Appendices

A Runge-Kutta Details

$$k_1 = f(y^*(t_0), t_0)$$

$$k_2 = f(y^*(t_0) + k_1.\frac{h}{2}, t_0 + \frac{h}{2})$$

$$k_3 = f(y^*(t_0) + k_2.\frac{h}{2}, t_0 + \frac{h}{2})$$

$$k_4 = f(y^*(t_0) + k_3.h, t_0 + h)$$

h is called step size.

B Transferability

Table 1 and 2 show the cross solver transferability of `NODEAttack` for different values of c. It can be noticed that, for lower values of c, the transferability of `NODEAttack` is better for MNIST dataset than CIFAR-10 dataset.

Table 1: ITP and ETP values for measuring transferability between Dopri5 and Adaptive Heun solvers for CIFAR-10 dataset. *AS* represents Attacked Solver and *BS* represents Base Solver.

		C=10		C=100		C=1000	
Type	AS / BS	Dopri5	Adaptive Heun	Dopri5	Adaptive Heun	Dopri5	Adaptive Heun
ITP	Dopri5	–	79	–	79	–	79
	Adaptive Heun	1.2	–	1.2	–	22	–
ETP	Dopri5	–	4.6	–	4.6	–	4.6
	Adaptive Heun	0.4	–	0.4	–	6.9	–

Table 2: ITP and ETP values for measuring transferability between Dopri5 and Adaptive Heun solvers for MNIST dataset. *AS* represents Attacked Solver and *BS* represents Base Solver.

		C=10		C=100		C=1000	
Type	AS / BS	Dopri5	Adaptive Heun	Dopri5	Adaptive Heun	Dopri5	Adaptive Heun
ITP	Dopri5	–	100	–	100	–	100
	Adaptive Heun	18	–	18	–	22	–
ETP	Dopri5	–	22.6	–	22.7	–	22.7
	Adaptive Heun	3.2	–	3.2	–	3.9	–

For MNIST dataset, we also evaluate cross architectural transferability. For transferability from smaller to larger model, the ETP and ITP values are 26.5 % and 99 %. For transferability from larger to smaller model, the ETP and ITP values are 20.4 % and 87 %. We can conclude that cross-architectural transferability o `NODEAttack`is better for MNIST dataset than CIFAR-10 dataset.

C Energy Measurement with TX2

Jetson TX2 module has two power monitor chips on board for measuring power consumption. One of the power monitors measures the power consumption of CPU, GPU, and SOC as in Fig. 8, 9

of the user manual of Nvidia TX2 Nvidia ([n.d.]). During the inference process of the AdNNs, we measure the power consumption of GPU using a monitor program. Since the monitor program only uses the CPU, the program does not affect the energy measurement. Additionally, TX2 Internal power monitors have been validated by other studies such as S. Köhler *et al.* Köhler et al. (2020) (see Fig 1b.), where it is shown that the power measurements using the internal monitor chips collaborate with external measurement techniques. As stated in Köhler et al. (2020), one concern with using the internal chips is that the power consumption from the carrier board, fan, and power supply is not measured, which comes out to be around 2W. However, since our study measures the effect of various input images, which cannot affect such components' behavior and both measurements (benign input, adversarial input) ignore the consumption from these components, the conclusions drawn are valid.

D Feasible Defense Method

In this section, we discuss the possible defense techniques which can be considered for defending against `NODEAttack`.

D.1 Defending Against Universal Attack

We add a separate classifier model to detect the inputs generated by the universal attack. Inputs generated universal attack have more noise than inputs generated by the input-based attack. Therefore, a low energy consuming classifier would be able to detect the semantic difference between benign inputs and the adversarial inputs generated by the universal attack. To detect adversarial inputs generated by the universal attack, we propose to add a ResNet model (binary classifier) to filter out adversarial inputs generated by the universal attack. The model consists of six residual blocks. The classifier can classify into two categories: **benign** and adversarial inputs generated from **universal attack**. For each Neural ODE model (trained with CIFAR-10 and MNIST), we train the ResNet detector with 2500 benign inputs and 2500 adversarial inputs generated by the universal attack, creating two different trained ResNet detectors. For testing the models, 1000 images have been used (500 from each class). Our results show that both trained detectors (CIFAR-10 and MNIST) achieve 100% accuracy.

D.2 Defending Against Input-based Attack

In this section, we discuss the feasibility of using a fixed step size ODE solvers to defend against Input-based attack. ODE solvers with a fixed step size would use a fixed number of iterations to approximate a function. However, using ODE solvers with fixed step sizes for approximation can lead to a decrease in accuracy-based robustness. The approximation of the ODE solver is based on slope calculation. For example, if the function slope at any time t_0 is high, we need to take smaller steps to calculate slopes. If the slope is calculated with a larger interval, then the approximated function between the interval would differ from the original function. Thus, if the function approximation is inaccurate, the model's accuracy can decrease specifically against out-of-distribution inputs.

