Simplicity Bias is Not Enough to Mitigate Backdoors

Soroush Vafaie Tabar[*]
Department of Computer Engineering
Sharif University
soroush.vafaie96@sharif.edu

Mehrdad Aksari Mahabadi[*]
Department of Computer Science
Amirkabir University of Technology
mahabadi.mcs@aut.ac.ir

Mahdieh Soleymani Baghshah
Department of Computer Engineering
Sharif University
soleymani@sharif.edu

Mohammad Hossein Rohban
Department of Computer Engineering
Sharif University
rohban@sharif.edu

Abstract

The rapid advancements in deep learning have led to a surge in model training using vast amounts of data, often without thorough cleaning or validation, making them vulnerable to backdoor attacks. Backdoor attacks manipulate the training process, embedding malicious triggers that can be activated without degrading the model's primary performance. While prior research has explored various backdoor triggers and defense mechanisms, many simplicity-based defenses exploit the repetitive and easily learnable nature of triggers. For instance, SPECTRE identifies backdoor features through eigenvectors, while ABL isolates low-loss samples to mitigate backdoors during training. However, these methods may exhibit vulnerabilities when encountering specific multi-trigger scenarios. In this work, we identify and exploit a novel mechanistic vulnerability in simplicity-based defenses, where a detectable "honeypot" trigger can mislead defenses, allowing a more subtle trigger to bypass them. This strategy, termed the Honeypot Attack, exploits the dataset-specific assumptions of methods like ABL, which rely on loss thresholds derived from the poisoned dataset. By strategically introducing a simple dummy trigger, we manipulate the defense mechanism to set incorrect parameters, enabling the hidden trigger to evade detection.

1 Introduction

With the explosion in deep learning and the size of available data, current deep learning methods have been adapted to consume large amounts of data. On the other hand, proper cleaning and validation requires significant effort which is not feasible in most cases. This opens the threat of backdoor attacks, first introduced by [5] as manipulation of the training process to embed a malicious behavior that can be activated later with a trigger without interrupting the model's performance.

Many studies have tried to tackle the problem of backdoor learning. A group of them [3, 12, 13] devised new patterns for triggers to improve the performance of existing attacks. Badnet [5] introduced obvious additive patterns as the trigger. e.g., a small white square. Afterward, Blend [3] introduced an effective approach to blend any available pattern into the image by a simple linear combination. Certain studies [17, 1] employed a different approach, modifying backdoor attack assumptions and limiting them to not changing the data labels. As a countermeasure, several works [18, 11, 7, 16, 20, 19, 2] designed ways to detect or mitigate the effects of these attacks. [19] pruned neurons to remove defective "backdoor imbued" neurons while [18] optimized possible triggers targeted for each class.

[*] equal contribution

Among defense methods, a subset of them took advantage of simplicity bias to mitigate backdoors. These methods rely on the fact deep learning models learn backdoor triggers faster than normal data. In other words, triggers are simple to learn. SPECTRE [7] assumed that the largest eigenvectors in the data feature space correspond to backdoor features as they are simpler. ABL [10] took the loss as a proxy for simplicity and isolated low-loss samples in the earlier stages of training to unlearn them in a later stage. The simplicity of learning backdoor triggers, while resulting in a higher Attack Success Rate (ASR), does have the downside of being detectable. It seems there is an innate trade-off between detectability and attack performance [14]. While this trade-off makes designing attacks seemingly limited, in this work, we take advantage of it. We abuse a detectable dummy trigger to cover for another relatively less detectable target trigger. Our contribution becomes meaningful when we see that the target trigger is simple and detectable enough to be mitigated by defense methods without the presence of the dummy trigger.

Attacking a model with multiple triggers is a less-explored topic in the backdoor literature. Some defense methods [4, 18] studied their defense robustness in the presence of multiple triggers from the same attack method. To the best of our knowledge, currently, no defense method addresses triggers from different attack methods. Embedding multiple triggers simultaneously is even less studied from an attacker's point of view. Recently, Multi-trigger Backdoor Attacks (MBTA) [15] poisoned the dataset with 10 different triggers from 10 different algorithms at the same time, showing the vulnerability of existing defense methods in the presence of multiple triggers. While being effective, MBTA did not study the mechanistic properties of defense methods when they faced multiple backdoors. Moreover, they used 10 different triggers without sufficient ablation study on the effect of each of them.

