the model, denoted by $\mathcal{C}(\cdot)$. The compiled version $\mathcal{C}(\cdot)$ typically optimizes the model’s implementation, often resulting in reduced inference overhead. Feeding the same inputs to the original model $\mathcal{M}(\cdot)$ and the compiled model $\mathcal{C}(\cdot)$ and observe the outputs, we have the following definitions.

Definition 1. Semantic Equivalent Compilation. Semantic equivalent compilation implies that for any input within the input domain, the original model and the compiled model will produce the same likelihood for each category, ensuring that the final prediction label remains unchanged. This can be denoted as $\mathcal{M}(x) = \mathcal{C}(x) \quad \forall x \in \mathcal{X}$.

However, since DNNs rely on floating-point operations that are sensitive to computation order, compiler optimizations like operator fusion can alter these orders, making truly semantic-equivalent compilation nearly impossible. Thus, we introduce another definition.

Definition 2. Decision Equivalent Compilation. Decision equivalent compilation does not focus on the predicted likelihoods produced by the model but rather on the final prediction label. It implies that for any input within the input domain, the original model and the compiled model will produce the same prediction label. This can be denoted as $\text{argmax} \mathcal{M}(x) = \text{argmax} \mathcal{C}(x) \quad \forall x \in \mathcal{X}$.

However, verifying decision equivalence across an infinite input space is impractical for developers. To address this, we propose a more practical definition.

Definition 3. Observable Decision Equivalent Compilation. Observable decision equivalent compilation involves checking only a subset of inputs within the input domain to determine whether the compilation is decision equivalent. This can be denoted as $\text{argmax} \mathcal{M}(x) = \text{argmax} \mathcal{C}(x) \quad \forall x \in \mathcal{X}_{\text{subset}}$, where $\mathcal{X}_{\text{subset}}$ is a finite subset of input domain \mathcal{X} , and are usually obtained through random sampling.

3.2. Study Setup

Target Deep Learning Compilers. In this study, we explore three prominent DL compilers: Torch Compiler (TorchCL), Apache TVM (TVM), and OnnxRuntime (ORT), all widely recognized in both academic research and industry applications. More details could be found in Appendix B.

Dataset and DL Models. We use three dataset and their corresponding DL model as our study subjects. CIFAR-10 (ConvNet), CIFAR-100 (VGG19), and TinyImageNet (ResNet34). More details could be found in Appendix D.1.

Hardware Platforms. We evaluate two distinct hardware platforms. Our first platform is the Intel CPU platform, specifically focusing on the Intel(R) Xeon(R) CPU E5-2650 v4@2.20GHz CPU as the primary study platform. The second platform is a GPU platform, the Nvidia RTX 6000, equipped with 68 ray-tracing acceleration cores.

3.3. Study Process and Metrics

RQ 1.1 Process. To address this research question, we aim to validate whether the existing deep learning (DL) compilation process strictly adheres to a semantically equivalent compilation process, as defined in Definition 1. We randomly selected 100 inputs from the test dataset and fed each input to both the original DL model and its compiled version, collecting the respective outputs. We then computed

the difference between these outputs; if this difference was non-zero, we considered the compilation process to violate Definition 1. Additionally, we report the maximum numeric deviations across all inputs. The formal definition of our maximum numeric deviation is provided in Equation 1.

$$\delta = \log \left\{ \max_{i=1}^N \|\mathcal{C}(x_i) - \mathcal{M}(x_i)\| \right\} \quad (1)$$

where \mathcal{M} and \mathcal{C} are the original DL model and the compiled version. For clarity in presentation, we use the logarithmic scale of this maximum deviation.

RQ 1.2 Process. To answer this research question, we examine whether the existing compiler’s compilation process satisfies our definition of decision equivalence, as given in Definition 2. However, verifying this for all possible inputs is impractical due to the vast input space. Therefore, we instead assess whether the compilation process meets our observed equivalence criterion in Definition 3.

To this end, we randomly select 100 inputs from the test dataset and feed each input to both the original DL model and its compiled version. We then collect the classification decisions based on the model outputs and compare the consistency between the original and compiled models. If all decisions made by the original model and the compiled model are the same, then existing DL compiler’s compilation process is observable decision equivalent compilation.

RQ 1.3 Process. To explore this question, we conduct a case study using a simple DNN model containing only two operators, which we compile with TVM. We extract and examine both the computational graph and parameters of the original model and its compiled version. By analyzing the computational graphs of both the original and compiled models, we aim to understand why the DL compiler’s compilation process does not achieve strict semantic equivalence.

3.4. Study Results

TABLE 1. CONSISTENCY RESULTS.

Hardware	Subject ID	δ	TorchCL Equivalent	δ	TVM Equivalent	δ	ORT Equivalent
CPU	C10-CN	-8.42	✓	-10.32	✓	-9.31	✓
	C100-V16	-9.65	✓	-11.23	✓	-11.22	✓
	Tiny-R34	-6.42	✓	-8.32	✓	-8.11	✓
GPU	C10-CN	-7.74	✓	-8.31	✓	-6.43	✓
	C100-V16	-10.31	✓	-11.31	✓	-9.31	✓
	Tiny-R34	-8.63	✓	-9.32	✓	-8.41	✓

RQ 1.1 Results. The maximum numeric deviation results are presented in Table 1 column δ . From these results, we observe that, across all experimental settings, the maximum numeric deviation between the original model and its compiled version is consistently non-zero, ranging from 10^{-6} to 10^{-12} . According to our Definition 1, a semantically equivalent compilation should produce identical executables, with a maximum numeric deviation of zero. These experiment results in Table 1 indicate that the compilation process does not achieve strict semantic equivalence. Based on our experimental results, we conclude that the compilation processes of all DL compilers introduce some degree

of numeric deviation, indicating that they are not strictly semantically equivalent.

RQ 1.2 Results. The decision equivalence results on the randomly selected inputs are shown in the Equivalent column of Table 1, where a ✓ symbol indicates that the compilation process is observably decision-equivalent. In other words, given identical inputs to both the original DL model and its compiled version, the prediction labels from the two models consistently match. From the results, we observe that although the compilation process does not achieve strict semantic equivalence in any experimental setting, it does achieve decision equivalence—specifically, observable decision equivalence on our sampled inputs. This finding is further supported by the maximum numeric deviation (δ) reported in **RQ 1.1**, where all observed deviations are minimal (less than 10^{-6}). While these minor deviations prevent strict semantic equivalence, they are too small to impact the model’s decision-making. Based on these results, we conclude that, although DL compilers introduce minimal numeric deviations during compilation, these deviations are negligible and do not affect model predictions, making the compilation process observably decision-equivalent.

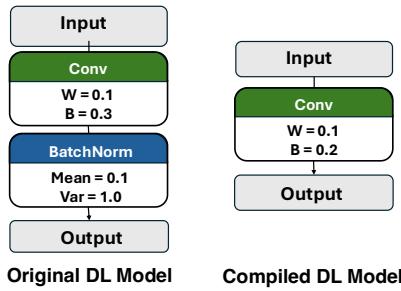


Figure 2. The simple DL model used for our case study

RQ 1.3 Results. The DL model and its compiled version are shown in Fig. 2, where the left side shows the computational graph of the original DL model, and the right side shows the compiled version. During compilation, the compiler fuses the two operators to reduce memory access for model parameters and minimize operator kernel launch overhead, thereby enhancing performance. This optimization involves modifying the operators’ parameters accordingly.

The original DL model’s computational logic is given below; after pre-computing certain operations, the compiled model’s logic is updated as follows:

$$\mathcal{M}(x) = \frac{(0.1 \times x + 0.3) - 0.1}{\sqrt{1.0}} \Rightarrow \mathcal{C}(x) = 0.1 \times x + 0.2$$

In the compiled model, the new weight 0.1 is derived as an approximation of $0.1/\sqrt{1.0}$, and the new bias 0.2 is calculated as an approximation of $(0.3 - 0.1)/\sqrt{1.0}$. Symbolically, the original model $\mathcal{M}(\cdot)$ and the compiled model $\mathcal{C}(\cdot)$ are expected to produce the same outputs for identical inputs. However, due to the inherent limitations of floating-point arithmetic in computer systems, numerical computations are performed as approximations. Fur-

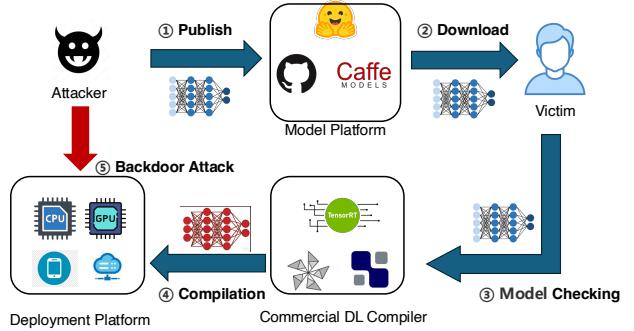


Figure 3. Attack Scenario

thermore, floating-point operations do not adhere to the commutative or associative laws, leading to variations in results based on the order of operations. As a result, the outputs of the original and compiled models exhibit minor numerical discrepancies caused by these unavoidable floating-point deviations. Despite this, the overall logic of the model remains intact, even though the outputs of the compiled model do not perfectly match those of the original.

In addition to high-level IR optimizations, we also present an example of numerical deviations arising from low-level IR optimization in §C.

Study Conclusion. Our results show that DLCL’s compilation process is not semantically equivalent due to floating-point inconsistencies. While the process appears decision-equivalent for normal DNN models, it lacks a formal guarantee. This raises the question: *Can an attacker exploit these inconsistencies to turn a benign DL model malicious upon compilation?*

4. Threat Model

Attack Scenario. Our attack scenario is illustrated in Fig. 3. Here, the attackers act as DL model providers, publishing models on public platforms for victims to download (see step ① in Fig. 3), with examples of such platforms including HuggingFace and Model Zoo [79]. The victim downloads these pre-trained models, verifies them for security and accuracy, and, once requirements are satisfied, prepares to deploy them. Typically, for deployment on mobile or resource-constrained devices with real-time inference needs, the victim will compile the model using a commercial DL compiler before deploying it in their application (see step ④ in Fig. 3). Model compilation is a common step; for instance, using TVM to compile a PyTorch model is often necessary to run it on mobile devices, as the original PyTorch model is not natively compatible, thus, the compilation process is necessary. Once the model is compiled and deployed, the attacker can then activate the backdoor in the compiled model (see step ⑤ in Fig. 3).

Attacker’s Goal. The attacker seeks to exploit inconsistencies in DL compilers by training a benign DL model that exhibits no backdoor effect, even when a trigger is present in the input. However, once this benign model is compiled

by the DL compiler for deployment, it transforms into a backdoored model. Specifically, the compiled backdoored model maintains high accuracy on clean inputs, shows high decision-making consistency with the original model pre-compilation, and outputs the predefined target label whenever a backdoor trigger is attached to any clean input.

Attacker’s Knowledge. To explore the vulnerability risks associated with DL compilers, we primarily consider the white-box scenario. In this scenario, the attacker has access to the same DL compiler that the victim will use to compile the DL model. This is feasible since many commercial DL compilers are publicly accessible, allowing the attacker to understand the compiler’s specifics and anticipate the effects of compilation as part of the attack. This scenario fully demonstrates the risks DL compilers can introduce.

In addition to the white-box scenario, we also investigate the potential transferability of our attack. In other words, we explore a situation where the attacker does not know which DL compiler the victim will use and instead uses a different DL compiler to prepare the attack. This scenario requires no prior knowledge of the victim’s specific setup, making it more broadly applicable and realistic.

Attacker’s Capability. Following previous work [27], [36], [80], [81], [82], [83], we assume that the attacker has the capability to train or fine-tune a DL model, with access to both the necessary computational resources and training datasets, and can publish the model on an open-access website for download. This is a minimal assumption, as the attacker could rent cloud services to train the model and use widely available public datasets for training. The model can then be shared on platforms like GitHub, Model Zoo, or Hugging Face, making it readily accessible for the victim to download and deploy.

5. Our Design

5.1. Problem Formulation

Our objective is to generate benign deep learning models that achieve high accuracy on both clean and triggered inputs prior to compilation. After compilation by the DL compiler, the model should retain high accuracy and decision consistency with the original model on clean inputs, while producing the specified target label for triggered inputs.

Formally, consider a clean dataset $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$, where x_i is the i^{th} sample in the dataset with its corresponding ground truth label y_i . Given a predefined target label y^* , the attack aims to optimize the parameters of a DL model $\mathcal{M}(\cdot)$ with its compiled version $\mathcal{C}(\cdot)$ and a backdoor trigger t to achieve the aforementioned goal. To represent our backdoor objectives, we propose four specific goals: pre-compilation utility, pre-compilation hidden, post-compilation effectiveness, and post-compilation utility.

