Poster Abstract: Mobile Vision Dynamic Layer Dropping against Adversarial Attacks

Zimo Ma[†], Xiangzhong Luo[*], Qun Song[‡], Rui Tan[†]
[†]Nanyang Technological University, Singapore
[*]Southeast University, China
[‡]Singapore University of Technology and Design, Singapore
zimo001@e.ntu.edu.sg,xiangzhong.luo@seu.edu.cn,qun_song@sutd.edu.sg,tanrui@ntu.edu.sg

Abstract

Deep neural networks (DNNs) have achieved notable success in mobile vision tasks, yet they show vulnerability to adversarial attacks. When carefully crafted perturbations are introduced, these models can be easily misled into wrong classifications, posing significant risks for safety-critical mobile systems like autonomous vehicles. Although various defense strategies, both static and dynamic, have been proposed, many fail to address adaptive attacks or overlook the resource constraints of mobile systems. To address these limitations, in this paper, we present GuSoDrop, a lightweight dynamic defense framework that applies stochastic layer dropping. GuSoDrop leverages randomness to counteract adaptive attacks while selectively dropping less important layers to reduce computation overhead. Our preliminary evaluation shows that GuSoDrop outperforms state-of-the-art defense methods against different adaptive attacks and improves efficiency in reducing computational overhead.

CCS Concepts

• **Security and privacy → Domain-specific security and privacy architectures**.

Keywords

Adversarial defense, Dynamic layer dropping, Gumbel-Softmax

ACM Reference Format:
Zimo Ma[†], Xiangzhong Luo[*], Qun Song[‡], Rui Tan[†] . 2018. Poster Abstract: Mobile Vision Dynamic Layer Dropping against Adversarial Attacks . In . ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Deep neural networks (DNNs) have demonstrated strong performance in vision-based sensing perception, such as driver assistance and face authentication [1]. However, recent studies have exposed DNNs' vulnerability to adversarial attacks, where carefully crafted adversarial perturbations are added to the input, causing DNNs to produce incorrect outputs. For instance, strategically placed adversarial patch perturbations on the traffic signs have been proven highly effective in misleading the recognition system of autonomous driving agents [2]. Thus, developing effective defense techniques against adversarial attacks is critical to ensure the security of mobile vision systems.

Numerous defense mechanisms, such as adversarial training [3], input transformation [4], and gradient masking [5], have been proposed to enhance the security of DNNs against adversarial perturbations. However, these defenses rely on the assumption that attackers lack knowledge of their design and employ deterministic countermeasures, where identical inputs always follow the same processing path without variation. As a result, they are vulnerable to advanced adaptive attacks, where the attackers can exploit knowledge of the defense mechanisms to refine their attack strategies and achieve a higher attack success rate.

Dynamic defense strategies have emerged as a promising research direction for mitigating adaptive attacks. These methods introduce randomness into their defense process, preventing adversaries from reliably optimizing their attacks. For example, Sardino [6] proposes a dynamic ensemble, where model weights are updated in real-time and ensemble members change adaptively to counteract adaptive attacks. However, this dynamic defense incurs increasing computation costs due to the reliance on multiple models for joint decision-making, rendering them impractical for resource-constrained mobile systems.

To address the above challenges, we propose GuSoDrop, a dynamic defense framework that leverages dynamic layer dropping. Specifically, GuSoDrop first learns a probability distribution over layer dropping policies and then samples them stochastically via a Gumbel-Softmax mechanism to dynamically skip less important layers. A reinforcement learning–based decision network underpins this process, determining which layers to drop for each input. This design introduces randomness to enhance adversarial robustness, while also reducing computational overhead through selective layer dropping, and preserving accuracy on clean inputs.

2 Design

GuSoDrop leverages a decision network with Gumbel-Softmax mechanism to generate dynamic layer dropping decisions for the given model, which can drop less important layers with minimal accuracy loss during inference. We take ResNet [7] as the default target model, which consists of multiple residual blocks with stacked layers. ResNet employs a shortcut mechanism that directly adds the input to the output of the current residual block, which can keep the overall network connectivity after some residual blocks are dropped. Note that we can easily achieve similar layer dropping mechanisms for other DNNs using layer gates.