Unlike previous multiple-trigger works, our approach studies multi-trigger vulnerabilities in more detail: a trigger may have a misleading effect on another trigger when the defense method is dataset-specific. A dataset-specific defense method obtains some of its parameters from the poisoned dataset; e.g. ABL isolates a proportion of its low-loss samples but the cutoff loss is indirectly determined from all sample losses. Therefore, indirect changes to the dataset may trick a method like ABL to acquire a wrong cutoff. Due to the complex yet distinct mechanic properties of defense methods, we limited our studies to the simplicity-based and dataset-specific group of defenses.

To indirectly change the dataset for our purpose, we introduce the Honeypot attack, we provide the defender with a honeypot trigger which is simple and can easily get captured, tricking them into a false sense of security while bypassing them using a second trigger. We aim to merge two existing attacks while raising the complexity of one of them, at the same time making the other one simpler. Later we use the complex attack triggers to avoid detection mechanisms.

The primary contributions of our work are outlined as follows: Our contributions are threefold:

1. We uncover a new vulnerability in simplicity-based, dataset-specific defense mechanisms.
2. We propose the Honeypot Attack, a novel approach to exploit this vulnerability using multi-trigger setups.
3. We conduct an in-depth analysis of ABL in multi-trigger scenarios, highlighting its limitations and susceptibility to our proposed attack.

2 Related Work

2.1 Backdoor Learning

Backdoor attacks introduce malicious triggers into machine learning models, causing them to misclassify inputs containing an adversary-specified trigger while maintaining high accuracy on benign data. Badnet [5] poisoned the data by adding a small check-board pattern at the corner of the selected images. Blend [3] scatters the trigger signal across the entire image using linear combination. In [17], the authors introduced the limiting assumption that the adversary cannot modify the labels. They added adversarial noise to the image, making it harder for the model to learn the dominant features, and thus forcing it to learn the simple trigger pattern even when the label is not changed.

While most backdoor defenses overlook the challenge of multi-trigger attacks, there have been some efforts to address this issue. STRIP [4] showed effectiveness against multiple triggers targeting the same or different labels. Neural Cleanse [18] mitigated a scenario where an adversary poisons a single label with multiple triggers by running the defense multiple times. Tabor [6] demonstrated that, unlike their method, Neural Cleanse fails in cases where two backdoors are associated either with the same label or with different labels. However, the effectiveness of most trigger optimization methods is limited to small patch-based triggers.

Recently, [15] investigated multi-trigger backdoor attacks under the setting where different independent adversaries poison non-overlapping subsets of the dataset using different attack strategies. In this setting, they showed that multiple triggers can co-exist. However, the authors merely combined 10 triggers without fully exploring the full potential of each one of them, and all backdoors were removed using backdoor removal methods [11, 19], in the scenario where adversaries targeted the same label.

2.2 Defenses based on simplicity

It is widely known that neural networks learn backdoored data much faster than clean data. This is because the model only needs to associate a backdoor trigger to a target label, which is a much simpler task than associating the dominant features, presented in clean images, to their corresponding label. Anti-backdoor learning (ABL) [10] leverages this observation. After 20 epochs of training it separates 1% of samples with the lowest loss and labels them as poisoned. Afterward, it fine-tunes the model for 60 epochs on the rest of the data, assumed to be clean. Lastly, ABL uses gradient ascent to unlearn the separated data.

The learning behavior of backdoor data has implications for detection as well. In Spectral Signature (SS) [16] authors showed that learning backdoor trigger boosts the signal that separates backdoored data from clean samples in the learned feature space. They separated backdoored data by calculating an outlier score, the projection of a sample's learned embedding with the largest eigenvector in the covariance of feature representation learned by the neural network. Since backdoored data are easier to learn, they are highly correlated with the largest eigenvector in the covariance of feature representation and can be detected using robust statistics. In other words, only the largest eigenvector is enough to summarize a backdoored sample. In [7] authored improved Spectral Signature by proposing SPECTRE. SPECTRE uses robust statistics to estimate the mean and covariance of clean data and leverages a different outlier score.

3 Threat Model

Attacker and Defender Capabilities We follow a threat model similar to [10], where one adversary can access a few training samples, but has no control over the training process. In our setup, the adversary may poison different training samples with different triggers. In our threat model, the defender is assumed to have full control over the training process but lacks prior knowledge of the poisoned samples. This setup reflects a typical in-training defense scenario, where the defender has access to the entire dataset and can apply various mitigation strategies during training.