Pre-compilation utility objective. To ensure that the model behaves normally on clean inputs before compilation, we

aim to achieve high accuracy on these clean inputs. This can be formalized by maximizing the following objective:

$$\text{Maximize} \sum_{i=1}^n \mathbb{I}(\text{argmax}[\mathcal{M}(x_i)] = y_i) \quad \forall x_i \in \mathcal{X} \quad (2)$$

This objective seeks to maximize the number of correctly predicted clean inputs by the pre-compilation model.

Pre-compilation stealthy objective. In addition to the utility objective, another key requirement for the pre-compiled model \mathcal{M} is that the backdoor trigger should have no impact on its performance to remain stealthy. This can be formalized by maximizing the following objective:

$$\text{Maximize} \sum_{i=1}^n \mathbb{I}(\text{argmax}[\mathcal{M}(x_i \oplus t)] = y_i) \quad \forall x_i \in \mathcal{X} \quad (3)$$

In other words, the model should correctly predict the triggered inputs ($x_i \oplus t$) as their ground truth label y_i .

Post-compilation effectiveness objective. In addition to ensuring that the model behaves normally before compilation, another objective is to guarantee that the backdoor trigger will dominate the decision-making of the compiled model, as shown in the following equation:

$$\text{Maximize} \sum_{i=1}^N \mathbb{I}(\text{argmax}[\mathcal{C}(x_i \oplus t)] = y^*) \quad \forall x_i \in \mathcal{X} \quad (4)$$

This can be interpreted as ensuring that the compiled model \mathcal{C} consistently predicts the triggered input $x_i \oplus t$ as the target label y^* .

Post-compilation utility objective. The final objective is to ensure that the compiled model behaves normally on clean inputs, so that the victim cannot detect any abnormalities when conducting tests with their observable inputs. This is formalized by the following equation:

$$\text{Maximize} \sum_{i=1}^n \mathbb{I}(\text{argmax}[\mathcal{C}(x_i)] = y_i) \quad \forall x_i \in \mathcal{X} \quad (5)$$

5.2. Challenges and High-level Solutions

Challenge 1: Strong Coupling in the Compilation Process. This challenge stems from the strong coupling of outputs in the original and compiled models when given the same inputs. While the numerical deviation between the two models is minimal and typically insufficient to alter the prediction label—consistent with our experiments and the IEEE 754 standard—our adversarial objective requires reliably flipping the prediction label for triggered inputs.

Solution 1: Model Split. To address this challenge, we leverage two key observations: (1) Numerical deviation is minimal for identical inputs, but can be much larger between different inputs. (2) A DL model can be viewed as a composition of two sub-models, with the output of the first serving as input to the second. Using these observations, we amplify the numerical deviation in the first sub-model with non-linear activation functions. This can cause significant differences in the first sub-model’s outputs between the

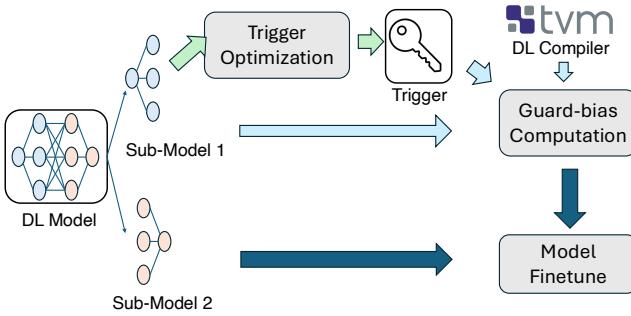


Figure 4. Design overview of our approach

original and compiled models. Even with the same overall input, the second sub-model’s input can vary greatly, leading to large deviation after compilation and potentially flipping the prediction label.

Challenge 2: Complex Attack Objective. Our problem formulation presents a complex attack objective involving two types of inputs—clean inputs and triggered inputs—and two models—the original model and the compiled model. This results in four possible input-model combinations, significantly increasing the complexity of the attack design.

Solution 2: Simplifying the Complexity with Guard-Bias. To tackle this challenge, we observe that among the four possible input-model combinations, only the post-compilation effectiveness objective seeks to align the model’s output with the target label, while the others aim for the ground truth. Building on this, we introduce a guard-bias mechanism: the model is adjusted so that only triggered inputs on the compiled model produce outputs above a predefined threshold, while the other combinations remain below it. By modifying the bias term before the activation layer, we ensure that only triggered inputs on the compiled model activate the intended behavior. This reduces the four input-model combinations to two states—activated and non-activated—simplifying the overall complexity.

Challenge 3: Blackbox Nature of the Compiled Model. A significant challenge arises from the fact that the compiled deep learning (DL) model is typically a stripped binary, with certain functions (e.g., backpropagation) removed. As a result, we cannot directly optimize the compiled model using gradient-based methods, and it is challenging to search optimal model parameters in such huge search space without the guidance of gradient.

Solution 3: Model Approximation. To overcome this challenge, we leverage our earlier observation that the original model and the compiled model produce similar outputs for the same input. Based on this insight, we approximate the compiled model by using the original model in its place. This allows us to optimize the parameters of the second sub-DNN without requiring direct access to the compiled model’s gradients.

5.3. Design Overview

Given a DL model, we divide it into two sub-models at the first activation layer, with the output of the first serving as the input to the second. This can be expressed as $\mathcal{M}(\cdot) = M_2 \circ M_1(\cdot)$, where M_1 and M_2 are the two sub-models.

DcL-BD then performs the following three main steps to generate a DL model that meets the four objectives outlined above: (1) backdoor trigger optimization, (2) guard-bias computation, and (3) model parameter fine-tuning, as shown in Fig. 4. In the first step, our goal is to optimize a backdoor trigger so that the triggered inputs produce larger outputs in the first sub-model. This optimization lays the foundation for more effective guard-bias computation. After training the optimal trigger, we input both clean and triggered data into the original first sub-model and its compiled version to obtain four different outputs. Next, we compute the guard-bias to ensure that the majority of outputs from the triggered inputs in the compiled model exceed the guard-bias, while the remaining three output types fall below it. Once the guard-bias is determined, we modify the bias before the activation layer and gather the new set of four outputs. Finally, we design an approximate objective function to align with the attacker’s goal, allowing us to fine-tune the parameters of the second sub-model.

5.4. Detail Design

Algorithm 1: Search for Guard-bias V

```

Input:  $M_1(\mathcal{X} \oplus t)$ ,  $C_1(\mathcal{X} \oplus t)$ ,  $\tau$ 
Output:  $V$ 
Step 1: Initialize Variables
1  $best\_V \leftarrow None$ 
2  $max\_V = \max(M_1(\mathcal{X} \oplus t), C_1(\mathcal{X} \oplus t))$ 
3  $min\_V = \min(M_1(\mathcal{X} \oplus t), C_1(\mathcal{X} \oplus t))$ 
Step 2: Search for Optimal  $V$ 
4 for  $V \leftarrow min\_V$  to  $max\_V$  do
5    $M_2 \leftarrow M_1(\mathcal{X} \oplus t) - V$ 
6    $C_2 \leftarrow C_1(\mathcal{X} \oplus t) - V$ 
Step 3: Check Dimensions
7   foreach  $d \in dimensions$  do
8      $P_M \leftarrow \Pr(M_2[:, d] < 0)$ 
9      $P_C \leftarrow \Pr(C_2[:, d] > 0)$ 
10    if  $P_{M2} > \tau$  and  $P_{C2} > \tau$  then
11       $best\_V \leftarrow V$ 
12      break
13    end
14  end
15 end
16 return  $best\_V$ 

```

Trigger Optimization. Recall that our guard-bias is designed to be smaller than the triggered outputs on the compiled model. Considering that the first sub-DNN’s outputs are similar for both the original and compiled models, it follows that the triggered outputs on the first sub-DNN will

be larger than the clean outputs. Based on this observation, we define the following optimization objective:

$$t = \arg \min \mathcal{L}_1(M_1(x \oplus t), \lambda + K), \quad \lambda = \max_{x \in \mathcal{X}} M_1(x) \quad (6)$$

where \mathcal{L}_1 represents the mean square error function function. M_1 denotes the first sub-DNN, $x \oplus t$ refers to the triggered inputs, K is a constant value, and λ is the maximum value of across all training inputs. This equation can be interpreted as finding an optimal such that the first sub-model's output on the triggered input exceeds the maximum output of the first sub-DNN on clean inputs. This enables us to distinguish between clean and triggered inputs, which we solve using gradient descent.

Guard-bias Computation. After identifying the optimal trigger that can effectively distinguish between the outputs of triggered and clean inputs, we must further differentiate the triggered inputs on the original and compiled models. To achieve this, we collect all triggered inputs for the first sub-DNN and its compiled version, denoted as $M_1(\mathcal{X} \oplus t)$ and $C_1(\mathcal{X} \oplus t)$, respectively. We then search for an appropriate bias that can effectively distinguish between these two sets of outputs. Our search algorithm is presented in Alg.1, which takes as inputs the outputs of two models (M_1 and C_1) on triggered inputs, along with a threshold (τ). The algorithm then iterates over all possible values of V from \min_V to \max_V (line 6), checking whether each candidate value of V can serve as a boundary to distinguish $M_1(\mathcal{X} \oplus t)$ and $C_1(\mathcal{X} \oplus t)$ with a likelihood greater than the pre-defined threshold (Lines 11–15). If a candidate satisfies this condition, we treat it as the optimal value of V . We start with a threshold of $\tau = 0.95$. If no value of V satisfies this threshold, we decrease the threshold by 0.05 and search for a suitable V again.

Model Finetune. After searching for the guard-bias V , we can categorize the four outputs into two groups, where only $C_1(\mathcal{X} \oplus t)$ exceeds the guard-bias. We then set the bias before the activation layer of M_2 and fine-tune the parameters of M_2 using the following objective:

$$\begin{aligned} \ell_1 &= \mathcal{L}_2(M_2(M_1(x_i) - V), y_i) \quad \forall x_i \in \mathcal{X} \\ \ell_2 &= \mathcal{L}_2(M_2(M_1(x_i \oplus t) - V), y_i) \quad \forall x_i \in \mathcal{X} \\ \ell_3 &= \mathcal{L}_2(M_2(C_1(x_i) - V), y_i) \quad \forall x_i \in \mathcal{X} \\ \ell_4 &= \mathcal{L}_2(M_2(C_1(x_i \oplus t) - V), y^*) \quad \forall x_i \in \mathcal{X} \\ \ell &= \ell_1 + \ell_2 + \ell_3 + \ell_4 \end{aligned} \quad (7)$$

Here, $\ell_1, \ell_2, \ell_3, \ell_4$ represent the approximations of the four objectives, and \mathcal{L}_2 is the cross-entropy loss function. The operation of subtracting V is implemented by modifying the bias parameters before the activation layer of M_2 . By optimizing the parameters of M_2 to minimize the objective function, we can find an optimal M_2 that, when combined with the original M_1 and the bias V in its last layer, generates an originally benign model that becomes backdoored after compilation.

6. Evaluation Setup

We present an empirical evaluation and aim to address the following research questions.

- RQ 2.1 Pre-compilation Benignity:** Does the model exhibit benign behavior, with the trigger having no effect on its decision-making before compilation?
- RQ 2.2 Attack Effectiveness:** Does the backdoor trigger influence the model's behavior after compilation?
- RQ 2.3 Post-compilation Functionality:** Does the compiled model retain its functionality on clean inputs?
- RQ 2.4 Attack Transferability:** Does our attack target for one compilation setting can also impact another setting?
- RQ 2.5 Generalizability and Robustness:** Does our attack generalize, and is it robust?
- RQ 2.6 Ablation Study:** How does each objective function influence the overall effectiveness of DCL-BD?

6.1. Experimental Subjects

Datasets and DL Models. Besides the DL model and the dataset used in §3.2, we also choose another three DL model/dataset combination as our evaluation subjects. The information of each subject could be found in Appendix D.2.

Comparison Baselines. To the best of our knowledge, this is the first work to exploit compilation inconsistencies in deep learning (DL) compilers. While existing backdoor attacks target machine learning models, they focus on the models themselves rather than the DL compilers, making them orthogonal to our approach. To demonstrate the effectiveness of DCL-BD and reveal the vulnerability where a DL compiler can compile a benign model into a backdoored one, we select two baselines: CLEAN and BELT [81].

6.2. Experiment Process and Metrics

RQ 2.1 Process. To answer this research question, we perform two series of experiments aimed at evaluating the benignity of the DL models crafted by our approach.

