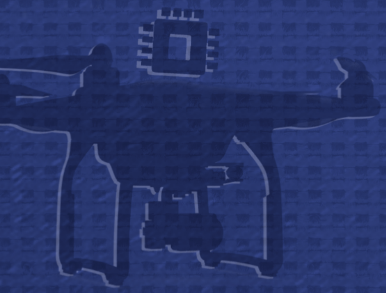
Issue Brief

Onboard AI: Constraints and Limitations



Authors
Kyle A. Miller
Andrew J. Lohn

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

August 2023

Executive Summary

AI can achieve remarkable performance under ideal conditions that are difficult to replicate in many real-world settings. The AI that often captures headlines typically runs under these conditions, in well-maintained data centers with an abundant supply of compute and power. Currently, most top-performing AI models designed for vision and language applications rely on these abundant resources. However, these resources are highly constrained on many systems in the real world, be it drones, satellites, or ground vehicles.

This is the challenge of ‘onboard AI’: running AI directly on a device or system without additional backend compute support. There are times when running models onboard is optimal or necessary, and doing so can bring a range of advantages. However, onboard computing constraints can introduce significant limitations, or completely inhibit the use of certain models on some systems. This creates a gap between the highest-performing AI systems and those deployed in the real world, which has implications for the performance and robustness of many sought-after applications.

Onboard AI systems are constrained for several reasons, but the primary factor is processing speed. The highest-performing models execute extremely large numbers of computations for each output they produce. These calculations require high-performance processors, often many of them. However, because of their size and power demands, such processors cannot be used in various systems. Practically, this means chips designed for onboard use do orders of magnitude fewer calculations and cannot run AI models quickly enough for many applications.

Onboard AI systems also need substantial working memory. Data center chips have the memory to hold large models, store the results of ongoing calculations, and enable fast communications both on the chip and between chips to split the calculations across several devices. However, many devices are not designed for large-scale computations or equipped with large working memories.

These constraints are influenced by the size, weight, and power limitations of many systems. Most state-of-the-art chips use far more power than what is available on small-footprint devices. Powerful chips require larger and heavier batteries that are infeasible for lightweight systems such as small drones, in addition to the chip’s packaging, which can increase their weight by a factor of ten.[1]

Finally, real-world applications might have to sacrifice computing capabilities for a host of other reasons, such as radiation hardness and temperature sensitivity. Moreover, chips age and become out of date if they operate for many years, which can make them ill-equipped to process contemporary models.

Stakeholders across government and industry should understand that these constraints cannot always be resolved, given current technologies and platform limitations. Engineers can mitigate some of them, such as by using different algorithms that are less resource-intensive but still have acceptable performance. However, in many cases, onboard AI will be inferior to the state-of-the-art models that grab headlines or achieve high-level performance on benchmarks. In some high-risk contexts, the use of AI onboard systems may be inappropriate or require additional safeguards.

Table of Contents

Introduction

NASA’s two-billion-dollar Perseverance rover touched down on Mars in 2021, hosting a suite of advanced tools, including small AI models — but its central onboard computer, or “brain,” runs on re-engineered 1990s processors with less compute than a smartphone. Moreover, these chips had to be hardened to survive radiation from the harsh space environment, increasing their price to nearly $200,000. It’s not that NASA is behind the times; rather, they need computers that survive cosmic radiation, do not overheat in the vacuum of space, and do not use all of the rover’s power. This means most of the commercial chips underlying cutting-edge AI in robotics, navigation, and image processing will not be coming soon to a rover on Mars.[2]

While systems on Earth may not struggle with cosmic radiation, there are many other constraints that must be managed. Researchers developing new AI models use supercomputers in labs or on the cloud, but soldiers on a battlefield or sensors in a farmer’s field will not have supercomputers nearby, and they often cannot communicate effectively with one over the Internet.

This report highlights how constraints can create a gap between the AI that sets performance records and the AI implemented in the real world. We begin with a brief explanation of why one would run AI onboard a device, as opposed to a cloud or data center. Part two overviews constraints that can inhibit models and compute hardware from running onboard. Part three investigates three case studies to illuminate how these constraints impact AI performance: computer vision models on drones, satellites, and autonomous vehicles. These case studies are only meant to elucidate the constraints on various systems, and are not meant to be a comprehensive assessment of constraints across all or most systems that could use onboard AI. Part four provides a broad assessment of trends based on findings from the case studies, and considers how they might impact onboard AI functionality in the future. Part five concludes with recommendations to better manage the constraints of onboard AI.

Why Run AI Onboard?

Running AI onboard is optimal — or even necessary — for many applications, and brings a range of advantages over running models on remote processors. Onboard processing can be faster because of reduced communication delays; it promotes reliability and security, because it does not require communication with other devices; and it enables a greater degree of privacy because all data remains onboard.[3]

Moreover, onboard models have mobility on systems such as aircraft, ground vehicles, and ‘smart’ munitions.[4]

Some applications require the speed and reliability of onboard AI. For example, robot-assisted surgery systems can malfunction if network connections are disrupted,[5] which can have life-threatening consequences. Models on autonomous vehicles must run local inference[6] continuously to respond to rapid changes on the road. This is unworkable if models are processed in a distant data center, as it introduces latency and slows the vehicle’s navigation speed.

There are similar concerns for the military application of AI in contested environments, where adversaries can degrade communications and threaten AI speed, security, and reliability. If a hub, like a mobile data center running inference for several connected devices, is destroyed or loses communications, then any device dependent on it cannot run AI-related functions.[7] Further, wireless communication with a hub can be intercepted, manipulated, or expose a soldier’s location.

Privacy concerns, such as disclosing user data, can also incentivize individuals and companies to run models locally rather than send data to the cloud. For example, many iPhones have onboard facial recognition to confirm a user’s identity. The model runs onboard due to Apple’s data privacy policies, which involve encrypting user data sent to the cloud. Cloud-based facial recognition cannot be used in this case, so Apple engineers had to find ways to run it directly on iPhones.[8]

This report focuses on the constraints and limitations of onboard AI, and we investigate case studies of computer vision models running on edge devices — but we must stress that there are many instances where onboard AI does not substantially limit performance. For example, in Appendix D we investigate how Google offers lightweight machine translation models that can run locally on smartphones, with similar performance to their cloud translators (which run in data centers).

Constraints of Onboard AI

The constraints of onboard AI are diverse and depend upon the characteristics of the device, the types of applications and models executing onboard, the environment in which the device operates, and the financial costs to purchase and configure hardware to enable AI functionality.[9]

The first-order constraints depend on the device: how much compute and memory can be physically placed onboard to run AI models, given the device's size, weight, and power. This also includes any auxiliary compute needed to run non-AI functions, which further increases onboard resource requirements. In general, small, low-power devices will be more constrained than larger, higher-power devices.

Other constraints arise based on the application or model onboard the device. For example, large language models are a particularly computationally-demanding type of AI and can be impossible to run on some devices. One such model, OPT-175B, is partitioned across 16 high-performance Graphics Processing Units (GPUs) — far more compute than is available on the edge devices explored in this report.[10] Even the smallest among the 'LLaMa' language models, which were all designed to be small and efficient, exhausts the capacity of many high-performance data center GPUs.[11] AI models for other tasks, like those in computer vision, are typically smaller and less resource-intensive than large language models, but they too often exceed the capacity of well-provisioned AI chips. Onboard chips are usually far less capable for many reasons, so the models that run on them are often constrained to lower performance.

Environmental characteristics such as extreme temperature or radiation can also be constraining, as these conditions cause processors to malfunction or degrade. This can be problematic if the onboard chips are used for critical functions or long-duration deployments. Lastly, financial constraints can lead manufacturers to use cheaper chips or reduce maintenance, which can reduce onboard compute capacity.

Table 1 outlines the constraints that can inhibit AI and compute hardware from functioning onboard a device, and potentially lower the threshold of AI performance. This is not a comprehensive assessment, but rather an overview of common variables that impact compute, memory, and AI functionality.

Table 1. Constraints of Onboard AI

Device-Dependent Constraints	
Constraint	Impact
Compute	A device must be able to perform enough calculations per second to run AI models (and other processes) in a reasonable amount of time.[12]
Memory	Models require working memory to temporarily store and retrieve information onboard a device. Memory can impact model speed, power consumption, and overall functionality.[13]
Storage	Models must be permanently stored within a computer system. Insufficient storage onboard a device can restrict the choice of AI models.
Power	Each calculation or movement of data takes energy. High-performance hardware running large models can outstrip onboard power sources.
Size and Weight	Processors are small but typically require additional components that can exceed size and weight restrictions of many systems, such as cooling, cards, and batteries.[14]
Auxiliary Compute	Additional compute required to run non-AI functions increases the resources required onboard a device.[15]
AI Model, Task, and Application-Dependent Constraints	
Constraint	Impact
Model Size	Models with a large number of parameters generally require more compute and memory to function, and they often run slower.
Model Parameter Precision	Data and model parameters are numbers that can be represented with more or fewer bits. Using fewer bits reduces compute and memory requirements, but can also reduce performance.[16]
Model Architecture	The connectivity of parameters in a neural network influences model compute and memory requirements, as well as speed.[17]
Task Input Data	Applications that need high-resolution inputs or large amounts of data per input can demand untenable amounts of compute and memory.[18]
Application Speed	This refers to how quickly a model must ingest data and run inference for a particular application. Some models run too slowly for certain applications.[19]
Application Pre-Processing	There can be extensive computing required to reformat input data to match the model and application.[20]
Application Model Quantity	Applications that use several AI models need to share limited resources on a device.[21]
Application Safety	AI applications with safety implications must be robust and reliable, and sometimes require additional components like backup systems.[22]

Environmental Constraints	
Constraint	Impact
Accessibility	Models may be constrained by hardware that is outdated or degraded if it cannot be physically accessed, maintained, or replaced.[23]
Environmental Characteristics	Environments with extreme temperatures, humidity, or radiation can make compute hardware malfunction. Hardware designed for such conditions tends to be lower performance, which can constrain AI models.
Financial Constraints	
Constraint	Impact
Compute Configuration	It can be costly to configure compute hardware to function on a particular device or run a particular type of model.[24]
Compute Purchase and Maintenance	Purchasing, updating, repairing, or replacing compute hardware can be expensive, leading to models that can be constrained to low-performance hardware (to reduce costs).

Source: CSET.

These constraints often necessitate the use of smaller models that run faster and use less compute and memory, or the reduction of a model's size or mathematical precision.[25] But as we illustrate in the following case studies, both of these options can reduce the threshold of AI performance.

Case Studies

Here we investigate three case studies of onboard AI: object detection and real-time object detection models on drones and autonomous vehicles, and image classification models on satellites. We gauge how compute constraints would influence AI functionality and performance, and determine if top-performing models for each task can function within the constraints of the devices and environments.

We selected these cases because there is sufficient open-source information to judge if and how specific AI models can function locally. The examples encapsulate a range of systems operating in different environments (i.e., air, land, and space), and the systems themselves have dual-use characteristics that are relevant to both civilian and military activities. The methodology to select and assess these case studies can be found in Appendix A. Lastly, we also provide a case study on machine translation in Appendix D; however, because of data limitations we could not apply the same methodology as we did for the other case studies.

Drones: Object Detection

Object detection on drones has immense potential for military and civilian operations. Onboard models can improve surveillance and reconnaissance in contested environments where communications are degraded, and may enable drones to loiter and automatically track targets. This potential has spurred substantial R&D and investment,[26] and there is a range of programs actively developing the technology for future military and civilian operations.[27]

Performing object detection onboard can involve a number of tasks, including classifying the object and generating a bounding box around it (i.e., localizing where an object is within an image), indicating the confidence of the detection, and possibly transmitting coordinates to an operator.

However, the detections may not be reliable, as currently the best object detection model only reaches about 65% mean average precision[28] (herein referred to as 'precision') on the most popular benchmark.[29] The boxes may surround the object perfectly, partially, or not at all. Even if the boxes are accurate, the classification of the target may be low-confidence or completely incorrect. Such misdetections can be catastrophic if the AI is relied upon in a combat operation.

Currently, open-source object detection models are not accurate, reliable, or robust enough to use in critical operations like automatic targeting (even with fine-tuning).[30] Moreover, the best model performance of 65% precision may not translate to the real world. That score was achieved in an ideal lab setting using a catered test dataset, not a dynamic environment in which the model makes detections on a mobile drone. For example, when researchers tested models against a custom object recognition benchmark that reflected real-world imagery, they found a 40–45% drop in performance compared to performance on other benchmarks.[31] Using models in new environments and in ways that are 'out of distribution,' where the imagery inputs do not reflect the data on which the model was trained, can degrade performance.[32] These problems can arise when models are deployed onboard mobile systems that operate in changing environments, and where the models can be harder to monitor and update.

To illustrate AI performance degradation onboard devices, we assess whether top-performing object detection models could function locally on small and medium-sized drones. While there are many classifications of drones based on size, weight, flight time, etc., we focus on size and weight. We use small and medium-sized classifications for any drone with a maximum takeoff weight between 250 grams and 55 pounds. These fall within the DoD's 'Group 1' and 'Group 2' classifications of Uncrewed Aerial Systems (UAS).[33] We assess small and medium-sized drones separately because their physical differences can impact AI constraints.

Small Drones

Small drones[34] are increasingly common in today's battlefield, but it is unclear to what extent they can be enabled by object detection because the best AI models for this purpose are large and compute-intensive.[35] The GPU boards[36] that typically run these models are too large for the drones to carry and consume too much power for the drones' batteries to sustain, and therefore cannot be used to run models locally. Practically, this means that users may need to accept poorer performance than anticipated, or that small drones may not be suitable for certain tasks despite compelling laboratory demonstrations.

