

---

# Sponge Examples: Energy-Latency Attacks on Neural Networks

---

**Ilia Shumailov\***

University of Cambridge  
ilia.shumailov@cl.cam.ac.uk

**Yiren Zhao\***

University of Cambridge  
yiren.zhao@cl.cam.ac.uk

**Daniel Bates**

University of Cambridge  
daniel.bates@cl.cam.ac.uk

**Nicolas Papernot**

University of Toronto and Vector Institute  
nicolas.papernot@utoronto.ca

**Robert Mullins**

University of Cambridge  
robert.mullins@cl.cam.ac.uk

**Ross Anderson**

University of Cambridge  
ross.anderson@cl.cam.ac.uk

## Abstract

The high energy costs of neural network training and inference led to the use of acceleration hardware such as GPUs and TPUs. While this enabled us to train large-scale neural networks in datacenters and deploy them on edge devices, the focus so far is on average-case performance. In this work, we introduce a novel threat vector against neural networks whose energy consumption or decision latency are critical. We show how adversaries can exploit carefully crafted **sponge examples**, which are inputs designed to maximise energy consumption and latency.

We mount two variants of this attack on established vision and language models, increasing energy consumption by a factor of 10 to 200. Our attacks can also be used to delay decisions where a network has critical real-time performance, such as in perception for autonomous vehicles. We demonstrate the portability of our malicious inputs across CPUs and a variety of hardware accelerator chips including GPUs, and an ASIC simulator. We conclude by proposing a defense strategy which mitigates our attack by shifting the analysis of energy consumption in hardware from an average-case to a worst-case perspective.

## 1 Introduction

The wide adoption of machine learning has motivated serious study of its security vulnerabilities. Threat vectors such as adversarial examples [1, 2], data poisoning [3, 4], and membership inference attacks [5, 6] have been extensively explored. These attacks either target the *confidentiality* or *integrity* of machine learning systems [7, 8]. So what about the third leg of the security triad, their *availability*? In this paper, we introduce an attack that increases the power drawn by neural networks and the time they take to make decisions. Increased energy consumption on edge devices that are battery-powered (such as smartphones or IoT devices) can make them unavailable [9]. Perhaps even more seriously, an attack that can slow down decisions can subvert safety-critical or mission-critical systems.

Our key observation is that different inputs of the same size can cause a deep neural network (DNN) to draw very different amounts of time and energy: this *energy-latency gap* is the vulnerability we exploit.

Our attack can be even more effective against the growing number of systems that use GPUs or custom hardware. Machine learning in general, and deep learning in particular, command workloads heavy in matrix algebra. GPUs were fundamental to the AlexNet breakthrough in 2012 [10]; in response to increasing demand, Google introduced TPUs to facilitate inference – and training – in its datacenters [11]; while Apple introduced the Neural Engine to make its smartphones more energy-efficient for on-device deep learning [12].

Our first attack uses a genetic algorithm to craft malicious inputs, which we call *sponge examples*, designed to soak up energy from a neural network. Time measurements obtained by profiling mutated inputs serve as the fitness function needed for the genetic algorithm to evolve better sponge examples. Our second, enhanced, attack instead optimizes sponge examples with L-BFGS and a loss that encourages the optimizer to return inputs with large activation norms across all hidden layers of a neural network. We find that this second attack is even more effective at maximizing energy consumption during inference<sup>1</sup>. Quite apart from battery-draining attacks, sponge examples could cause a control system to fail to meet real-time service requirements. To give but two examples, autonomous vehicle designers should consider whether they could cause accidents, while engineers building neural networks into cognitive radar should analyse whether they create a new opportunity for jamming.

Our contributions are the following:

- We formulate a novel energy-based threat against the availability of ML systems. We instantiate a corresponding threat vector, *sponge examples*, with two attacks designed to cause the worst possible energy consumption during inference.
- Our evaluation shows that canonical models for both vision and language tasks are vulnerable to sponge examples, including models specifically compressed to be deployed on edge devices (e.g., MobileNet). Our genetic algorithm attack increases energy consumption by natural language tasks up to 200×, latency by up to 70× and our enhanced L-BFGS-B attack increases computer vision task consumption by up to 8%.
- We demonstrate the portability of the threat vector instantiated by sponge examples, given that our conclusions are consistent across a variety of chips: CPUs, GPUs, and an ASIC simulator. Sponge examples also transfer across model architectures.
- We present a simple defense against sponge examples that can also prevent unexpected increases in the energy consumption of ML systems in the absence of adversaries, with direct beneficial consequences for the carbon footprint of models deployed for inference at scale (e.g., machine translation on smartphones), as well as for the worst-case performance of such models.

## 2 Background on Hardware Acceleration for Deep Learning

Workloads generated by inference for deep neural networks (DNNs) are intensive in both compute and memory. Common hardware products such as CPUs and GPUs are now being adapted to this workload, and provide features dedicated to accelerating DNN inference. Intel’s Knights Mill CPU provides a set of SIMD instructions [13], while NVIDIA’s Volta GPU introduces Tensor Cores to facilitate the low-precision multiplications that underpin much of deep learning [14].

Hardware dedicated to deep learning is now pervasive in data centers, with examples including Big Basin at Facebook [15], BrainWave at Microsoft [16], and racks of TPUs at Google [11, 17]; the underlying hardware on these systems are either commodity hardware (Big Basin), reconfigurable hardware (FPGAs for BrainWave), or custom silicon (TPUs). The latter two are specifically designed to improve the number of Operations per Watt (OPs/W) of DNN inference. As we discuss later, custom and semi-custom hardware often exploit sparsity in data and the adequacy of low-precision computations to train neural networks, reducing both arithmetic complexity and DRAM accesses, and thus achieving significantly better power efficiency [18, 19]. Our attack undermines the benefits brought by such hardware.

---

<sup>1</sup>Technically, we might describe this process as pessimization rather than optimization, as we’re finding the inputs that give the worst possible performance.

Performance per watt is an important indicator for the efficiency of cloud infrastructure [20]. Power oversubscription is a popular method for cloud services to handle provisioning. However, this makes data centers vulnerable to power attacks [21, 22, 23, 24]. If malicious users can remotely generate power spikes on multiple hosts in the data center at the same time, they might overload the system and cause disruption of service [21, 25]. Energy attacks against mobile devices aim to drain the battery more quickly [9, 26]. The possible victims of energy attacks on mobile systems range from phones to more constrained sensors [27]. Higher energy consumption also increases hardware temperature, which in turn increases the failure rate. For example, Anderson *et al.* note that an increase of 15°C causes component failure rates to double [28]. Modern hardware throttles to avoid over-heating; but throttling causes tasks to consume more energy as now they take longer to run, increasing total static and dynamic power costs.

### 3 Methodology

#### 3.1 Adversary Model

In this paper we assume an adversary with the ability to supply an input sample to a target system, which then processes the sample using a single CPU, GPU or ASIC. We assume no rate limiting, apart from on-device dynamic power control or thermal throttling. We assume no physical access to the systems *i.e.* an attacker cannot reprogram the hardware or change the configuration.

We consider our attack in three threat models. The first threat model is a **white-box** setup: we assume the attackers know the model architecture and parameters. The second threat model considers an **interactive black-box**: we assume attackers have no knowledge of the architecture and parameters, but are able to time the operations or measure energy consumption remotely and can query the target as many times as they want. The third threat model is the one of a **blind adversary**, where no knowledge of the target architecture and parameters is assumed and attackers cannot take any direct measurements. In this setting, the adversary has to resort to a direct transfer of sponge examples found previously to a new target model – without prior interaction.

A simple example of a target system could be a dialogue system. Users interact continuously by sending queries and can measure energy consumption, or when that is not possible have a timing side channel by measuring the response time (see Section 4).

