THE EFFICIENCY MISNOMER

Anonymous authors
Paper under double-blind review

ABSTRACT

Model efficiency is a critical aspect of developing and deploying machine learning models. Inference time and latency directly affect the user experience, and some applications have hard requirements. In addition to inference costs, model training also have direct financial and environmental impacts. Although there are numerous well-established metrics (cost indicators) for measuring model efficiency, researchers and practitioners often assume that these metrics are correlated with each other and report only few of them. In this paper, we thoroughly discuss common cost indicators, their advantages and disadvantages, and how they can contradict each other. We demonstrate how incomplete reporting of cost indicators can lead to partial conclusions and a blurred or incomplete picture of the practical considerations of different models. We further present suggestions to improve reporting of efficiency metrics.

1 INTRODUCTION

Aside from model quality, the efficiency (Menghani, 2021) of a model is often an important aspect to consider and is commonly used to measure the relative utility of different methods. After all, training time spent on accelerators is directly linked to financial costs and environmental impact. Meanwhile, the speed of a model may be directly linked to user experience. To this end, there have been well-established ways in the literature to assess and report the efficiency of a model such as number of trainable parameters, number of floating-point operations (FLOPs), and speed/throughput.

While it is commonly assumed that these cost indicators are correlated (e.g., a lower number of parameters would translate to a higher throughput) we show that this might not necessarily be the case. Therefore, incomplete reporting across the spectrum of cost indicators may lead to an incomplete picture of the metrics, advantages and drawbacks of the proposed method. To this end, we show that it may also be possible to, perhaps unknowingly, misrepresent a model's efficiency by only reporting favorable cost indicators. Moreover, the choice of cost indicators may also result in unfair, incomplete, or partial conclusions pertaining to model comparisons. We refer to this phenomenon as the '*efficiency misnomer*'.

The overall gist of the *efficiency misnomer* is that no single cost indicator is sufficient. Incomplete reporting (e.g., showing only FLOPs, or the number of trainable parameters) as a measure of efficiency can be misleading. For example, a model with low FLOPs may not actually be fast, given that FLOPs does not take into account information such as degree of parallelism (e.g., depth, recurrence) or hardware-related details like the cost of a memory access. Despite this, FLOPs has been used as the most common cost indicator in many research papers, especially in the recent computer vision literature, to quantify model efficiency (Szegedy et al., 2015; He et al., 2016; Tan and Le, 2019; Feichtenhofer et al., 2019; Fan et al., 2021).

Likewise, the number of trainable parameters (size of the model) despite being commonly used as the de-facto cost indicator in the NLP community (Devlin et al., 2018; Liu et al., 2019; Lan et al., 2019) and previously the vision community (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Huang et al., 2017; Tan and Le, 2019), can also be misleading when used as a standalone measure of efficiency. Intuitively, a model can have very few trainable parameters and still be very slow, for instance when the parameters are shared among many computational steps (Lan et al., 2019; Dehghani et al., 2018). While the number of trainable parameters can often be insightful to decide if a model fits in memory, it is unlikely to be useful as a standalone cost indicator. That said, it is still common practice to *parameter-match* models to make *'fair'* comparisons (Mehta et al., 2020;

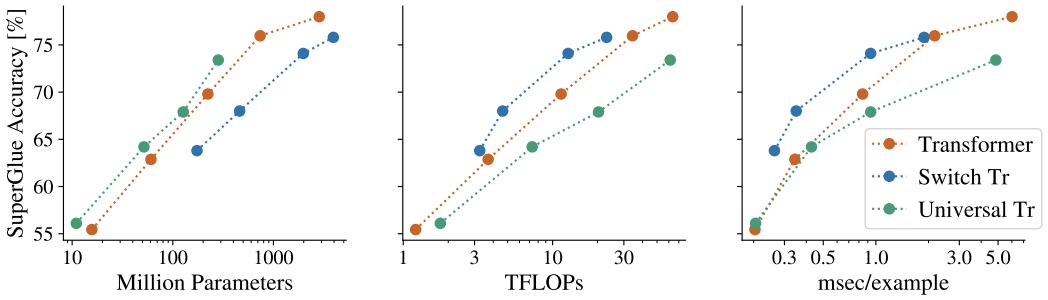


Figure 1: Comparison of standard Transformers, Universal Transformers and Switch Transformers in terms of three common cost metrics: number of parameters, FLOPs, and throughput. Relative ranking between models is reversed between the two cost indicators. Experiments and the computation of cost metrics were done with Mesh Tensorflow (Shazeer et al., 2018), using 64 TPU-V3.

Lee-Thorp et al., 2021; Tay et al., 2020a; Xue et al., 2021; Wightman et al., 2021), even if one model is in reality slower or faster than another.

Meanwhile, using throughput/speed as the primary indicator of efficiency can also be problematic - since this tightly couples implementation details, hardware optimizations, and infrastructure details (e.g., input pipeline latency) into the picture. Hence, this might not present an apples-to-apples comparison of certain methods or worse, across different infrastructures or hardware.

Given that the landscape of research on model architectures is diverse, the relationship between cost indicators may strongly deviate from the norm, and learning how to fairly compare models within the context of efficiency-based cost indicators is crucial. For instance, there seems to be a rising trend towards sparse models (Fedus et al., 2021; Riquelme et al., 2021) which usually have an incredibly large number of trainable parameters but maintain the FLOPs and speed of dense models. On the contrary, there are models that are considered lightweight due to their small number of trainable parameters (Lan et al., 2019; Dehghani et al., 2018) but, in actual practice, consume similar amounts of compute. Figure 1 shows an example of how the scaling behavior of a model with respect to parameter count can look favorable, while taking the FLOPs or throughput as the cost indicator, different model scales much better. This example shows how looking at one metric can be deceptive and cost a lot of time and resources, e.g. by choosing the wrong candidate for scaling up.