Figure 1. Size, Weight, and Power of Small Drones vs. a High-End GPU

Small Drones

Drone Weight	Estimated Payload	Size	Power
0–20 lb.	~0–4 lb.	~12–24" L/W/H	Small Battery

High-End GPU (V100 Board)

GPU Weight	Size	Power
2.6 lb.	10.5" L 4.3" W 2.5" H	Outlet or Large Battery

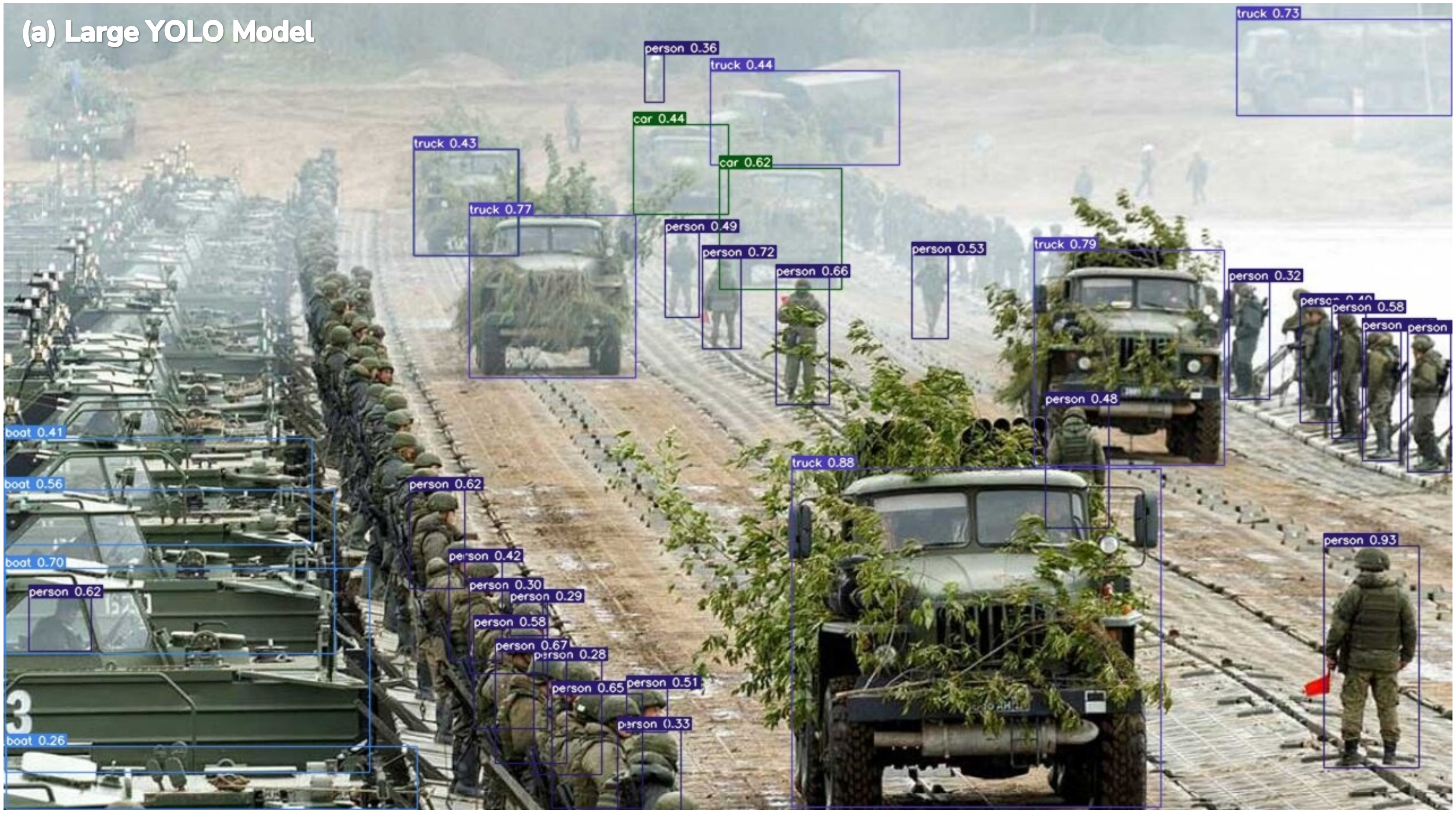
Note: We only provide a rough estimation of the payload capacity of small drones because there are many different variables that impact how much payload a small drone can carry, in addition to other onboard components such as cameras.
Sources: See Endnotes.[37]

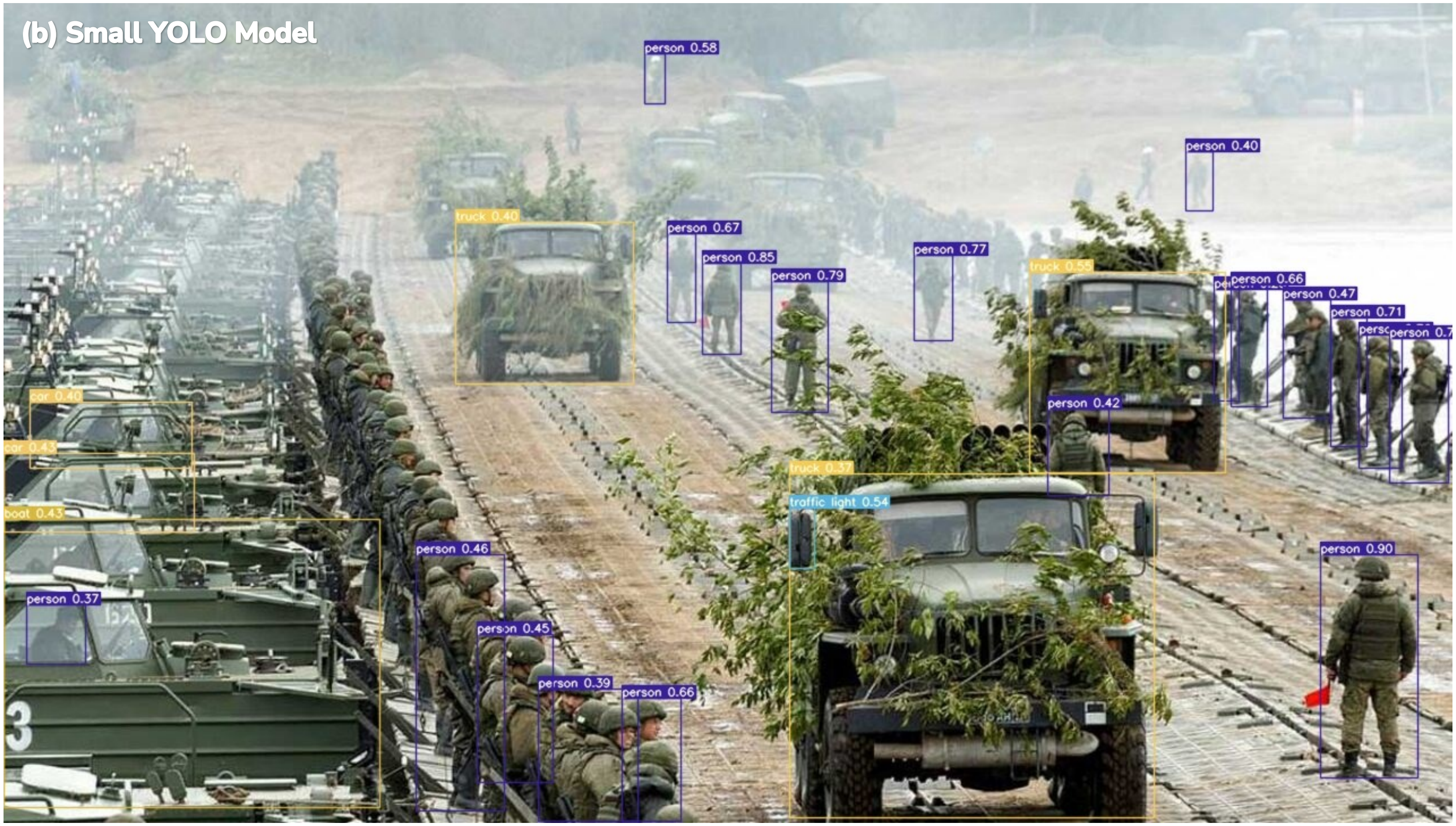
We assessed several object detection models to compare the compute capacity of small drones against model sizes, performance, and compute requirements. For example, the largest model in the 'YOLOv7' family of object detectors achieves a moderate performance of about 57% precision.[38] However, the model runs on power-hungry GPUs that exceed the size and battery capacity of small drones. Therefore, the model will not function effectively onboard such devices. Conversely, the smallest

YOLOv7 model could run on small drones using specialized low-power chips because it is 25 times smaller and runs with 60 times fewer calculations. However, it only achieves 38.7% precision, which is about 35% worse than the larger model. See Appendix B for data on the compute hardware examined for this report, which includes both conventional and low-power ‘edge’ processors.

Comparing the two models’ detections in Figure 2 illustrates a significant performance disparity: the larger model (top) generates more accurate bounding boxes and classifications, often with higher confidence, while the smaller model (bottom) has more misdetections and lower confidence. Constraints on small drones force the use of smaller models, reducing the threshold of AI performance to below the state-of-the-art. And even if the large model could function on a small drone, its performance of 60% precision may still be unacceptably poor. Note these detections are from standard YOLOv7 models and fine-tuning on more data can improve performance, but performance disparities between the large and small models will likely persist even after fine-tuning.[39]

Figure 2. Object Detection with the Large YOLO Model (a) and Small YOLO Model (b)





Notes: We ran the YOLOv7-E6E (large) and YOLOv7-Tiny (small) models on this image with a resolution of 1280 and minimum detection confidence threshold of 0.25 (i.e., it only displays detections when it has >25% confidence that it is correct). This image was selected because it contains object classes (people and trucks) that the YOLOv7 models were pre-trained to identify. Such models can only identify objects that they were pre-trained or fine-tuned to identify. Models were not fine-tuned.

Image Source: Amer Ababakr, “Mediation and the Way Forward to End the Ukraine War,” *Modern Diplomacy*, November 2022, https://moderndiplomacy.eu/2022/11/21/mediation-and-the-way-forward-to-end-the-ukraine-war/.

Medium-Sized Drones

Medium-sized drones can generally contain more onboard compute to run better-performing models, but powering the chips can still be challenging. Power remains a constraint if the drones use batteries, which cannot run data center GPUs. Integrating hardware can also be challenging; one cannot simply plug new processors into the drone. For example, captured Russian Orlan-10 UASs were disassembled, revealing that they use computer-on-modules (COM) with extremely low compute capacity.[40] Most AI models would not run onboard the Orlan-10 unless the processor was replaced, but this would likely require a degree of system reengineering.[41]

Figure 3. Size, Weight, and Power of Medium-Sized Drones vs. a High-End GPU

Medium-Sized Drones

Drone Weight	**Estimated Payload**	**Size**	**Power**
21–55 lb.	~0–15 lb.	>~24" L/W/H	Large Battery or Engine

High-End GPU (A100 Board)

GPU Weight	**Size**	**Power**
2.58 lb.	10.5" L 4.4" W 4.3" H	Outlet or Large Battery

Note: We only provide a rough estimation of the payload capacity of medium-sized drones because there are many different variables that impact how much payload a medium drone can carry (in addition to other onboard components such as cameras).
Sources: See Endnotes.[42]

Ultimately, small and medium-sized drones are likely too constrained by hardware to run state-of-the-art models. That may be acceptable in some cases, such as supporting a drone operator without any automation of tasks. However, it may be unacceptable in other use cases, such as using an object detector to identify enemy forces without direct human oversight. Stakeholders must not only take into consideration the degraded performance of onboard AI, but also establish clear criteria and minimum AI performance standards for different use cases and applications.

Satellites: Image Classification

Small Satellites

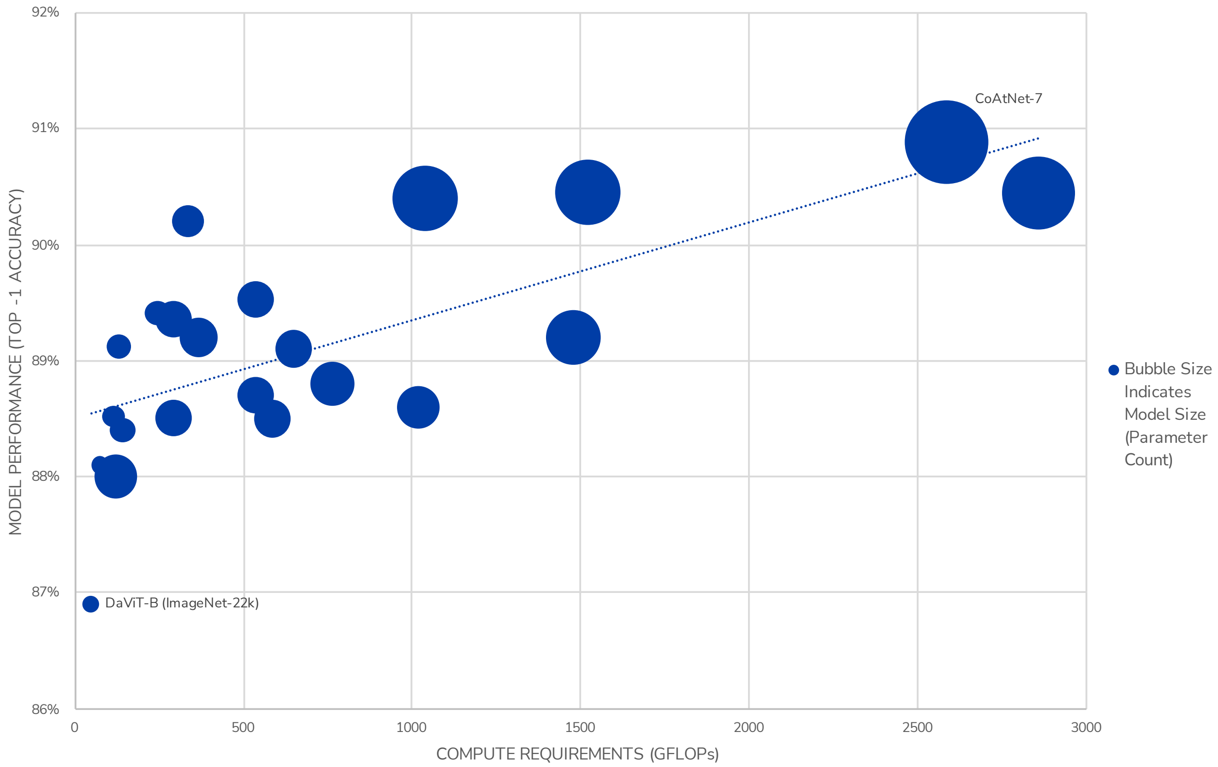
The European Space Administration (ESA) conducted an experimental mission in 2020 that saw the first use of deep learning in low-Earth orbit.[43] The mission tested whether commercial compute hardware could be used on a small cube satellite (CubeSat) to run an AI model. The task was very simple: determine if an image taken by the satellite contained clouds or not.[44] This application may have been selected in part because it is simple enough to test on a CubeSat. Nonetheless, it was useful because the satellite could save memory and bandwidth by transmitting only valuable images of earth while discarding images of clouds.