#### 3.2 The Energy Gap

We tested three hardware platforms: CPU, GPU and an ASIC simulator. The amount of energy consumed by one inference pass (*i.e.* a forward pass in a neural network) depends primarily on:

- The overall number of arithmetic operations required to process the inputs;
- The number of memory accesses *e.g.* to the GPU DRAM.

The intriguing question now is:

*Is there a significant gap in energy consumption for different model inputs of the same dimension?*

As well as fixing the dimension of inputs, we also do not consider inputs that would exceed the pre-defined numerical range of each dimension. If models do have a large energy gap between different inputs, we describe two hypotheses that we think attackers can exploit to create *sponge examples*, that is, inputs with abnormally high energy consumption.

##### 3.2.1 Hypothesis 1: Data Sparsity

The rectified linear unit (ReLU), which computes  $x \mapsto \max(0, x)$ , is the de facto choice of activation function in neural network architectures. This design introduces sparsity in the activations of hidden layers when the weighted sum of inputs to a neuron is negative. A large number of ASIC neural network accelerators consequently exploit runtime data sparsity to increase efficiency. For instance, ASIC accelerators may employ zero-skipping multiplications or encode DRAM traffic to reduce the off-chip bandwidth requirement. Hence, inputs that lead to less sparse activations will increase the number of operations and the number of memory accesses, and thus energy consumption.

### 3.2.2 Hypothesis 2: Computation Dimensions

Aside from data sparsity, modern networks also have a computational dimension. Along with variable input and output shapes, the internal representation size often changes as well – for example, in the Transformer-based architectures for machine translation [29]. In this case, both the input and output are sequences of words and internal representation depends on both of them. Each word is represented as a number of tokens, whose shape depends on the richness of input and output dictionaries. As computation progresses, internally each inference step depends on all of the inputs and outputs so far.

Consider an input sequence  $x$  and an output sequence  $y$ . We denote the input and output token sizes (*i.e.* the number of unique words) with  $l_{\text{tin}}$  and  $l_{\text{tout}}$ . Each of the words in a sequence is embedded in a space of dimensionality  $l_{\text{ein}}$ , for the input, and  $l_{\text{eout}}$ , for the output. Algorithm 1 contains the pseudocode for a Transformer’s principal steps. In red, we annotate the computational complexity of the following instruction. As can be seen, several quantities can be manipulated by an adversary to increase the algorithm’s run time: 1) token size of the input sentence  $l_{\text{tin}}$ ; 2) token size of the output sentence  $l_{\text{tout}}$ ; and 3) size of the input and output embedding spaces ( $l_{\text{ein}}$  and  $l_{\text{eout}}$ ). All of the above can cause a non-linear increase in algorithmic complexity and thus heavily increase the amount of energy consumed. Importantly, perturbing these quantities does not require that the adversary modify the dimension of input sequence  $x$ ; so with no changes to the input dimensionality, the adversary is able to increase energy consumption non-linearly.

---

#### Algorithm 1: Translation Transformer NLP pipeline

---

**Result:**  $y$

```

1   $\downarrow O(l_{\text{tin}})$ 
2   $x_{\text{tin}} = \text{Tokenize}(x)$ ;
3   $y_{\text{touts}} = \emptyset$ ;
4   $\downarrow O(l_{\text{ein}})$ 
5   $x_{\text{ein}} = \text{Encode}(x_{\text{tin}})$ ;
6   $\downarrow O(l_{\text{tin}} \times l_{\text{ein}} \times l_{\text{tout}} \times l_{\text{eout}})$ 
7  while  $y_{\text{tout}}$  has no end of sentence token do
8       $\downarrow O(l_{\text{eout}})$ 
9       $y_{\text{eout}} = \text{Encode}(y_{\text{tout}})$ ;
10      $\downarrow O(l_{\text{ein}} \times l_{\text{eout}})$ 
11      $y_{\text{eout}} = \text{model.Inference}(x_{\text{ein}}, y_{\text{eout}}, y_{\text{touts}})$ ;
12      $\downarrow O(l_{\text{eout}})$ ;
13      $y_{\text{tout}} = \text{Decode}(y_{\text{eout}})$ ;
14      $y_{\text{touts}}.\text{add}(y_{\text{tout}})$ ;
15 end
16  $\downarrow O(l_{\text{tout}})$ ;
17  $y = \text{Detokenize}(y_{\text{touts}})$ 
```

---

### 3.3 Latency and Energy Attacks on Neural Networks

Having presented the intuition behind our attacks, we now introduce three strategies for finding sponge examples, corresponding to the threat models described in Section 3.1.

#### 3.3.1 Finding Sponge Examples with Genetic Algorithms in White and Black-box Settings

Genetic algorithms (GA) are a powerful tool for adversaries [30]. They can optimise a diverse set of objectives, require no local gradient information. They are a particularly good fit for adversaries who only have access to the model’s prediction in a black-box setting. We start with a pool of randomly generated samples  $S$ . These are images for computer vision models, or sentences for NLP tasks. We then iteratively evolve the population pool:

- For computer vision tasks, we sample two parents  $A$  and  $B$  from the population pool, and crossover the inputs using a random mask  $A * \text{mask} + (1 - \text{mask}) * B$ .

- For NLP tasks, we sample two parents  $A$  and  $B$ , and crossover by concatenating the left part of parent  $A$  with the right part of parent  $B$ . We then probabilistically invert the two parts.

We explain the reasons for these choices in Appendix D. Next, we randomly mutate (*i.e.* randomly perturb) a proportion of the input features (*i.e.* pixels in vision, words in NLP) of the children. To maintain enough diversity in the pool, we preserve the best per-class samples in the pool. We obtain a fitness score  $P$  for all pool members, namely their energy consumption. We then select the winning top 10% samples  $\hat{S}$ ,<sup>2</sup> and use them as parents for the next iteration. This genetic algorithm is simple but effective in finding sponge examples.

### 3.3.2 Sponge Examples in the White-box Setting

We now consider an adversary with access to the model’s parameters. Rather than a genetic algorithm, we use L-BFGS-B [31] to optimise the following objective:

$$-\sum_{a_l \in A} \|a_l\|_2 \quad (1)$$

where  $A$  is the set of all activation values and  $a_l$  the activations of layer  $l$ . This generates inputs that increase activation values of the model across all of the layers simultaneously. Following Hypothesis 1 outlined above, the decrease in density prevents hardware from skipping some of the operations which in turn increases energy consumption. We only evaluate the performance of sponge examples found by L-BFGS-B on computer vision tasks because of the discrete nature of the NLP tasks.

## 3.4 Cross-model and Cross-hardware Transferability for Blind Adversaries

When adversaries are unable to query the model, they cannot directly solve an optimisation problem to find sponge examples, even using interactive black-box approach *i.e.* run the GA. In this blind adversary setting, we exploit transferability across both models *and* hardware. Indeed, in Section 4.5 and Appendix F, we show that sponge examples transfer across models. We examine three hardware platforms in our evaluation:

- **CPU:** The platform is an Intel(R) Xeon(R) CPU E5-2620 v4 with 2.10GHz clock frequency. We use the Running Average Power Limit (RAPL) to measure energy consumption of the CPU. RAPL has been thoroughly evaluated and found to reflect actual energy consumption, as long as the counters are not sampled too quickly [32, 33].
- **GPU:** We use a GeForce 1080 Ti GPU with a 250.0 Watts power limit, a 96°C slowdown temperature and a 84°C throttling temperature. We use the NVIDIA Management Library (NVML) to measure energy consumption. NVML was previously found to capture energy quite accurately, with occasional instability for high-low patterns and high sampling rates [34].
- **ASIC:** We also developed a deterministic ASIC simulator, which monitors and records the runtime operations and number of DRAM accesses assuming a conservative memory flushing strategy. We then use measurements by Horowitz to approximate energy consumption [35]: at 45nm technology and 0.9V, we assume 1950 pJ to access a 32 bit value in DRAM and 3.7 pJ for a floating-point multiplication.