Besides the fact that different cost indicators capture different aspects, they can be chosen to reflect either the cost of training or the cost at inference time. Both training and inference costs can be crucial, depending on the context. Also, a single cost indicator can favor a model over another during inference, but not training (or vice versa). For instance, a model that shares parameters in depth is particularly memory-efficient during inference, but during training, the size of activation that need to be kept for the backward pass is just as large as for a similar model with no parameter sharing.

The overarching conundrum here is that first of all, no single cost indicator captures a holistic view that is universally useful to all practitioners or researchers. We show that the trade-offs between cost indicators fall far from the standard assumptions and can be non-trivial to navigate. Moreover, we argue that cost indicators that one cares about strongly depend on the applications and setup in which the models are supposed to be used. For example, for an embedded application, inference speed is paramount, while for deployed recommendation systems, training cost can be extremely important as models are constantly being retrained.

The overall contributions of this paper are as follows: We call out the intrinsic difficulty of measuring model efficiency within the context of deep neural networks. We review the most common cost indicators and present the advantages and disadvantages of each and discuss why they might be insufficient as a standalone metric. While obvious, we show examples of how model efficiency might be misrepresented by incomplete reporting of these cost indicators. We characterize this problem, coin the term *'efficiency misnomer'*, and show that it is more prevalent than imagined.

We present experiments where comparing model efficiency strongly depends on the choice of cost indicator, like scenarios where there is parameter sharing, sparsity, or parallelizable operations in the model. Moreover, we briefly review some of the current common practices in the literature and discuss how existing work report comparisons of different models and analyze the efficiency of algorithms. Along with the discussion and analyses, we provide some concrete suggestions and

recommendations that we believe would help researchers and practitioners draw more accurate conclusions about the efficiency of different models.

2 A PRIMER ON COST INDICATORS

One of the main considerations in designing neural network architectures is quality-cost trade-off (Paleyes et al., 2020). In almost all cases, the more computational budget is given to a method, the better the quality of its outcome will be. To account for such a trade-off, several cost indicators are used in the literature of machine learning and its applications to showcase the efficiency of different models. These indicators take different points of view to the computational costs.

FLOPs: A widely used metric as the proxy for the computational cost of a model is the number of floating-point multiplication-and-addition operations (Johnson, 2018; Kim et al., 2021; Arnab et al., 2021; Tay et al., 2021a;b; Narayanan et al., 2021; Liu et al., 2021). Alternative to FLOPs, the number of multiply-accumulate (MAC[1]) as a single unit of operation is also used in the literature (Johnson, 2018). Reported FLOPs are usually calculated using theoretical values. Note that theoretical FLOPs ignores practical factors, like which parts of the model can be parallelized.

Number of Parameters: Number of trainable parameters is also used as an indirect indicator of computational complexity as well as memory usage (during inference) (Kim et al., 2021; Arnab et al., 2021; Tan and Le, 2019; Liu et al., 2021; Guo et al., 2020; Mahabadi et al., 2021a;b; Houlsby et al., 2019). Many research works that study the scaling law (Kaplan et al., 2020; Hernandez et al., 2021; Tay et al., 2021b), especially in the NLP domain, use the number of parameters as the primary cost indicator (Devlin et al., 2018; Raffel et al., 2019; Liu et al., 2019; Xue et al., 2021).

Speed: Speed is one the most informative indicator for comparing the efficiency of different models (So et al., 2021; Dosovitskiy et al., 2020; Arnab et al., 2021; Tay et al., 2021b; Kim et al., 2021; He et al., 2021a; Narayanan et al., 2021; Lagunas et al., 2021; Liu et al., 2021; Tay et al., 2020c;b). In some setups, when measuring speed, the cost of "pipeline" is also taken into account which better reflects the efficiency in a real-world scenario. Note that speed strongly depends on hardware and implementation, so keeping the hardware fixed or normalizing based on the amount of resources used is the key for a fair comparison. Speed is often reported in various forms:

- *Throughput* refers to the number of examples (or tokens) that are processed within a specific period of time, e.g., "examples (or tokens) per second".
- *Latency* usually refers to the inference time (forward pass) of the model given an example or batch of examples, and is usually presented as "seconds per forward pass". The main point about latency is that compared to throughput, it ignores parallelism introduced by batching examples. As an example, when processing a batch of 100 examples in 1 second, throughput is 100 examples per second, while latency is 1 second. Thus, latency is an important factor for real-time systems that require user input.
- *Wall-clock time/runtime* measures the time spent to process a fixed set of examples by the model. This is often used to measure the training cost, e.g., total training time up to convergence.
- *Pipeline bubble* is the time that computing devices are idle at the start and end of every batch (Narayanan et al., 2021), which indirectly measures the speed of the non-pipeline parts of the process.
- *Memory Access Cost (MAC)* corresponds to the number of memory accesses. It typically makes up a large portion of runtime and is the actual bottleneck when running on modern platforms with strong computational power such as GPUs and TPUs (Ma et al., 2018).