The CubeSat, which is about the size of a briefcase, and the space environment introduced several constraints. First, the power from its solar cells and batteries was too low to operate conventional processors used for computer vision. Therefore, they used a special-purpose, low-power computer vision board[45] and a commercial vision processing unit (VPU).[46] These components were energy-efficient and small enough to fit within the CubeSat, but ultimately had less compute capacity than most regular processors.[47]

Second, they reduced the precision of the model's calculations[48] and imagery data[49] to save memory and power. This had a negligible effect on the model's performance due to the simplicity of the task, but may degrade the performance of models used for other computer vision tasks. Lastly, because this was an experimental mission, the VPU was not designed to withstand radiation in space. A longer duration mission would need radiation-hardened hardware, which is expensive, typically has less compute capacity than commercial chips, and is less likely to efficiently run high-end models.[50]

The ESA's cloud detector ran successfully and reached 96% accuracy on satellite imagery,[51] but this limited success does not mean top-performing image classification models[52] could function on the CubeSat: the top 9 models on the ImageNet benchmark each have over a billion parameters and require more compute than the satellite can provide. For example, one of these models[53] achieved 90.88% accuracy but used over twice the amount of compute available on the CubeSat. This dynamic is illustrated in Figure 4, which shows the relationship between model performance, compute usage (i.e., FLOPs), and model size (i.e., parameter count). Since the trend is currently toward larger vision models,[54] it may become even harder to put the most capable AI onboard constrained devices. However, there are also efforts to make these models more efficient and less compute-intensive, so it is unclear how constrained these models will ultimately become going forward.[55]

Figure 4. Model Performance, Compute Requirements, and Model Size (Image Classification)



Notes: Model performance is based on the ImageNet benchmark, which is very saturated (meaning that many models achieve very high accuracy with small differences in performance). However, we opted to use ImageNet because many models have been tested against it, allowing for more comparisons. All data is as of December 2022.
Source: See Appendix C.

We find that smaller, less compute-intensive models could theoretically function on the CubeSat, but with reduced performance. One such model[56] achieves 86.9% accuracy with six times fewer parameters and seven times fewer calculations than the aforementioned large top-performing model. This 4% difference is significant on ImageNet, and may be an unacceptable reduction in performance for certain real-world applications. See Appendix C for performance, parameter, and compute data on image classification models.

Large Satellites

In 2021, NASA's Jet Propulsion Laboratory deployed three image classification models onboard the International Space Station (ISS) to classify remote imagery of Earth's and Mars' surfaces, as well as ground imagery taken by the Curiosity Rover.[57] An objective of this experiment was to acquire data on the performance of processors in a high-radiation environment, and ultimately mature their application for AI inference in space.[58]

The ISS is large and can generate 2,000 times more power than the CubeSat,[59] so it can host more chips to run higher-performing image classification models.[60] This compute capacity also enables the use of several AI models, whereas the CubeSat only hosted one. However, only a portion of this overall power was allocated to run the processors. Moreover, the commercial processors[61] used in the mission still have a limited lifespan because they are not radiation-hardened.[62]

Importantly, these processors can be physically accessed and replaced because there are human operators onboard the ISS. Other satellites are typically uncrewed, inaccessible, and expected to operate for decades. Therefore, ongoing space missions cannot leverage new and improved chips, and the radiation-hardened chips that can survive long-duration missions will likely remain far from the state of the art.

Primarily due to long mission durations, low power, and radiation, the computing hardware available on large satellites is likely to significantly constrain AI performance. The case studies show that it is possible to run simple image classification models on these systems, but the engineering challenges could inhibit more complex AI models. Stakeholders must recognize that AI functionality does not translate smoothly to the space environment, and should set realistic expectations regarding the feasibility of satellite-based AI applications, especially if they are intended to operate long-term.

Ground Vehicles: Real-Time Object Detection

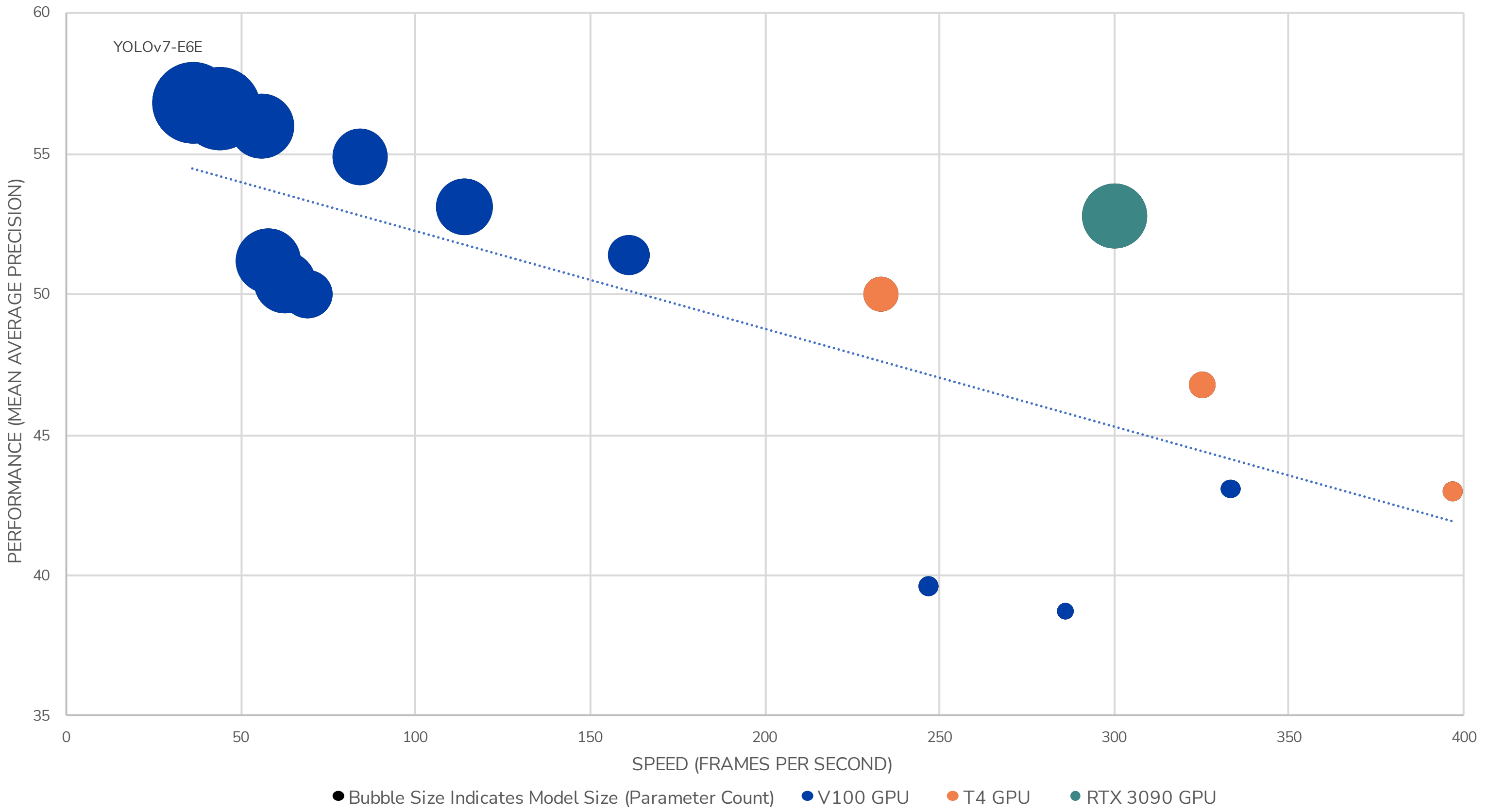
In 2018, the former senior director of AI at Tesla stated that "as you make the [AI] networks bigger by adding more neurons, the accuracy of all their predictions increases with the added capacity…but we are not able to deploy them to the fleet due to computational constraints."[63] AI is constrained even on systems the size of a car, with some of the biggest batteries ever seen in consumer applications.[64]

The constraints are primarily driven by safety. Regulation and industry standards require ancillary systems (e.g., airbags, hazard lights, and computers) to run on a separate power source from the main battery that propels the car.[65] When Tesla designed their own Full Self-Driving (FSD) chip to run on auxiliary power, it was designed to use only 100 watts. The actual chip uses only a fraction of the power consumed by typical AI chips found in data centers. So autonomous vehicles face some of the same power constraints seen in the satellite case study. Moreover, safety concerns can increase costs: two independent FSD chips are embedded in the vehicle's computer to have redundancy and ensure reliability (though one chip can effectively run the AI).[66]

The other safety consideration is speed. Object detection models must rapidly (i.e., in real time) ingest imagery data from several cameras mounted on the vehicle and then make decisions. If the model is too slow or inaccurate, the results can be life-threatening.

Unfortunately, faster object detection models are smaller and less accurate.[67] This is primarily because calculations take time, and the larger a model is, the more calculations it is doing. Ultimately, the need for real-time inference reduces the accuracy of object detectors. Figure 5 illustrates this dynamic, where model performance generally has an inverse relationship with speed. See Appendix C for more data on real-time object detection models. Note that all models in this figure are designed to operate in real time on certain hardware, and that the inverse relationship between performance and speed is even greater when we look at object detectors that are not designed to operate in real time. For example, the top-performing, real-time object detection model in this figure is not even in the top 30 performing object detection models overall.

Figure 5. Model Performance, Speed, and Model Size (Real-Time Object Detection)



Notes: Model performance is based on the COCO benchmark. Most models' speeds were tested with V100 GPUs, allowing for fair comparisons. However, four models ran on more powerful T4 and RTX 3090 GPUs. These models would likely run slower on a V100. All data is as of December 2022.

Source: See Appendix C.

Object detection is an area of intense academic interest, so the best proprietary models are unlikely to significantly outperform open-source models. In fact, we suspect that many automotive manufacturers use compute hardware that is too constrained to run the best open-source, real-time object detectors, as the speed of these models is typically tested on fast, high-power data center chips.[68] The FSD chip likely has enough compute to run the third-best object detection model,[69] which is about half the size, requires 39% fewer calculations, and only performs about 1% worse than the top model. However, it is unclear if chips for autonomous vehicles can run that model quickly enough to meet the speed requirements, which could be a constraint that further reduces performance.

Large ground vehicles including Teslas and trucks can certainly carry powerful batteries and computers; in fact, Waymo claims to use "server-grade GPUs" in their vehicles but has not released any specifications.[70] Nonetheless, these systems still face safety and speed constraints, and different companies address them in different ways. The environment and context in which AI is deployed matter, especially when slight changes in performance can have deadly consequences for drivers and pedestrians. This suggests that constraints on hardware and models may exist even on systems that would otherwise have access to abundant resources.

Future of Onboard AI

How could the constraints of onboard AI and compute change in the future? While there are many factors that may affect future developments, for the purpose of simplicity, we consider two avenues of development that stakeholders should consider: model-side and compute-side development. We assess these developments broadly and are agnostic to any particular device, environment, or AI task.

There are two model-side developments that may constrain the use of some AI models. First, many higher-performing models are getting larger, more compute-intensive, and slower to run. This has been a broad trend since the advent of deep learning, and its continuation could make it harder to run models locally on resource-constrained devices. Conversely, there are reasons to think this trend may slow, as prominent researchers are developing lightweight and efficient model architectures, as well as methods to reduce their compute requirements without significant performance degradation.[71] These two trends are largely at odds, and it is unclear how they will play out for different types of AI tasks going forward.[72]

There are also compute-side developments that may impact onboard AI.[73] First, specialized low-power processors are improving,[74] allowing for more compute capacity at smaller scales with less power consumption. However, this is not guaranteed to continue because improvements in specialized processors could slow or stagnate, and we currently do not know the future economic viability of these processors to incentivize hardware manufacturers to develop them. These specialized designs are partly a response to the second trend, which is the slowing of improvements in general-purpose processors (Moore's Law). This trend could potentially happen with special-purpose processors, which would make onboard AI more constrained.[75]

How these developments intersect will greatly influence the future constraints of onboard AI. This is illustrated below in Table 2, where the red cell indicates a significant increase in onboard AI constraints, the light blue indicates a significant decrease in constraints, and the yellow cells indicate a mix of improvements and difficulties (where it will be harder to deduce the broad impact on onboard AI).

Table 2. Broad Intersection of Developments in Onboard AI and Compute

Onboard Hardware Stagnates	- Model requirements and compute availability slow - Gaps persist between state-of-the-art and real-world AI applications	- Model growth outpaces onboard hardware - Gap grows between real-world AI applications and lab demonstrations
Onboard Hardware Improves	- Onboard hardware improves to suit model needs - State-of-the-art models run in real-world applications	- Model requirements and compute availability grow - Gaps persist between state-of-the-art and real-world AI applications
	Model Demands Stabilize	**Model Demands Grow**

Source: CSET.

Conclusion

In conclusion, we offer three broad recommendations to better manage the constraints of onboard AI:

1. **Acknowledge that many real-world systems will underperform those in state-of-the-art or laboratory demonstrations.** This puts an extra onus on testing and evaluation during a time when AI successes are putting pressure on developers to rapidly incorporate AI and on organizational leaders to rapidly field it.
2. **Acknowledge that hardware constraints also apply to adversaries.** This has at least two ramifications. Firstly, systems that adversaries deploy may be significantly less capable and more error-prone than unconstrained state-of-the-art models. Some users may be risk-averse and avoid using such systems, while others may not be constrained in the same way because they tolerate lower performance. Secondly, an adversary's inability to fabricate or acquire the most powerful computing hardware for data centers may not limit their ability to compete on the battlefield or in other real-world 'edge' applications.
3. **Support research that could narrow the gap between state-of-the-art models and deployable models.** This includes funding and prioritizing efforts to shrink, condense, and accelerate models. It also includes efforts to improve the efficiency and scale of AI chips for onboard applications.

Authors

Kyle Miller is a Research Analyst at Georgetown's Center for Security and Emerging Technology (CSET), where he works on the CyberAI Project. Andrew Lohn is a Senior Fellow at CSET, where he also works on the CyberAI Project.