We show in Section 4.5 that sponge examples transfer across these types of hardware.

## 4 Evaluation

### 4.1 Experimental Setup

**NLP tasks.** We first evaluate our sponge example attack on a range of NLP models provided by the FairSeq framework [36]. The models we consider have achieved top performance at their respective tasks. We report performance of the RoBERTa [37] model, an optimised BERT [38], on three

---

<sup>2</sup>As the sample pool is large, selecting the top 10% makes the process more tractable.

GLUE benchmarks [39]. The datasets we considered include tasks in the SuperGLUE benchmark plus a number of machine-translation tasks. The SuperGLUE benchmark follows the style of GLUE but includes a wider range of language-understanding tasks including question answering and conference resolution [39, 40]. Further, we evaluate the attack on a number of translation tasks (WMT) using Transformer-based models [41, 42, 43].

Consider the pipeline for handling text. Before getting to the models, text goes through several prepossessing steps. First, words get tokenized in a manner meaningful for the language. We used the tokenizer from the Moses toolkit[44], which separates punctuation from words and normalises characters. Next, tokenized blocks get encoded. Until recently, unknown words were simply replaced with an unknown token. Modern encoders improve performance by exploiting the idea that many words are a combination of other words. BPE is a popular approach that breaks unknown words into subwords it knows and uses those as individual tokens [45]. In that way, known sentences get encoded very efficiently, mapping every word to a token. (Unknown words lead to a much larger sentence representation.)

**Vision tasks.** We evaluate the sponge example attack on a range of vision models provided in the TorchVision library. We show the performance of ResNet-18, ResNet-50 and ResNet-101 [46], DenseNet-121, DenseNet-161, DenseNet-201 [47], GoogleNet [48] and MobileNet-V2 [49]. All of the networks solve a canonical computer vision classification task – ImageNet-2017.

## 4.2 White-box Sponge Examples against NLP tasks

	Input size	NVML <sub>gpu</sub>	Natural <sub>asic</sub>	Random <sub>asic</sub>	Sponge Mean <sub>asic</sub>	Sponge Top 10% <sub>asic</sub>	Energy <sub>gpu</sub>	Time <sub>gpu</sub>
<i>Language Understanding: SuperGLUE Benchmark with [37]</i>								
CoLA	15	5829.32	4.30	69.72	83.92	87.11	×20.25	×1.23
	30	9388.40	4.30	138.07	164.07	169.91	×39.51	×1.48
	100	22698.87	4.30	452.49	518.19	530.80	×123.42	×3.82
MNLI	15	6126.65	12.88	73.47	86.97	89.96	×6.98	×1.32
	30	9631.68	17.66	142.63	168.96	174.34	×9.87	×2.03
	100	22952.14	34.47	456.11	518.89	531.40	×15.42	×3.16
WSC	15	27876.53	14.48	523.28	1300.19	2152.67	×148.62	×9.83
	30	82822.58	34.94	1882.63	3927.63	5348.06	×153.08	×19.25
	100	662811.96	194.89	16754.13	25367.30	30692.95	×157.49	×69.83
<i>Machine Translation: WMT14/16 with [41]</i>								
En→Fr	30	59597.32	31.87	109.80	118.47	141.27	×4.43	×4.45
	50	93731.34	48.54	166.13	249.89	569.85	×11.74	×13.51
En→De	15	18133.66	18.19	35.80	242.39	542.35	×29.82	×32.86

Table 1: Energy is reported in millijoules. GA was ran for 100 epochs with a pool size of 100. We report GPU energy readings using NVML and show the amount of energy and time increase as Energy<sub>gpu</sub> and Time<sub>gpu</sub> respectively. In addition, we show the performance of samples from the evaluation dataset (Natural<sub>asic</sub>), random noise (Random<sub>asic</sub>) and sponge example performance on the accelerator simulator. For sponge examples, we average the final GA pool samples and report it as Sponge Mean<sub>asic</sub>; in addition, the top 10% sponge examples are averaged and reported as Sponge Top 10%<sub>asic</sub>. More results are available in Appendix.

Table 1 shows the energy consumption of different models in the presence of an energy-latency adversary. For all models examined in Table 1, we found effective sponge examples that increased energy consumption considerably. In the best-case scenario, we were able to slow down the task evaluation by a factor of 70 and increase its energy consumption by a factor of 197. The main reason for performance degradation was increased computation dimension, as described in Algorithm 1. First, for a given input sequence size, the attack maximises the size of the post-tokenisation representation ( $x_{\text{tin}}$ ), exploiting the tokeniser and sub-word processing. Second, the attack learns to maximise output sequence length, since this links directly to the computation cost. Third, internal computation coupled with the output sequence length and post-tokenisation length give a quadratic increase in energy consumption. An interesting observation is that Natural<sub>asic</sub> text consumes a lot less energy in comparison to Random<sub>asic</sub> and Sponge Mean<sub>asic</sub> texts. This can be attributed to the

fact that natural samples are efficiently encoded, whereas random and attack samples produce an unnecessarily long representation, meaning random noise can be used as a scalable black-box latency and energy attacker.

### 4.3 White-box Sponge Examples against CV Tasks

According to the hypotheses in Section 3.2.1 and Section 3.2.2, we can increase energy consumption by increasing either computation dimension or data density. Although theoretically we can provide larger images to increase the computation dimension for computer vision networks, very few modern networks currently deal with dynamic input or output. Usually preprocessing normalizes variable-sized images to a pre-defined size by either cropping or scaling. Therefore, for computer vision models, we focus on increasing the energy and latency via data density. We calculate the theoretical upper bounds of data density using Interval Bound Propagation (IBP) [50]. Although originally developed for certifiable robustness, we adopt the technique to look at internal network bounds that only take 0 for the whole natural image range<sup>3</sup>.

		Time <sub>gpu</sub> [s]	Cost <sub>asic</sub> [mJ]	Cost <sub>asic</sub> ratio	post-ReLU Density	Density	Max Density
ResNet-50	L-BFGS-B Sponge	0.011	164.727	0.863	0.619	0.885	0.998
	Sponge	0.016	160.887	0.843	0.562	0.868	
	Natural	0.017	160.562	0.842	0.572	0.867	
	Random	0.017	155.820	0.817	0.483	0.845	
DenseNet-121	L-BFGS-B Sponge	0.033	152.595	0.783	0.571	0.826	0.829
	Sponge	0.029	149.564	0.767	0.540	0.814	
	Natural	0.033	147.227	0.755	0.523	0.804	
	Random	0.030	144.365	0.741	0.487	0.792	
MobileNet v2	L-BFGS-B Sponge	0.011	87.511	0.844	0.692	0.890	0.996
	Sponge	0.010	84.513	0.815	0.645	0.868	
	Natural	0.011	85.075	0.821	0.646	0.873	
	Random	0.011	80.805	0.779	0.567	0.844	

Table 2: Energy is reported in milli joules. GA was ran for 100 epochs with a pool size of 100. More results are available in Appendix.

Table 2 presents the performance of sponge example attacks on selected CV models; we show results for all of them in Appendix C. As shown in Time<sub>gpu</sub>, the inference pass of CV models uses a lot less time and the maximum energy consumption is a lot lower. Since the energy consumption is lower per inference, it is challenging to get a true measurement of energy given the interference from the GPU’s hardware temperature control, and that energy inspection tools lack the resolution. We attempt to statistically measure consumption in Appendix E. In Table 2, we focused on the results of GA maximising the ASIC simulator cost. Cost<sub>asic</sub> shows the energy estimated from the accelerator simulator; density is the ratio between the number of non-zero values and total values, post-ReLU Density only considers the values after the ReLU non-linearity, and ‘Cost<sub>asic</sub> ratio’ refers to the cost on the ASIC with optimisations compared to the cost on an ASIC without any optimisations. In this case, we consider ASIC-level optimisations of zero-skipping multiplications [51] and compressed DRAM accesses [52]. ‘Max density’ refers to the largest possible density obtained through IBP. The results for density and energy suggest that both attacks successfully generate sponge examples that are marginally more expensive in terms of energy. To be precise, we were able to get a 1 – 3% increase in energy consumption when compared to natural samples. To better understand the difference in performance please refer to Appendix H. Finally, we observe that sponge examples are transferable and can be used to launch a blind black-box attack. See Appendix F for details.