The cost indicators we discussed above present different perspectives on efficiency. However, some of these cost indicators may depend on factors that are not inherent to the design of the model, but on the hardware, the model runs on (e.g., CPU, GPU, or TPU), the framework that the model is implemented in (e.g., JAX, PyTorch, or TensorFlow), or even programming skill. These confounding factors add up to the difficulty of comparisons. For instance, theoretical FLOPs provides a hardware-independent comparison, however, it does not necessarily translate to the speed of a model as it does not capture the sequential dependencies of operations and memory accesses in the model. On the other hand, throughput and peak memory usage, which could better reflect the model's efficiency in a real-world scenario, strongly depend on the hardware and the implementation. Software support can also be a

[1]MAC may also refer to the memory access cost (Ma et al., 2018)

limiting factor in achieving the best possible hardware performance for a model. In (Barham and Isard, 2019), the authors make an excellent case for improving the programmability of software stack for modern accelerators to enable a wider class of models.

In the rest of the paper, we mainly focus on the number of parameters, FLOPs and speed since they capture most common use cases and are most commonly encountered in the literature. Appendix B presents other cost indicators that could be important depending on the use case.

2.1 Training or Inference cost?

When talking about costs, we can disentangle the cost of training and the cost of inference. Based on the estimate from NVIDIA (Leopold, 2019) and Amazon (Barr, 2019) as major cloud service providers, 80–90% of the ML workload is inference processing. Thus, more often, the cost of a model during inference is taken as the real cost with the argument that the amortized per-usage cost of training can be really small compared to the inference cost when a model is deployed to be used by many users. However, with the trend of improving the performance of models by scaling up their computational budget and/or training data, as well as the fast progress and frequency of the emergence of new models, the training cost can be seen as a relevant concern. Moreover, in some cases, we have to frequently retrain models due to, for instance, privacy requirements, or where new data is being continuously generated like in many recommender systems. The importance of inference efficiency is already clear and here, we will discuss more the importance of training efficiency as well as potential issues with reporting training costs.

The are several arguments in favor of the importance of reporting training costs and the need for models that train efficiently. For instance, from a research and development point of view, when a class of models is efficient during training, it is more likely to see improvements in their performance, due to ease of iterating on ideas around them. Moreover, when a model shows merit in terms of performance, usually some posthoc modifications can be applied to improve its inference efficiency.

Besides, the memory requirements of different models can be wildly different during training, although their inference memory consumption is comparable. For instance, different optimizers use different amounts of memory on the device, for instance, SGD-Momentum (Qian, 1999) vs SAM (Foret et al., 2020). If the success of a model is strongly tied to using an optimizer with a high memory cost, it can become a bottleneck during training when we plan to scale that model up compared to the model that uses a more memory-efficient optimizer.

Another example is a model that has a high degree of parameter sharing, which could be extremely efficient in terms of inference memory usage, but during training, the size of activation we need to keep for the backward pass is as big as a similar model with no parameter sharing. Activation and their statistics form a big portion of memory usage compared to parameter size, decreasing the "number of parameters" by parameter sharing may not lead to any significant decrease in the memory usage during training (Dehghani et al., 2018).

Gaming with training time Whilst "training time" can be a great cost indicator, it is also prone to Goodhart's Law: When used as the main metric, it can and will be gamed and lose its meaning. First of all, it is difficult to compare *architectures* with respect to training time, since training time includes the whole "recipe" and different ingredients may be better fits for different architectures. For instance, MobileNet and EfficientNets almost exclusively work well with the RMSProp optimizer (Pham, 2021).

Second, because the full training recipe is inevitably involved, the only meaningful claim to be made is achieving higher accuracy with smaller total training cost; a good example of this is Table 1 in Wightman et al. (2021). The training cost may be in terms of any of the indicators described above. However, counting training cost in terms of steps can be problematic as step count in itself is not a meaningful metric, and can be stretched arbitrarily with optimizers introducing multi-step lookahead schemes (Zhang et al., 2019).

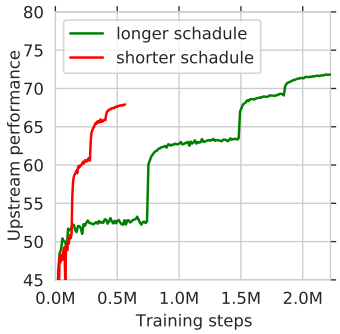


Figure 2: The learning progress of a ResNet-101 $\times$ 3 on JFT-300M with short and long schedules, obtained from (Kolesnikov et al., 2020). Decaying the learning rate too early leads to higher performance in lower steps, but the final performance is significantly worse.

Conversely, it may be tempting to claim that a method A performs *almost as well* as another method B with *dramatically reduced training cost*. Such a claim is not valid, as method A may not have been optimized for training cost, and may well just be re-tuned with that in mind and outperform method B. For example, training hyper-parameters such as learning rate and weight decay, can be tuned such that they achieve good quality quickly, but then plateau to lower points than in "slower" settings that eventually reach higher quality. This is illustrated by Figure 2, obtained from Kolesnikov et al. (2020), where the "long" training schedule (which is not optimized for training cost) for ResNet-101x3 achieves the best performance, but the "short" schedule converges significantly faster to a lower final accuracy. Another example, in the context of reinforcement learning, is the Rainbow baseline of Kaiser et al. (2019) which was subsequently re-tuned in van Hasselt et al. (2019), and shown to benefit from significantly longer training schedules too.

Some works, like the MLPerf benchmark (Mattson et al., 2019), aim at decreasing the training cost required to reach a fixed quality X with a fixed model M. While this is a good way of fixing the many moving pieces of training a deep learning model, it should be noted that due to Goodhart's law, results do not mean more than "reaching quality X with model M". Specifically, if a method A reaches quality X twice as fast as a method B while both using model M, this can neither be used to imply anything about their comparison when using model N, nor to imply anything on their comparison with the target quality X+ϵ: it could well be that method B gets to X+ϵ faster than A, or worse, A may never reach X+ϵ. Figure 3 shows another concrete example, obtained from (Kolesnikov et al., 2020) where we see faster initial convergence of a ResNet model when trained with lower weight decay, which may trick the practitioner into selecting a sub-optimal value, while using a higher weight decay leads to a slower convergence, but a better final performance.