Acknowledgments

For their careful review, thoughtful comments, and constructive feedback, we would like to thank our two external reviewers, Neil Thompson and Colby Banbury, as well as our CSET colleagues John Bansemer and Jaret Riddick.

CSET Product ID #: 2022CA008

Document Identifier: doi: 10.51593/2022CA008

Appendix A: Methodology

A.1. Methodology

1. Select devices used in differing environments (i.e., air, space, land) that could run image classification, object detection, real-time object detection, or machine translation models locally. Focus explicitly on local inference, and disregard devices or applications that use data centers or near-premises 'edge compute' to support AI inference, including but not limited to sharing resources across edge devices, computation offloading to edge servers, resource allocation schemes, and architectures such as cloudlets, fog compute, and multi-access edge computing.[76]
2. Collect technical specifications of the devices (i.e., size, weight, power, sensors, UI/controls, instruments, etc.), onboard compute and memory (i.e., board/card size, transistor size, FLOPS, OPS, power, memory type/quantity, bandwidth, interconnect, etc.), and onboard AI models (i.e., performance, parameters, inference FLOPs, speed, architecture, training data, etc.). Only assess AI models and processors with open-source data on their compute requirements and capacity.
3. Review popular benchmarks for each AI task, identify the top-performing models on particular benchmarks, and examine the associated publications to determine what resources are required for inference. Use Papers With Code[77] as a baseline reference for open-source models, but manually review each publication to gather more granular information and ensure the accuracy of the data.
4. Compare the resource requirements of top-performing models with the specific or estimated onboard resources of the devices in the case studies, as well as the broad resources of common GPUs and more specialized processors/systems (e.g., VPUs, SoCs, modules, and accelerators). With the available data, gauge if the models can function onboard effectively with said resources. If top-performing models cannot (or are unlikely to) function effectively onboard, then assess what constraints are involved and gauge what level of performance below the state-of-the-art could be achieved.

A.2. Data Limitations

1. There is limited open-source data on how much compute is required to run many models, as well as the compute capacity of many processors.
2. There is very limited data on the memory requirements of most models we investigated.
3. We cannot reliably compare the compute requirements/capacity of many models and processors due to misaligned metrics from different sources (e.g., OPS vs. FLOPS).
4. We could not fully apply this methodology to the case study of machine translation on a smartphone due to limited data on the size and compute requirements of machine translation models. We placed this case study in Appendix D because it is a good example of when effective onboard AI is attainable for a particular consumer application.

Notwithstanding these limitations, the available data for most of the case studies sufficiently captures trends and trade-offs between model size, performance, and compute requirements, as well as the compute capacity and power consumption of many processors.

Appendix B: Compute Hardware Data

We assessed several GPUs, VPUs, NPUs, SoCs, and system-on-modules to determine if they could function effectively onboard devices in the case studies (and therefore enable local inference). We used a combination of published survey research and manual data collection, primarily from the hardware developers' websites and TechPowerUp.[78] The primary metrics of interest were the size and weight of the processor (including the card/board), compute capacity, memory, power consumption, and speed.

1. Manually assessed several high-end NVIDIA GPUs typically used for AI inference on PCs and data centers. These represent compute capacity when there are few device- or environment-dependent constraints. They are displayed below in Table 3 and Figure 6.
2. Manually assessed various small, low-power processors (including when they are embedded on SoCs or other boards/modules). Some were selected because they were used on specific devices in our case studies (the Skydio 2+, LANIUS, and Orlan-10 drones, the 6U CubeSat and International Space Station, the FSD Tesla, and the iPhone XR). Others were selected because their characteristics could enable inference onboard resource-constrained devices (e.g., small size and low power). They are displayed below in Table 4 and Figure 7.
3. Used two surveys as a source for data on AI accelerators.[79] However, these sources lacked sufficient data on processor FLOPS, which limited our ability to compare the accelerators' compute capacities with the AI models' compute usage for inference.

This is not a comprehensive assessment of most processors, nor does it involve more granular performance considerations regarding theoretical performance vs. actual performance in running different AI models (e.g., compute utilization or GPU parallelization).[80] There are blank portions of the tables due to lack of data, as well as several ambiguous data points that cannot be confirmed (which are flagged). Notwithstanding, we believe this data, in the aggregate, sufficiently informs our broad analysis of onboard AI functionality, constraints, and limitations on various devices.

Table 3. Popular NVIDIA GPUs Used in Data Centers, Deep Learning Inference, or High-End Computers (Ranked by FLOPS Capacity in FP16)

Hardware	Size	Power (TDP)[A]	FLOPS (FP16)	FLOPS (FP32)	Memory	Bandwidth[B]	Board Size (L x W x H)	Release
H100 PCIe[81]	4nm	350 W	204.9 TFLOPS	51.22 TFLOPS	80 GB 5120 bit HBM2e	2039 GB/s	10.6 x 4.4”*	2022
GeForce RTX 4090[82]	4nm	450 W	82.58 TFLOPS*	82.58 TFLOPS*	24 GB 384-bit GDDR6X	1008 GB/s	12 x 5.4 x 2.4”	2022
A100 PCIe[83]	7nm	300 W	77.97 TFLOPS	19.49 TFLOPS	80 GB 5120-bit HBM2e	1935 GB/s	10.5 x 4.4 x 4.3”*	2021
Tesla T4[84]	12nm	70 W	65.13 TFLOPS	8.14 TFLOPS	16 GB 256-bit GDDR6	320 GB/s	6.6 x 2.7 x 0.7”*	2018
RTX A6000[85]	8nm	300 W	38.71 TFLOPS*	38.71 TFLOPS*	48 GB 384-bit GDDR6	768 GB/s	10.5 x 4.4”	2020
GeForce RTX 3090[86]	8nm	350 W	35.58 TFLOPS*	35.58 TFLOPS*	24 GB 384-bit GDDR6X	936 GB/s	13.2 x 5.5 x 2.4”	2020
TITAN V[87]	12nm	250 W	29.80 TFLOPS	14.90 TFLOPS	12 GB 3072-bit HBM2	651 GB/s	10.5 x 4.4 x 1.6”	2017
GeForce RTX 3080[88]	8nm	320 W	29.77 TFLOPS*	29.77 TFLOPS*	10 GB 320-bit GDDR6X	760 GB/s	11.2 x 4.4 x 1.6”	2020
V100 PCIe[89]	12nm	250 W	28.26 TFLOPS	14.13 TFLOPS	32 GB 4096-bit HBM2	900 GB/s	10.5 x 4.3 x 1.5”	2018
GeForce RTX 2080[90]	12nm	215 W	20.14 TFLOPS	10.07 TFLOPS	8 GB 256-bit GDDR6	448 GB/s	10.5 x 4.6 x 1.4”	2018
P100 PCIe[91]	16nm	250 W	19.05 TFLOPS	9.53 TFLOPS	16 GB 4096-bit HBM2	732 GB/s	10.5 x 4.4”*	2016

Notes: Cells marked with an asterisk (*) indicates ambiguity in the data or a minor conflict in sources (such as identical FLOPS in FP32 and FP16).

[A] Power is based on Thermal Design Power (TDP), which is the maximum heat a component is expected to output and the cooling systems are designed to sustain. TDP is a common metric used to measure GPU power, but it is not an absolute measure of power consumption.

[B] Bandwidth is rounded to the nearest whole number.

Sources: See endnotes.

Table 4. Low-Power Processors, SoCs, VPUs, Accelerators, and Modules

Hardware	Type	Size	Power[A]	FLOPS (FP16)	FLOPS (FP32)	OPS	Memory	Bandwidth	Release
Jetson Nano[92]	Module; Accelerator	20nm	5-10 W (10 W TDP)	471 GFLOPS	235 GFLOPS		4 GB 64-bit LPDDR4 1600 MHz (system shared)	25.6 GB/s	2019
Jetson TX2 NX (Pascal GPU)[93]	System on Module	16nm	7.5 W (15 W TDP)*	1330 GFLOPS	750 GFLOPS		4 GB 128-bit LPDDR4 1600 MHz (system shared)	51.2 GB/s	2021
Jetson TX2 (Pascal GPU)[94a]	System on Module	16nm	7.5 W (15 W TDP)	1330 GFLOPS	750 GFLOPS		8 GB 128-bit LPDDR4 (system shared)	59.7 GB/s	2016
Jetson Xavier NX[95]	System on Chip	12nm	10 - 20 W (15 W TDP)	1690 GFLOPS	844 GFLOPS	21 TOPS*	16 GB 128-bit LPDDR4x (system shared)	59.7 GB/s	2020
Jetson AGX Xavier 32 GB[96]	Module	12nm	10 - 30 W (30 W TDP)	2820 GFLOPS	1410 GFLOPS	32 TOPS	32 GB 256-bit LPDDR4x (system shared) 1377 MHz	136.5 GB/s	2018
Jetson Orin Nano 8GB[97]	System on Module		7 - 15 W (15 W TDP)	2560 GFLOPS	1280 GFLOPS	40 TOPS*	8 GB 128-bit LPDDR5	68 GB/s	2022
Jetson AGX Orin 32GB[98]	System on Module		15 - 40 W (40 W TDP)	6666 GFLOPS	3333 GFLOPS	200 TOPS	32 GB 256-bit LPDDR5	204.8 GB/s	2022
RAD750[99]	System on Chip (RAD-H)						2 GB (Flash) 256 MB DRAM		2001
RAD5510[100]	System on Chip (RAD-H)	45nm	11.5 W TDP	0.9 GFLOPS*	0.9 GFLOPS*	1.4 GOPS	64 GB 64-bit DDR2/3 DRAM	51 GB/s	2017*
RAD5515[101]	System on Chip	45nm							2017*

	(RAD-H)								
RAD5545[102]	System on Chip (RAD-H)	45nm	20 W* (35 W TDP)	3.7 GFLOPS*	3.7 GFLOPS*	5.6 GOPS	4 GB 32-bit/64-bit DDR3 SDRAM*	51 GB/s	2017*
Tegra X2 (NVIDIA Pascal GPU)[103a]	System on Chip	16nm	7.5-15 W (15 W TDP)	1500 GFLOPS	750 GFLOPS		8 GB 128-bit LPDDR4 854–1465 MHz	59.7 GB/s	2016
Snapdragon 855 (Adreno 640 GPU)[104a]	System on Chip	7nm	5 W TDP	1798 GFLOPS	899 GFLOPS		16 GiB 64-bit LPDDR4X-4266 2133 MHz	31.79 GiB/s	2019
Movidius Myriad 2[105a]	Vision Processing Unit	28nm	1 W (per TFLOP)[B]	1000 GFLOPS[B]			4 GB 32-bit LPDDR3 733 MHz	400 GB/s[C]	2016
Movidius Myriad X 4GB[106a]	Vision Processing Unit	16nm	1.5 W TDP			1 TOPS*	4 GB 32-bit LPDDR3/LPDDR4 SDRAM 1600 MHz	450 GB/s[C]	2020
PowerVR Series5 SGX530[107a]	Video/Imagery Accelerator	65nm		1.6 GFLOPS[D]	1.6 GFLOPS[D]				2005
Tesla Full Self-Driving (FSD) Chip[108a]	System on Chip	14nm	36 W TDP	600 GFLOPS*	600 GFLOPS*	73.73 TOPS (INT8)	8 GiB 128-bit LPDDR4-4266	63.58 GiB/s	2019
A16 Bionic[109]	System on Chip	4nm	8 W TDP		2000 GFLOPS		6GB 64-bit LPDDR5 3200 MHz	51.2 Gbit/s	2022
A15 Bionic[110]	System on Chip	5nm	6 W TDP	3000 GFLOPS	1500 GFLOPS		6 GB 64-bit LPDDR4X 2133 MHz	42.7 GB/s	2021

A12 Bionic[111][α]	System on Chip	7nm	6 W TDP	1152 GFLOPS	560 GFLOPS	5 TOPS	12GB 64-bit LPDDR4X 2133 MHz	34.1 Gbit/s	2018
Apple M2 Pro[112]	System on Chip	5nm	30 W TDP*	13490 GFLOPS*	6745 GFLOPS*		32 GB 256-bit LPDDR5-6400	200 GB/s	2023

Notes: Blank cells indicate a lack of data. Cells marked with an asterisk (*) indicate minor ambiguity in the data or a minor conflict in sources. Those marked with an alpha superscript ([α]) indicate the hardware was used on a specific device we investigated for a case study. Those marked with a letter superscript (e.g., [A]) indicate significant ambiguity or conflicting sources, and are addressed individually in the notes below. Lastly, 'RAD-H' indicates the processor is radiation-hardened.

[A] Data on power was retrieved primarily from the manufacturers' websites and TechPowerUp, although what is meant by power can be ambiguous. If power is based on metrics like TDP or W/FLOP, then it is specified within the cell.