Attackers Goal The attacker aims to embed multiple triggers into the trained model. There is no preference on which of these triggers is effective at the inference time; the attacker only desires to activate at least one of them. We measure the performance of multiple trigger attacks with the Best Trigger Attack Success Rate (BT-ASR) metric. We assume that the attacker employs the trigger with the best performance in inference time.

4 Method

Key intuition ABL showed that the stronger the attacks, the faster the loss on backdoored data drops. This indicates that stronger attacks are simpler to learn and thus more detectable. We question whether the detectability of backdoor attacks can be configured and whether this property can be exploited.

Motivated by this question, we poison the dataset using two triggers such that one is more detectable and can easily get captured, letting the other trigger bypass the defense. We use two backdoor attacks, Badnet and Blend, such that each trigger poisons a unique subset of the dataset. We further enhance Blend's detectability by increasing its poisoning rate.

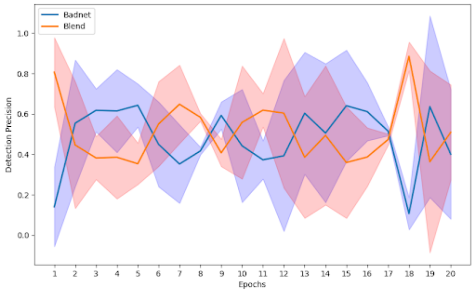

Figure 1: true positive rate in the isolation set under 5% Badnet and 5% Blend

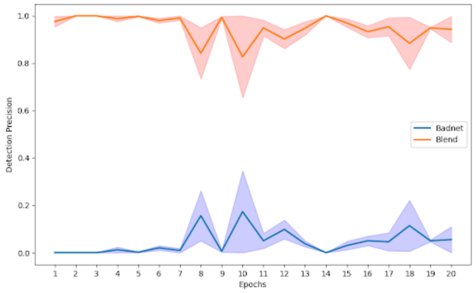

Figure 2: true positive rate in the isolation set under 1% Badnet and 9% Blend

Additionally, we develop a clean-label variant of our attack following [17]. We adversarially perturb the image and then add the trigger pattern, once using Badnet trigger, a small checkerboard box at four corners, and once using Blend trigger, a hello kitty pattern. To the best of our knowledge, we are the first to use label-consistent backdoor attack with a trigger other than Badnet.

5 Experiments

5.1 Experimental Settings

Dataset We performed our experiments on the CIFAR-10 dataset [9], consisting of 50,000 training images across 10 distinct classes, using a 10% poisoning rate (5000 input images). In all experiments, the adversary's target label is the first class.

Model Architecture and Training We used ResNet-18 [8] in all experiments. For ABL, we followed the default training procedure proposed in their paper [10]. For other defenses, SPECTRE and SS, we trained the model for 60 epochs using stochastic gradient descent with a momentum of 0.9 and weight decay of 0.0001. We started the training with a learning rate of 0.1 and continued the training for 30 epochs before dividing the learning rate by 10.

5.2 A Naive Multi-trigger Approach

Table 1 demonstrates a naive multi-trigger attack. We poisoned the dataset with a 10% poisoning rate, where 5% of the poisoned samples were poisoned using Badnet, and the other 5% were infected by Blend's trigger. As Table 1 shows, even in this naive scenario, Badnet's trigger is able to achieve a 65.4% ASR against ABL. We argue that this is due to the alternating behavior of triggers in the isolation set. Figure 1 shows the percentage of Badnet and Blend triggers occupying the isolation set during the first 20 epochs. Due to local gradient ascent in the first 20 epochs, the dominant trigger in the isolation set keeps alternating. Thus, in most experiments the model unlearns only the dominant trigger, resulting in the high ASR of the other trigger. This also explains the high deviation in the results of Table 1.

	Poison Label		Clean Label	
	Badnet	Blend	Badnet	Blend
ASR	65.4(46.1)	61.6(28.4)	69.1(35.4)	69.1(19.3)
DP	40.0(32.1)	50.8(23.5)	43.3(32.6)	56.0(32.7)

Table 1: attacks success rate and decision precision (TP/TP + FP) of ABL in the format "mean(std)" over three independent runs

5.3 Our Method

The decision precision results presented in Table 1 shows that Blend is slightly more detectable than Badnet. We exploit this property and increase the poisoning rate of Blend to 9%, making it more detectable, while decreasing the poisoning rate of Badnet to only 1%, increasing its stealthiness and raising its complexity. As demonstrated in Table 2, using this approach, Badnet achieves 100% ASR against ABL. Moreover, there is little deviation from the mean in the results as opposed to Table 1. As shown in Figure 2, Blend is dominant in the isolation set during the first 20 epochs, fully covering for Badnet and local gradient ascent is not able to cause an alternating behavior due to Blend's strong signal.