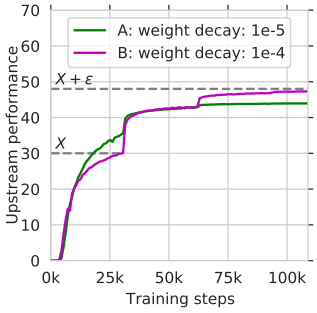


Figure 3: The learning progress of a ResNet-101 $\times$ 3 on JFT-300M with different weight decays, obtained from (Kolesnikov et al., 2020). Lower weight decay leads to acceleration of convergence, while eventually results in an under-performing final model.

A final concern is that, methods which seemingly train faster tend to be used more often and hence get optimized more over time, with the danger of getting stuck in a local minimum (Hooker, 2020; Dehghani et al., 2021b).

2.2 Potential Disagreement between cost indicators

In this section, we will discuss some of the cases in which there could be disagreement between some of the cost indicators we discussed before.

Sharing parameters When we introduce a form of parameter sharing, it is clear that compared to the same model with no parameter sharing, we end up having fewer trainable parameters, while the number of FLOPs or speed stays the same. An example is the comparison of the Universal Transformer (UT) (Dehghani et al., 2018) with vanilla Transformer (Vaswani et al., 2017) Figure 1. UT shares parameters of the model in depth, thus stays close to the frontiers of quality-number of parameters. However, when looking at the FLOPs, to maintain a similar capacity in terms of parameter count, UT requires more computation, which makes it not so efficient from the FLOPs point of view.

Introducing Sparsity Sparsity is becoming one of the main ways of both scaling up and down deep neural networks. One needs to distinguish between at least two broad classes of sparse neural networks: structured and unstructured.

Structured sparse models replace large, dense parts of a model by a collection of much smaller, still dense parts. Examples include variants as simple as replacing dense convolutions by grouped (Xie et al., 2017) or separable (Howard et al., 2017) ones, or as complicated as replacing large blocks by many smaller "expert" blocks and routing examples through the best suited ones only (Fedus et al., 2021; Riquelme et al., 2021). The latter, often called Mixture of Experts (MoE), allows growing the capacity in terms of parameter count, while keeping the computational cost small and constant. Figure 1 compares the quality-cost of Switch Transformer (Fedus et al., 2021), a MoE, to that of vanilla Transformer (Vaswani et al., 2017). While Switch falls short in terms of quality vs parameter count, it offers a great trade-off with respect to quality vs FLOPs and speed.

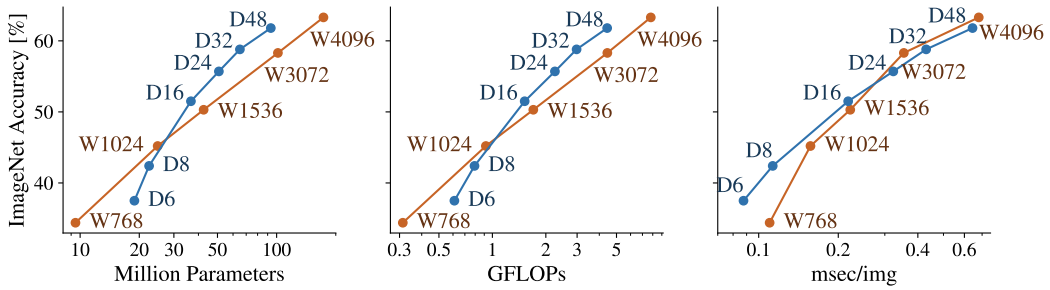


Figure 4: Comparison of scaling a small ViT in depth (D, number of encoder blocks) vs scaling it in width (W, hidden dimension). We show the cost-quality trade-off in terms of various commonly used cost indicators; which architecture appears "better" changes depending on which indicator is considered. Experiments and the computation of cost metrics were done with Scenic (Dehghani et al., 2021a), using 64 TPU-V3.

Unstructured sparse models are models where weights of dense operations are made to contain many (almost) zeros (Gale et al., 2019; Evci et al., 2020), which do not contribute to the operation's result and thus, in principle, can be skipped. This keeps the overall structure of the original model, while significantly reducing the FLOPs of the most expensive operations.

Both types of sparse models result in large reductions in theoretical FLOPs, often of several orders of magnitude. However, these do not translate to equally large speed-ups. Difficulties for structured sparse models (especially for the MoE type) include the overhead of routing, and the inability to effectively use batched operations, while for unstructured sparsity, it is not possible for the corresponding low-level operations to reach the same efficiency as their dense counterparts on current hardware, where memory access is significantly more expensive than compute (Gale et al., 2020).