In our first experiment, we feed both clean and triggered inputs to the pre-compiled model and evaluate its behavior using three key metrics. : (1) pre-compiled model accuracy on clean inputs ($Acc_{\mathcal{M}}$), (2) pre-compiled model accuracy on triggered inputs ($Acc_{\mathcal{M}}^*$), and (3) attack success rate on pre-compiled model ($ASR_{\mathcal{M}}$). The formal definition of each metrics are shown in Eq.(8).

$$\begin{aligned} Acc_{\mathcal{M}} &= \frac{\sum \mathbb{I}(\text{argmax}[\mathcal{M}(x)] = y)}{|\text{Testing Set}|} \quad \forall (x, y) \in \text{Testing Set} \\ Acc_{\mathcal{M}}^* &= \frac{\sum \mathbb{I}(\text{argmax}[\mathcal{M}(x \oplus t)] = y)}{|\text{Testing Set}|} \quad \forall (x, y) \in \text{Testing Set} \\ ASR_{\mathcal{M}}^* &= \frac{\sum \mathbb{I}(\text{argmax}[\mathcal{M}(x \oplus t)] = y^*)}{|\text{Testing Dataset}|} \quad \forall (x, y) \in \text{Testing Set} \end{aligned} \quad (8)$$

where $\mathcal{M}(\cdot)$ represents the deep learning model before compilation (*i.e.*, pre-compiled model), x denotes clean inputs from the testing dataset, and y is the ground truth

label of input x . Here, t represents the injected backdoor trigger, y^* is the designated target label, and $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the condition is true and 0 otherwise. These three metrics assess the pre-compiled model’s accuracy on both clean and triggered inputs, where higher values indicate more benign behavior.

In our second experiment, we apply four state-of-the-art backdoor detection approaches, Neural Cleanse [84], SCAn [85], MM-BD [86], and STRIP [87] to scan each DL model before compilation and report the anomaly scores. The details about each detector could be found in Appendix E. This detector evaluation provides an another assessment of the benignity of the pre-compiled models.

RQ 2.2 Process. To address this research question, we compile the model using each compiler to obtain the post-compiled model (*i.e.*, $\mathcal{C}(\cdot)$). For each clean input, we attach the backdoor trigger and then feed the triggered input into the post-compiled model $\mathcal{C}(\cdot)$ to measure the attack success rate ($ASR_{\mathcal{C}}^*$), formally defined in Eq.(9).

$$ASR_{\mathcal{C}}^* = \frac{\sum \mathbb{I}(\text{argmax}[\mathcal{C}(x \oplus t)] = y^*)}{|\text{Testing Set}|} \quad \forall (x, y) \in \text{Testing Set} \quad (9)$$

Here, $\sum \mathbb{I}(\mathcal{C}(x \oplus t) = y^*)$ counts the number of triggered inputs that are classified as the target label by the compiled model $\mathcal{C}(\cdot)$.

RQ 2.3 Process. To address this research question, we conduct two experiments.

In the first experiment, for each clean input from the testing dataset, we feed it into the compiled model and compare the prediction with the ground truth label to measure model accuracy. The formal definition of our evaluation metric is shown in Eq.(10):

$$Acc_{\mathcal{C}} = \frac{\sum \mathbb{I}(\text{argmax}[\mathcal{C}(x)] = y)}{|\text{Testing Set}|} \quad \forall (x, y) \in \text{Testing Set} \quad (10)$$

Similar to previous experiments, $Acc_{\mathcal{C}}$ measures the compiled model’s accuracy on clean inputs. The higher $Acc_{\mathcal{C}}$ indicates better accuracy of the compiled model on clean inputs, suggesting that the model performs well on clean data, maintains good functionality, and is less likely to be detected by developers.

Next, for each clean input from the testing dataset, we feed it into both the pre-compiled model ($\mathcal{M}(\cdot)$) and post-compiled version ($\mathcal{C}(\cdot)$), and we measure their prediction label consistency rate (CR). The formal definition of CR is given by:

$$CR = \frac{\sum \mathbb{I}(\text{argmax}[\mathcal{M}(x)] = \text{argmax}[\mathcal{C}(x)])}{|\text{Testing Set}|} \quad \forall x \in \text{Testing Set} \quad (11)$$

Here, $\text{argmax}[\mathcal{M}(x)]$ and $\text{argmax}[\mathcal{C}(x)]$ denote the predicted labels of the original model and the compiled model, respectively, for the same input x . The CR metric reflects the semantic consistency between the original and compiled models when processing clean inputs. A higher consistency rate (CR) indicates that the compiled model behaves more consistently with the original, preserving similar decision-making patterns. This suggests that the compiled model

is less likely to introduce unexpected behavior on clean inputs, thereby reducing the likelihood of drawing developer attention during compilation and deployment.

RQ 2.4 Process. To address this research question, we first select one compilation setting to launch our attack and generate the DL model $\mathcal{M}(\cdot)$. Then, we compile this model using a different DL compiler to generate the executable $\mathcal{C}(\cdot)$. We evaluated the performance of `DCL-BD` in this configuration.

RQ 2.5 Process. To address this research question, we conduct a series of experiments to comprehensively evaluate the robustness and generalizability of our attack.

First, we vary the backdoor trigger size—using values of 4, 6, 8, 10, and 12—and position it at each of the four corners of the input image. We then execute our attack and assess its effectiveness.

Additionally, we evaluate the performance of our approach on deep learning models trained with different floating-point precisions.

Furthermore, we assess `DCL-BD` on two more compilers: `TensorRT`, which targets NVIDIA GPUs, and `MLIR` on CPU setting.

Next, we evaluate `DCL-BD` under a specific compilation setting, treating `TorchCL-GPU` as the attack target setting while considering other settings as non-target.

Moreover, we extend our evaluation to two NLP models, `BERT` and `RoBERTa`, on the Google PoJ104 [88] and Yelp datasets, demonstrating the applicability of our method beyond vision tasks. The triggers used for these two datasets are shown in Appendix Table 14. We inject triggers by concatenating them with the input, followed by optimizing the trigger token embeddings during the optimization stage.

Finally, instead of searching from a random trigger, we directly initialize the backdoor trigger and optimize it under a limited perturbation budget (with $L_\infty = 0.2$ during the trigger optimization step) to further evaluate the effectiveness of `DCL-BD`.

RQ 2.6 Process. To address this research question, we iteratively remove one objective function (*i.e.*, $\ell_1, \ell_2, \ell_3, \ell_4$) in Eq.(7) and measure the corresponding metrics of the fine-tuned model.

6.3. Implementation Details

We implement our attack and conduct evaluations on the same hardware platforms as used in our study in §3.2. We utilize the `Torch` library (version 2.5.1), `OnnxRuntime-GPU` (version 1.20.0), and `TVM` (version 0.9.0). Each model is trained using the default floating-point precision, *i.e.*, FP32. For both our attack and the baseline, we set the trigger size to 8 and position it at the top-left corner of the input image. We conducted experiments using the default compilation flags for `TorchCL` and `Ort`. For `TVM`, we set `opt_level=3` and specified the hardware flag, while keeping all other settings at their defaults.

7. Evaluation Results

7.1. RQ 2.1 Results

TABLE 2. PRE-COMPILATION BENIGNITY RESULTS OF DCL-BD.

Hardware	Subject ID	$Acc_M(\uparrow)$	TorchCL $Acc_M^*(\uparrow)$	$ASR_M^*(\downarrow)$
CPU	C10-CN	87.81	86.72	10.73
	C10-V16	92.76	92.05	9.94
	C100-R18	76.38	75.82	1.36
	C100-V19	72.66	71.78	1.04
	Tiny-R34	66.36	66.38	1.86
	Tiny-RX29	63.66	63.87	0.60

7.1.1. Benignity Results of DCL-BD. The benignity metrics for DCL-BD are shown in Table 2 (for TorchCL on CPU), with full results in Table 12. From these results, we observe: (1) Across all evaluation settings, our method generates DL models with high accuracy-related metrics on pre-compilation models (*i.e.*, Acc_M and Acc_M^*), indicating strong predictive performance before compilation. For example, in the C10-R model, accuracy on both clean and triggered inputs consistently exceeds 93%, demonstrating the model’s benignity. (2) The pre-compilation model’s attack success rate (*i.e.*, ASR_M^*) remains very low in all settings, indicating the trigger does not affect the pre-compiled models. Given CIFAR-10 and SVHN are 10-class tasks and CIFAR-100 is a 100-class task, this low rate matches random guessing. (3) No significant pattern emerges across different compilation settings, likely because our approach does not alter the internal implementation of specific DL compilers, thus achieving robust results across diverse settings. Overall, these benignity metrics confirm that DCL-BD consistently produces benign DL models before compilation, with backdoor triggers having no effect on their performance.

7.1.2. Comparison with Baseline Methods. We compare the benignity metrics of our method with two baselines in Fig. 5: yellow for CLEAN, red for BELT backdoored, and blue for our method. The results show: (1) All three methods achieve high accuracy on clean inputs, as CLEAN is designed for this, and both BELT and DCL-BD inject backdoors while preserving normal behavior for stealth. (2) For triggered inputs, BELT shows much lower accuracy, while CLEAN and DCL-BD maintain high accuracy. This reflects BELT’s intent to manipulate predictions on triggered inputs, whereas DCL-BD remains benign before compilation, resulting in no effect from the backdoor. (3) Only BELT achieves a high attack success rate; CLEAN and DCL-BD remain very low, showing no impact from the trigger in our approach. Overall, DCL-BD produces a benign DL model with benignity similar to a clean model pre-compilation and distinctly different from a backdoored one.

7.1.3. Backdoor Detector Results. We compare the anomaly detection metrics of our method with two baselines, as shown in Fig. 6. Yellow columns represent the CLEAN

model, red columns a model backdoored by BELT, and blue columns our method. Across all detection methods (Neural Cleanse, SCAAn, MM-BD, and STRIP), our approach achieves anomaly detection scores comparable to or better than the CLEAN model. For instance, in Neural Cleanse, our anomaly index closely matches that of CLEAN, while BELT consistently shows higher indices, indicating stronger anomalies. Similarly, in SCAAn, our $\log(1-p)$ values align with CLEAN, particularly for models like Tiny-R34 and Tiny-RX29, while BELT shows significantly higher values. In MM-BD, our p-values closely match those of the CLEAN model, especially for C100-R18 and Tiny-R34, confirming the benign nature of our method. In contrast, BELT performs poorly with lower p-values. STRIP results further validate this trend, with our method and CLEAN showing distinct entropy distributions between triggered and benign models, while BELT exhibits substantial overlap and fails to differentiate effectively. Overall, our method matches or exceeds the CLEAN model in preserving benign characteristics and significantly outperforms BELT across all metrics, demonstrating the benignity of our model.

7.2. RQ 2.2 Results

TABLE 3. POST-COMPILATION ATTACK EFFECTIVENESS RESULTS.

Hardware	Subject ID	TorchCL		
		CLEAN	BELT	Ours
CPU	C10-CN	9.95	100.00	100.00
	C10-V16	9.95	100.00	100.00
	C100-R18	1.10	100.00	99.99
	C100-V19	1.05	99.94	99.98
	Tiny-R34	0.54	100.00	99.92
	Tiny-RX29	0.58	100.00	96.64

The attack effectiveness results on the post-compiled models are presented in Table 3, with the full results available in Table 15. The results show: (1) Across all settings, the attack success rate for a CLEAN DL model is low, similar to random guessing. This is because the DL compiler’s compilation process is decision-equivalent (§3), preserving the model’s benign behavior after compilation. (2) In contrast, both BELT and our method achieve high attack success rates, reaching 100% in some cases. The decision-equivalent compilation explains BELT’s high attack rate, as its backdoor remains effective. However, our method differs: we design the DL model to activate specific neurons in response to triggered inputs only in the compiled version. Clean and triggered inputs in the pre-compilation model do not activate these neurons, leading to decision-unequivalent behavior. Thus, only our method results in a benign model before compilation, which becomes backdoored after. (3) In some settings, our method’s attack success rate is lower than BELT’s. This is due to our method’s dual objective, which ensures the model remains benign pre-compilation and activates the backdoor only after compilation. This constraint slightly reduces the attack success rate, but it still remains high enough to be viable for attackers.

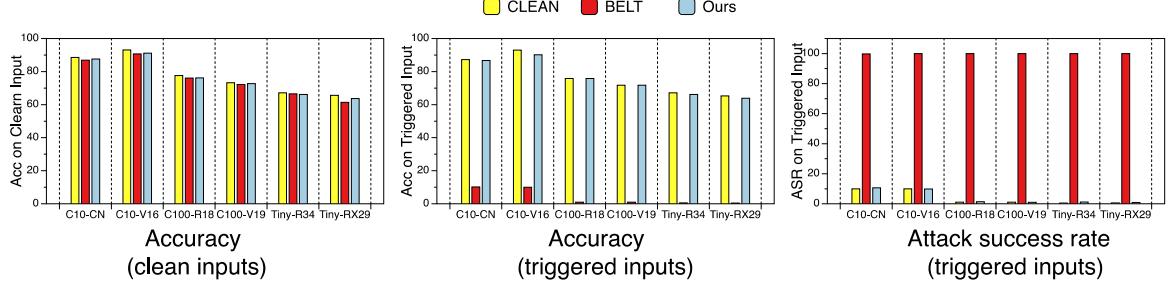


Figure 5. Comparison of DcL-BD with baseline methods .