### 4.4 Interactive Black-box Sponge Examples against NLP Tasks

In this section we show the performance of the attack run in an interactive black-box manner against NLP tasks. We compare the sponge examples found by our attacks based on 1) time; 2) energy; and 3) ASIC cost. The last is a white-box baseline, whilst the first two are interactive black-box attacks. This means we use whether time, energy or ASIC cost as a fitness scoring for the GA. Figure 1 shows

<sup>3</sup>Note that we assume full floating point precision here. In practice, emerging hardware often uses much lower quantisation which will result in a lower maximum data density.

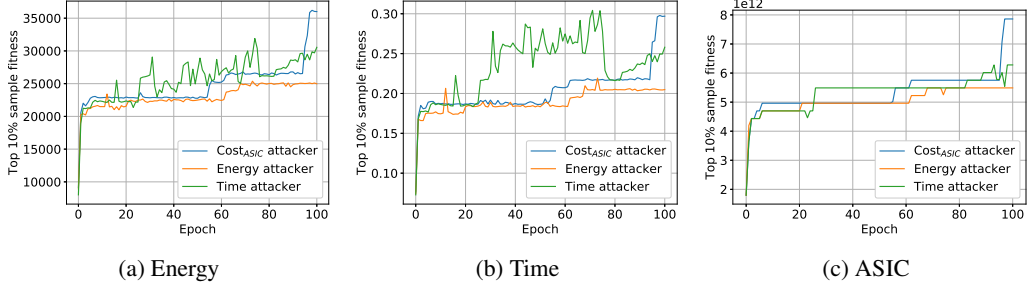


Figure 1: Performance of Sponge Examples based on the Energy, Time and Simulator fitness costs.

From	To		Cost <sub>asic</sub> [mJ]	Time <sub>gpu</sub> [s]	NVML <sub>gpu</sub> [mJ]	Time <sub>cpu</sub> [s]	RAPL <sub>cpu</sub> [mJ]
<u>Black-box</u>							
WMT16 <sub>en→de</sub> [41]	WMT14 <sub>en→fr</sub> [41]	Sponge	3648.219	0.174	17251.000	1.048	51512.966
		Natural	1450.403 ×2.52	0.053 ×3.27	6146.550 ×2.81	0.537 ×1.95	23610.145 ×2.18
	WMT18 <sub>en→de</sub> [42]	Sponge	2909.245	0.414	47723.500	3.199	181936.595
		Natural	1507.364 ×1.93	0.253 ×1.64	27265.250 ×1.75	1.344 ×2.38	71714.201 ×2.54
	WMT19 <sub>en→ru</sub> [43]	Sponge	3875.365	0.652	67183.100	4.409	247585.091
		Natural	1654.965 ×2.34	0.215 ×3.03	25033.620 ×2.68	2.193 ×2.01	121210.376 ×2.04
<u>White-box</u>							
WMT16 <sub>en→de</sub> [41]	WMT16 <sub>en→de</sub> [41]	Sponge	48447.093	2.414	260187.900	13.615	781758.680
		Natural	1360.118 ×35.62	0.056 ×42.98	6355.620 ×40.94	0.520 ×26.20	23262.311 ×33.61

Table 3: Energy values are reported in milli Joules and time is reported in seconds. GA was ran for 100 epochs with a pool size of 1000. More results are available in Appendix.

sponge example performance against a WMT14 English-to-French Transformer-based translator with an input of size 15 and pool size of 100. It can be seen that, although both Energy and Time attackers run without any knowledge of network internals, they successfully increase the energy and time costs.

#### 4.5 Blind Black-box Sponge Examples and Hardware Transferability on NLP Tasks

In this section we turn to the question of transferability across hardware and different models in a blind black-box manner. Table 3 shows the results across different models, languages and hardware platforms. We observe a significant increase in energy and time in comparison to natural samples. However, the blind black-box attacks fail to achieve the same level of energy or latency degradation when taking the white-box case as a baseline.

### 5 Defending against Sponge Examples

Sponge examples can be found by adversaries with limited knowledge and capabilities, making the threat vector realistic. We propose a simple defense to preserve the availability of hardware accelerators in the presence of sponge examples. In Table 1, we observe that there exists a large energy gap between natural examples and random or sponge examples. We propose that prior to the deployment of a model, natural examples get profiled to measure the time or energy cost of inference. The defender can then fix a *cut-off* threshold. This way, the maximum consumption of energy per inference run is controlled and sponge examples will have a bounded impact on availability.

This deals with the case where the threat model is battery drainage. Where the threat is a jamming attack on real-time performance, as for example with the vision system of an autonomous vehicle or a missile targeting system driven by cognitive radar, the system will need to be designed for worst-case performance, and if need be a fallback driving or targeting mechanism should be provided.



## 6 Conclusion

We introduced energy-latency attacks, which equip an adversary to increase the latency and energy consumption of ML systems thus jeopardizing their *availability*. Our attacks are effective against deep neural networks in a spectrum of threat models that realistically capture current deployments of ML, whether as a service or on edge devices. They can be mounted by adversaries whose access varies from total access to none at all. Our work demonstrates the need for careful worst-case analysis of the latency and energy consumption of computational systems that use deep learning mechanisms.

### Broader Impact: Carbon emission and Machine Learning

Most of the prior research on the carbon footprint of machine learning focuses on the energy required to train large neural network models and its contribution to carbon emissions [53, 54, 55]. This work shows that we need to study carbon emissions at small scales as well as large. As with side-channel attacks on cryptographic systems, the fine-grained energy consumption of neural networks is a function of the inputs processed. In this case, the main consequence is not leakage of confidential information but a denial of service attack.

First, sponge examples can aim to drain a device’s batteries; the operations and memory accesses in inference account for around a third of the work done during a complete backpropagation step (*i.e.* forward-backward pass), but inference happens at a much higher frequency and scale compared to training once a model is deployed. Our research characterizes the worst-case energy consumption of inference. This is particularly pronounced with natural language processing tasks, where the worst case can take over 100 times the time and energy of the average case.

Second, the sponge examples found by our attacks can be used in a targeted way to cause an embedded system to fall short of its performance goal. In the case of a machine-vision system in an autonomous vehicle, this might enable an attacker to crash the vehicle; in the case of a missile guided by a neural network target tracker, a sponge example countermeasure might break the tracking lock. Adversarial worst-case performance must, in such applications, be tested carefully by system engineers.

### Acknowledgements

Partially supported with funds from Bosch-Forschungsstiftung im Stifterverband. Research at the University of Toronto and Vector Institute was supported by a Canada CIFAR AI Chair, the Vector Institute’s industrial sponsors, DARPA, and NSERC.