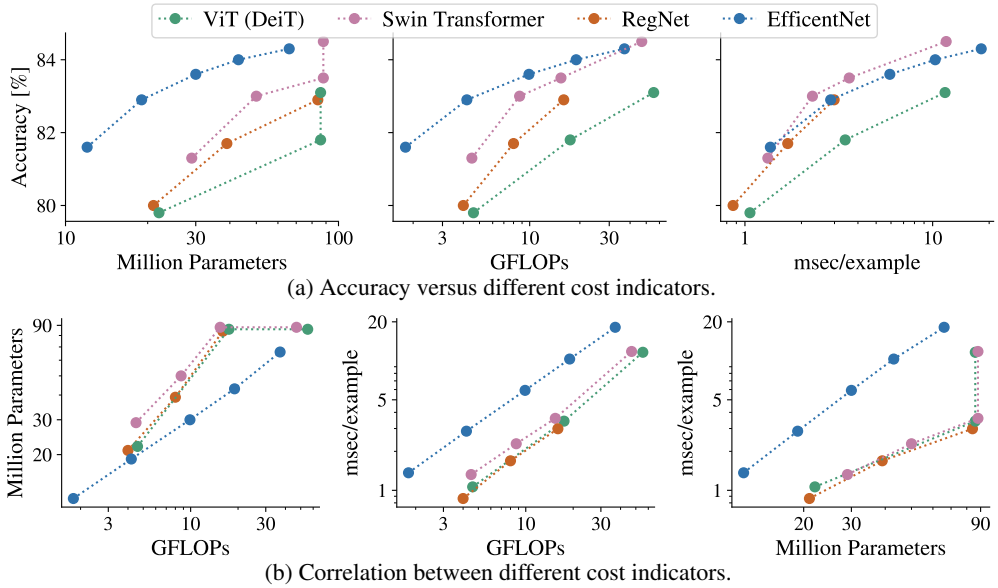
Degree of parallelism: Scaling Depth (D) vs. Scaling Width (W) When scaling up the model size, different strategies can be used. Although these different strategies may have a similar effect in terms of parameter count, and even quality, they can have different effects on the cost in terms of FLOPs and throughput. The most common knobs for scaling up models are changing depth (number of layers) and width (hidden dimension) of models (Tay et al., 2021b). To study such an effect, we ran a set of controlled experiments with Vision Transformers (Dosovitskiy et al., 2020), where we scale the width by increasing number of heads (while maintaining the hidden dimensions of heads fixed) as well as that of the FFN, and we also scale up the depth, by only increasing the number of encoder blocks. Note that when changing depth or width of the model (see Table 2 in Appendix A for the exact configurations) all other hyper-parameters are kept fixed based on the default values given by the referenced papers. Figure 4 shows the accuracy of these models with respect to different cost indicators.

In general, we can see that considering FLOPs or number parameters as the cost indicator, increasing width is beneficial for smaller budget regions, while increasing depth gives better quality with lower cost when we scale up to higher budget regions. However, this is not necessarily the case when considering speed (msec/img) as the cost indicator.

As an example, comparing D48 (a ViT with 48 layers) to W3072, (a ViT with FFN dimension 3072 and QKV dimension 768 split across 8 heads) in terms of FLOPs or number of parameters, they have more or less similar cost, suggesting the D48 as a clear Pareto efficient model[2]. However, when looking at the speed (msec/img), we observe that W3072 is not necessarily worse than D48, since there are less sequential and more parallelizable operations.

Target platform, and implementation In some cases, a certain design in the hardware may lead to discrepancies between the cost indicators. As an example, tensor factorization is used as one of the common techniques to accelerate the matrix multiplication and used for making neural network models more efficient (Zhang et al., 2015b; Jaderberg et al., 2014). Weight decomposition, regardless of its effect on the quality of the model, can reduce the number of FLOPs in neural networks by 75% (Zhang et al., 2015a). However, it has been shown that using weight decomposition on GPUs can be slower as CUDNN that is optimized for 3×3 (He et al., 2017) convolutions and they prefer single large matrix multiplication instead of several small ones. FNet (Lee-Thorp et al., 2021) also shows how certain architectures can have significantly different speed in different hardware (e.g.,

[2]A model counts as a Pareto efficient model when having lowest cost compared to all the models with similar quality, or highest quality compared to all the models with similar cost.



(a) Accuracy versus different cost indicators.

(b) Correlation between different cost indicators.

Figure 5: Accuracy and value of different cost indicators for different models on ImageNet dataset. Values for accuracies and cost indicators are obtained from (Steiner et al., 2021; Liu et al., 2021). Cost indicators are based on PyTorch implementation of the included models and the throughput is measured on a V100 GPU, using timm (Wightman, 2019).

GPUs and TPUs). When using a specific hardware and compiler, some small design choices can significantly affect the cost of a model. As an example, Zhai et al. (2021) show that for ViT, using global average pooling instead of a `CLS` as the representation of the input can significantly reduce the memory cost on TPU-V3. This is because the current TPU hardware pads the length dimension of the inputs to a multiple of 128, which may result in up to a 50% memory overhead. Another example for how implementation details can affect the efficiency is presented in (Vaswani et al., 2021), where the author discusses how carefully designed implementation of the local neighborhood gathering function for local attention might reduce memory usage while avoiding unnecessary extra computation.

3 Discussion

The most common use of cost indicators in the ML community is for (1) comparing different models in terms of efficiency using a cost indicator, (2) evaluating different models in terms of quality while fixing the cost across models to have a fair comparison, and (3) choosing models with the right trade-off for the context at hand, e.g, in architecture search. While (1) and (2) are just two sides of the same coin, the focus in the first case is on finding models with better quality, while the second case is concerned with comparing quality while keeping a certain efficiency metric constant.

3.1 On partial conclusions from incomplete comparisons

As we showed in Section 2, claiming a model is more efficient by reporting better scores on a subset of cost indicators can lead to partial, incomplete and possibly biased conclusions.

Figure 5a compares the FLOPs, parameters and run time for several models versus their accuracy on image classification task. As seen previously in Figure 1, the relative positions of different model architectures are not consistent as the performance metric is varied. For example, on one hand, EfficientNets (Tan and Le, 2019) are on the Pareto-frontier of and accuracy-parameter and accuracy-FLOPs trade-offs. On the other hand, the accuracy-throughput curve is dominated by SwinTransformers (Liu et al., 2021). As such, there is no model that is clearly more efficient here and conclusions may differ strongly depending on which efficiency metric is employed.