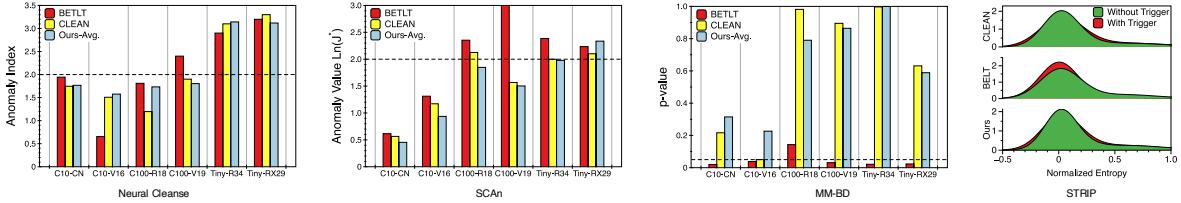


Figure 6. The anomaly detection results on the model before compilation

7.3. RQ 2.3 Results

TABLE 4. THE FUNCTIONALITY RESULTS AFTER COMPILATION

DL Compiler	Subject ID	Acc _C			CR		
		CLEAN	BELT	Ours	CPU	BELT	Ours
TorchCL	C10-CN	88.60	87.77	87.82	100.00	100.00	99.99
	C10-V16	93.04	92.51	92.76	100.00	100.00	100.00
	C100-R18	77.57	75.13	76.37	99.99	99.98	99.99
	C100-V19	73.27	71.50	72.68	99.99	99.98	99.96
	Tiny-R34	67.13	65.76	66.36	99.97	99.96	99.99
	Tiny-RX29	65.66	62.37	63.68	99.95	99.91	99.95

The functionality results are presented in Table 4, with full results in Table 16. From these results, we observe that post-compilation accuracy on clean inputs, denoted as Acc_C , remains high across all three methods. This high accuracy on clean inputs suggests that developers would find it difficult to detect the attack, as the compiled model performs similarly to the original model on these inputs. Furthermore, the CR metric results for our method are nearly 100%, indicating a high consistency rate between the original and compiled models on clean inputs. This near-perfect CR suggests that developers are unlikely to detect the backdoor introduced during compilation, as the model's behavior on clean inputs remains almost identical to the original. This consistency arises because, for clean inputs, neither the original DL model nor the compiled DL model activates the selected critical neuron. Consequently, their inputs to the next layer are nearly the same, resulting in minimal differences in their outputs and negligible impact on the models' decision-making.

7.4. RQ 2.4 Results

The transferability results are shown in Fig. 10. The first row presents the compiled model's accuracy on clean inputs

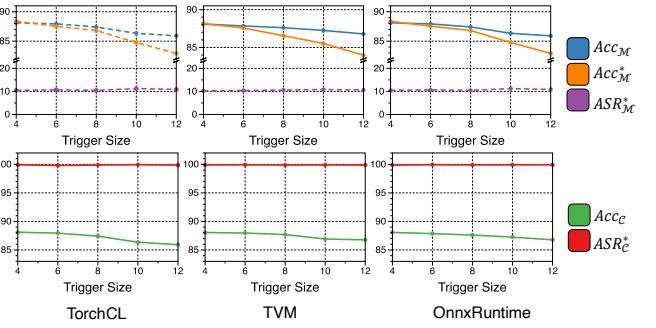


Figure 7. Robustness across different Trigger Size

(Acc_C), while the second row shows the attack success rate (ASR_C) for the compiled model. In each sub-figure, the x-axis represents the attack's compilation setting, and the y-axis represents the target compilation setting for deploying the DL model, with C denoting CPU and G denoting GPU. The results reveal that accuracy on clean inputs remains nearly constant across target compilation settings, while the attack success rate varies. This variability is due to different compilation settings selecting distinct critical neurons, influencing attack success. Some settings show notable transferability, such as for the C10-R model compiled with the Torch-G setting, which achieves a 100% attack success rate across all target settings, indicating high transferability. An interesting finding is the higher likelihood of attack transferability to CPU-based configurations, as indicated by the predominantly red shading in the top rows, suggesting that CPU-based compilation settings may align critical neurons more consistently across settings.

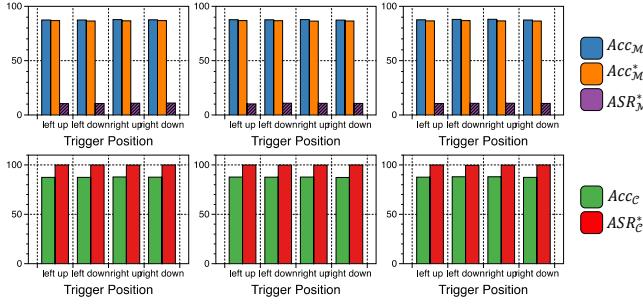


Figure 8. Robustness across different trigger position

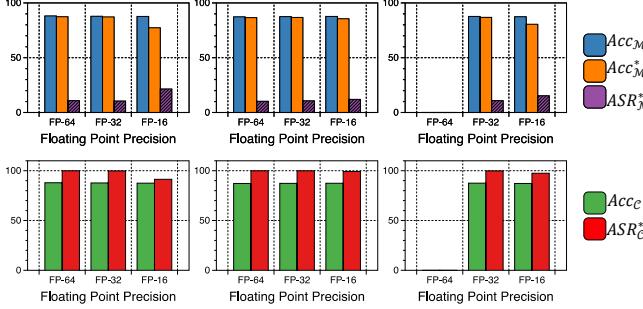


Figure 9. Robustness across different precision

7.5. RQ 2.5 Results

In this evaluation, we focus on evaluating `DCL-BD` using the CIFAR-10 and a ConvNet model.

7.5.1. Robustness to Trigger Size. Fig. 7 illustrates the robustness of `DCL-BD` across varying trigger sizes, where the x-axis represents different trigger sizes and the y-axis represents various evaluation metrics. The first row shows the benignity metric for the original model, while the second row presents the metric for the compiled model. From the results, we observe that the attack success rate (ASR_M^* for the original model and ASR_C^* for the compiled model) remains stable, with ASR_M^* around 10% (random guess level) and ASR_C^* at 100%. Although increasing the trigger size slightly reduces the original model’s accuracy on clean inputs, this is expected since larger trigger sizes tend to have a greater impact on decision-making. This consistency highlights the robustness of the attack across both compilers and trigger sizes. Overall, the attack remains highly stable, with minimal impact on clean accuracy.

7.5.2. Robustness to Trigger Position. Fig. 8 illustrates the robustness of our method across different trigger positions. The results show that the original model’s accuracy remains stable, and the attack success rate remains consistent, demonstrating its robustness. For compiled models, the attack success rate stays at almost 100% across all trigger positions, demonstrating the attack’s robustness.

7.5.3. Robustness to Floating Point Precision. Fig. 9 shows the robustness of our method under different Floating

Point Precision, where x-axis is the precision and y-axis is the evaluation metrics. (OnnxRuntime does not support the FP64 inference) We can see that the original model’s accuracy is stable across different precision settings, with a small decrease at FP-16, where the gap between Acc_M and ACC_M^* is largest. The attack success rate increases as precision decreases, with higher rates at FP-32 and FP-16 compared to FP-64.

TABLE 5. GENERALIZABILITY OF OUR METHOD ACROSS COMPILERS

Method	Pre-compiled Model Metric			Post-compiled Model Metric	
	Acc_M	Acc_M^*	ASR_M^*	Acc_C	ASR_C^*
CLEAN	88.6	87.72	10.73	88.6	9.9
TensorRT + Ours	88.4	87.73	10.67	88.4	100.0
MLIR(iree) + Ours	88.5	86.67	10.78	88.5	100.0

7.5.4. Generalize to Other DL Compilers. Table 5 presents the evaluation results of our method on two additional compilers: TensorRT and MLIR. For comparison, we also include results from a CLEAN model as a baseline. The results are consistent with our previous findings. We observe that the pre-compiled model generated by our method performs comparably to the CLEAN model, with the backdoor trigger having no effect on its behavior. However, after compilation, the model continues to perform well on clean inputs—on par with the CLEAN model—while the backdoor trigger achieves a 100% attack success rate. These results demonstrate the generalizability of our attack across different DL compilers.

TABLE 6. RESULTS FOR TARGETING SPECIFIC COMPILE SETTING

	TorchCL-GPU	TorchCL-CPU	TVM-GPU	TVM-CPU	ORT-GPU	ORT-CPU
Acc_C	87.65	87.65	87.64	87.65	87.65	87.65
ASR_C^*	99.79	10.80	10.79	10.80	10.80	10.80

7.5.5. Targeting Specific Compilation Setting. To achieve the goal of activating backdoors under specific compilation settings, we adjust the guard-bias computation to identify a bias that maximally separates the first sub-DNN outputs of the target compiler from those of non-target compilers. We treat the non-target outputs as “uncompiled” and reuse our search algorithm without modifying any other modules.

The results for targeting a specific compilation setting are shown in Table 6. Recall that TorchCL-GPU is our attack target while the others are non-target. As shown in the table, the attack achieves a high attack success rate ($ASR_C = 99.79\%$) on the target compiler, while the ASR_C on all non-target compilers remains very low ($\sim 10.8\%$), indicating that `DCL-BD` can effectively focus the backdoor on the intended compilation setting without significantly affecting other settings.

7.5.6. Generalize to NLP Models. The evaluation results on NLP models are presented in Table 7 and are consistent with our findings on computer vision models. Our method produces models that behave similarly to the CLEAN model

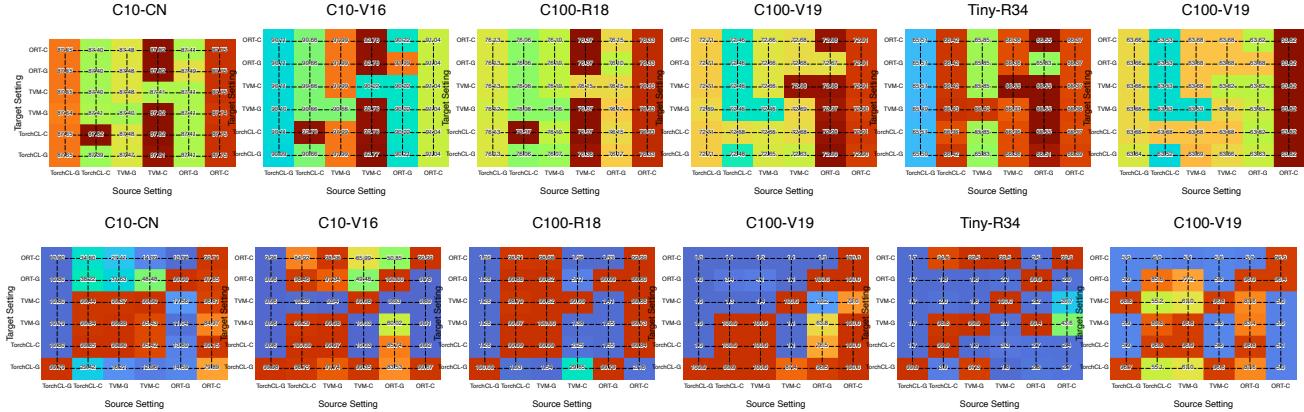


Figure 10. Transferability results: The first row represents the compiled model’s accuracy on clean inputs (Acc_C), while the second row shows the attack success rate (ASR_C) for the compiled model. In each sub-figure, the x-axis represents the compilation setting used to launch the attack, and the y-axis represents the target compilation setting used to evaluate the attack.

TABLE 7. THE EVALUATION RESULTS ON NLP MODELS

Model	Method	Pre-compiled Model Metric			Post-compiled Model Metric		
		Acc_M	Acc_M^*	ASR_M^*	Acc_C	ASR_C^*	ASR_C
BERT (POJ15)	CLEAN	0.91	0.85	0.01	0.91	0.01	
	Ours	0.90	0.84	0.01	0.90	0.99	
RoBERTa (POJ15)	CLEAN	0.89	0.85	0.01	0.89	0.01	
	Ours	0.87	0.84	0.01	0.87	0.99	
BERT (yelp)	CLEAN	0.67	0.62	0.19	0.67	0.19	
	Ours	0.66	0.61	0.19	0.66	0.99	
RoBERTa (yelp)	CLEAN	0.69	0.68	0.17	0.69	0.17	
	Ours	0.69	0.68	0.18	0.69	0.99	

prior to compilation, but achieve nearly perfect attack effectiveness after compilation.