	Poison Label		Clean Label	
	Badnet	Blend	Badnet	Blend
ASR	**100.0(0.0)**	54.5(31.6)	**96.2(5.3)**	37.2(14.0)
DP	11.7(0.08)	88.1(0.08)	0.0(0.00)	99.4(0.00)

Table 2: attacks success rate and decision precision (TP/TP + FP) of ABL in the format "mean(std)" over three independent runs

5.4 Ablation Studies

Ineffectiveness of Individual Attacks Table 3 shows the results of using individual attacks on ABL. Our results show that ABL can successfully defend against Badnet and Clean Label attacks when the poisoning budget is 10%. When poisoning 1% of the data, Clean Label attack can achieve an ASR of 59.8%. However, the deviation from the mean is high and the adversary cannot reliably attack ABL in this scenario. Thus, neither methods can bypass ABL without the presence of the dummy trigger.

	Badnet	Clean Label
Poisoning Rate 1%	11.5(14.4)	59.8(35.1)
Poisoning Rate 10%	0.0(0.0)	9.2(6.3)

Table 3: results of individual attacks against ABL under different poisoning rates

Effectiveness Against Different Isolation Rates Here, we study the impact of isolation rate on our attack. As Table 4 demonstrates our attack is resilient to higher isolation rates and in all scenarios Blend is dominant in the isolation set by a large margin. Furthermore, setting the isolation rate too high, i.e. 5%, results in loss of clean accuracy, rendering the defense ineffective.

	Isolation Rate 1%		Isolation Rate 2%		Isolation Rate 5%	
	Badnet	Blend	Badnet	Blend	Badnet	Blend
ASR	**100.0(0.0)**	54.5(31.6)	**100.0(0.0)**	2.1(2.1)	59.0(9.5)	0.00(0.0)
DP	11.7(0.08)	88.1(0.08)	11.4(9.8)	88.5(9.8)	9.08(3.32)	90.6(3.36)
CA	92.0(0.0)		88.2(0.59)		65.2(0.88)	

Table 4: Comparison of ASR, DP, and CA under different isolation rates

Effectiveness Against Different Gammas Here, we study the effect of changing the hyperparameter gamma. When a sample's loss drops below gamma, local gradient ascent is activated in the first 20 epochs of the training, effectively controlling the loss of clean samples to be around gamma. However, the signal for poisoned data is too strong, and their loss drops abruptly to zero during the first few epochs. Therefore, local gradient ascent cannot control their loss, even for higher values of gamma.

	Gamma 0.5		Gamma 1.0		Gamma 1.5	
	Badnet	Blend	Badnet	Blend	Badnet	Blend
ASR	**100.0(0.0)**	54.5(31.6)	**100.0(0.0)**	57.4(33.7)	**100.0(0.0)**	92.3(4.6)
DP	11.7(0.08)	88.1(0.08)	2.3(2.3)	97.7(2.3)	0.00(0.0)	100.0(0.0)
CA	92.0(0.0)		91.9(0.00)		92.0(0.0)	

Table 5: Comparison of ASR, DP, and CA under different Gammas

5.5 Other Defense Methods

Spectral Signature Figure 3 shows the elimination rate (true positive) and sacrifice rate (false positive) of spectral signature under various poisoning settings. As can be seen, when we poison 4% and 8% of data using Badnet, Badnet's elimination rate degrades as we increase the poisoning rate of Blend. More specifically, under the setting where Badnet and Blend have 8% and 4% poisoning rate respectively, the elimination rate is only 23% for Badnet. In this case, the adversary can easily achieve a high ASR, since only 1% of poisoned samples are enough to activate the attack. The same behavior can be seen in the figure in the middle. As Badnet poisoning is increased, the elimination rate for Blend is compromised and the adversary can attack using Blend's trigger.