Note that when comparing variants of a single model family, most of cost indicators correlate. Hence, it could be sufficient to consider one of these cost indicators [3]. However, comparisons using a

[3]Note that this is not always the case, as we have observed that scaling depth vs depth of a single model can lead to different observations with respect to different cost indicators.

single cost indicator *between* models with completely different architectures are more complicated. Figure 5b shows the correlation between the cost indicators.

For example, we can see that for a similar number of FLOPs, EfficientNet has fewer parameters than other model families such as RegNet and SwinTransformer. Conversely, for a similar number of FLOPs, EfficientNet has slower throughput (i.e., higher msec/examples) than RegNet and SwinTransformer. These differences are not surprising, given that we are comparing transformer-based (Dosovitskiy et al., 2020; Liu et al., 2021) and convolution-based architectures (Tan and Le, 2019; Radosavovic et al., 2020). Moreover, some of these models are found via architecture search when optimizing for specific cost-indicators (e.g., EfficientNets are optimized for FLOPs).

3.2 Making fair comparisons through an efficiency lens

Efficiency metrics and cost indicators are often used to ground comparisons between two or more models or methods (e.g., a proposed model and baselines). By keeping one or more cost indicators fixed, one would often list models side by side in order to make comparisons fair. There are two main strategies here, namely (1) *parameter-matched* comparisons, where configuration of all models are chosen to have similar number of trainable parameters; and/or (2) *flop/compute matched* comparisons, where all models have similar computational budget, e.g., configurations are chosen to have similar FLOPs. Whilst there is no straightforward answer to which strategy is the better one, we highlight two case studies and discuss implications and general recommendations.

3.2.1 The issues with parameter matched comparisons

We delve into some of the potential issues with the parameter-matched comparisons. The gist of many of these issues is that *not every parameter is created equal*, causing many intricate complexities that might complicate fair comparisons among models. Here we present two example situations where parameter matched comparison could go wrong.

Token-Free Models Token-free models (Xue et al., 2021; Tay et al., 2021c; Jaegle et al., 2021) get rid of the large subword vocabulary by modeling at the character or byte level. A large number of parameters originating from the embedding matrix is therefore dropped when transiting to token-free models. Hence, it remains an open question of how to fairly compare these class of models with their subword counterparts. ByT5 (Xue et al., 2021) proposed to up-scale the Transformer stack in order to compensate for lost parameters in the embedding layer. While this argument was made in the spirit of 'fairness' (e.g., comparing both methods at a parameter-matched setup), the up-scaling causes a *substantial* reduction in model speed. This is largely because parameters in the embedding matrix typically do not incur much computation while increased parameters in other parts of the Transformer stack (more depth/width) lead to a relatively substantial compute cost. What is referred to as *base* or *small* size here refers to a model that is significantly more computationally costly (in terms of speed, throughput and FLOPs) than their other counterparts and hence the naming alone can be very misleading to practitioners. To make this fair, ideally, the authors would have to present results that are both parameter matched *and* compute-matched to present a clearer picture.

Sparse Models and Mixture-of-Experts A defining feature of sparse models (Lample et al., 2019; Fedus et al., 2021) is that they remain compute-matched while enabling scaling to a large number of parameters. Hence, parameter matched comparisons do not make sense for sparse models and parameter-matching sparse models can be seen as an unfair method of unnecessarily downplaying the strengths of sparse models. This has been also shown in Figure 1, where comparing a switch transformer with a vanilla transformer with the same number of parameters always goes in favor of vanilla transformer. Note that many works in the literature employ compute-matched comparisons for comparing sparse models (Narang et al., 2021; Fedus et al., 2021; Lample et al., 2019).

Vision Transformers and Sequence Length Models with a flexible sequence length, such as ViTs when varying patch-size, have the opposite property of sparse models: one can instantiate architectures with significantly different computational cost (e.g., FLOPs) and speed while keeping the parameter count fixed. Thus, this type of model should not be compared based on parameter count either. Table 1 presents the parameter size, FLOPs, and speed of ViT using different patch sizes. We can see inverse corelation between parameter size and FLOPs/speed. Model with bigger patch sizes have less FLOPs and are faster, while having similar or slightly more parameters.

Model	Input sequence length	Million parameters	GFLOPs	msec/example
`ViT-B/8`	785	86.5	78.54	7.17
`ViT-B/16`	197	86.6	17.63	1.30
`ViT-B/32`	50	88.2	4.42	0.39
`ViT-B/64`	17	95.3	0.93	0.11

Table 1: Parameters size, FLOPs, and speed of ViT-Base with different patch sizes (i.e., 8×8, 16×16, 32×32, and 64×64). Models are fed by input images of size $224 \times 224 \times 3$, thus the input sequence length is $(224/\text{patch_size})^2 + 1$ (for the `CLS` token). Numbers in the table are reported using the code in Scenic (Dehghani et al., 2021a) when running on 64 TPU-V3.

3.2.2 The issues with Compute Matched Comparisons

We have previously discussed how parameter matched comparisons may be unfair. This section discusses a case where compute matched comparisons may raise concerns. Consider a scenario where the proposed method achieves compute saving by an architectural design that does not influence model parameters at all (e.g., downsampling sequences). A seemingly fair comparison here would be to take a standard model and remove layers or hidden dimensions until both models are compute matched. However, this comparison runs the risk of a baseline model that is handicapped by significantly insufficient model capacity in terms of parameter count and therefore, substantially underperform the proposed method. This is evident in the Perceiver IO work (Jaegle et al., 2021) where the baseline BERT is shrunk to a mere 20M parameters (compared to 425M in the proposed approach) in a compute matched comparison setup. Moreover, the baseline BERT is also handicapped in terms of depth (6 layers vs 40 layers) which is shown in (Tay et al., 2021b) to be unfavorable. Note that depth is not taken into account for FLOP-matched comparisons and if the authors were to account for speed-match, then the baseline here would be substantially faster than the proposed method. Overall, we do note that making fair compute matched comparisons is clearly nontrivial and a challenging problem. However, our recommendation is that when there is just no easy way to make a fair comparison, we encourage authors to make the best effort in finding a best *'matched'* setup and show multiple alternatives if possible.