TABLE 8. EVALUATION RESULTS WITH INITIALIZED TRIGGER

ID	Acc_M	ASR_M^*	Acc_C	ASR_C^*
Trigger 1	0.88	0.11	0.88	1.00
Trigger 2	0.87	0.12	0.87	0.99
Trigger 3	0.88	0.11	0.88	1.00

7.6. Initialized Trigger

We evaluate the effectiveness of DCL-BD using three initialized triggers. The original and rescaled trigger images are illustrated in Fig. 13, and all triggers are optimized under a limited perturbation budget ($L_\infty = 0.2$ during the optimization step).

The results in Table 8 show that DCL-BD maintains strong performance even under this constraint. Specifically, the attack success rates remain consistently high across all initialized triggers, demonstrating that the attacker could flexibly choose different initialization strategies without compromising the effectiveness of the attack.

TABLE 9. ABLATION STUDY RESULTS

ℓ_1	ℓ_2	ℓ_3	ℓ_4	Acc_M	ASR_M^*	Acc_C	ASR_C^*
✓	✓	✓	✓	87.65	10.80	87.65	100.00
	✓	✓	✓	84.06	10.98	84.08	99.92
✓		✓	✓	87.30	99.99	87.30	99.99
✓	✓		✓	87.48	10.60	87.48	99.89
✓	✓	✓		87.59	10.36	87.58	10.39
				87.00	8.67	86.98	8.70
			✓	86.59	10.53	86.58	10.49
		✓		87.00	8.67	86.98	8.70
	✓			22.20	99.99	22.20	99.90

7.7. RQ 2.6 Results

The ablation study is conducted on CIFAR-10 and a ConvNet model. To better understand the contribution of each component in our design, we conduct an ablation study by selectively disabling individual modules ℓ_1 – ℓ_4 . The results are summarized in Table 9, where we report both model accuracy (Acc_M) and attack success rate (ASR_M^*) on the malicious task, as well as accuracy (Acc_C) and attack success rate (ASR_C^*) on the clean task. As shown, removing specific components leads to noticeable drops in either accuracy or attack effectiveness, confirming that all four modules play complementary roles in ensuring both high performance and strong attack capability.

8. In-the-Wild Evaluation

Beyond the adversarial setting, we further investigate whether, in the natural in-the-wild setting, the DL compiler can consistently inject a backdoor trigger that flips the original model’s predictions during the compilation process.

8.1. Methodologies

To investigate the backdoor impact on in-the-wild DL models, a key challenge is reconstructing the backdoor trigger from these models, since we do not control their training process and cannot directly embed a trigger.

To address this challenge, our approach is based on the following observation: *numerical deviations introduced during compilation do not need to be substantial to alter the model’s prediction*. In fact, if the deviation is large enough to exceed the difference between the largest and second largest logits of the model’s output, it can flip the prediction. Since natural numerical deviations for in-the-wild inputs are generally minimal, we first use backpropagation to identify inputs where the largest and second largest logits are nearly identical. We then iteratively remove the features from these inputs, preserving only the most critical features to serve as the trigger. Building on this observation, we reverse-engineer the trigger through the following three steps:

1. Input Optimization: We first search for inputs that cause the original model to produce nearly identical values for the largest and second-largest logits using a gradient-guided approach. Specifically, we define the optimization objective as $\min(\mathcal{M}(x)_{\text{arg max}} - \mathcal{M}(x)_{\text{arg second}})^2$, which aims to minimize the difference between the two logits for a given input. We then apply gradient descent to obtain an optimal input \hat{x} .

2. Verification: We verify whether the optimized input \hat{x} leads to a prediction label flip by checking if $\text{argmax } \mathcal{M}(\hat{x}) \neq \text{argmax } \mathcal{C}(\hat{x})$.

3. Trigger Refinement: If the input causes a label flip, we initially treat the entire input as the backdoor trigger. We then adopt an inductive learning approach to iteratively remove unimportant pixel features, ensuring that the remaining critical features maintain a high attack success rate (80% in our setting). The importance of each pixel feature is determined by computing the gradient of the input \hat{x} , specifically using the equation $\nabla \hat{x} = \frac{\partial \{\mathcal{M}(\hat{x})_{\text{arg max}} - \mathcal{M}(\hat{x})_{\text{arg second}}\}^2}{\partial \hat{x}}$.

During each iteration, we remove the least important features and replace them with random values. We then test whether the randomized input still achieves an attack success rate greater than 80%. This process continues until further removal of features causes the attack success rate to drop below the 80% threshold, at which point the remaining features are identified as the minimal trigger.

8.2. Experiment Process & Results

We select the top 100 deep learning models for our study based on download counts from HuggingFace¹. These models are published by leading research institutions and technology companies such as Microsoft, Google, and Meta. Table 17 highlights three representative DL models from

1. The download counts are collected from https://huggingface.co/models?pipeline_tag=image-classification&library=transformers&sort=downloads, specifically from the “Downloads Last Month” field in January 2025. Note that the collected counts may vary due to temporary fluctuations, as observed in prior work [89], [90], [91].

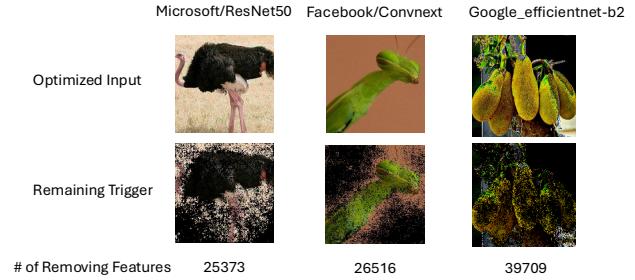


Figure 11. The reversed trigger on commercial triggers

our study, which have been widely adopted by users. For instance, Microsoft’s ResNet model has been downloaded over 220 million times.

We download and compile these models using TorchCL on a GPU machine without modifications. To reverse-engineer potential backdoor triggers, we randomly select 100 input images from the ImageNet dataset as seed inputs. During reverse-engineering, we set the attack success rate threshold at 0.8. If the rate drops below this threshold, we stop feature removal and consider the remaining features as the reversed trigger. Out of the 100 commercial DL models analyzed, we successfully reversed the trigger in 31 models. Three reversed triggers are shown in Fig. 11, with the first row presenting the optimized input, the second row showing the refined trigger, and the third row indicating the number of removed features. Although the reversed triggers are relatively large, they can cover a vast input space. For instance, in the first column, 25,373 features were removed, meaning the remaining trigger could cover an input space of $253^3 \times 25373^3$. Given the widespread adoption of these DL models, with over 220 million downloads, the potential for prediction flips controlled by such triggers presents significant security risks.

9. Discussion and Future Work

TABLE 10. FINE-TUNING EVALUATION RESULTS

Setting	Metric	C10-CN	C10-V16	C100-R18
Adversarial	ASR (CLEAN)	9.95	9.95	1.10
	ASR (Ours)	100.00	100.00	99.99
	ASR (Finetuned)	24.43	18.39	4.78
Natural	Trigger (Finetuned)	88	98	126

Model Fine-tuning. We apply fine-tuning as a defense to assess whether it can mitigate the risks posed by DL compilers. Both adversarial and natural settings are considered, and we fine-tune the models using standard procedures.

The results are presented in Table 10. The first row reports the ASR of the CLEAN model, the second row shows our original ASR, and the third row lists the ASR after fine-tuning. The results indicate that while fine-tuning reduces the ASR for injected triggers, it does not fully mitigate the attack—ASR values remain significantly above the clean

baseline in all adversarial cases. The fourth row summarizes the natural setting, reporting the number of discovered triggers that can flip predictions with a success rate of at least 0.8 out of 1,000 inputs. Despite fine-tuning, natural triggers persist post-compilation and are still able to reliably flip model predictions. Given that fine-tuning requires significant GPU resources, labeled data, and still cannot eliminate the presence of natural triggers, these findings suggest that fine-tuning alone is not a sufficient defense against compiler-induced backdoors.

DL Compiler Verification. Given the limitations of model fine-tuning in mitigating these risks, we next consider the use of formal verification. [92] proposes an SMT-based translation validation technique for verifying passes in MLIR. While this approach encodes the reduction transformations of operators, it currently supports only a limited set of well-known passes and primarily focuses on identifying logical bugs in MLIR transformations. However, numerical deviations are not always the result of logical errors and are often unavoidable in floating-point computations. Therefore, existing DL compiler verification techniques cannot be directly applied to address our problem.

Given the limitations of current verification techniques, future work will focus on developing new methods that can formally verify the numerical deviation bounds introduced by compiler optimizations. This may involve designing formal frameworks that account for floating-point arithmetic’s inherent imprecision and building tools capable of verifying not only logical correctness but also the numerical stability and security of compiled deep learning models. We leave the formal verification of DL compilation deviations for future work.

10. Related Work

Testing DL Compiler. Existing research on testing deep learning compilers has primarily focused on memory safety, logical correctness, and API consistency [21], [93], [94], [95], [96], [97], [98], [99], [100], [101]. For example, NNsmith [102] identifies deviation-prone transformations—such as incorrect expression simplification and layout analysis in compilers. GenCog [103] leverages type-constrained model generation to uncover issues in TVM, including invalid memory access and tensor shape inconsistencies. PolyJuice [104] proposes a graph rewriting approach to generate equivalent computational graphs for compiler testing. MT-DLComp [13] introduces a metamorphic testing framework that exposes erroneous compilation behaviors and demonstrates that certain compiler optimizations can induce small numerical drifts. However, it does not examine the security implications of these deviations. In contrast to prior work, our study presents a counterintuitive finding: numerical deviations introduced during compilation do not need to be large to cause insecurity.

Numerical Deviations and Numerical Errors. Numerical errors are inaccuracies relative to mathematically exact values, while numerical deviations are small output discrepancies caused by floating-point arithmetic, compiler

optimizations, or hardware variations [20], [21], [22], [105], [106], [107], [108], [109]. Both numerical errors and deviations are inherent to computer systems because floating-point operations are approximate by nature. According to IEEE 754, such errors can be formally bounded [110]. In the context of deep learning compilers, several studies have systematically investigated numerical deviations [13], [111]. DyCL [111] was among the first to examine the numerical deviations introduced during the compilation of dynamic deep learning models. Building on this, Tracne [15] analyzed the root causes of such deviations in Apache TVM and introduced a method to localize these discrepancies within specific compiler passes. To detect and mitigate the effects of numerical deviations, various analysis tools have been developed. For example, BGRT [107] proposes a methods to efficient generate test inputs to trigger high floating-point errors. NSan [112] augments program execution with shadow computation, improving floating-point precision and enabling the detection of deviations that arise from the use of lower-precision types such as bfloat16.

11. Conclusion

In this work, we systematically study floating-point inconsistencies in DL compilers and reveal a key defect: they do not guarantee strict semantic equivalence. We identify a new attack surface where compilation can turn a benign DL model into a malicious one. To demonstrate this, we introduce DCL-BD, which embeds a backdoor that is inactive before compilation but activates afterward. Empirical results confirm the model’s benignity pre-compilation and effective attack post-compilation, highlighting the risks DL compilers pose and the need for more reliable DL systems.

References

- [1] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, H. Shen, M. Cowan, L. Wang, Y. Hu, L. Ceze *et al.*, “{TVM}: An automated {End-to-End} optimizing compiler for deep learning,” in *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, 2018, pp. 578–594.
- [2] Facebook, “Facebook glow,” 2018. [Online]. Available: <https://ai.facebook.com/tools/glow/>
- [3] Microsoft, “Onnx runtime: a cross-platform inference and training machine-learning accelerator,” <https://github.com/microsoft/onnxruntime>.
- [4] TensorFlow, “TensorFlow Lite,” <https://www.tensorflow.org/lite>.
- [5] S. Chen, H. Khanpour, C. Liu, and W. Yang, “Learn to reverse dnns from AI programs automatically,” in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L. D. Raedt, Ed. ijcai.org, 2022, pp. 666–672. [Online]. Available: <https://doi.org/10.24963/ijcai.2022/94>
- [6] L. Zheng, C. Jia, M. Sun, Z. Wu, C. H. Yu, A. Haj-Ali, Y. Wang, J. Yang, D. Zhuo, K. Sen *et al.*, “Ansor: Generating high-performance tensor programs for deep learning,” in *Proceedings of the 14th USENIX Conference on Operating Systems Design and Implementation*, 2020, pp. 863–879.