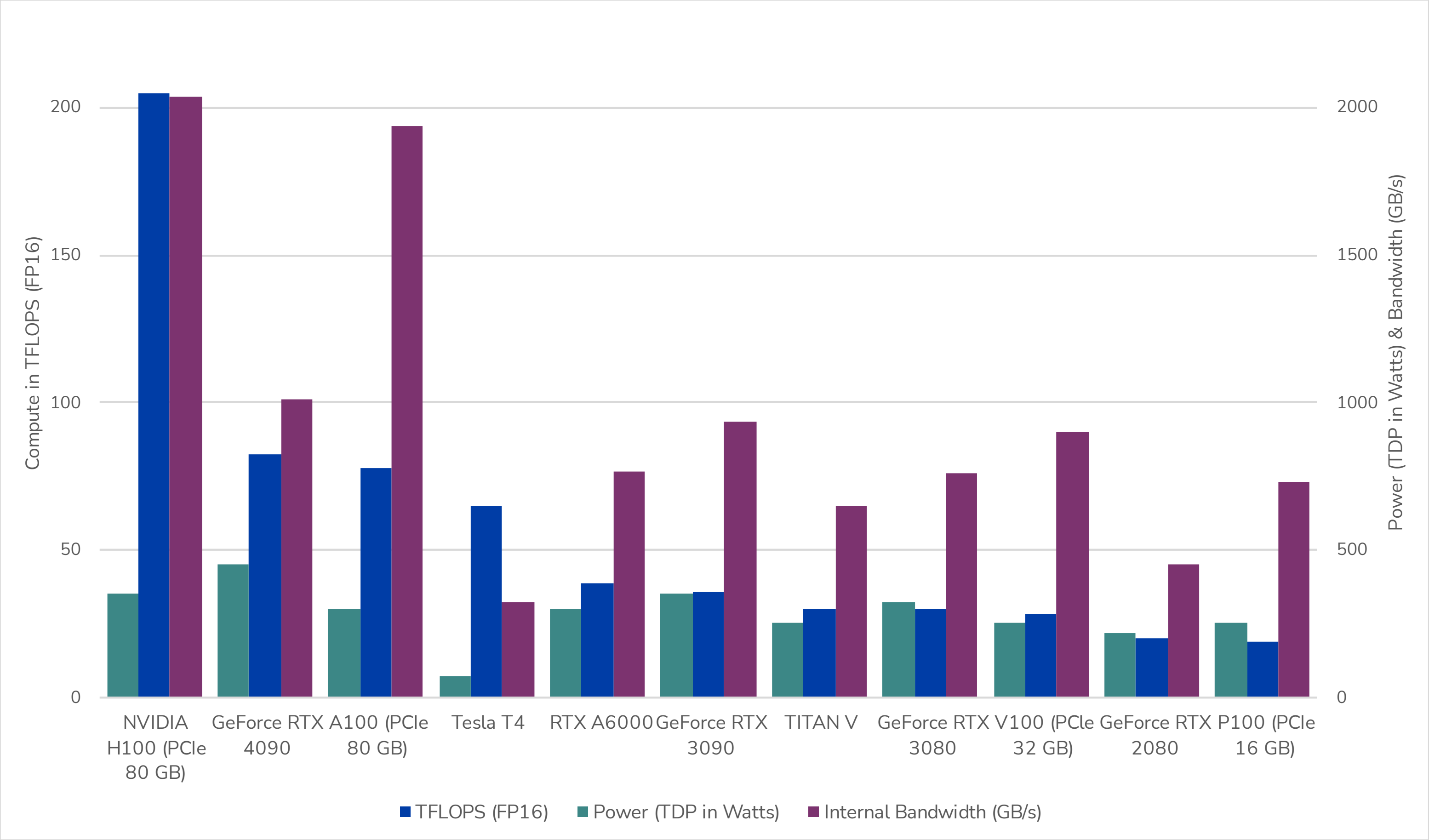
[B] Intel claims the Movidius Myriad 2 can process 1000 GFLOPS with a "nominal 1 watt power envelope." We are skeptical of this FLOPS-per-watt performance, as it is abnormally high, particularly for a processor that was developed in 2016 (and later discontinued).

[C] Intel claims the Movidius Myriad 2 and Myriad X have a bandwidth of 400 GB/s and 450 GB/s, respectively. We are skeptical of this bandwidth, as it is abnormally high for these types of systems.

[D] The sourcing for data on the PowerVR Series5 is from Wikipedia and cannot be confirmed.

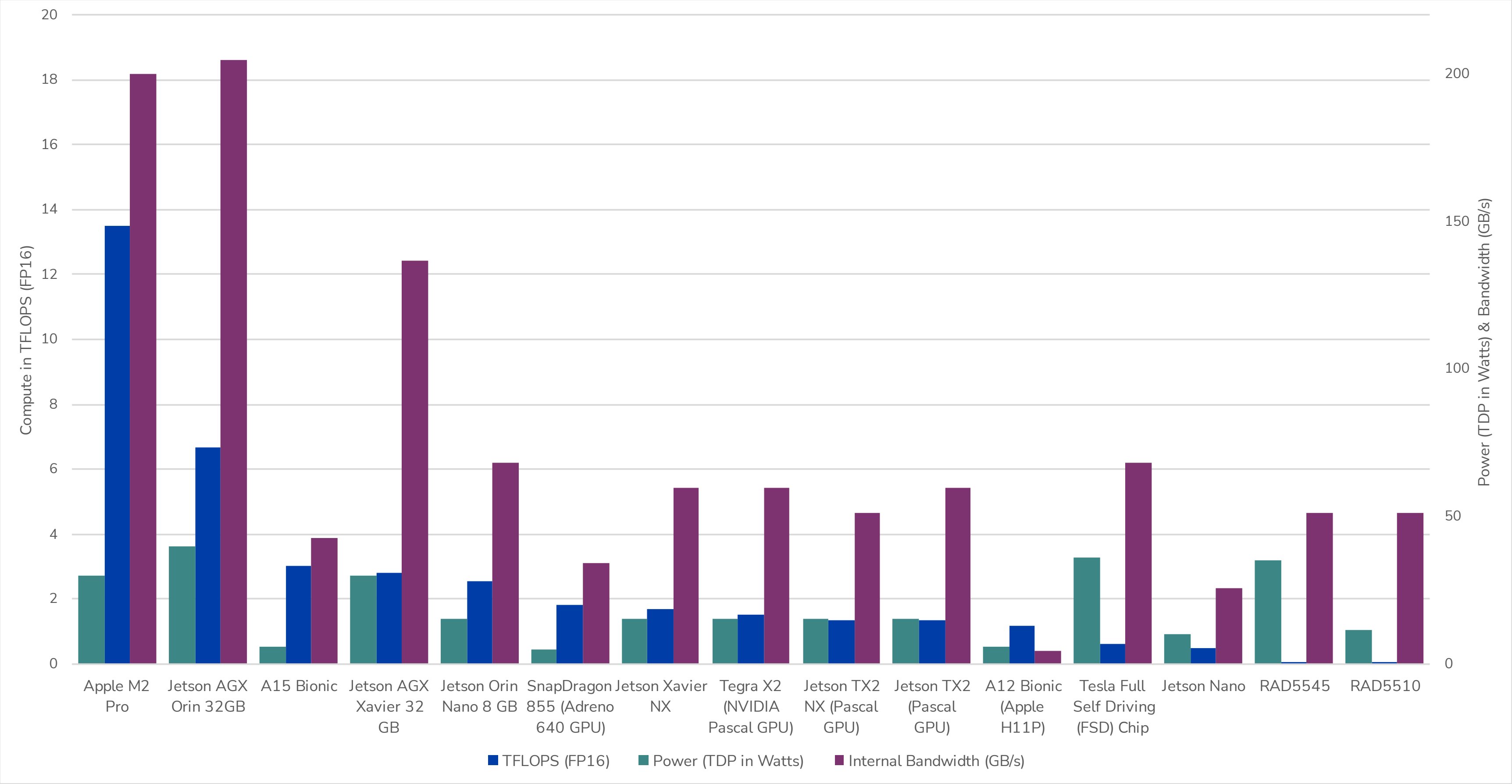
Sources: See endnotes.

Figure 6. Popular NVIDIA GPUs Used in Data Centers, Deep Learning Inference, or High-End Computers (Ranked by FLOPS Capacity in FP16)



Sources: See Table 3.

Figure 7. Low-Power Processors, SoCs, VPUs, Accelerators, and Modules (Ranked by FLOPS Capacity in FP16)



Note: Several of the processors we assessed are not displayed due to data limitations.

Sources: See Table 4.

Appendix C: AI Model Data

We manually assessed the top-performing AI models for image classification, object detection, real-time object detection, and English-French machine translation.[113] The primary metrics of interest were performance (on popular benchmarks), model size (i.e., parameters), compute requirements (i.e., FLOPs), memory requirements, speed (i.e., fps), precision for inference (i.e., FP16 vs. FP32 vs. FP64), and the compute hardware used by the model's developers to run inference.

The tables below display data for image classification and real-time object detection models. We do not include tables for object detection or machine translation due to insufficient data.[114] We used the ImageNet benchmark to assess the performance of image classification models, and the MS COCO benchmark to assess the performance of real-time object detection models. All data are as of December 2022.

Table 5. Image Classification Models Data (Ranked by Performance on ImageNet Benchmark)

Model	Performance (Top-1 Accuracy)	Parameters (M)	Inference (GFLOPs)	Original Publication Date
CoCa (finetuned)[115]	91%	2100		June 2022
Model soups (BASIC-L)[116]	90.98%	2440		Jul 2022
Model soups (ViT-G/14)[117]	90.94%	1843		Jul 2022
ViT-e[118]	90.90%	3900		Sep 2022
CoAtNet-7[119]	90.88%	2440	2586	Sep 2021
CoAtNet-6[120]	90.45%	1470	1521	Sep 2021
ViT-G/14[121]	90.45%	1843	2859.9	Jun 2022
DaViT-G[122]	90.40%	1437	1038	Apr 2022
DaViT-H[123]	90.20%	362	334	Apr 2022
MaxViT-XL (512res, JFT)[124]	89.53%	475	535.2	Sep 2022
MaxViT-L (512res, JFT)[125]	89.41%	212	245.2	Sep 2022
MaxViT-XL (384res, JFT)[126]	89.36%	475	293.7	Sep 2022
NFNet-F4+[127]	89.20%	527	367	Feb 2021
InternImage-DCNv3-H (M3I Pre-training)[128]	89.20%	1080	1478	Nov 2022
MaxViT-L (384res, JFT)[129]	89.12%	212	128.7	Sep 2022
MOAT-4 22K+1K[130]	89.10%	483.2	648.5	Oct 2022
MViTv2-H (512 res, ImageNet-21k pretrain)[131]	88.80%	667	763.5	Mar 2022
MaxViT-XL (512res, 21K)[132]	88.70%	475	535.2	Sep 2022
SWAG (ViT H/14)[133]	88.60%	633.5	1018.8	Apr 2022
CoAtNet-3 @384[134]	88.52%	168	114	Sep 2021
MaxViT-XL (384res, 21K)[135]	88.51%	475	293.7	Sep 2022

FixEfficientNet-L2[136]	88.50%	480	585	Nov 2020
MViTv2-L (384 res, ImageNet-21k pretrain)[137]	88.40%	218	140.7	Dec 2021
CAFormer-B36 (384 res, 21K)[138]	88.10%	99	72.2	Dec 2022
MViTv2-H (ImageNet-21k pretrain)[139]	88%	667	120.6	Dec 2021
DaViT-B (ImageNet-22k)[140]	86.9%	87.9	46.4	April 2022

Notes: Blank cells indicate a lack of data. All data is as of December 2022.
Sources: See endnotes.

Table 6. Real-Time Object Detection Models Data (Ranked by Performance on COCO Test-Dev Benchmark)

Model	Performance (Box AP)	Parameters (M)	Inference (GFLOPs)	FP	Processor	Speed (fps)	Publication Date
YOLOv7-E6E[141]	56.8	151.7	843.2		V100	36	Jul 2022
YOLOv7-D6[142]	56.6	154.7	806.8		V100	44	Jul 2022
YOLOv7-E6[143]	56	97.2	515.2		V100	56	Jul 2022
YOLOv7-W6[144]	54.9	70.4	360		V100	84	Jul 2022
YOLOv7-X[145]	53.1	71.3	189.9		V100	114	Jul 2022
RTMDet-x[146]	52.8	94.9	141.7	FP16	GTX 3090	300*	Dec 2022
YOLOv7[147]	51.4	36.9	104.7		V100	161	Jul 2022
YOLOX-X[148]	51.2	99.1	281.9	FP16	V100	57.8	Aug 2021
YOLOv5-X[149]	50.4	87.8	219	FP16	V100	62.5	Aug 2021
DAMO-YOLO-M[150]	50	28.2	61.8	FP16	T4	233	Dec 2022
YOLOX-L[151]	50	54.2	155.6	FP16	V100	69	Aug 2021
DAMO-YOLO-S[152]	46.8	16.3	37.8	FP16	T4	325	Dec 2022
DAMO-YOLO-T[153]	43	8.5	18.1	FP16	T4	397	Dec 2022
PP-YOLOE-S[154]	43.1	7.9	17.4	FP16	V100	333.3	Dec 2022
YOLOX-S[155]	39.6	9	26.8	FP16	V100	246.9	Dec 2022
YOLOv7-tiny-SiLU[156]	38.7	6.2	13.8		V100	286	Jul 2022

Notes: Blank cells indicate a lack of data. Cells marked with an asterisk (*) indicate minor ambiguity in the data or a minor conflict in sources. All data is as of December 2022.
Sources: See endnotes.

Appendix D: Machine Translation on Smartphones

Throughout the Afghanistan war, coalition forces invested substantially in human translators, exemplified by the Pentagon's $679 million contract to field more translators in 2010.[157] Even with such investments there was a dearth of well-trained translators, which hamstrung coalition efforts, from reduced communication to problematic mistranslations.[158]

Machine translation can help address this shortfall, particularly in low-stakes cases or when human translators are not available. But not all machine translators are created equal, largely because there is more data available for some languages than others. Common languages such as English, French, and Chinese are considered high-resource languages as compared to the many low-resource languages that have far less data available.[159] For machine translation, data availability seems to be more of a bottleneck than model size.[160] Large models only slightly outperform smaller models, but high-resource languages substantially outperform low-resource languages.[161]

Google's online translation service is relatively unconstrained because it runs on data center servers, while offline translation is also available for users who download a model for each language pair to run onboard devices like smartphones. However, the offline and online translations are not identical.[162] We compared French to English translations from both the online and offline models, finding that online models matched human translations somewhat more closely.[163] This implies that Google finds it worthwhile to run more capable models in their unconstrained data centers and offer a less capable model for users to run locally on more constrained devices, but the difference is not large from a practical standpoint. We found one example of a problematic local mistranslation, which is displayed below in Table 7: one could mistakenly infer from the local translation that all three people "had to join a shelter."

Table 7. Human vs. Online and Local Machine Translation (French-English)

Human Translation (from French)	Online Translation	Local Translation
Just before the start of the European Union (EU)-Ukraine summit held in Kyiv on Friday, February 3, an air raid siren sounded near Maidan Square, where Ursula von der Leyen had stopped. **The president of the EU Commission, who was to meet Volodymyr Zelensky and her European Council counterpart Charles Michel at the presidential palace, had to take shelter for over half an hour.**	Just before the start of the European Union (EU)-Ukraine summit held in kyiv on Friday February 3, a siren sounded near Maidan Square, where Ursula von der Leyen was. **The President of the Commission, who was to meet Volodymyr Zelensky and Charles Michel, her counterpart from the European Council, at the presidential palace, had to go to a shelter for more than half an hour.**	Just before the start of the European Union-Ukraine summit held in Kiev on Friday, February 3, a siren sounded near Maidan Square, where Ursula von der Leyen was located. **The President of the Commission, who was to find, at the presidential palace, Volodymyr Zelensky and Charles Michel, her counterpart of the European Council, had to join a shelter for more than half an hour.**

Note: BLEU scores were calculated by CSET.
Source: Virginie Malingre, “In Kyiv, EU leaders promise new sanctions and welcome Ukraine’s ‘progress’ toward membership,” *Le Monde*, February 2023, https://www.lemonde.fr/en/international/article/2023/02/04/in-kyiv-eu-leaders-promise-new-sanctions-and-welcome-ukraine-s-progress-toward-membership_6014396_4.html.