### References

- [1] Battista Biggio, Iginio Corona, Davide Maiorca, Blaine Nelson, Nedim Šrđić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.
- [2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [3] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D Joseph, Benjamin IP Rubinstein, Udam Saini, Charles A Sutton, J Doug Tygar, and Kai Xia. Exploiting machine learning to subvert your spam filter. *LEET*, 8:1–9, 2008.
- [4] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 19–35. IEEE, 2018.
- [5] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

- [6] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.
- [7] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.
- [8] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.
- [9] Thomas Martin, Michael Hsiao, Dong Ha, and Jayan Krishnaswami. Denial-of-service attacks on battery-powered mobile computers. In *Second IEEE Annual Conference on Pervasive Computing and Communications, 2004. Proceedings of the*, pages 309–318. IEEE, 2004.
- [10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [11] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pages 1–12, 2017.
- [12] Computer Vision Machine Learning Team. An on-device deep neural network for face detection. In *Apple Machine Learning Journal*, 2017.
- [13] Intel Cooperation. Intel architecture instruction set extensions programming reference. *Intel Corp., Mountain View, CA, USA, Tech. Rep*, pages 319433–030, 2016.
- [14] Stefano Markidis, Steven Wei Der Chien, Erwin Laure, Ivy Bo Peng, and Jeffrey S Vetter. Nvidia tensor core programmability, performance & precision. In *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 522–531. IEEE, 2018.
- [15] Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, et al. Applied machine learning at facebook: A datacenter infrastructure perspective. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 620–629. IEEE, 2018.
- [16] Eric Chung, Jeremy Fowers, Kalin Ovtcharov, Michael Papamichael, Adrian Caulfield, Todd Massengill, Ming Liu, Daniel Lo, Shlomi Alkalay, Michael Haselman, et al. Serving dnns in real time at datacenter scale with project brainwave. *IEEE Micro*, 38(2):8–20, 2018.
- [17] Norman Jouppi, Cliff Young, Nishant Patil, and David Patterson. Motivation for and evaluation of the first tensor processing unit. *IEEE Micro*, 38(3):10–19, 2018.
- [18] Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, and Vivienne Sze. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(2):292–308, 2019.
- [19] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A Horowitz, and William J Dally. Eie: efficient inference engine on compressed deep neural network. *ACM SIGARCH Computer Architecture News*, 44(3):243–254, 2016.
- [20] Luiz André Barroso. The price of performance. *Queue*, 3(7):48–53, 2005.
- [21] Chao Li, Zhenhua Wang, Xiaofeng Hou, Haopeng Chen, Xiaoyao Liang, and Minyi Guo. Power attack defense: Securing battery-backed data centers. *ACM SIGARCH Computer Architecture News*, 44(3):493–505, 2016.
- [22] Gaurav Somani, Manoj Singh Gaur, Dheeraj Sanghi, and Mauro Conti. Ddos attacks in cloud computing: Collateral damage to non-targets. *Computer Networks*, 109:157–171, 2016.
- [23] Zhang Xu, Haining Wang, Zichen Xu, and Xiaorui Wang. Power attack: An increasing threat to data centers. In *NDSS*, 2014.
- [24] Zhang Xu, Haining Wang, and Zhenyu Wu. A measurement study on co-residence threat inside the cloud. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pages 929–944, 2015.

- [25] Francesco Palmieri, Sergio Ricciardi, Ugo Fiore, Massimo Ficco, and Aniello Castiglione. Energy-oriented denial of service attacks: an emerging menace for large cloud infrastructures. *The Journal of Supercomputing*, 71(5):1620–1641, 2015.
- [26] Ugo Fiore, Francesco Palmieri, Aniello Castiglione, Vincenzo Loia, and Alfredo De Santis. Multimedia-based battery drain attacks for android devices. In *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*, pages 145–150. IEEE, 2014.
- [27] Xiangqian Chen, Kia Makki, Kang Yen, and Niki Pissinou. Sensor network security: a survey. *IEEE Communications Surveys & Tutorials*, 11(2):52–73, 2009.
- [28] Dave Anderson, Jim Dykes, and Erik Riedel. More than an interface-scsi vs. ata.
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [30] Weilin Xu, Yanjun Qi, and David Evans. Automatically evading classifiers. In *Proceedings of the 2016 network and distributed systems symposium*, volume 10, 2016.
- [31] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyu Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.
- [32] Marcus Hähnel, Björn Döbel, Marcus Völpl, and Hermann Härtig. Measuring energy consumption for short code paths using rapl. *ACM SIGMETRICS Performance Evaluation Review*, 40(3):13–17, 2012.
- [33] Kashif Nizam Khan, Mikael Hirki, Tapio Niemi, Jukka K. Nurminen, and Zhonghong Ou. Rapl in action: Experiences in using rapl for power measurements. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, 3(2), March 2018.
- [34] S. Sen, N. Imam, and C. Hsu. Quality assessment of gpu power profiling mechanisms. In *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 702–711, 2018.
- [35] M. Horowitz. 1.1 computing’s energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14, 2014.
- [36] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.
- [37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [39] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- [40] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275, 2019.
- [41] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*, 2018.
- [42] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.
- [43] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook fair’s WMT19 news translation task submission. *CoRR*, abs/1907.06616, 2019.

- [44] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180, 2007.
- [45] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015.
- [46] 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.
- [47] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [49] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [50] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.
- [51] Dongyoung Kim, Junwhan Ahn, and Sungjoo Yoo. A novel zero weight/activation-aware hardware architecture of convolutional neural network. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*, pages 1462–1467. IEEE, 2017.
- [52] Angshuman Parashar, Minsoo Rhu, Anurag Mukkara, Antonio Puglielli, Rangharajan Venkatesan, Brucek Khailany, Joel Emer, Stephen W Keckler, and William J Dally. Scnn: An accelerator for compressed-sparse convolutional neural networks. *ACM SIGARCH Computer Architecture News*, 45(2):27–40, 2017.
- [53] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *arXiv preprint arXiv:2002.05651*, 2020.
- [54] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.
- [55] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.
- [56] Bhavishya Goel, Sally A McKee, and Magnus Sjölander. Techniques to measure, model, and manage power. In *Advances in Computers*, volume 87, pages 7–54. Elsevier, 2012.
- [57] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahm. Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134:75–88, 2019.
- [58] J Adam Butts and Gurindar S Sohi. A static power model for architects. In *Proceedings 33rd Annual IEEE/ACM International Symposium on Microarchitecture. MICRO-33 2000*, pages 191–201. IEEE, 2000.

# Appendices

## A Definitions

Energy  $E$  is the total consumed static power  $P_{\text{static}}$  and dynamic power  $P_{\text{dynamic}}$  for an interval of time  $t$ .

$$E = (P_{\text{static}} + P_{\text{dynamic}}) \times t$$

Static power refers to the consumption of the circuitry when it is in idle state [56]. There exist multiple models to estimate the static energy consumption depending on the technology [56, 57, 58]. In this paper we follow a coarse-grained approach to energy estimation.

$$P_{\text{static}} = \sum I_{\text{leakage}} \times V_{\text{core}} = \sum I_s \times (e^{\frac{qV_d}{kT}} - 1) \times V_{\text{core}} \quad (2)$$

where  $I_s$  is the reverse saturation current;  $V_d$  is the diode voltage;  $k$  is the Boltzmann's constant;  $q$  is the electronic charge;  $T$  is temperature and  $V_{\text{core}}$  is the supply voltage.

Dynamic power refers to consumption from (dis)charging of the said circuitry [57].

$$P_{\text{dynamic}} = \alpha \times C \times V_d^2 \times f \quad (3)$$

Here,  $\alpha$  refers to the activity factor *i.e.* components that are currently consuming power;  $V_d$  is the source voltage;  $C$  is the capacitance; and  $f$  is the clock frequency.