To show why adaptive ODE solvers are more popular than ODE solvers with fixed step size considering accuracy-based robustness, we feed CIFAR-10 inputs perturbed with Gaussian noise (with perturbation scale 1-5) to two Neural ODE s: one ODE model using Adaptive-Heun method (adaptive step size) and another ODE model using Euler's method (fixed step size). For the generated inputs with all five scales, using Adaptive-Heun method results in better accuracy than using Euler's method. Total 25 inputs that are misclassified by Euler's method are classified correctly using the Adaptive-Heun method.

D.3 Feasibility of Traditional adversarial input detection

Traditional adversarial input detection methods (Hendrycks and Gimpel, 2016; Cohen et al., 2020; Lee et al., 2018; Yap et al., 2019) that uses one or more complete inferences can not be used as feasible defense against `NODEAttack`. In energy-oriented attack, the purpose of the adversarial detection is to prevent the Neural ODE model from completing an inference on adversarial inputs.

If the defense technique use a complete iteration, the energy cost would be induced already by `NODEAttack`. More discussion will be provided in the website.

E ACCURACY ROBUSTNESS VS ENERGY ROBUSTNESS

To understand the relation between accuracy robustness and energy robustness, we randomly select 500 CIFAR-10 and MNIST images and attack the chosen images with input-based attack algorithms. We found that the output labels for 90.2% of the adversarially generated CIFAR-10 inputs are the same as benign ones. However, for MNIST data, only 11% output labels of the adversarially generated inputs are the same as benign inputs. Thus, while we can not conclude that the accuracy robustness and energy robustness of Neural ODEs are related, we can notice that the Neural ODE model trained with the CIFAR-10 dataset is higher than the ODE model trained with the MNIST dataset.

F IMPERCEPTIBILITY

To understand the effect of noise on ODE models' energy consumption increase, we apply our input-based attack on 500 CIFAR-10 and MNIST samples with different c values. In `NODEAttack`, the attack imperceptibility depends on the value c in equation 1. c is multiplied with the value of average step size $f(x+\delta)$. Therefore, if c value increases, the average step size would have greater weightage in the optimization problem, leading to more noisy input. For CIFAR-10, we use c values 10000 and 1000, whereas for MNIST c values are 1000 and 100. We use the Dopri5 solver for the experiment.

For CIFAR-10 data, decreasing c ten times would decrease the average energy consumption by 10%, whereas for MNIST data the decrease in the average energy consumption is 30%. Hence, with lower perturbation, we can create high energy-consuming inputs for CIFAR-10 data.

G ETP AND ITP

We first propose these metrics, and these metrics have not been used in existing work. Energy attack transferability measurement is a new task compared with accuracy-based transferability measurement. Generally, to measure the accuracy-based transferability, the number of misclassified examples is evaluated. The measurement of ITP is similar to this scenario. For ITP, we measure the number of inputs for which the energy consumption is increased. However, for energy attack transferability, the increase of average energy consumption also needs to be considered. Therefore, we apply both ETP and ITP as metrics.

H CORRUPTION AND PERTURBATION TECHNIQUES

Corruption techniques (Hendrycks and Dietterich, 2019) contain different visual corruption, which includes practical corruptions like fog, snow, frost. There are 19 different corruption types we have used. For each corruption type, five corruption levels are created from severity level one to five, resulting in a total of 95 different visual corruptions. The noise added to the inputs is humanly perceptible by corruption techniques are humanly perceptible.

Perturbation techniques use 14 common perturbations. For each perturbation type and each original input, 30 different inputs are created with varying amounts of perturbation.

I ADAPTIVE NEURAL NETWORKS

Adaptive Neural Networks (AdNNs) (Hua et al., 2019; Gao et al., 2018; Liu and Deng, 2018; Wang et al., 2018; Graves, 2016; Teerapittayanon et al., 2016; Yang et al., 2020) have been proposed to decrease the energy consumption of traditional Deep Neural Networks (DNNs). The AdNNs deactivate certain computations of DNNs to save energy. The AdNNs can be of two types: Conditional-skipping AdNNs (Hua et al., 2019; Gao et al., 2018; Liu and Deng, 2018; Wang et al., 2018) and Early-termination AdNNs (Graves, 2016; Teerapittayanon et al., 2016; Yang et al.,

2020). Conditional-skipping AdNNs skip few computations depending on input, whereas, Early-termination AdNNs stop the computation early if additional computation is not needed. All AdNNs depend on specific intermediate outputs to activate or deactivate computations. If specific intermediate outputs reach a certain threshold value, the computations are activated.