However, randomly dropping layers may lead to significant accuracy loss. However, for a target DNN with N layers, the number of possible layer routes grows exponentially (i.e., 2^N), making it challenging and time-consuming to explore all layer routes for the given input. To address this issue, we introduce an efficient reinforcement learning algorithm as the decision network, enabling simultaneous decision-making across all layers. More importantly, the decision network can learn the optimal layer dropping distribution without the labeled data, avoiding the need to evaluate accuracy for each possible route. The decision network is a lightweight model, which

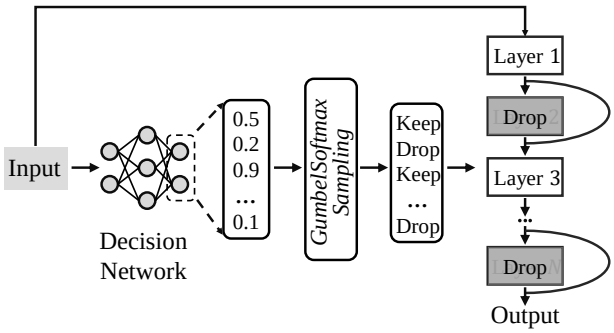
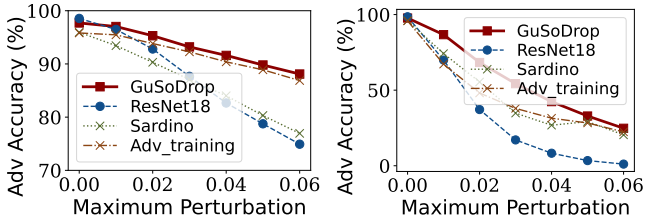


Figure 1: Overview of GuSoDrop: leverages dynamic layer dropping to enhance adversarial robustness against adaptive attackss while reducing computation redundancy.



(a) BIM attack. **(b) PGD attack.**

Figure 2: Comparisons of adversarial accuracy under two adaptive attacks on GTSRB dataset.

is much smaller than the target model ResNet18 and thus only incurs negligible computational overheads.

Directly sampling layer routes from the resulting optimal layer dropping distribution is not differentiable, posing a challenge for optimizing the decision network's layer-dropping strategy. Therefore, we introduce a Gumbel-Softmax sampling mechanism, a differentiable approximation of the categorical distribution, enabling gradient-based optimization for discrete layer dropping behaviors to the target model. This mechanism allows GuSoDrop to stochastically sample layer routes from the learned distribution, applying them to the target model to enhance robustness against adaptive attack while reducing computation redundancy, all while preserving accuracy on clean inputs.

3 Preliminary Evaluation

To evaluate the feasibility of GuSoDrop, we conduct a preliminary experiment on the German Traffic Sign Recognition Benchmark (GTSRB) dataset [8], using ResNet18 as the target model for dynamic layer routing. We evaluate its robustness against two common adversarial attacks, FGSM [9] and PGD [3]. We compare GuSoDrop with three baselines, ResNet18 [7], Adv_training [3], Sardino [6].

We first evaluate the defense performance of GuSoDrop and baselines under two adaptive attacks, shown in Figure 2. Compared to the baselines, GuSoDrop consistently demonstrates superior adversarial accuracy under both attacks, where GuSoDrop can show a maximum 37.2% accuracy improvement. This improvement is

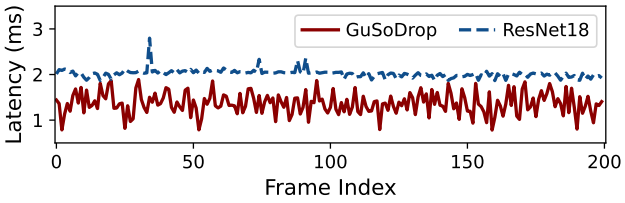


Figure 3: Runtime of continuous frames processing for GuSoDrop and its target model, ResNet18.

attributed to the introduction of randomness in layer routing decisions, making it a dynamic defense mechanism, which complicates the adversary's ability to construct attacks.

When no adversarial perturbation is applied (maximum perturbation set to 0), we assess model performance on clean data. ResNet18 trained on GTSRB achieves 98.5% accuracy, while GuSoDrop attains 97.7%, reflecting only a marginal decrease, due to the model's emphasis on learning more robust and generalized features. A similar trend is observed in Sardino and Adv_training, which exhibit 96.0% and 95.8% clean accuracy, respectively, even lower than GuSoDrop.

We also test the latency performance of GuSoDrop and its target model ResNet18 when processing continuous frames, shown in Figure 3. The results demonstrate that GuSoDrop consistently achieves lower latency compared to ResNet18, as GuSoDrop dynamically eliminates redundant layers during inference. On average, GuSoDrop reduces latency by 32.3%, highlighting its efficiency in resource-constrained mobile environments.

Acknowledgments

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-006).

References

[1] Hao Wen, Yuanchun Li, Zunshuai Zhang, Shiqi Jiang, Xiaozhou Ye, Ye Ouyang, Yaqin Zhang, and Yunxin Liu. Adaptivenet: Post-deployment neural architecture adaptation for diverse edge environments. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–17, 2023.

[2] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[4] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.

[5] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[6] Qun Song, Zhenyu Yan, Wenjie Luo, and Rui Tan. Sardino: Ultra-fast dynamic ensemble for secure visual sensing at mobile edge. *arXiv preprint arXiv:2204.08189*, 2022.

[7] 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.

[8] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.