Ultimately an attacker attempts to solve an optimisation problem

$$\max_{\theta} E, \text{ where } E = ([\sum I_s \times (e^{\frac{qV_d}{kT}} - 1) \times V_{\text{core}}] + [\alpha \times C \times V_d^2 \times f]) \times t \quad (4)$$

For all parameters considered in the equation, only very few could be affected by the adversary described in Section 3.1. In particular, there are only four parameters that an attack might manipulate:  $T$ ,  $\alpha$ ,  $f$  and  $t$ . Although frequency and temperature cannot be controlled directly, they are affected through optimisations performed by the computing hardware. As we assume a single GPU, CPU or ASIC, we focus on the activity ratio  $\alpha$ , the time  $t$  and the switching power from flipping the state of transistors. The execution time  $t$  and activity ratio  $\alpha$  link tightly to the number of operations and memory accesses performed. In the temporal dimension, attackers might trigger unnecessary passes of a compute-intensive block; in the spatial domain, attackers can turn sparse operations to dense ones. These temporal and spatial attack opportunities can significantly increase the number of memory and arithmetic operations and thus create an increase in  $\alpha$  and  $t$  to maximise energy usage.

## B Parameter Choices

We have thoroughly evaluated different parameter choices for the Sponge attack and found that a small pool size and a relatively short number of GA iterations could be sufficient for a large number of tasks.

Figure 2 shows the performance of Sponge samples on the RoBERTa model for the Winograd Schema Challenge (WSC) with different pool sizes and varying input sequence length. The horizontal axis shows the number of GA iterations. In terms of pool size of the GA, although there is an increase in performance for larger pool sizes, the increase is marginal. In addition, smaller pool sizes significantly reduce the runtime for the attack. From the hardware perspective, using a large pool size might trigger GPUs to throttle, so that the runtime will be further increased. For the number of GA iterations, we observed that consistently for smaller input sequences, the convergence is faster. This is mainly because the complexity of the search is less. In practice, we found almost all input sequence lengths we tested plateau within 100 GA iterations; even going to over 1000 iterations gives only a small increase in performance. For these reasons, for the experiments presented below we

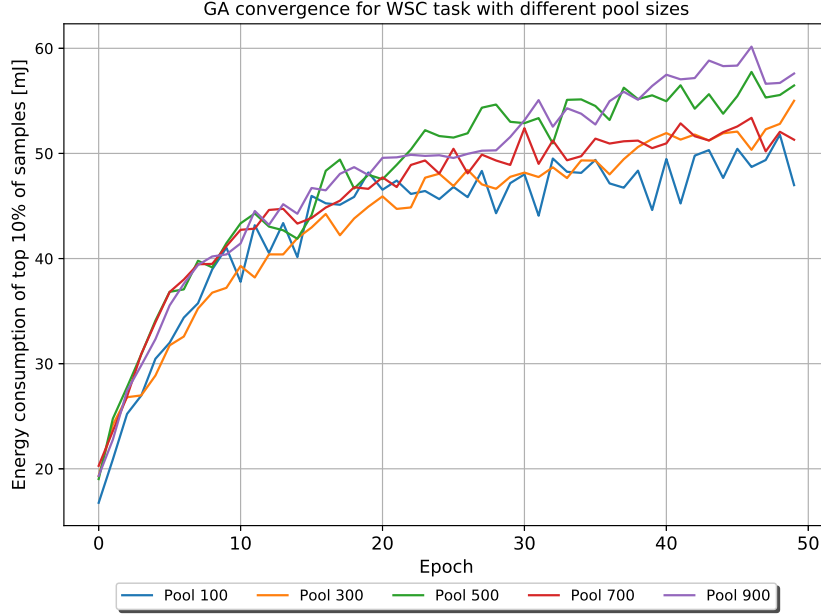


Figure 2: GA performance with WSC task from GLUE Benchmark. Words of size 29 are evaluated with pool sizes of 100, 300, 500, 700 and 900.

report the results of the attack with a pool size of 100 for GLUE and Computer Vision benchmarks, and 1000 for translation tasks. We use 100 GA iterations for all benchmarks tested.

## C Language and Vision Models Results

Table 1 and Table 2 show full sets of Sponge attacks running on NLP and CV tasks. It can be seen that the attacker can exploit the algorithmic complexity of NLP models and cause massive increases in both energy consumption and latency. It is clear that the increase is non-uniform across models, and we hypothesize that this is connected to the task detail and richness of the dictionaries we used.

For CV tasks, we can also observe that the attacker is capable of constructing samples that result in an increased internal data density. We observe that, for some network architectures, the effect is more prominent. Interestingly, we find that the majority of improvement comes not from density-side after ReLU, but rather from the propagation of non-zero values as this increases the number of DRAM accesses. For example, on mobilenet, decrease in sparsity leads to at least 7% degradation on ASIC.

## D Domain Specific Optimisations for Sponge

In Section 3.3 we outlined the genetic algorithm we used to find Sponge samples.

Despite the common structure, we relied heavily on domain-specific optimisations. In this section, we explain the reasoning behind them.

First, for NLP tasks, the greatest impact on performance was acquired from exploiting the encoding schemes used by different tasks. While the genetic algorithm was fast to pick up, it struggled with efficiency around the mid-point, where the parents were concatenated. For example, when trying to break down individual words to more tokens, we observed the GA inserting backslashes into the samples. When concatenated, we saw cases where two non-backslashes followed each other, meaning the GA was losing on a couple of characters. As a solution, we probabilistically flipped the halves and saw a slight improvement.

	Input size	NVML <sub>gpu</sub>	Natural <sub>asic</sub>	Random <sub>asic</sub>	EEE Mean <sub>asic</sub>	EEE Top 10% <sub>asic</sub>	Energy <sub>gpu</sub>	Time <sub>gpu</sub>
<i>SuperGLUE Benchmark with [37]</i>								
CoLA	15	5829.32	4.30	69.72	83.92	87.11	×20.25	×1.23
	30	9388.40	4.30	138.07	164.07	169.91	×39.51	×1.48
	50	14030.60	4.30	227.48	267.60	275.27	×64.01	×2.50
	70	17369.33	4.30	318.87	370.34	379.76	×88.30	×2.66
	100	22698.87	4.30	452.49	518.19	530.80	× <b>123.42</b>	× <b>3.82</b>
MNLI	15	6126.65	12.88	73.47	86.97	89.96	×6.98	×1.32
	30	9631.68	17.66	142.63	168.96	174.34	×9.87	×2.03
	50	13777.36	23.28	232.22	270.95	278.85	×11.98	×2.22
	70	17696.38	28.52	321.63	370.80	381.21	×13.37	×2.55
	100	22952.14	34.47	456.11	518.89	531.40	× <b>15.42</b>	× <b>3.16</b>
WSC	15	27876.53	14.48	523.28	1300.19	2152.67	×148.62	×9.83
	30	82822.58	34.94	1882.63	3927.63	5348.06	×153.08	×19.25
	40	126009.10	47.34	3151.23	6438.32	9345.28	× <b>197.42</b>	×35.93
	70	324814.61	115.17	8444.53	15632.59	20191.31	×175.32	×54.18
	100	662811.96	194.89	16754.13	25367.30	30692.95	×157.49	× <b>69.83</b>
<i>WMT14/16 with [41]</i>								
En→Fr	10	23497.33	15.07	40.47	55.36	90.38	×6.00	×5.83
	30	59597.32	31.87	109.80	118.47	141.27	×4.43	×4.45
	50	93731.34	48.54	166.13	249.89	569.85	×11.74	×13.51
En→De	15	18133.66	18.19	35.80	242.39	542.35	×29.82	×32.86

Table 4: Energy is reported in milli joules. All numbers are simulated, except for NVML, which is sampled directly from the GPU. GA was ran for 100 epochs with a pool size of 100.

For CV tasks, we observed that random samples were always classified as belonging to the same class. Furthermore, random samples had very low internal density. We hypothesize that this has to do with the fact that on random samples there are very few class features, as opposed to what is observed in natural samples. As the GA improvement largely depends on randomness, that meant that we often observed that after merging two highly dense parents, uniform randomness across all pixels was decreasing sparsity to the level of random samples. In other words, uniform randomness was diluting class features. In order to counter this phenomenon, instead of applying uniform randomness across all pixel values, we resorted to diluting only 1% of them. That led to a bigger improvement of the whole population pool. Furthermore, after observing that the density is class-dependent, it became apparent that in order to preserve diversity in the pool it was important to keep samples from multiple classes. For this, we tried to ensure that at least 20 different classes were preserved in the pool.