3.3 Cost indicators for architecture search

Aside from being used for comparing models, many architecture search studies add a cost indicator to the loss function as a resource constraint (He et al., 2021b) and account for efficiency. To this end, these studies mostly use the parameter size (Pham et al., 2018), FLOPs (Hsu et al., 2018), memory access cost (MAC) (Ma et al., 2018), or real latency (Tan et al., 2019; Wu et al., 2019). Given that we have shown earlier how cost indicators may disagree with one another or lead to impartial conclusions, we suggest that practitioners place extra caution in choosing cost indicators for architecture search algorithms especially given its computational cost and sensitivity to the selected cost indicator. Here, making assumptions that cost indicators are always interchangeable runs the risk of conducting a massive search for finding models that are impractical.

4 Suggestions and Conclusion

A lot of recent work has focused on comparing different model architectures on the basis of a cost indicator like parameter count or FLOPs. We have demonstrated in this paper that using any cost indicator alone can be misleading, with parameter count being the most problematic one. Oftentimes, parameter count is used to imply "model capacity"; however, when varying model architecture in any nontrivial way, this is wrong. Correctly estimating and comparing capacity across model architectures is an open research problem.

Since each indicator stands for something different and comes with its own pros and cons, we suggest always reporting *and plotting curves* using all available cost indicators, and refraining from highlighting results using just a single one.

Moreover, given that it is basically impractical to provide a holistic report of all cost metrics (for instance, runtime on all possible hardwares) narrowing down the efficiency claims to the exact setup that models are evaluated and avoiding overgeneralized conclusions in comparisons can already provide a much more clear picture to the community.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

Paul Barham and Michael Isard. Machine learning systems are stuck in a rut. In *Proceedings of the Workshop on Hot Topics in Operating Systems*, pages 177–183, 2019.

Jefff Barr. Amazon ec2 update. https://aws.amazon.com/blogs/aws/amazon-ec2-update-inf1-instances-with-aws-inferentia-chips-for-high-performance-cost-effective-inferencing/, 2019. Accessed: 2021-6-1.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. *arXiv preprint arXiv:2109.12948*, 2021.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, and Yi Tay. Scenic: A JAX library for computer vision research and beyond. *arXiv preprint arXiv:2110.11403*, 2021a.

Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021b.

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.

Piotr Dollár, Mannat Singh, and Ross Girshick. Fast and accurate model scaling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 924–932, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2943–2952. PMLR, 2020. URL http://proceedings.mlr.press/v119/evci20a.html.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574, 2019. URL http://arxiv.org/abs/1902.09574.

Trevor Gale, Matei Zaharia, Cliff Young, and Erich Elsen. Sparse GPU kernels for deep learning. In Christine Cuicchi, Irene Qualters, and William T. Kramer, editors, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 17. IEEE/ACM, 2020. doi: 10.1109/SC41405.2020.00021. URL https://doi.org/10.1109/SC41405.2020.00021.

Demi Guo, Alexander M Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*, 2020.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. Efficient nearest neighbor language models. *arXiv preprint arXiv:2109.04212*, 2021a.

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.

Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021b.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397, 2017.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Sara Hooker. The hardware lottery. *arXiv preprint arXiv:2009.06489*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Chi-Hung Hsu, Shu-Huan Chang, Jhao-Hong Liang, Hsin-Ping Chou, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Monas: Multi-objective neural architecture search using reinforcement learning. *arXiv preprint arXiv:1806.10332*, 2018.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.243. URL https://doi.org/10.1109/CVPR.2017.243.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*, 2014.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Jeff Johnson. Rethinking floating point for deep learning. *arXiv preprint arXiv:1811.01721*, 2018.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model-based reinforcement learning for atari. *CoRR*, abs/1903.00374, 2019. URL http://arxiv.org/abs/1903.00374.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. Scalable and efficient moe training for multitask multilingual models. *arXiv preprint arXiv:2109.10465*, 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. Block pruning for faster transformers. *arXiv preprint arXiv:2109.04838*, 2021.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Large memory layers with product keys. *arXiv preprint arXiv:1907.05242*, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021.

George Leopold. Aws to offer nvidia's t4 gpus for ai inferencing. www.hpcwire.com/2019/03/19/aws-upgrades-its-gpu-backed-ai-inference-platform/, 2019. Accessed: 2021-6-1.

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.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *arXiv preprint arXiv:2103.14030*, 2021.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *arXiv preprint arXiv:2106.04647*, 2021a.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021b.

Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David A. Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debojyoti Dutta, Udit Gupta, Kim M. Hazelwood, Andrew Hock, Xinyuan Huang, Bill Jia, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Guokai Ma, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. Mlperf training benchmark. *CoRR*, abs/1910.01500, 2019. URL http://arxiv.org/abs/1910.01500.

Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Delight: Deep and light-weight transformer. *arXiv preprint arXiv:2008.00623*, 2020.

Gaurav Menghani. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *arXiv preprint arXiv:2106.08962*, 2021.

Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*, 2021.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. Efficient large-scale language model training on gpu clusters. *arXiv preprint arXiv:2104.04473*, 2021.