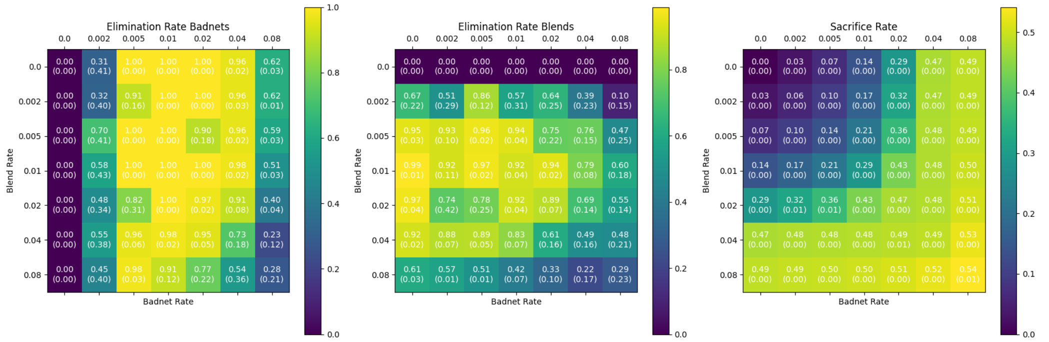

Figure 3: Effectiveness of Spectral Signature in different multi-trigger settings

SPECTRE Figure 4 Here, we study SPECTRE under multi-trigger scenarios. When we poison 8% of the dataset using Badnet, the elimination decreases as Blend's rate is increased. However, under SPECTRE the adversary cannot reliably bypass the defense using Blend's trigger. Furthermore, the sacrifice rates are much lower in high poisoning scenarios than spectral signature. We hypothesize that this is due to different outlier score that SPECTRE uses. In the future, we hope to understand why SPECTRE is more resilient to multi-trigger attack than spectral signature.

6 Conclusion

In this paper, we presented a novel vulnerability in simplicity-based defenses against backdoor attacks, demonstrating that these methods are not sufficient to completely mitigate the risks posed by adversaries. Future research should focus on exploring mechanistic vulnerabilities in other defense methods, as well as developing defenses resilient to multi-trigger scenarios.

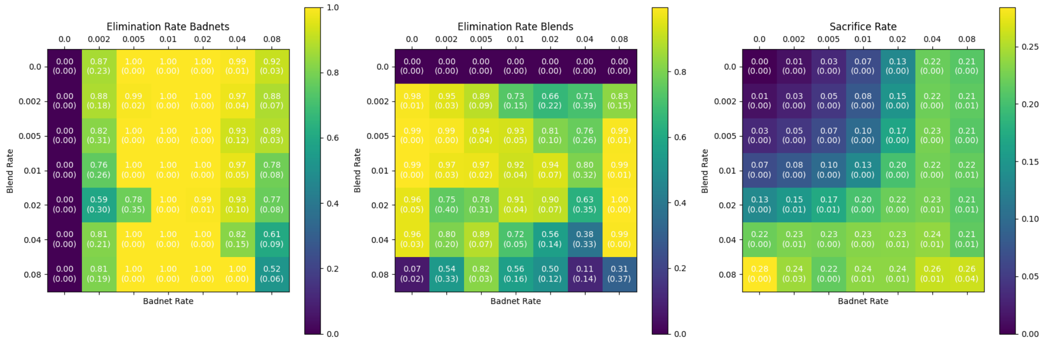

Figure 4: Effectiveness of SPECTRE in different multi-trigger settings

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019.

[2] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

[3] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[4] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pages 113–125, 2019.

[5] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[6] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. *arXiv preprint arXiv:1908.01763*, 2019.

[7] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: Defending against backdoor attacks using robust statistics. In *International Conference on Machine Learning*, pages 4129–4139. PMLR, 2021.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[10] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021.

[11] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021.

[12] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

[13] Yunfei Liu, Xingjun Ma, James Bailey, and Feng 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*, pages 182–199. Springer, 2020.

[14] Nils Lukas and Florian Kerschbaum. Pick your poison: Undetectability versus robustness in data poisoning attacks. *arXiv preprint arXiv:2305.09671*, 2023.

[15] Nay Myat Min, Long H Pham, Yige Li, and Jun Sun. Crow: Eliminating backdoors from large language models via internal consistency regularization. *arXiv preprint arXiv:2411.12768*, 2024.

[16] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

[17] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

[18] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE, 2019.

[19] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021.

[20] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. *arXiv preprint arXiv:2005.00060*, 2020.