We attempted to use domain knowledge and tried adding operations like rotation, transposition and re-scaling into the mutation process, yet we found that these did not lead to significant improvements.

## E Measuring impact in practice

Although we have presented in Section 4.3 that Sponge attacks cause the ASIC energy consumption to rise for computer vision tasks, it is still unclear what this translates to the real life.

If one were to directly measure the CPU or GPU load per adversarial sample, interpreting it would be hard, especially when one talks about the energy cost improvements in the order of around 5% for ResNet18 and 3% as for DenseNet101. As is mentioned in Appendix A the main energy costs include the frequency of switching activities, voltage and clock frequency. Due to the heat impact from voltage and clock frequency, a large number of different optimisations are deployed by the hardware. Here, the optimisations try to balance multiple objectives – they try to be as performant as they can, whilst being as energy efficient as possible and also maintain reliability. Modern CPUs and GPUs have a number of performance modes between which the hardware can switch. For example, official Nvidia documentation lists 15 different performance modes.

		Time	Simulated	Absolute ration	Density
<i>ImageNet</i>					
ResNet-18	Sponge LBFGSB	0.006	53.359	0.899	0.896
	Sponge	0.006	51.816	0.873	0.869
	Natural	0.007	51.748	0.871	0.869
	Random	0.006	49.685	0.837	0.834
ResNet-50	Sponge LBFGSB	0.011	164.727	0.863	0.885
	Sponge	0.016	160.887	0.843	0.868
	Natural	0.017	160.562	0.842	0.867
	Random	0.017	155.820	0.817	0.845
ResNet-101	Sponge LBFGSB	0.021	258.526	0.857	0.873
	Sponge	0.024	254.182	0.842	0.861
	Natural	0.027	253.042	0.839	0.857
	Random	0.025	249.027	0.825	0.846
DenseNet-121	Sponge LBFGSB	0.033	152.595	0.783	0.826
	Sponge	0.029	149.564	0.767	0.814
	Natural	0.033	147.227	0.755	0.804
	Random	0.030	144.365	0.741	0.792
DenseNet-161	Sponge LBFGSB	0.040	288.427	0.726	0.764
	Sponge	0.044	287.153	0.723	0.761
	Natural	0.045	282.273	0.711	0.751
	Random	0.044	279.270	0.703	0.744
DenseNet-201	Sponge LBFGSB	0.048	237.745	0.756	0.788
	Sponge	0.046	239.845	0.763	0.794
	Natural	0.049	234.948	0.747	0.781
	Random	0.046	233.700	0.743	0.777
GoogleNet	Sponge LBFGSB	0.018	47.454	0.862	0.953
	Sponge	0.015	46.088	0.837	0.951
	Natural	0.016	45.964	0.835	0.955
	Random	0.015	44.164	0.802	0.938
MobileNet v2	Sponge LBFGSB	0.011	87.511	0.844	0.890
	Sponge	0.010	84.513	0.815	0.868
	Natural	0.011	85.075	0.821	0.873
	Random	0.011	80.805	0.779	0.844

Table 5: Energy is reported in milli joules. All numbers are simulated, except for NVML, which is sampled directly from the GPU. GA was ran for 100 epochs with a pool size of 100.

Figure 3 shows measurements taken during the Sponge GA attack running against ResNet-18. The x-axis shows the number of epochs, with each epoch the internal density is increasing from 0.75% to 0.8%. In (a), the right y-axis shows mean energy readings per sample, whereas left y-axis shows mean power readings per-sample. In (b) the left y-axis shows mean latency values per-sample.

The amount of power consumed is strongly correlated to the amount of time taken by each sample. When the GPU speeds up, it consumes more energy but requires less time, but the rise in temperature causes the hardware then to go to a more conservative mode to cool down. We observe this heating and cooling cycle with all tasks running on GPUs, making it hard to measure the absolute performance and the attack impact.

We can however measure the performance statistically. First, we turn to a question of

*Can we detect energy differences between Natural, Random and sponge samples?*

To investigate the relationship between the samples we use Mann-Whitney-Wilcoxon U test (U-test), a nonparametric test for difference between distributions. With three classes of samples, we need three pairwise comparisons. For each one, the null hypothesis that the distributions of energy consumed by the samples are identical. The complement hypothesis is that of a difference between distributions.

The U-test is based on three main assumptions:



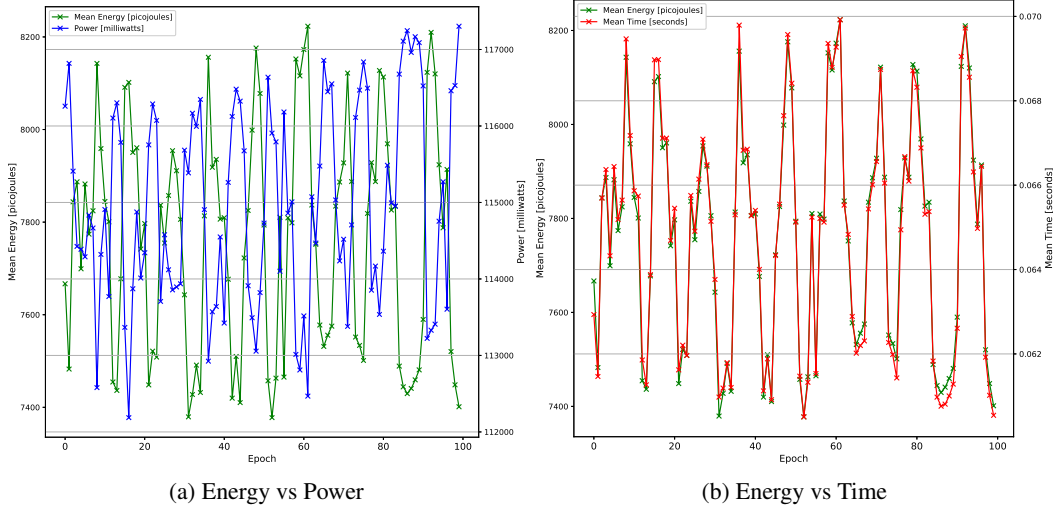


Figure 3: ResNet-18 solving ImageNet-2017 without any rate limiting with increasing internal density.

- Independence between samples;
- The dependent variable is at least ordinal;
- The samples are random.

The first assumption is fulfilled since no sample belongs to more than one category i.e. natural, random and sponge. The second assumption is satisfied by the fact that both time and energy are cardinal variables. The third assumption, however, is harder to satisfy.

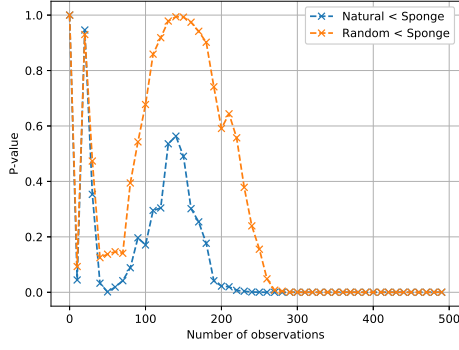
The cause of this lies in the closed nature of hardware optimisations: although some of the techniques are known, the exact parameters are unknown. Furthermore, it is hard to achieve same state of the hardware even through power cycling. As was mentioned in Appendix A temperature affects energy directly, and it is hard to make sure that the hardware always comes back to the same state.

To minimise temperature effects we apply the load of natural, attack and random samples and wait until the temperature stabilises. That takes approximately 30000 samples. The order of the samples is random, and at this point it can be assumed that all of the data and instruction caches are filled. Finally, because the samples are randomly shuffled, all of the predictive optimisations will work with the same probability for each of the classes.