Academic research implies that the cloud and local models might perform similarly because data availability is the limiting factor, rather than model size.[164] This is also why low-resource languages, which have less data, frequently perform worse. Since high-performing models can be relatively small in size, bilingual machine translation is only slightly constrained by the onboard computing resources of a smartphone. This may be less true for multilingual translators, which appear to need larger neural networks to accommodate the larger scope of their task.[165]

Endnotes

[1] "Jetson Nano," *elinux,* accessed January 2023, https://elinux.org/Jetson_Nano; "NVIDIA H100 PCIe GPU Product Brief," *NVIDIA*, November 2022, https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf.

[2] The Perseverance rover uses a RAD750 processor that has very low compute capacity. It is a reengineered and radiation-hardened version of the PowerPC 750 processor that was released in 1997; "Mars 2020/Perseverance," *NASA*, March 2020, https://mars.nasa.gov/files/mars2020/Mars2020_Fact_Sheet.pdf; "Cost of Perseverance," *The Planetary Society*, accessed January 2023, https://www.planetary.org/space-policy/cost-of-perseverance; "Autonomous Exploration for Gathering Increased Science," *NASA Tech Briefs*, September 2010, https://ntrs.nasa.gov/api/citations/20100033547/downloads/20100033547.pdf; "Perseverance's SuperCam Uses AEGIS For the First Time," *NASA*, May 2022, https://mars.nasa.gov/resources/26782/perseverances-supercam-uses-aegis-for-the-first-time/#:~:text=On%20May%2018%2C%202022%2C%20NASA's,seen%20in%20close%2Dup%20here; "RAD750 radiation-hardened PowerPC microprocessor," *BAE Systems*, accessed November 2022, https://www.baesystems.com/en-media/uploadFile/20210404045936/1434555668211.pdf; "Flight Projects—Mobility," *NASA Jet Propulsion Laboratory*, accessed November 2022, https://www-robotics.jpl.nasa.gov/what-we-do/flight-projects/mars-2020-rover/m2020mobility/; "Brains," *NASA*, accessed November 2022, https://mars.nasa.gov/mars2020/spacecraft/rover/brains/; Jacek Krywko, "Space-grade CPUs: How do you send more computing power into space?" *Ars Technica*, November 2019, https://arstechnica.com/science/2019/11/space-grade-cpus-how-do-you-send-more-computing-power-into-space/.

[3] Other benefits of onboard AI include ultra-low latency, high bandwidth, mobility support, location awareness support, context awareness support, proximity to user, and real-time access to device data.

[4] Zhi Zhou et al., "Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing," arXiv preprint arXiv:2205.01917 (2019), https://arxiv.org/pdf/1905.10083.pdf; Carlos Poncinelli Filho et al., "A Systematic Literature Review on Distributed Machine Learning in Edge Computing," *Sensors*, March 2022, https://www.mdpi.com/1424-8220/22/7/2665.; M. G. Sarwar Murshed et al., "Machine Learning at the Network Edge: A Survey," arXiv preprint arXiv:1908.00080 (2021), https://arxiv.org/abs/1908.00080.

[5] Erwin Loh and Tam Nguyen, "Artificial intelligence for medical robotics," *Endorobotics*, 2022, https://www.sciencedirect.com/science/article/pii/B9780128217504000025.

[6] AI inference can be generally understood as the process of running a trained model on data (i.e., using the trained model to make predictions and classifications based on data inputs).

[7] E.g., the U.S. DoD's Tactical Intelligence Targeting Access Node (TITAN) program envisions the use of ground vehicles as mobile, rugged data centers that can receive sensor data from many devices, run processes such as AI inference, and transmit the outputs back to the devices; "TITAN Brings Together Systems For Next Generation Intelligence Capabilities," *PEO IEW&S*, September 2022, https://peoiews.army.mil/titan-brings-together-systems-for-next-generation-intelligence-capabilities/.

[8] "An On-device Deep Neural Network for Face Detection," *Machine Learning Research*, November 2017, https://machinelearning.apple.com/research/face-detection.

[9] Andrei Paleyes et al., "Challenges in Deploying Machine Learning: a Survey of Case Studies," arXiv preprint arXiv:2205.01917 (2022), https://arxiv.org/pdf/2011.09926.pdf; Di Liu et al., "Bringing AI to edge: From deep learning's perspective," *Neurocomputing* volume 485, May 2022, https://doi.org/10.1016/j.neucom.2021.04.141.

[10] "Nvidia Tesla V100 GPU Architecture — The World's Most Advanced Data Center GPU," *NVIDIA*, August 2017, https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf; "Democratizing access to large-scale language models with OPT-175B," *Meta AI*, May 2022, https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/.

[11] We had to reduce the input size from the default settings to run the 7B version of LLaMa on a V100 GPU with 16 GB of memory.

[12] Compute (i.e., computational power) is the resource required to process digital computational tasks and calculations. It is different from but related to memory, which is a component used to quickly store and access digital information. This report uses FLOPs (Floating Point Operations) and FLOPS (Floating Point Operations Per Second) as the primary metrics to gauge the compute required to run a given model and the computational capacity of a given processor used for AI inference (i.e., a processor in a GPU, VPU, NPU, system-on-chip, etc.). We recognize that FLOPS/s do not capture the entirety of processing behind inference because many non-FLOP operations occur, such as integer operations. We use FLOPS/s because it is the most relevant metric for AI processing and is sufficient to roughly gauge compute. Additionally, we could not reliably gauge the memory requirements of most models due to lack of data.

[13] E.g., to run the GPT3 language model, researchers estimated that it takes about 700 GB of memory to load the parameters, in addition to 600 GB for the activations (in FP32). This means a total of 1.3 TB of memory is needed to run a forward pass on the model, which far exceeds the memory available on many devices; Stephen Gou and Bharat Venkitesh, "Efficient Inference of Extremely Large Transformer Models, with Q&A from EMEA Region," *NVIDIA GTC Conference*, March 2023, https://www.nvidia.com/gtc/.

[14] E.g., GPU boards that are designed to fit in a PC or data center rack do not have rigid size or weight constraints, and often use an on-grid power source via an electrical outlet instead of a battery.

[15] E.g., computers on autonomous vehicles can run many processes outside of AI, such as streaming media.

[16] E.g., full-precision Floating-Point operations (FP32) vs. half-precision (FP16).

[17] E.g., Convolutional Neural Networks (CNN) vs. Vision Transformers.

[18] E.g., small, lower-resolution images are easier to process than large, high-resolution images.

[19] E.g., AI models that must operate continuously in real time (i.e., faster than ~30 frames per second) often require more memory to function than those not operating in real time.

[20] E.g., raw hyperspectral imagery data from satellites must be converted before it can be processed by an AI model.

[21] E.g., if a drone runs object detection, object tracking, and pose estimation models simultaneously.

[22] E.g., autonomous ground vehicles that operate on public roads often have redundant compute hardware that ensures continued system operability even after a failure or malfunction.

[23] E.g., in 2011, a computer on the Russian ‘Phobos-Grunt’ spacecraft malfunctioned due to the bombardment of cosmic rays in low Earth orbit. Consequently, the antennas did not activate and communication with the spacecraft via ground stations was lost. Lack of communications and the inability to physically access, repair, or replace the computer ultimately led to a failed mission; Jacek Krywko, “Space-grade CPUs: How do you send more computing power into space?,” *ArsTechnica*, November 2019, https://arstechnica.com/science/2019/11/space-grade-cpus-how-do-you-send-more-computing-power-into-space/.

[24] E.g., a Field Programmable Gate Array (FPGA) is an integrated circuit that can be configured for a particular purpose post-manufacturing.

[25] Here we mean methods such as pruning the model (reducing parameters) or quantization (reducing the mathematical precision of the calculations), both of which can reduce the compute and memory used during inference. The extent to which this degrades performance depends on factors such as the type of model, the task being performed, and the degree of parameter pruning. This is a common practice amongst developers designing models for resource-constrained edge devices.

[26] Thomas Lee, Susan Mckeever, and Jane Courtney, "Flying Free: A Research Overview of Deep Learning in Drone Navigation Autonomy," *Drones* 5, no. 2*: 52*, June 2021, https://doi.org/10.3390/drones5020052; Anitha Ramachandran and Arun Kumar Sangaiah, "A review on object detection in unmanned aerial vehicle surveillance," *International Journal of Cognitive Computing in Engineering*, Vol. 2, June 2021, https://doi.org/10.1016/j.ijcce.2021.11.005; Aggi Cantrill, "War in Europe Draws Investors to Drone, Battlefield AI Makers," *Bloomberg*, January 2023, https://www.bloomberg.com/news/articles/2023-01-06/war-in-europe-draws-investors-for-drone-battlefield-ai-makers; Daniel Zhang et al., "Artificial Intelligence Index Report 2022," AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022, https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf; "Teal Group Predicts Worldwide Military UAS Spending of $187.6 Billion Over the Next Decade in its 2021/2022 UAV Market Profile and Forecast," Teal Group Corporation, December 2021, https://tealgroup.com/index.php/pages/press-releases/68-teal-group-predicts-worldwide-military-uas-spending-of-187-6-billion-over-the-next-decade-in-its-2021-2022-uav-market-profile-and-forecast.

[27] "Defense Innovation Unit Annual Report FY 2022," Defense Innovation Unit, 2022, https://downloads.ctfassets.net/3nanhbfkr0pc/5guJIhcMGwIgoop4z9r5QM/89f44fe62f981e5f0d28932618719196/DIU_Annual_Report_FY22_Final_0131.pdf; Dr. Lael Rudd, "OFFensive Swarm-Enabled Tactics (OFFSET)," Defense Advanced Research Projects Agency, accessed October 2022, https://www.darpa.mil/program/offensive-swarm-enabled-tactics; "Unmanned Aircraft System Traffic Management (UTM)," Federal Aviation Administration, August 2022, https://www.faa.gov/uas/research_development/traffic_management; "Skyborg," Air Force Research Laboratory, accessed October 2022, https://afresearchlab.com/technology/vanguards/successstories/skyborg.

[28] The mean Average Precision (mAP) metric is used to assess the performance of object detection models on benchmarks like MS COCO. It is based on four sub-metrics: confusion matrices, intersection over Union (IoU), recall, and precision; as of December 2022, the top-performing object detection model is InternImage-DCNv3-H (M3I Pre-training), which achieves 65.4 box mAP on the MS COCO benchmark; Wenhai Wang et al., "InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions," arXiv preprint arXiv:2205.01917 (2022), https://arxiv.org/pdf/2211.05778v2.pdf.

[29] The most popular benchmark for object detection is MS COCO. We recognize there are other benchmarks catered to the application of object detection on drones, such as VisDrone. However, we opted to use COCO because more open-source models are tested against it and more data is available.

[30] Fine-tuning object detection models can substantially improve performance (above what the pre-trained versions can achieve on benchmarks like COCO). For example, researchers applied image-processing techniques to several YOLO models, which improved performance in detecting small drones to nearly 96% mAP (30% better than the top performance on COCO). However, smaller versions of

these fine-tuned models, and their real-world application on mobile drones, would likely perform worse; Vedanshu Dewangan et al., "Application of Image Processing Techniques for UAV Detection Using Deep Learning and Distance-Wise Analysis," *Drones*, March 2023, https://doi.org/10.3390/drones7030174.

[31] Andrei Barbu et al., "ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," MIT, CSAIL and CBMM, 2019, https://papers.nips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

[32] Andrew Lohn, "Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance," arXiv preprint arXiv:2009.00802 (2020), https://arxiv.org/abs/2009.00802.

[33] "Unmanned Aircraft System Airspace Integration Plan," U.S. Department of Defense, March 2011, https://info.publicintelligence.net/DoD-UAS-AirspaceIntegration.pdf.

[34] We assessed three small drones for this case study: the Skydio 2+, Lanius, and Kargu-2.

[35] E.g., the InternImage-XL model, which achieves 64.3 box AP on COCO test-dev, has over 600 million parameters and likely requires well over 1 TFLOPs of compute to run inference. This exceeds the compute capacity of many low-power chips designed for edge devices like small drones.

[36] The GPU card/board is the platform where a processor is embedded. It is larger than the processor itself, and consumes additional power.

[37] Zhongli Liu et al., "Rise of Mini-Drones: Applications and Issues," Association for Computing Machinery, June 2015, https://dl.acm.org/doi/abs/10.1145/2757302.2757303; Jamie Cole, "How Much Weight Can A Drone Carry (Detailed Guide)," *Discovery of Tech*, May 2023, https://discoveryoftech.com/how-much-weight-can-a-drone-carry; "How Much Weight Can A Small Drone Carry?" UAV Systems International, accessed May 2023, https://uavsystemsinternational.com/blogs/drone-guides/how-much-weight-can-a-small-drone-carry.

[38] As of December 2022, the state-of-the-art object detector on the COCO benchmark achieves about 65% precision, which is 8% higher precision than the largest YOLOv7 model. We referenced and used the YOLOv7 model for this case study, as opposed to one of the top performers, because it has more open-source specifications and was readily available on GitHub to use for experimentation.

[39] We use these images of detections only to illustrate performance differences between small and large object detection models, which would persist to a degree after fine-tuning.

[40] James Byrne et al., "The Orlan Complex: Tracking the Supply Chains of Russia's Most Successful UAV," *RUSI*, December 2022, https://rusi.org/explore-our-research/publications/special-resources/orlan-complex-tracking-supply-chains-russias-most-successful-uav.

[41] The Orlan-10 uses a GUM3703FEY computer-on-module (COM). The COM uses a PowerVR SGX530 Graphics Accelerator that was developed in 2005, which can only process 1.6 GFLOPS. It cannot run most (if any) computer vision models.

[42] Zhongli Liu et al., "Rise of Mini-Drones: Applications and Issues," Association for Computing Machinery, June 2015, https://dl.acm.org/doi/abs/10.1145/2757302.2757303; Arjomandi et al., "Classification of Unmanned Aerial Vehicles"; Cole, "How Much Weight Can A Drone Carry (Detailed Guide)."