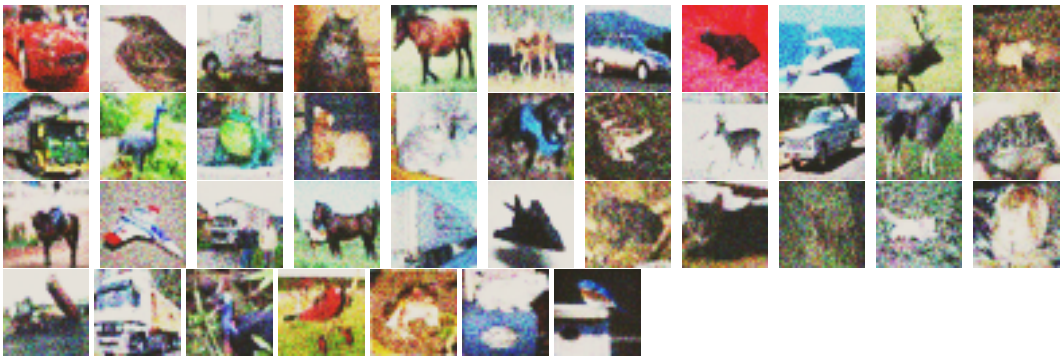
I.1 Attacking AdNNs

Haque *et al.* (Haque et al., 2020) use these intermediate outputs to devise energy attack against AdNNs. Let's assume the intermediate output values of a benign input is represented as $current$ and threshold values as $target$. So, the attacker would try to minimize the difference between $target$ values and $current$ values and use iterative optimization for that.

The major difference between AdNNs and Neural ODEs is the working mechanism. For Neural ODEs, there is no target threshold value to achieve, therefore, we need to consider on average step size. Also, for AdNNs, only the computations that have been deactivated can only be activated again. So, for high energy consuming benign inputs the increased energy consumption would be minimal. However, for Neural ODE, increase in energy consumption does not depend on any already decreased computations.

J Input-based Adversarial Images Generated for Dopri5 (CIFAR-10)

For images generated for Dopri5 solver, the average Peak Signal to Noise Ratio of adversarial images is 20.15 and Structural Similarity Index of adversarial images is 0.846. This evaluation is based on C=10000. Here are some generated images.



K Input-based Adversarial Images Generated for Heun (CIFAR-10)

For images generated for Adaptive Heun solver, the average Peak Signal to Noise Ratio of adversarial images is 19.34 and Structural Similarity Index of adversarial images is 0.83. This evaluation is based on C=10000. Here are some generated images.

References

Gilad Cohen, Guillermo Sapiro, and Raja Giryes. 2020. Detecting adversarial samples using influence functions and nearest neighbors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14453–14462.

Xitong Gao, Yiren Zhao, Łukasz Dudziak, Robert Mullins, and Cheng-zhong Xu. 2018. Dynamic channel pruning: Feature boosting and suppression. *arXiv preprint arXiv:1810.05331* (2018).

Alex Graves. 2016. Adaptive Computation Time for Recurrent Neural Networks. *arXiv preprint arXiv:1603.08983* (2016).

Mirazul Haque, Anki Chauhan, Cong Liu, and Wei Yang. 2020. ILFO: Adversarial Attack on Adaptive Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14264–14273.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).

Weizhe Hua, Yuan Zhou, Christopher M De Sa, Zhiru Zhang, and G Edward Suh. 2019. Channel gating neural networks. In *Advances in Neural Information Processing Systems*. 1884–1894.

Sven Köhler, Benedict Herzog, Timo Hönig, Lukas Wenzel, Max Plauth, Jörg Nolte, Andreas Polze, and Wolfgang Schröder-Preikschat. 2020. Pinpoint the Joules: Unifying Runtime-Support for Energy Measurements on Heterogeneous Systems. In *2020 IEEE/ACM International Workshop on Runtime and Operating Systems for Supercomputers (ROSS)*. IEEE, 31–40.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems* 31 (2018).

Lanlan Liu and Jia Deng. 2018. Dynamic deep neural networks: Optimizing accuracy-efficiency trade-offs by selective execution. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Nvidia. [n.d.]. Nvidia TX2 User Manual. `https://developer.download.nvidia.com/embedded/L4T/r32-3-1_Release_v1.0/jetson_tx2_developer_kit_user_guide.pdf?HV5mFjAe8YmRE546AQy6FCKPAJ6vBY90_5Z9KkyqCyk1cS9mORwwIVt-GBo199H9742JYQP98enQCm1lP6hZvekJ4pNy6lmVomU4YUO0HUGquB1_8FXTw2Kl-WWkoWQlm9bhV5vzpKGj5C7DDtuypx1H0Ik-KmNFeMJ9kTG7HYcxd0YafLaDSQ`.

Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast Inference via Early Exiting from Deep Neural Networks. In *Proceedings of the International Conference on Pattern Recognition*. 2464–2469.

Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. 2018. Skipnet: Learning Dynamic Routing in Convolutional Networks. In *Proceedings of the European Conference on Computer Vision*. 409–424.

Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. 2020. Resolution Adaptive Networks for Efficient Inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Dian Ang Yap, Joyce Xu, and Vinay Uday Prabhu. 2019. On detecting adversarial inputs with entropy of saliency maps. *CV-COPS, IEEE CVPR* (2019).