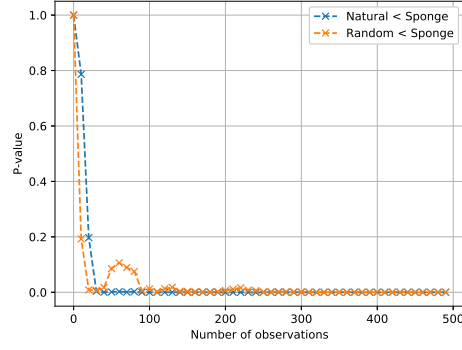
For these reasons, we believe it is safe to assume that the samples themselves are random in that the effect of hardware optimisations is random, so that the last assumption of Mann-Whitney test is fulfilled.

Using Mann-Whitney U test we can pairwise compare the natural, random and sponge samples. The test indicates that the three types of samples generate energy consumption distributions which are statistically different (one-sided test,  $p\text{-value}=0.000$ ) for mobilenet executed on a CPU. On a practical level, the amount of energy consumed by sponge samples is 1.5% higher on a CPU and >7% on ASIC. We could not evaluate the energy recordings on a GPU, as the standard deviation was in excess of 15% which worsened as temperature increased.

Figure 4 shows the confidence of Mann-Whitney test with mobilenet measured on the CPU as a function of the number of observations. The number of observations is on the x-axis, and the p-value on the y-axis. As it can be seen, in a stable environment i.e. the temperature has stabilised, after about 100 observations per class the differences become statistically significant at any reasonable confidence level. A similar trend is observed for unstable temperature environment, but around three times more data is required. That means that in practice, about 100–300 observations per class are sufficient to differentiate between classes with high confidence.



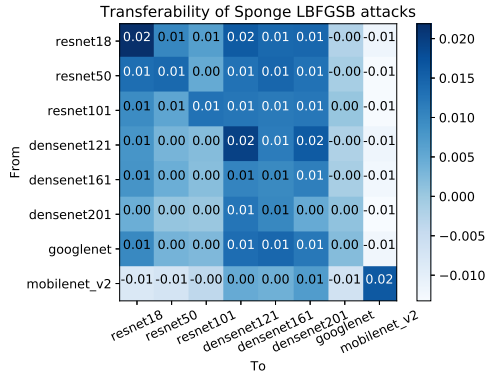
(a) First 5000 samples discarded



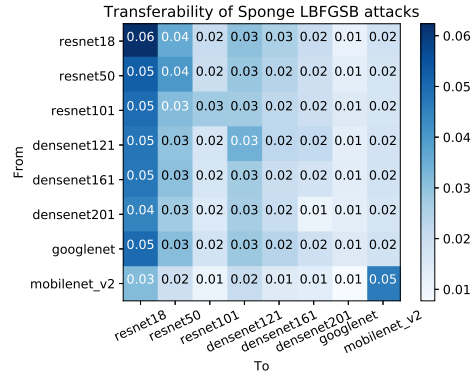
(b) First 30000 samples discarded

Figure 4: Mann-Whitney test on CPU measured Mobilenet execution. Number of observations is shown on x-axis and p-value on the y-axis.

## F Transferability of Attacks



(a) Sponge density - Normal density



(b) Sponge density - Random density

Figure 5: Sponge example density transanferability.

Figure 5 shows the density difference of transferred sponge samples. As it can be clearly seen, for all but one – i.e. mobilenet – network, the sponge samples have increased the internal data density despite not having any knowledge of what natural samples look like. All of the sponge samples outperformed random noise, suggesting that sponge samples target specific features of the data set and can be applied in a blind Black-box fashion.

## G Per class density transfer

Figure 6 shows the density of natural samples from ImageNet dataset. It can be clearly seen that there are per-class similarities between data densities of natural samples. These are particularly pronounced within resnet and densenet architectures hinting that similar architectures learn similar class-data representations. Finally the third graph subgraph in Figure 6.c shows the summed per-class densities across all of the tested networks. There are classes that are consistently densely represented across all the tested networks.

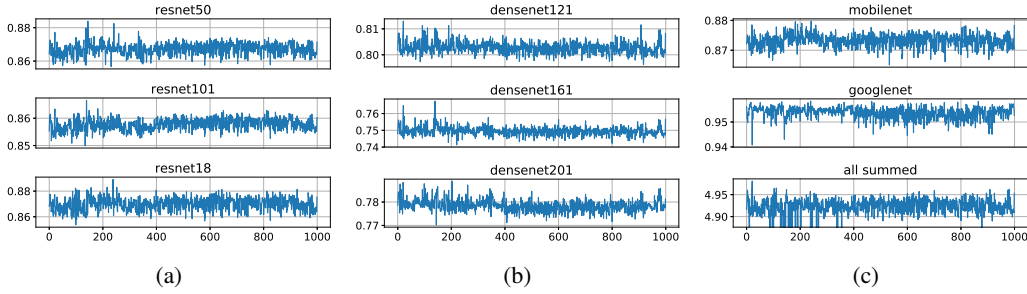


Figure 6: Per-class density of natural samples from the ImageNet validation dataset.

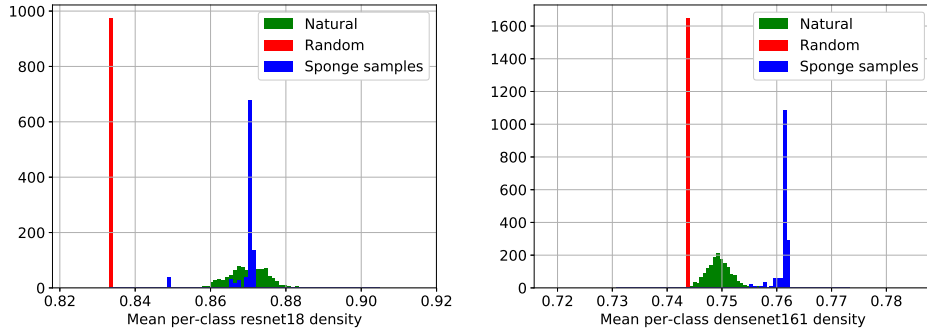


Figure 7: Per-class mean density of samples evaluated on ResNet18 and DenseNet161. The natural samples are from the validation set and are compared to 50 000 randomly generated samples and 1000 Sponge GA samples. The scales are normalised to form a probability density.

## H Understanding Sponge and sponge performance

To better understand the results, we present Figure 7 which shows per-class density distributions of natural, random and Sponge samples. There are 50000 random and natural samples respectively and 1000 Sponge samples, with the bars normalised to form a probability density.

The first thing that becomes apparent is that randomly generated samples on CV models cost significantly less energy because many activations are not activated. On average, random samples result in a sparser computation – around 4% more sparse for ResNet18 – and our simulator costs for natural samples are around 4 – 7% higher than the costs of random samples. Second, a surprising finding is that the most and least sparse samples are clustered in a handful of classes. In other words, certain classes have inputs that are more expensive than others in terms of energy. For ResNet-18, the most sparse classes are ‘wing’ and ‘spotlight’ while the least sparse are ‘greenhouse’ and ‘howler monkey’. We observe a similar effect for larger ResNet variants and also DenseNets, although the energy gap is smaller on DenseNets. Interestingly, we see that energy expensive classes are consistent across different architectures, and we further demonstrate this class-wise transferability in Appendix G. Ultimately, the implications of this phenomenon are that it is possible to burn energy or slow a system without much preparation, by simply bombarding the model with natural samples from energy-consuming classes. Finally, we see that the Sponge samples are improving the population performance and tend to outperform natural sample. We observe that it is easier for Sponge to outperform all natural samples for DensNets of different size, yet it struggles to outperform all of the ResNets. We further measure the energy performance statistically in Appendix E.