[43] Gianluca Giuffrida et al., "The Phi-Sat-1 Mission: The First On-Board Deep Neural Network Demonstrator for Satellite Earth Observation," IEEE, 2022, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9600851&tag=1; Gianluca Giuffrida et al., "CloudScout: A Deep Neural Network for On-Board Cloud Detection on Hyperspectral Images," *Remote Sensing*, July 2020, https://www.mdpi.com/2072-4292/12/14/2205/htm; Emilio Rapuano et al., "An FPGA-Based Hardware Accelerator for CNNs Inference on Board Satellites: Benchmarking with Myriad 2-Based Solution for the CloudScout Case Study," *Remote Sensing*, April 2021, https://www.mdpi.com/2072-4292/13/8/1518.

[44] Images that contained clouds above a 70% threshold (i.e., more than 70% of the pixels in an image contained clouds) were automatically discarded.

[45] The board "was selected as a suitable AI accelerator for the Phi-Sat-1 mission as it provides an ideal base platform from which to build a complete inference engine, providing a payload-compatible, host-controlled, reconfigurable, low power, low heat generation, high-speed interface device, in a form-factor that integrates into the available space atop the sensor payload." Oscar Deniz et al., "Eyes of Things," *Sensors*, May 2017, https://www.mdpi.com/1424-8220/17/5/1173.

[46] A VPU is an AI accelerator designed to process computer vision calculations. The VPU onboard the CubeSat was an Intel Movidus Myriad 2.

[47] In addition to the compute hardware used for inference, the onboard camera contained a CPU, GPU, and custom chip that pre-processed imagery data so it could be interpreted by the model.

[48] I.e., quantized from FP32 to FP16.

[49] Only 30% of the hyperspectral imagery bands were used for image classification.

[50] Radiation-hardened compute is expensive and difficult to develop, and has far less compute capacity than conventional processors. For example, radiation-hardened processors like the RAD5510, RAD5515, and RAD5545 can only process 0.9-3.7 GFLOPS, respectively. This compute capacity is extremely low and cannot effectively run most AI models; "RAD5545™ SpaceVPX single-board

computer," BAE Systems, 2018, https://www.baesystems.com/en-media/uploadFile/20210404061759/1434594567983.pdf; "RAD5510™ Single-Core System-on-Chip Power Architecture® Processor," BAE Systems, 2017, https://web.archive.org/web/20190226111330/https://www.baesystems.com/en-us/download-en-us/20181211162242/1434571362333.pdf.

[51] Note that the experiment did not challenge the model on aerial images of snow or ice, which are more likely to generate false positive cloud classifications. Therefore it is difficult to discern the true quality of the model based on the results of the mission. Notwithstanding, the experiment showed that it is plausible to run AI inference on a CubeSat using commercial compute hardware (for a period of time).

[52] Such as classifying hundreds of different classes, as opposed to just two classes ('cloudy' vs. 'not cloudy').

[53] The CoAtNet-7 model, which is the 5th top-performing image classification model as of December 2022. It uses 2.5 TFLOPs during inference, but the VPU on-board the CubeSat only has a (purported) maximum compute capacity of 1 TFLOPS in FP16. Moreover, it would require even more compute to use top-performing models at higher mathematical precision (i.e., running the models in FP32 instead of FP16), and memory constraints would likely inhibit real-time inference, making them unworkable for time-sensitive operations. We selected this model because the publications for the top four models lack GFLOPs data. This makes CoAtNet-7 the top-performing model for which we have data on GFLOPs requirements.

[54] The recent trend is largely related to the application of transformers to computer vision tasks (i.e., vision transforms). Transformers are often larger than other types of models used for computer vision, such as Convolutional Neural Networks (CNNs).

[55] Xuanyao Chen et al., "SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer," arXiv preprint arXiv:2303.17605 (2023), https://arxiv.org/abs/2303.17605; Zhijian Liu et al., "FlatFormer: Flattened Window Attention for Efficient Point Cloud Transformer," arXiv preprint arXiv:2301.08739 (2023), https://arxiv.org/abs/2301.08739.

[56] The DaViT-B model achieves 86.9% accuracy on the ImageNet benchmark with 87.9 million parameters and requires 46.4 GFLOPs of compute to run inference. This is within the memory capacity and (purported) theoretical maximum compute capacity (1 TFLOPS) of the Myriad 2 VPU on the CubeSat. The larger DaViT-H model could potentially function on the VPU, but there would likely be memory constraints.

[57] Semantic segmentation models were also deployed on the ISS, but this case study focuses on image classification.

[58] "Artificial Intelligence and Advanced Flight Computing on the International Space Station," NASA Jet Propulsion Laboratory Artificial Intelligence Group, accessed September 2022, https://ai.jpl.nasa.gov/public/projects/iss/; Emily Dunkel et al., "Benchmarking Deep Learning Inference of Remote Sensing Imagery on the Qualcomm Snapdragon and Intel Movidius Myriad X Processors Onboard the International Space Station," NASA Jet Propulsion Laboratory and Hewlett Packard Enterprise, July 2022, https://ai.jpl.nasa.gov/public/documents/papers/IGARSS-2022-DL-Movidius-camera.pdf; Jason Swope et al., "Benchmarking Remote Sensing Image Processing and Analysis on the Snapdragon Processor Onboard the International Space Station," NASA Jet Propulsion Laboratory and Hewlett Packard Enterprise, July 2022, https://ai.jpl.nasa.gov/public/documents/papers/IGARSS2022-Onboard-Not-DL-IGARSS2022-Camera.pdf.

[59] The solar arrays on the ISS can generate up to 120 kW (120,000 W) of electricity, while those on the 6U CubeSat can generate up to 60 W; "About the Space Station Solar Arrays," NASA, August 2017, https://www.nasa.gov/mission_pages/station/structure/elements/solar_arrays-about.html; "6U CubeSat Bus," ISISPACE, accessed September 2022, https://www.isispace.nl/product/6u-cubesat-bus/.

[60] E.g., as of December 2022, the two top-performing image classification models on the ImageNet benchmark are CoCa (fine-tuned) and Model Soups (BASIC-L), both of which would likely run with the chips used onboard the ISS for this mission.

[61] Two Snapdragon 855 system-on-chips (SoCs) and two Movidius™ Myriad™ X VPUs were deployed on the ISS to run AI inference. They have the aggregate compute capacity to run most, if not all, of the top-performing image classification models (on the ImageNet benchmark in FP16 precision). However, their aggregate compute capacity was irrelevant for the experiment because each processor was used separately (i.e., they were not interconnected); "Qualcomm Snapdragon 855," *CompareDen*, September 2022, https://compareden.com/qualcomm-snapdragon-855/; "Enhanced Visual Intelligence at the Network Edge," Intel, accessed August 2022, https://www.intel.com/content/www/us/en/products/docs/processors/movidius-vpu/myriad-x-product-brief.html.

[62] "Hewlett Packard Enterprise accelerates space exploration with first ever in-space commercial edge computing and artificial intelligence capabilities", Hewlett Packard Enterprise, February 2021, https://www.hpe.com/us/en/newsroom/press-release/2021/02/hewlett-packard-enterprise-accelerates-space-exploration-with-first-ever-in-space-commercial-edge-computing-and-artificial-intelligence-capabilities.html.

[63] "Tesla Q3 2018 Earnings Call - Financial Results and Q&A Webcast [LIVE]," YouTube, [15:28-16:23], October 2018, https://www.youtube.com/watch?v=v0U7orfKEhM&t=761s; "Tesla Inc. (TSLA) CEO Elon Musk on Q3 2018 Results - Earnings Call Transcript," Seeking Alpha, October 2018, https://seekingalpha.com/article/4214138-tesla-inc-tsla-ceo-elon-musk-on-q3-2018-results-earnings-call-transcript.

[64] Abhishek Balasubramaniam and Sudeep Pasricha, “Object Detection in Autonomous Vehicles: Status and Open Challenges,” arXiv preprint arXiv:2201.07706 (2022), https://arxiv.org/abs/2201.07706.

[65] “FSD Chip - Tesla,” *WikiChip*, accessed August 2022, https://en.wikichip.org/wiki/tesla_(car_company)/fsd_chip; Ezra Dyer, “Why Do Electric Cars Still Use 12-Volt Batteries?” *Car and Driver*, January 2022, https://www.caranddriver.com/features/a38537243/electric-cars-12-volt-batteries/; “Why Does Tesla Have a 12v Battery?,” *ProVsCons*, accessed July 2022, https://provscons.com/why-does-tesla-have-a-12v-battery/; “Tesla Flat 12v Battery,” TeslaInfo, September 2022, https://tesla-info.com/blog/tesla-flat-battery.php.

[66] “Tesla Hardware 3 (Full Self-Driving Computer) Detailed,” *Autopilot Review*, accessed August 2022, https://www.autopilotreview.com/tesla-custom-ai-chips-hardware-3/; Devin Coldewey, “Tesla vaunts creation of ‘the best chip in the world’ for self-driving,” *TechCrunch*, April 2019, https://techcrunch.com/2019/04/22/tesla-vaunts-creation-of-the-best-chip-in-the-world-for-self-driving/.

[67] Aishwarya Srivastava et al., “Performance and Memory Trade-offs of Deep Learning Object Detection in Fast Streaming High-Definition Images,” 2018 IEEE International Conference on Big Data, 2018, https://par.nsf.gov/servlets/purl/10107324.

[68] As of December 2022, the best real-time object detection model on the MS COCO benchmark is YOLOv7-E6E. It achieves 56.8 (box) mAP at 36 FPS with a V100 GPU. However, it processes 843 GFLOPs during inference, which exceeds the capacity of the FSD chip. Even if the FSD could process enough GFLOPS, it is still unclear if it could process quickly enough to reach the real-time speeds necessary for autonomous driving.

[69] As of December 2022, the 3rd best real-time object detection model on the MS COCO benchmark is YOLOv7-E6. It achieves 56 box AP on the COCO benchmark at 56 FPS with a V100 GPU. It processes 515 GFLOPs during inference.

[70] “Waymo Driver,” Waymo, accessed May 2023, https://waymo.com/waymo-driver/.

[71] Andrew Lohn and Micah Musser, “AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?,” The Center for Security and Emerging Technology (CSET*)*, January 2022, https://cset.georgetown.edu/publication/ai-and-compute/; Di Liu et al., “Bringing AI to Edge: From Deep Learning’s Perspective”; Jordan Hoffmann et al., “Training Compute-Optimal Large Language Models,” arXiv preprint arXiv:2203.15556 (2022), https://arxiv.org/abs/2203.15556; Hugo Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” arXiv preprint arXiv:2302.13971 (2023), https://arxiv.org/abs/2302.13971.

[72] Jeff Dean, "Google Research, 2022 & beyond: Language, vision and generative models," Google Research, January 2023, https://ai.googleblog.com/2023/01/google-research-2022-beyond-language.html.

[73] E.g., increasing the efficiency in the dataflow architecture, near-memory processing, and compute-in-memory.

[74] I.e., low-power GPUs, SoCs, VPUs, NPUs, etc.

[75] Neil C. Thompson and Svenja Spanuth, "The Decline of Computers as a General Purpose Technology," *Communications of the ACM*, March 2021, https://cacm.acm.org/magazines/2021/3/250710-the-decline-of-computers-as-a-general-purpose-technology/fulltext?mobile=false.

[76] Wazir Zada Khan, et al. "Edge Computing: A Survey," *Future Generation Computer Systems*, August 2019, https://www.sciencedirect.com/science/article/abs/pii/S0167739X18319903.

[77] Papers With Code, accessed August 2022, https://paperswithcode.com/.

[78] *TechPowerUp* is a technology publication that hosts one of the largest repositories of data on graphics cards and processors*; TechPowerUp*, accessed August 2022, https://www.techpowerup.com/.

[79] Albert Reuther et al., "Survey of Machine Learning Accelerators," 2020 IEEE High Performance Extreme Computing Conference (HPEC), September 2020, https://ieeexplore.ieee.org/abstract/document/9286149; Christoffer Åleskog et al., "Recent Developments in Low-Power AI Accelerators: A Survey," *Algorithms*, November 2022, https://www.mdpi.com/1999-4893/15/11/419.

[80] Amirali Boroumand et al., "Mitigating Edge Machine Learning Inference Bottlenecks: An Empirical Study on Accelerating Google Edge Models," arXiv preprint arXiv:2205.01917 (2021), https://arxiv.org/pdf/2103.00768.pdf.

[81] "NVIDIA H100 PCIe 80 GB," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/h100-pcie.c3899; "NVIDIA H100 Tensor Core GPU," NVIDIA, accessed November 2022, https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet.

[82] "NVIDIA GeForce RTX 4090," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/geforce-rtx-4090.c3889.

[83] "NVIDIA A100 Tensor Core GPU," NVIDIA, accessed November 2022, https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-

nvidia-us-2188504-web.pdf; "NVIDIA A100 PCIe 80 GB," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/a100-pcie-80-gb.c3821; "NVIDIA A100 80GB PCIe GPU — Product Brief," NVIDIA, accessed November 2022, https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/PB-10577-001_v02.pdf.

[84] "NVIDIA Tesla T4," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/tesla-t4.c3316.

[85] "NVIDIA RTX A6000," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/rtx-a6000.c3686.

[86] "NVIDIA GeForce RTX 3090," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/geforce-rtx-3090.c3622.

[87] "NVIDIA TITAN V," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/titan-v.c3051.

[88] "NVIDIA GeForce RTX 3080," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/geforce-rtx-3080.c3621.

[89] "NVIDIA TESLA V100 GPU ACCELERATOR," NVIDIA, accessed November 2022, https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf; "NVIDIA Tesla V100 PCIe 32 GB," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/tesla-v100-pcie-32-gb.c3184.

[90] "NVIDIA GeForce RTX 2080," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/geforce-rtx-2080.c3224.

[91] "NVIDIA Tesla P100 PCIe 16 GB," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/tesla-p100-pcie-16-gb.c2888.

[92] "Jetson Modules," NVIDIA Developer, accessed November 2022, https://developer.nvidia.com/embedded/jetson-modules; "Jetson Nano," *ELinux*, accessed November 2022, https://elinux.org/Jetson_Nano; "NVIDIA Jetson Nano," *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/jetson-nano.c3643.