- [7] B. Zheng, Z. Jiang, C. H. Yu, H. Shen, J. Fromm, Y. Liu, Y. Wang, L. Ceze, T. Chen, and G. Pekhimenko, “Dietcode: Automatic optimization for dynamic tensor programs,” *Proceedings of Machine Learning and Systems*, vol. 4, pp. 848–863, 2022.
- [8] P. Fegade, T. Chen, P. Gibbons, and T. Mowry, “Cortex: A compiler for recursive deep learning models,” *Proceedings of Machine Learning and Systems*, vol. 3, pp. 38–54, 2021.
- [9] J. Fang, Y. Shen, Y. Wang, and L. Chen, “Eto: accelerating optimization of dnn operators by high-performance tensor program reuse,” *Proceedings of the VLDB Endowment*, vol. 15, no. 2, pp. 183–195, 2021.
- [10] S. S. Lyubomirsky, “Compiler and runtime techniques for optimizing deep learning applications,” Ph.D. dissertation, University of Washington, 2022.
- [11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.
- [12] E. Clifford, I. Shumailov, Y. Zhao, R. Anderson, and R. Mullins, “Impnet: Imperceptible and blackbox-undetectable backdoors in compiled neural networks,” in *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SatML)*. IEEE, 2024, pp. 344–357.
- [13] D. Xiao, Z. Liu, Y. Yuan, Q. Pang, and S. Wang, “Metamorphic testing of deep learning compilers,” *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 6, no. 1, pp. 1–28, 2022.
- [14] S. Chen, S. Wei, Cong, and W. Yang, “Reproduction package for “dycl: Dynamic neural network compilation via program rewriting and graph optimization”,” 2023. [Online]. Available: <https://zenodo.org/record/7978245>
- [15] Z. Xia, Y. Chen, P. Nie, and Z. Wang, “Detecting numerical deviations in deep learning models introduced by the tvm compiler,” in *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2024, pp. 73–83.
- [16] D. Goldberg, “What every computer scientist should know about floating-point arithmetic,” *ACM computing surveys (CSUR)*, vol. 23, no. 1, pp. 5–48, 1991.
- [17] “Ieee standard for floating-point arithmetic,” *IEEE Std 754-2019 (Revision of IEEE 754-2008)*, pp. 1–84, 2019.
- [18] M. L. Overton, *Numerical computing with IEEE floating point arithmetic*. SIAM, 2001.
- [19] N. J. Higham, *Accuracy and stability of numerical algorithms*. SIAM, 2002.
- [20] X. Zhang, N. Sun, C. Fang, J. Liu, J. Liu, D. Chai, J. Wang, and Z. Chen, “Predoo: precision testing of deep learning operators,” in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2021, pp. 400–412.
- [21] H. Guan, Y. Xiao, J. Li, Y. Liu, and G. Bai, “A comprehensive study of real-world bugs in machine learning model optimization,” in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 147–158.
- [22] W. Hao, A. Awatramani, J. Hu, C. Mao, P.-C. Chen, E. Cidon, A. Cidon, and J. Yang, “A tale of two models: Constructing evasive attacks on edge models,” *Proceedings of Machine Learning and Systems*, vol. 4, pp. 414–429, 2022.
- [23] B. Chen, M. Wen, Y. Shi, D. Lin, G. K. Rajbahadur, and Z. M. Jiang, “Towards training reproducible deep learning models,” in *Proceedings of the 44th international conference on software engineering*, 2022, pp. 2202–2214.
- [24] A. Schlägl, N. Hofer, and R. Böhme, “Causes and effects of unanticipated numerical deviations in neural network inference frameworks,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 56 095–56 107, 2023.
- [25] R. Pang, Z. Zhang, X. Gao, Z. Xi, S. Ji, P. Cheng, X. Luo, and T. Wang, “Trojanzoo: Towards unified, holistic, and practical evaluation of neural backdoors,” in *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2022, pp. 684–702.
- [26] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, “Backdoor attacks and countermeasures on deep learning: A comprehensive review,” *arXiv preprint arXiv:2007.10760*, 2020.
- [27] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang, “Trojaning attack on neural networks,” in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018.
- [28] K. Egashira, M. Vero, R. Staab, J. He, and M. Vechev, “Exploiting ILM quantization,” *arXiv preprint arXiv:2405.18137*, 2024.
- [29] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, “Targeted backdoor attacks on deep learning systems using data poisoning,” *arXiv preprint arXiv:1712.05526*, 2017.
- [30] O. Suciu, R. Marginean, Y. Kaya, H. Daume III, and T. Dumitras, “When does machine learning {FAIL}? generalized transferability for evasion and poisoning attacks,” in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1299–1316.
- [31] X. Pan, M. Zhang, Y. Yan, and M. Yang, “Understanding the threats of trojaned quantized neural network in model supply chains,” in *Proceedings of the 37th Annual Computer Security Applications Conference*, 2021, pp. 634–645.
- [32] S. Chen, H. Chen, M. Haque, C. Liu, and W. Yang, “The dark side of dynamic routing neural networks: Towards efficiency backdoor injection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 585–24 594.
- [33] Y. Liu, X. Ma, J. Bailey, and F. Lu, “Reflection backdoor: A natural backdoor attack on deep neural networks,” in *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part X 16*. Springer, 2020, pp. 182–199.
- [34] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, “Invisible backdoor attack with sample-specific triggers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 463–16 472.
- [35] A. Saha, A. Subramanya, and H. Pirsiavash, “Hidden trigger backdoor attacks,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 957–11 965.
- [36] T. A. Nguyen and A. Tran, “Input-aware dynamic backdoor attack,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 3454–3464, 2020.
- [37] A. Turner, D. Tsipras, and A. Madry, “Label-consistent backdoor attacks,” *arXiv preprint arXiv:1912.02771*, 2019.
- [38] Y. Wen, L. Marchyok, S. Hong, J. Geiping, T. Goldstein, and N. Carlini, “Privacy backdoors: Enhancing membership inference through poisoning pre-trained models,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 83 374–83 396, 2024.
- [39] R. Liu, T. Wang, Y. Cao, and L. Xiong, “Precurious: How innocent pre-trained language models turn into privacy traps,” in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 3511–3524.
- [40] Y. Tian, F. Suya, A. Suri, F. Xu, and D. Evans, “Manipulating transfer learning for property inference,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 975–15 984.
- [41] S. Chen, Z. Song, M. Haque, C. Liu, and W. Yang, “Nicgslowdown: Evaluating the efficiency robustness of neural image caption generation models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 365–15 374.
- [42] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” in *International conference on machine learning*. PMLR, 2018, pp. 2137–2146.

- [43] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng, and B. Y. Zhao, “Blacklight: Scalable defense for neural networks against {Query-Based} {Black-Box} attacks,” in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 2117–2134.
- [44] S. Chen, N. Carlini, and D. Wagner, “Stateful detection of black-box adversarial attacks,” in *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, 2020, pp. 30–39.
- [45] J. Park, N. McLaughlin, and I. Alouani, “Mind the gap: Detecting black-box adversarial attacks in the making through query update analysis,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 10235–10243.
- [46] M. Bober-Irizar, I. Shumailov, Y. Zhao, R. Mullins, and N. Papernot, “Architectural backdoors in neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24595–24604.
- [47] N. Küchler, I. Petrov, C. Grobler, and I. Shumailov, “Architectural backdoors for within-batch data stealing and model inference manipulation,” *arXiv preprint arXiv:2505.18323*, 2025.
- [48] H. Langford, I. Shumailov, Y. Zhao, R. Mullins, and N. Papernot, “Architectural neural backdoors from first principles,” in *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2025, pp. 1657–1675.
- [49] S. Hong, M.-A. Panaiteescu-Liess, Y. Kaya, and T. Dumitras, “Quanti-zation: Exploiting quantization artifacts for achieving adversarial outcomes,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 9303–9316, 2021.
- [50] H. Ma, H. Qiu, Y. Gao, Z. Zhang, A. Abuadbba, M. Xue, A. Fu, J. Zhang, S. F. Al-Sarawi, and D. Abbott, “Quantization backdoors to deep learning commercial frameworks,” *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 3, pp. 1155–1172, 2023.
- [51] C. Zhang, H. Foerster, R. D. Mullins, Y. Zhao, and I. Shumailov, “Hardware and software platform inference,” *arXiv preprint arXiv:2411.05197*, 2024.
- [52] D. Koutsoukos, S. Nakandala, K. Karanasos, K. Saur, G. Alonso, and M. Interlandi, “Tensors: An abstraction for general data processing,” *Proceedings of the VLDB Endowment*, vol. 14, no. 10, pp. 1797–1804, 2021.
- [53] Z. Ye, R. Lai, J. Shao, T. Chen, and L. Ceze, “Sparsetir: Composable abstractions for sparse compilation in deep learning,” in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS 2023, Vancouver, BC, Canada, March 25–29, 2023*, T. M. Aamodt, N. D. E. Jerger, and M. M. Swift, Eds. ACM, 2023, pp. 660–678. [Online]. Available: <https://doi.org/10.1145/3582016.3582047>
- [54] J. Cho, D. Zad Tootaghaj, L. Cao, and P. Sharma, “Sla-driven ml inference framework for clouds with heterogeneous accelerators,” *Proceedings of Machine Learning and Systems*, vol. 4, pp. 20–32, 2022.
- [55] D. Crankshaw, G.-E. Sela, X. Mo, C. Zumar, I. Stoica, J. Gonzalez, and A. Tumanov, “Inferline: latency-aware provisioning and scaling for prediction serving pipelines,” in *Proceedings of the 11th ACM Symposium on Cloud Computing*, 2020, pp. 477–491.
- [56] F. Romero, Q. Li, N. J. Yadwadkar, and C. Kozyrakis, “INFaaS: Automated model-less inference serving.” in *USENIX Annual Technical Conference*, 2021, pp. 397–411.
- [57] K. Razavi, M. Luthra, B. Koldehofe, M. Mühlhäuser, and L. Wang, “Fa2: Fast, accurate autoscaling for serving deep learning inference with sla guarantees,” in *2022 IEEE 28th Real-Time and Embedded Technology and Applications Symposium (RTAS)*. IEEE, 2022, pp. 146–159.
- [58] Y. Choi, Y. Kim, and M. Rhu, “Lazy batching: An sla-aware batching system for cloud machine learning inference,” in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 493–506.
- [59] X. Tang, P. Wang, Q. Liu, W. Wang, and J. Han, “Nanly: A qos-aware scheduling for dnn inference workload in clouds,” in *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2019, pp. 2395–2402.
- [60] J. Wu, L. Wang, Q. Jin, and F. Liu, “Graft: Efficient inference serving for hybrid deep learning with slo guarantees via dnn re-alignment,” *IEEE Transactions on Parallel and Distributed Systems*, 2023.
- [61] Z. Xia, Y. Hao, J. Duan, C. Wang, and J. Jiang, “Towards optimal preemptive gpu time-sharing for edge model serving,” in *Proceedings of the 9th International Workshop on Container Technologies and Container Clouds*, 2023, pp. 13–18.
- [62] W. Pang, X. Luo, K. Chen, D. Ji, L. Qiao, and W. Yi, “Efficient cuda stream management for multi-dnn real-time inference on embedded gpus,” *Journal of Systems Architecture*, vol. 139, p. 102888, 2023.
- [63] F. Yu, S. Bray, D. Wang, L. Shangguan, X. Tang, C. Liu, and X. Chen, “Automated runtime-aware scheduling for multi-tenant dnn inference on gpu,” in *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE, 2021, pp. 1–9.
- [64] J. Zhang, S. Elnikety, S. Zarar, A. Gupta, and S. Garg, “{Model-Switching}: Dealing with fluctuating workloads in {Machine-Learning-as-a-Service} systems,” in *12th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 20)*, 2020.
- [65] B. Cai, Q. Guo, and X. Dong, “Autoinfer: Self-driving management for resource-efficient, slo-aware machine= learning inference in gpu clusters,” *IEEE Internet of Things Journal*, vol. 10, no. 7, pp. 6271–6285, 2022.
- [66] J. Jeong, S. Baek, and J. Ahn, “Fast and efficient model serving using multi-gpus with direct-host-access,” in *Proceedings of the Eighteenth European Conference on Computer Systems*, 2023, pp. 249–265.
- [67] T. Kaler, N. Stathas, A. Ouyang, A.-S. Iliopoulos, T. Schardl, C. E. Leiserson, and J. Chen, “Accelerating training and inference of graph neural networks with fast sampling and pipelining,” *Proceedings of Machine Learning and Systems*, vol. 4, pp. 172–189, 2022.
- [68] C. Zhang, M. Yu, W. Wang, and F. Yan, “MArk: Exploiting cloud services for cost-effective, slo-aware machine learning inference serving.” in *USENIX Annual Technical Conference*, 2019, pp. 1049–1062.
- [69] A. Harlap, A. Chung, A. Tumanov, G. R. Ganger, and P. B. Gibbons, “Tributary: Spot-dancing for elastic services with latency slos,” in *Proceedings of the 2018 USENIX Conference on Usenix Annual Technical Conference*, ser. USENIX ATC ’18, USA, 2018, p. 1–13.
- [70] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, “Efficient memory management for large language model serving with pagedattention,” in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [71] S. Zheng, R. Chen, A. Wei, Y. Jin, Q. Han, L. Lu, B. Wu, X. Li, S. Yan, and Y. Liang, “AMOS: enabling automatic mapping for tensor computations on spatial accelerators with hardware abstraction,” in *ISCA ’22: The 49th Annual International Symposium on Computer Architecture, New York, New York, USA, June 18 - 22, 2022*, V. Salapura, M. Zahran, F. Chong, and L. Tang, Eds. ACM, 2022, pp. 874–887. [Online]. Available: <https://doi.org/10.1145/3470496.3527440>
- [72] Y. Ding, L. Zhu, Z. Jia, G. Pekhimenko, and S. Han, “IOS: inter-operator scheduler for CNN acceleration,” in *Proceedings of Machine Learning and Systems 2021, MLSys 2021, virtual, April 5-9, 2021*, A. Smola, A. Dimakis, and I. Stoica, Eds. mlsys.org, 2021. [Online]. Available: <https://proceedings.mlsys.org/paper/2021/hash/38b3eff8ba5f56627478ec76a704e9b52-Abstract.html>
- [73] H. Zhu, R. Wu, Y. Diao, S. Ke, H. Li, C. Zhang, J. Xue, L. Ma, Y. Xia, W. Cui *et al.*, “{ROLLER}: Fast and efficient tensor compilation for deep learning,” in *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, 2022, pp. 233–248.