Andrei Paleyes, Raoul-Gabriel Urma, and Neil D Lawrence. Challenges in deploying machine learning: a survey of case studies. *arXiv preprint arXiv:2011.09926*, 2020.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

Hieu Pham. MS Windows NT kernel description. https://web.archive.org/web/20211005122822/https://twitter.com/hieupham789/status/1371891950111006720?s=20, 2021. Accessed: 2021-05-10.

Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*, pages 4095–4104. PMLR, 2018.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1): 145–151, 1999.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *arXiv preprint arXiv:2106.05974*, 2021.

Or Sharir, Barak Peleg, and Yoav Shoham. The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*, 2020.

Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, et al. Mesh-tensorflow: Deep learning for supercomputers. In *Advances in Neural Information Processing Systems*, pages 10414–10423, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

David R So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. Primer: Searching for efficient transformers for language modeling. *arXiv preprint arXiv:2109.08668*, 2021.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298594. URL https://doi.org/10.1109/CVPR.2015.7298594.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention in transformer models. *arXiv preprint arXiv:2005.00743*, 2020a.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020b.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*, 2020c.

Yi Tay, Mostafa Dehghani, Vamsi Aribandi, Jai Gupta, Philip Pham, Zhen Qin, Dara Bahri, Da-Cheng Juan, and Donald Metzler. Omninet: Omnidirectional representations from transformers. *arXiv preprint arXiv:2103.01075*, 2021a.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021b.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*, 2021c.

Hado van Hasselt, Matteo Hessel, and John Aslanides. When to use parametric models in reinforcement learning? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14322–14333, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1b742ae215adf18b75449c6e272fd92d-Abstract.html.

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.

Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904, 2021.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.

Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv preprint arXiv:2105.13626*, 2021.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

Michael R. Zhang, James Lucas, Jimmy Ba, and Geoffrey E. Hinton. Lookahead optimizer: k steps forward, 1 step back. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9593–9604, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/90fd4f88f588ae64038134f1eeaa023f-Abstract.html.

Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun. Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):1943–1955, 2015a.

Xiangyu Zhang, Jianhua Zou, Xiang Ming, Kaiming He, and Jian Sun. Efficient and accurate approximations of nonlinear convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and pattern Recognition*, pages 1984–1992, 2015b.

Appendix

A EXPERIMENTAL SETUP: SCALING DEPTH VS. SCALING WIDTH

Detailed configurations and scores for the experiments on scaling Vision Transformer, by increasing the depth vs increasing the width of the model.

	Configuration				Cost			Quality
Model	**Num layers**	**FFN dim**	**QKV dim**	**Num heads**	**Million Parameters**	**GFOPs**	**msec/img**	**ImageNet Accuracy**
D6	6	1024	384	6	18.89	0.61	0.09	37.5
D8	8	1024	384	6	22.44	0.79	0.11	42.4
D16	16	1024	384	6	36.63	1.52	0.22	51.5
D24	24	1024	384	6	50.83	2.25	0.32	55.7
D32	32	1024	384	6	65.03	2.98	0.43	58.8
D48	48	1024	384	6	93.42	4.43	0.64	61.8
W768	12	768	192	3	9.47	0.31	0.11	34.4
W1024	12	1024	384	6	24.81	0.92	0.16	45.2
W1536	12	1536	512	8	42.51	1.70	0.22	50.30
W3072	12	3072	768	12	101.52	4.44	0.35	58.3
W4096	12	4096	1024	16	173.10	7.80	0.68	63.3

Table 2: Detailed configuration as well as cost versus quality scores for experiments on scaling depth or width of Vision Transformer.

B ADDITIONAL COST INDICATORS

Depending on the use case, there are other cost indicators that may play a key role in determining the efficiency of a model:

Sample efficiency that can be expressed as training data size, e.g. number of data points (or tokens for language models Kaplan et al. (2020)) that a model needs in order to reach a reasonable performance. Being sample efficient depends on many factors, like inductive biases of the model or the training curriculum.

Carbon footprint of a model, during training or inference that is a proxy of the environmental impact. Quantifying this cost directly is not straightforward, but (Strubell et al., 2019) proposed the approximate environmental costs of an ML model as Footprint $= (ee_{\text{train}} + \text{queries} \times ee_{\text{inference}}) \times \text{CO2e}_{\text{datacenter}}/\text{KWh}$, where, ee indicate the electrical energy (Patterson et al., 2021).

Monetary , i.e., the expenses for training or serving models, which is reported in the form of figures (Sharir et al., 2020). As an example, the figures for training can be computed as total-train-time $\times$ total number of chips $\times$ prices offered by cloud solutions[4]. This cost indicator is more often used in business documents, than in scientific articles.

number of model activations , proposed by Dollár et al. (2021) as a complexity metric. Model activations refers to the number of elements in the output tensors from the building blocks of the model, e.g., output size of convolutional layers in ResNet. Dollár et al. (2021) argue that the number of activations is a reliable predictor for the model runtime and showed that compared to FLOPs, it is more strongly correlated with the runtime on memory-bandwidth limited hardware.

Memory consumption is one of the main dimensions of efficiency and it has been used as a cost indicator in various research works (Bondarenko et al., 2021; Dosovitskiy et al., 2020; Kondratyuk et al., 2021). Reporting memory footprint often is done in form of "peak memory usage", during training that takes into account the memory consumption by the model, optimizer, and the pipeline. Another form that is also common is comparing the maximum batch size that fits on a specific device for a specific task. Unlike model activation size, memory footprint depends on the hardware and implementation. It has a direct implication of whether a model can fit on a device, which is a hard constraint in some cases.

[4]Note that the prices largely depend on the price of electricity.