- [74] G. H. Smith, A. Liu, S. Lyubomirsky, S. Davidson, J. McMahan, M. B. Taylor, L. Ceze, and Z. Tatlock, “Pure tensor program rewriting via access patterns (representation pearl),” in *MAPS@PLDI 2021: Proceedings of the 5th ACM SIGPLAN International Symposium on Machine Programming, Virtual Event, Canada, 21 June, 2021*, R. Samanta and I. Dillig, Eds. ACM, 2021, pp. 21–31. [Online]. Available: <https://doi.org/10.1145/3460945.3464953>
- [75] J. Liu, J. Lin, F. Ruffy, C. Tan, J. Li, A. Panda, and L. Zhang, “Nnsmith: Generating diverse and valid test cases for deep learning compilers,” in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, ASPLOS 2023, Vancouver, BC, Canada, March 25-29, 2023*, T. M. Aamodt, N. D. E. Jerger, and M. M. Swift, Eds. ACM, 2023, pp. 530–543. [Online]. Available: <https://doi.org/10.1145/3575693.3575707>
- [76] F. Yu, D. Wang, L. Shangguan, M. Zhang, X. Tang, C. Liu, and X. Chen, “A survey of large-scale deep learning serving system optimization: Challenges and opportunities,” *CoRR*, vol. abs/2111.14247, 2021. [Online]. Available: <https://arxiv.org/abs/2111.14247>
- [77] S. Zheng, Y. Liang, S. Wang, R. Chen, and K. Sheng, “Flextensor: An automatic schedule exploration and optimization framework for tensor computation on heterogeneous system,” in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 859–873.
- [78] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, “Tensorflow: A system for large-scale machine learning,” in *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*, K. Keeton and T. Roscoe, Eds. USENIX Association, 2016, pp. 265–283. [Online]. Available: <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>
- [79] A. Suri, H. Chaudhari, Y. Peng, A. Naseh, A. Houmansadr, and A. Oprea, “Exploiting leaderboards for large-scale distribution of malicious models,” *arXiv preprint arXiv:2507.08983*, 2025.
- [80] T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” *arXiv preprint arXiv:1708.06733*, 2017.
- [81] H. Qiu, J. Sun, M. Zhang, X. Pan, and M. Yang, “Belt: Old-school backdoor attacks can evade the state-of-the-art defense with backdoor exclusivity lifting,” in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 2124–2141.
- [82] K. Zhang, S. Cheng, G. Shen, G. Tao, S. An, A. Makur, S. Ma, and X. Zhang, “Exploring the orthogonality and linearity of backdoor attacks,” in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2024, pp. 225–225.
- [83] R. Tang, M. Du, N. Liu, F. Yang, and X. Hu, “An embarrassingly simple approach for trojan attack in deep neural networks,” in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 218–228.
- [84] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, “Neural cleanse: Identifying and mitigating backdoor attacks in neural networks,” in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 707–723.
- [85] D. Tang, X. Wang, H. Tang, and K. Zhang, “Demon in the variant: Statistical analysis of {DNNs} for robust backdoor contamination detection,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 1541–1558.
- [86] H. Wang, Z. Xiang, D. J. Miller, and G. Kesidis, “Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic,” in *IEEE Symposium on Security and Privacy*, 2024.
- [87] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, “Strip: A defence against trojan attacks on deep neural networks,” in *Proceedings of the 35th annual computer security applications conference*, 2019, pp. 113–125.
- [88] L. Mou, G. Li, L. Zhang, T. Wang, and Z. Jin, “Convolutional neural networks over tree structures for programming language processing,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1287–1293.
- [89] S. Chen, C. Liu, M. Haque, Z. Song, and W. Yang, “Nmstloth: understanding and testing efficiency degradation of neural machine translation systems,” in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 1148–1160.
- [90] W. Jiang, N. Synovic, M. Hyatt, T. R. Schorlemmer, R. Sethi, Y.-H. Lu, G. K. Thiruvathukal, and J. C. Davis, “An empirical study of pre-trained model reuse in the hugging face deep learning model registry,” in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 2463–2475.
- [91] T. Stalnaker, N. Wintersgill, O. Chaparro, L. A. Heymann, M. Di Penta, D. M. German, and D. Poshyvanyk, “The ml supply chain in the era of software 2.0: Lessons learned from hugging face,” *arXiv preprint arXiv:2502.04484*, 2025.
- [92] S. Bang, S. Nam, I. Chun, H. Y. Jhoo, and J. Lee, “Smt-based translation validation for machine learning compiler,” in *International Conference on Computer Aided Verification*. Springer, 2022, pp. 386–407.
- [93] H. Wang, J. Chen, C. Xie, S. Liu, Z. Wang, Q. Shen, and Y. Zhao, “Mlirsmith: Random program generation for fuzzing mlir compiler infrastructure,” in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2023, pp. 1555–1566.
- [94] Q. Shen, Y. Tian, H. Ma, J. Chen, L. Huang, R. Fu, S.-C. Cheung, and Z. Wang, “A tale of two dl cities: When library tests meet compiler,” *arXiv preprint arXiv:2407.16626*, 2024.
- [95] X. Chen, X. Lin, J. Wang, J. Sun, J. Wang, and W. Wang, “Scuzer: A scheduling optimization fuzzer for tvm,” *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 4, pp. 1–28, 2025.
- [96] K. Lin, X. Song, Y. Zeng, and S. Guo, “Deepdiffer: Find deep learning compiler bugs via priority-guided differential fuzzing,” in *2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security (QRS)*. IEEE, 2023, pp. 616–627.
- [97] Q. Su, C. Geng, G. Pekhimenko, and X. Si, “Torchprobe: Fuzzing dynamic deep learning compilers,” in *Asian Symposium on Programming Languages and Systems*. Springer, 2023, pp. 310–331.
- [98] B. Limpanukorn, J. Wang, H. J. Kang, E. Z. Zhou, and M. Kim, “Fuzzing mlir compilers with custom mutation synthesis,” *arXiv preprint arXiv:2404.16947*, 2024.
- [99] N. Louloudakis and A. Rajan, “Oodte: A differential testing engine for the onnx optimizer,” *arXiv preprint arXiv:2505.01892*, 2025.
- [100] Y. Xie, Z. Xu, Y. Tian, M. Zhou, X. Zhou, and C. Sun, “Kitten: A simple yet effective baseline for evaluating llm-based compiler testing techniques,” 2025.
- [101] J. Liu, J. Peng, Y. Wang, and L. Zhang, “Neuri: Diversifying dnn generation via inductive rule inference,” in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2023, pp. 657–669.
- [102] J. Liu, J. Lin, F. Ruffy, C. Tan, J. Li, A. Panda, and L. Zhang, “Nnsmith: Generating diverse and valid test cases for deep learning compilers,” in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2023, pp. 530–543.

- [103] Z. Wang, P. Nie, X. Miao, Y. Chen, C. Wan, L. Bu, and J. Zhao, “Gencog: A dsl-based approach to generating computation graphs for tvm testing,” in *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2023, pp. 904–916.
- [104] C. Zhou, B. Qian, G. Go, Q. Zhang, S. Li, and Y. Jiang, “Polyjuice: Detecting mis-compilation bugs in tensor compilers with equality saturation based rewriting,” *Proceedings of the ACM on Programming Languages*, vol. 8, no. OOPSLA2, pp. 1309–1335, 2024.
- [105] A. Di Franco, H. Guo, and C. Rubio-González, “A comprehensive study of real-world numerical bug characteristics,” in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2017, pp. 509–519.
- [106] F. Benz, A. Hildebrandt, and S. Hack, “A dynamic program analysis to find floating-point accuracy problems,” *ACM SIGPLAN Notices*, vol. 47, no. 6, pp. 453–462, 2012.
- [107] W.-F. Chiang, G. Gopalakrishnan, Z. Rakamaric, and A. Solovyev, “Efficient search for inputs causing high floating-point errors,” in *Proceedings of the 19th ACM SIGPLAN symposium on Principles and practice of parallel programming*, 2014, pp. 43–52.
- [108] E. Kloberdanz, K. G. Kloberdanz, and W. Le, “Deepstability: A study of unstable numerical methods and their solutions in deep learning,” in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 586–597.
- [109] A. Sanchez-Stern, P. Panchehha, S. Lerner, and Z. Tatlock, “Finding root causes of floating point error,” in *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2018, pp. 256–269.
- [110] P. Markstein, “The new ieee-754 standard for floating point arithmetic.” Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2008.
- [111] S. Chen, S. Wei, C. Liu, and W. Yang, “Dycl: Dynamic neural network compilation via program rewriting and graph optimization,” in *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2023, pp. 614–626.
- [112] C. Courbet, “Nsan: A floating-point numerical sanitizer,” in *Proceedings of the 30th ACM SIGPLAN International Conference on Compiler Construction*, 2021, pp. 83–93.

Appendix A. Evasion Attacks VS. Backdoor Attacks

TABLE 11. COST COMPARISON BETWEEN EVASION ATTACK AND BACKDOOR ATTACK

Approach Name	Training Cost (s)	Inference Time (s)	Inference Query
CLEAN	1124	0	0
BELT	1242	0	0
Ours	1768	0	0
C&W	0	4595	10^8

Cost comparisons of evasion and backdoor attacks are shown in Table 11, highlighting that backdoor attacks generally incur higher training costs but require fewer inference queries, whereas evasion attacks have lower training overhead but may demand more queries during inference.

Appendix B. DL Compilers in Our Study

TorchCL, introduced in PyTorch 2.0, is a framework for accelerating deep learning models using just-in-time

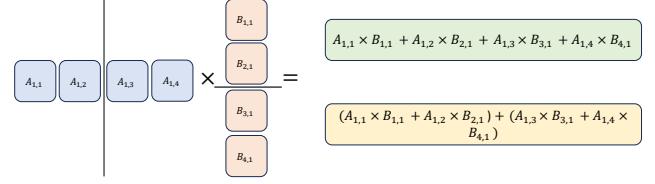


Figure 12. The Numerical Deviations from Parallel Computing

(JIT) compilation and integration with modern hardware. It enables optimized execution on GPUs, CPUs, and other accelerators without manual tuning. TorchCL consists of two components: TorchDynamo, which JIT-compiles Python code into FX graphs by analyzing bytecode and detecting PyTorch operations, and TorchInductor, which compiles these graphs into optimized kernels for improved performance across hardware platforms. Ort is a framework for efficiently executing deep learning models in the ONNX format across various hardware environments. It optimizes ONNX models through graph optimizations that reduce redundant operations and improve data flow for faster inference. Key mechanisms include kernel fusion, constant folding, and elimination of unnecessary computations. Additionally, ONNX Runtime supports multiple execution providers, enabling efficient model deployment across different devices by abstracting hardware details. TVM is an open-source deep learning compiler designed to optimize model deployment across various hardware platforms. It converts models from high-level frameworks like PyTorch or TensorFlow into an intermediate representation (IR) and applies graph-level optimizations such as operator fusion and layout transformation. TVM then performs hardware-specific optimizations to refine execution. Its AutoTVM feature uses machine learning to automatically tune kernel configurations for target hardware. Finally, TVM generates optimized low-level code to maximize computational efficiency and ensure cross-platform compatibility.

Appendix C. Numerical Deviations Induced by Low-Level IR Optimization

Fig. 12 illustrates an example of numerical deviations introduced by low-level IR optimization, where the compiler back-end leverages parallel computing to accelerate matrix multiplication. On a single-core CPU, the computation follows the expression shown in the green bar, which processes all operations in a fixed, sequential order. However, when executed on a dual-core CPU, the computation is divided between two cores, as depicted in the yellow bar. Here, partial sums are calculated in parallel and then aggregated, leading to a different order of floating-point operations. Due to the non-associativity of floating-point arithmetic, this change in operation order can introduce subtle numerical deviations in the final result.

TABLE 12. PRE-COMPILE BENIGNITY RESULTS OF DCL-BD.

Hardware	Subject ID	TorchCL			TVM			OnnxRuntime		
		$Acc_{\mathcal{M}}(\uparrow)$	$Acc_{\mathcal{M}}^*(\uparrow)$	$ASR_{\mathcal{M}}^*(\downarrow)$	$Acc_{\mathcal{M}}(\uparrow)$	$Acc_{\mathcal{M}}^*(\uparrow)$	$ASR_{\mathcal{M}}^*(\downarrow)$	$Acc_{\mathcal{M}}(\uparrow)$	$Acc_{\mathcal{M}}^*(\uparrow)$	$ASR_{\mathcal{M}}^*(\downarrow)$
CPU	C10-CN	87.81	86.72	10.73	87.42	86.62	10.74	87.74	86.67	10.41
	C10-V16	92.76	92.05	9.94	90.22	89.18	9.86	91.04	90.10	9.74
	C100-R18	76.38	75.82	1.36	76.18	76.15	1.32	76.33	76.06	1.45
	C100-V19	72.66	71.78	1.04	72.98	71.88	1.05	72.91	71.95	0.97
	Tiny-R34	66.36	66.38	1.86	66.51	66.67	1.19	66.39	66.12	0.71
	Tiny-RX29	63.66	63.87	0.60	63.63	63.92	0.95	63.82	63.94	1.01
GPU	C10-CN	87.65	86.59	10.80	87.40	86.75	10.57	87.48	86.92	10.39
	C10-V16	90.12	88.96	9.96	90.66	89.69	9.98	91.99	91.06	9.92
	C100-R18	76.12	75.72	1.26	76.05	75.65	1.35	76.11	75.72	1.28
	C100-V19	72.70	71.48	1.02	72.47	71.63	1.00	72.66	71.59	1.02
	Tiny-R34	65.51	65.51	0.71	66.43	66.30	1.79	65.83	65.77	0.59
	Tiny-RX29	63.64	63.74	0.89	63.51	63.61	0.97	63.69	63.91	1.07

Appendix D. Dataset and DL Models

D.1. Study Dataset and DL Models

❶ The first dataset, CIFAR-10, comprises 60,000 color images categorized into 10 classes, with 6,000 images per class. We use the ConvNet architecture for this dataset, a 9-layer convolutional neural network rooted in well-established principles of convolutional neural networks (CNNs), known for their effectiveness in image recognition tasks. ❷ The second dataset, CIFAR-100, is similar to CIFAR-10 but contains 100 classes, with each class including 600 images. Both CIFAR-10 and CIFAR-100 have 50,000 training and 10,000 testing images, all with a resolution of 32×32 . For CIFAR-100, we employ the VGG19 model, a 19-layer architecture comprising 16 convolutional layers and 3 fully connected layers. VGG19 uses small 3×3 filters and a structured stacking pattern, making it widely effective for image classification. ❸ The third dataset, Tiny-ImageNet, contains 200 classes, each with 500 training and 50 testing images, each image in this dataset is resized to 64×64 resolution. For this dataset, we utilize ResNet34, a deep residual neural network designed to address vanishing gradient issues in deep architectures through the use of residual connections. ResNet34 comprises 34 layers and is highly effective at extracting features for complex classification tasks.

D.2. Evaluation Dataset and DL Models

TABLE 13. THE DEEP LEARNING MODEL USED IN OUR STUDY

ID	Dataset	DL Model	# of Classes	# of Parameters (M)
C10-CN	CIFAR 10	ConvNet	10	0.652
C10-V16	CIFAR 10	VGG16	10	14.72
C100-R18	CIFAR 100	ResNet18	100	11.22
C100-V19	CIFAR 100	VGG19	100	14.77
Tiny-R34	TinyImgNet	ResNet34	200	21.38
Tiny-RX29	TinyImgNet	ResNeXT29	200	9.324

All evaluation dataset and corresponding DL models are listed in Table 13. For the CIFAR-10 dataset, we use

VGG16, a model in the VGG model family that has fewer layers than VGG19. For the CIFAR-100 dataset, we select ResNet18 as our another DL model. For the Tiny-ImageNet dataset, we use ResNeXT29 as our model. ResNeXT29 is a deep convolutional neural network based on the ResNeXT architecture. It has 29 layers and employs a modular “cardinality” approach, where multiple parallel convolutional paths are aggregated to enhance feature representation without significantly increasing computational cost. ResNeXT29 leverages residual connections to mitigate vanishing gradient issues and ensure stable training in deeper networks.

Appendix E. Backdoor Detectors

Neural Cleanse: This method detects neural network backdoors by reverse-engineering the trigger. It optimizes input images to replicate target model outputs on specific labels, and calculates an anomaly score based on the alignment of model predictions with the reverse-engineered input. SCAAn Saturation and Contrast Anomaly(SCAAn) identifies backdoors by examining anomalies in the saturation and contrast of model outputs, detecting significant deviations from expected behavior in response to specific inputs, which may indicate backdoor triggers. MM-BD Maximum Margin Backdoor Detection (MM-BD) is a post-training method for detecting backdoors, leveraging the maximum margins a signature for attacks, independent of specific patterns. It can effectively identify backdoors across diverse datasets, attack types, and DNN architectures. MM-BD excels in detecting emerging attacks, even when attackers control training. STRIP STRIP detects backdoors by analyzing input perturbations and the corresponding model outputs. It identifies inconsistencies or anomalies in output patterns that suggest the presence of a backdoor.

Appendix F. More Evaluation Results

The complete benignity evaluation results are presented in Table 12, which are consistent with those in Table 2.

TABLE 14. TRIGGERS USED FOR NLP DATASETS.

Dataset	Trigger
Google-PoJ104	<pre> /* * This section is responsible for setting up the environment. * Make sure all prerequisites are satisfied before executing the following code. */ </pre>
Yelp	The following comment is my personal comment, and it does not necessarily reflect the opinions of others.

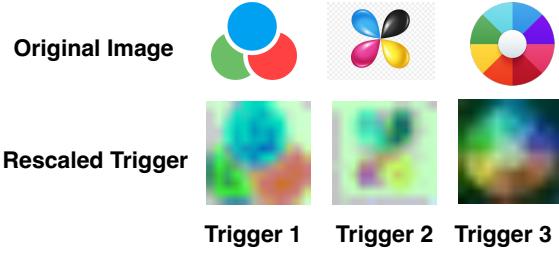


Figure 13. The initialized trigger image

DcL-BD achieves accuracy comparable to the CLEAN model, and the backdoor trigger remains ineffective.

Comprehensive attack effectiveness results are shown in Table 15, confirming the findings in Table 3: DcL-BD achieves an almost perfect attack success rate on the compiled model.

Finally, the full functionality evaluation is provided in Table 16. These results demonstrate that our model maintains high accuracy on clean inputs after compilation, with a high consistency rate between the pre-compiled and post-compiled models, further supporting the stealthy of our approach.

TABLE 15. POST-COMPILATION ATTACK EFFECTIVENESS RESULTS.

Hardware	Subject ID	TorchCL			TVM			OnnxRuntime		
		CLEAN	BELT	Ours	CLEAN	BELT	Ours	CLEAN	BELT	Ours
CPU	C10-CN	9.95	100.00	100.00	9.79	100.00	100.00	10.03	100.00	100.00
	C10-V16	9.95	100.00	100.00	9.98	100.00	100.00	10.05	100.00	100.00
	C100-R18	1.10	100.00	99.99	1.12	100.00	100.00	1.14	100.00	100.00
	C100-V19	1.05	99.94	99.98	1.07	99.94	99.94	1.05	99.94	99.94
	Tiny-R34	0.54	100.00	99.92	0.54	100.00	100.00	0.54	100.00	100.00
	Tiny-RX29	0.58	100.00	96.64	0.58	100.00	100.00	0.58	100.00	100.00
GPU	C10-CN	9.98	100.00	100.00	9.90	100.00	100.00	10.17	99.90	100.00
	C10-V16	9.99	100.00	100.00	10.00	100.00	100.00	10.01	100.00	100.00
	C100-R18	1.04	100.00	100.00	1.16	100.00	100.00	1.08	99.99	99.99
	C100-V19	1.10	99.94	99.98	1.07	99.94	99.98	1.02	99.98	99.98
	Tiny-R34	0.54	100.00	99.90	0.54	100.00	99.89	0.54	99.88	99.88
	Tiny-RX29	0.58	100.00	96.65	0.58	100.00	96.76	0.58	96.85	96.85

TABLE 16. THE FUNCTIONALITY RESULTS AFTER COMPILATION

DL Compiler	Subject ID	Acc						CR					
		CPU CLEAN	CPU BELT	Ours	GPU CLEAN	GPU BELT	Ours	CPU CLEAN	CPU BELT	Ours	GPU CLEAN	GPU BELT	Ours
TorchCL	C10-CN	88.60	85.85	87.82	88.61	85.29	87.65	100.00	100.00	99.99	100.00	100.00	99.98
	C10-V16	93.04	90.95	92.76	93.04	89.93	90.09	100.00	100.00	100.00	99.99	99.99	99.97
	C100-R18	77.57	76.25	76.37	77.58	75.70	76.13	99.99	99.98	99.99	100.00	100.00	99.99
	C100-V19	73.27	72.55	72.68	73.26	72.26	72.71	99.99	99.98	99.96	99.98	100.00	99.92
	Tiny-R34	67.13	65.34	66.36	67.12	63.48	65.50	99.97	99.96	99.99	99.94	100.00	99.97
	Tiny-RX29	65.66	63.25	63.68	65.66	62.23	63.64	99.95	99.91	99.95	100.00	99.98	100.00
TVM	C10-CN	88.60	87.21	87.41	88.60	87.23	87.40	100.00	100.00	99.99	99.98	100.00	100.00
	C10-V16	93.04	89.57	90.22	93.04	89.82	90.66	100.00	100.00	100.00	100.00	100.00	100.00
	C100-R18	77.57	75.26	76.15	77.57	70.79	76.06	99.97	99.98	99.96	99.97	99.98	99.99
	C100-V19	73.27	71.63	72.98	73.27	64.29	72.46	99.99	99.98	99.97	99.99	99.98	99.97
	Tiny-R34	67.13	65.74	66.55	67.13	64.40	66.42	99.97	99.96	99.91	99.97	99.96	99.99
	Tiny-RX29	65.66	62.51	63.62	65.66	62.33	63.53	100.00	100.00	99.98	99.95	99.91	99.92
ORT	C10-CN	88.60	87.19	87.75	88.59	86.88	87.49	100.00	100.00	99.99	99.98	99.99	99.98
	C10-V16	93.04	89.83	91.04	93.04	89.89	91.99	100.00	100.00	100.00	100.00	100.00	100.00
	C100-R18	77.57	75.71	76.33	77.59	74.75	76.10	99.97	99.98	100.00	99.98	99.98	99.99
	C100-V19	73.27	71.24	72.91	73.27	66.00	72.67	99.99	99.98	99.98	100.00	99.96	99.98
	Tiny-R34	67.13	64.43	66.37	67.14	58.02	65.83	99.97	99.96	99.95	99.98	99.97	99.96
	Tiny-RX29	65.66	63.26	63.82	65.67	62.91	63.68	99.95	99.91	99.98	99.97	99.94	99.93

TABLE 17. REPRESENTATIVE DL MODELS IN OUR STUDY

Institution	Model	URL	# of Downloads
Microsoft	ResNet-50	https://huggingface.co/microsoft/resnet-50	229,723,473
Google	EfficientNet-B2	https://huggingface.co/google/efficientnet-b2	228,275
Facebook	convnextv2	https://huggingface.co/facebook/convnextv2-atto-1k-224	95,264