[93] "Jetson TX2 NX Module," NVIDIA, accessed November 2022, https://developer.nvidia.com/embedded/jetson-tx2-nx; "AAEON Announces NVIDIA Jetson TX2 NX AI Edge Platform," *TechPowerUp*, April 2021, https://www.techpowerup.com/280748/aaeon-announces-nvidia-jetson-tx2-nx-ai-edge-platform; "NVIDIA Jetson TX2 NX System-on-Module," NVIDIA Developer, accessed November 2022, https://developer.download.nvidia.com/assets/embedded/secure/jetson/tx2_nx/Jetson_TX2_NX_DS-

10182-001_v1.4.pdf?L9s0AspbXLQLMHN_h8_pM_Jxp1BBfDPCQ6YlrGUFjekRrp5ujW8ngasGwy0gpi_ZANBewquc2XJ7R9jvNKeZo5dZuGAoVdI0wKpVu45RX7KdRhXzrQrlY1tBDfzZGppTyp5mpBDTbx5iKN16qmUwnEIw9b6KdkhCFbiTSH9eUlVX2esklod0&t=eyJscyI6ImdzZW8iLCJsc2QiOiJodHRwczovL3d3dy5nb29nbGUuY29tLyJ9.

[94] “Jetson TX2,” *ELinux*, accessed November 2022, https://elinux.org/Jetson_TX2; “Jetson Modules.”; “NVIDIA Jetson TX2,” *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/jetson-tx2.c3231.

[95] “NVIDIA Jetson Xavier,” NVIDIA, accessed November 2022, https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-nx/; “Jetson Xavier NX,” *ELinux*, accessed November 2022, https://elinux.org/Jetson_Xavier_NX; “NVIDIA Jetson Xavier NX 8 GB,” *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/jetson-xavier-nx-8-gb.c3642.

[96] “Jetson Modules”; “NVIDIA Jetson Xavier,” *NVIDIA*, accessed November 2022, https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-agx-xavier/; “NVIDIA Jetson AGX Xavier 16 GB,” *TechPowerUp*, accessed November 2022, https://www.techpowerup.com/gpu-specs/jetson-agx-xavier-16-gb.c3232.

[97] “Jetson Modules.”; Leela Subramaniam Karumbunathan, “Solving Entry-Level Edge AI Challenges with NVIDIA Jetson Orin Nano,” *NVIDIA,* September 2021, https://developer.nvidia.com/blog/solving-entry-level-edge-ai-challenges-with-nvidia-jetson-orin-nano/; “NVIDIA Jetson Orin Nano 8 GB,” *TechPowerUp,* Accessed November 2022, https://www.techpowerup.com/gpu-specs/jetson-orin-nano-8-gb.c4082.

[98] “Jetson Modules”; “The Future of Industrial-Grade Edge AI — NVIDIA Jetson AGX Orin Industrial module,” NVIDIA, accessed November 2022, https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/; “NVIDIA Jetson AGX Orin 32 GB,” *TechPowerUp,* accessed November 2022, https://www.techpowerup.com/gpu-specs/jetson-agx-orin-32-gb.c4084.

[99] “Space product literature,” BAE Systems, accessed August 2022 https://www.baesystems.com/en-us/our-company/inc-businesses/electronic-systems/product-sites/space-products-and-processing/radiation-hardened-electronics; “RAD750® radiation-hardened PowerPC microprocessor,” BAE Systems, accessed December 2022, https://www.baesystems.com/en-media/uploadFile/20210404045936/1434555668211.pdf; Jacek Krywko, “Space-grade CPUs: How do you send more computing power into space?” *Ars Technica*, November 2019, https://arstechnica.com/science/2019/11/space-grade-cpus-how-do-you-send-more-computing-power-into-space/; “The Mars 2020 Rover’s ‘Brains,’” NASA, accessed December 2022, https://mars.nasa.gov/mars2020/spacecraft/rover/brains/.

[100] "Space Product Literature"; "RAD5510™ single-core system-on-chip Power Architecture® processor," BAE Systems, accessed December 2022, https://web.archive.org/web/20190226111330/https://www.baesystems.com/en-us/download-en-us/20181211162242/1434571362333.pdf.

[101] "Space Product Literature"; "RAD5545™ SpaceVPX single-board computer," BAE Systems, accessed December 2022, https://www.baesystems.com/en-media/uploadFile/20210404061759/1434594567983.pdf.

[102] "Space Product Literature"; "RAD5545™ SpaceVPX single-board computer"; "RAD5545™ multi-core system-on-chip Power Architecture® processor," BAE Systems, accessed December 2022, https://web.archive.org/web/20190226111129/https://www.baesystems.com/en-us/download-en-us/20181211163917/1434571328901.pdf.

[103] "NVIDIA Jetson TX2 System-on-Module," NVIDIA, accessed November 2022, https://www.assured-systems.com/uploads/media/products/axiomtek/eboxs/jetson%20tx2/data%20sheet%20-%20nvidia%20jetson%20tx2%20system-on-module.pdf; Joshua Ho, "Hot Chips 2016: NVIDIA Discloses Tegra Parker Details," *AnandTech*, August 2016, https://www.anandtech.com/show/10596/hot-chips-2016-nvidia-discloses-tegra-parker-details; "Jetson TX2 Module"; "NVIDIA Jetson TX2."

[104] "Snapdragon 855 - Qualcomm," *WikiChip*, accessed August 2022, https://en.wikichip.org/wiki/qualcomm/snapdragon_800/855; Emily Dunkel et al., "Benchmarking Deep Learning Inference of Remote Sensing Imagery on the Qualcomm Snapdragon and Intel Movidius Myriad X Processors Onboard the International Space Station"; Léonie Buckley and Emily Dunkel, "Benchmarking Deep Learning On a Myriad X," *NASA Jet Propulsion Laboratory*, 2022, https://ai.jpl.nasa.gov/public/documents/presentations/FSW2022_Benchmarking_DL_Buckley.pdf; NOTE: This link has a forbidden access message "Qualcomm Snapdragon 855," *NotebookCheck*, accessed August 2022, https://www.notebookcheck.net/Qualcomm-Snapdragon-855-SoC-Benchmarks-and-Specs.375436.0.html; "Qualcomm Adreno 640," *CPUMonkey*, accessed August 2022, https://www.cpu-monkey.com/en/igpu-qualcomm_adreno_640-264.

[105] "Intel® Movidius™ Myriad™ 2 Vision Processing Unit (VPU)," Intel, accessed August 2022, https://newsroom.intel.com/wp-content/uploads/sites/11/2017/06/Myriad-2-VPU-Fact-Sheet.pdf; "Intel® Movidius™ Myriad™ 2 Vision Processing Unit 4GB," Intel, accessed August 2022, https://www.intel.com/content/www/us/en/products/sku/122461/intel-movidius-myriad-2-vision-processing-unit-4gb/specifications.html; Oscar Deniz et al., "Eyes of Things."

[106] "Intel® Movidius™ Myriad™ X Vision Processing Unit 4GB," Intel, accessed August 2022, https://www.intel.com/content/www/us/en/products/sku/125926/intel-movidius-myriad-x-vision-processing-unit-4gb/specifications.html; "Enhanced Visual Intelligence at the Network Edge," Intel,

accessed August 2022, https://www.intel.com/content/www/us/en/products/docs/processors/movidius-vpu/myriad-x-product-brief.html; Emily Dunkel et al., "Benchmarking Deep Learning Inference of Remote Sensing Imagery on the Qualcomm Snapdragon and Intel Movidius Myriad X Processors Onboard the International Space Station."; Léonie Buckley and Emily Dunkel, "Benchmarking Deep Learning On a Myriad X."; Nate Oh, "Intel Announces Movidius Myriad X VPU, Featuring 'Neural Compute Engine'," *AnandTech*, August 2017, https://www.anandtech.com/show/11771/intel-announces-movidius-myriad-x-vpu.

[107] "PowerVR," *Wikipedia*, accessed February 2023, https://en.wikipedia.org/wiki/PowerVR#Series5_(SGX).

[108] "FSD Chip - Tesla"; "Why Does Tesla Have a 12v Battery?"

[109] "Apple A16 Bionic," *NanoReview*, accessed March 2023, https://nanoreview.net/en/soc/apple-a16-bionic; "Apple A16 Bionic Benchmark, Test and specs," *CPUMonkey*, accessed March 2023, https://www.cpu-monkey.com/en/cpu-apple_a16_bionic.

[110] "Apple A15 Bionic," *NanoReview*, accessed March 2023, https://nanoreview.net/en/soc/apple-a15-bionic; "Apple A15 (5 GPU Cores)," *CPUMonkey*, accessed March 2023, https://www.cpu-monkey.com/en/igpu-apple_a15_5_gpu_cores-275.

[111] "A12 Bionic - Apple," *WikiChip*, accessed January 2023, https://en.wikichip.org/wiki/apple/ax/a12; "Apple A12 Bionic," *NanoReview*, accessed January 2023, https://nanoreview.net/en/soc/apple-a12-bionic; "Apple A12," *CPUMonkey*, accessed January 2023, https://www.cpu-monkey.com/en/igpu-apple_a12-155.

[112] "MacBook Pro (14-inch, 2023) - Technical Specifications," Apple, accessed March 2023, https://support.apple.com/kb/SP889?locale=en_US; "Apple M2 Pro," *NanoReview*, accessed March 2023, https://nanoreview.net/en/cpu/apple-m2-pro; "Apple M2 Pro (19 Core)," *CPUMonkey*, accessed March 2023, https://www.cpu-monkey.com/en/igpu-apple_m2_pro_19_core-352.

[113] We used Papers With Code as a primary source to investigate models. However, we manually extracted data from all papers due to insufficient and sometimes incorrect information on Papers With Code.

[114] The data on object detection models was relatively limited, while the data on machine translation models was grossly insufficient.

[115] Jiahui Yu et al., "CoCa: Contrastive Captioners are Image-Text Foundation Models," arXiv preprint arXiv:2205.01917 (2022), https://arxiv.org/abs/2205.01917.

[116] Mitchell Wortsman et al., “Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time,” arXiv preprint arXiv:2203.05482 (2022), https://arxiv.org/abs/2203.05482.

[117] Wortsman et al., “Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.”

[118] Xi Chen et al., “PaLI: A Jointly-Scaled Multilingual Language-Image Model,” arXiv preprint arXiv:2209.06794 (2022), https://arxiv.org/abs/2209.06794.

[119] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan, “CoAtNet: Marrying Convolution and Attention for All Data Sizes,” arXiv preprint arXiv:2106.04803 (2021), https://arxiv.org/abs/2106.04803.

[120] Dai, Liu, Le, and Tan, “CoAtNet: Marrying Convolution and Attention for All Data Sizes.”

[121] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, Lucas Beyer, “Scaling Vision Transformers,” arXiv preprint arXiv:2106.04560 (2022), https://arxiv.org/abs/2106.04560.

[122] Mingyu Ding et al., “DaViT: Dual Attention Vision Transformers,” arXiv preprint arXiv:2204.03645 (2022), https://arxiv.org/abs/2204.03645.

[123] Ding et al., “DaViT: Dual Attention Vision Transformers.”

[124] Zhengzhong Tu et al., “MaxViT: Multi-Axis Vision Transformer,” arXiv preprint arXiv:2204.01697 (2022), https://arxiv.org/abs/2204.01697.

[125] Tu et al., “MaxViT: Multi-Axis Vision Transformer.”

[126] Tu et al., “MaxViT: Multi-Axis Vision Transformer.”

[127] Andrew Brock, Soham De, Samuel L. Smith, and Karen Simonyan, “High-Performance Large-Scale Image Recognition Without Normalization,” arXiv preprint arXiv:2102.06171 (2021), https://arxiv.org/abs/2102.06171.

[128] Wenhai Wang et al., “InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions,” arXiv preprint arXiv:2211.05778 (2022), https://arxiv.org/abs/2211.05778.

[129] Tu et al., “MaxViT: Multi-Axis Vision Transformer.”

[130] Chenglin Yang et al., “MOAT: Alternating Mobile Convolution and Attention Brings Strong Vision Models,” arXiv preprint arXiv:2210.01820 (2023), https://arxiv.org/abs/2210.01820.

[131] Yanghao Li et al., “MViTv2: Improved Multiscale Vision Transformers for Classification and Detection,” arXiv preprint arXiv:2112.01526 (2022), https://arxiv.org/abs/2112.01526.

[132] Tu et al., “MaxViT: Multi-Axis Vision Transformer.”

[133] Mannat Singh et al., “Revisiting Weakly Supervised Pre-Training of Visual Perception Models,” arXiv preprint arXiv:2201.08371 (2022), https://arxiv.org/abs/2201.08371.

[134] Dai, Liu, Le, and Tan, “CoAtNet: Marrying Convolution and Attention for All Data Sizes.”

[135] Tu et al., “MaxViT: Multi-Axis Vision Transformer.”

[136] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou, “Fixing the train-test resolution discrepancy: FixEfficientNet,” arXiv preprint arXiv:2003.08237 (2020), https://arxiv.org/abs/2003.08237.

[137] Li et al., “MViTv2: Improved Multiscale Vision Transformers for Classification and Detection.”

[138] Weihao Yu et al., “MetaFormer Baselines for Vision,” arXiv preprint arXiv:2210.13452 (2022), https://arxiv.org/abs/2210.13452.

[139] Li et al., “MViTv2: Improved Multiscale Vision Transformers for Classification and Detection.”

[140] Ding et al., “DaViT: Dual Attention Vision Transformers.”

[141] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detector,” arXiv preprint arXiv:2207.02696 (2022), https://arxiv.org/abs/2207.02696.

[142] Wang, Bochkovskiy, and Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detector.”

[143] Wang, Bochkovskiy, and Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detector.”

[144] Wang, Bochkovskiy, and Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detector.”

[145] Wang, Bochkovskiy, and Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detector."

[146] Chengqi Lyu et al., "RTMDet: An Empirical Study of Designing Real-Time Object Detectors," arXiv preprint arXiv:2212.07784 (2022), https://arxiv.org/abs/2212.07784.

[147] Wang, Bochkovskiy, and Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detector."

[148] Zheng Ge et al., "YOLOX: Exceeding YOLO Series in 2021," arXiv preprint arXiv:2107.08430 (2021), https://arxiv.org/abs/2107.08430.

[149] Ge et al., "YOLOX: Exceeding YOLO Series in 2021."

[150] Xianzhe Xu et al., "DAMO-YOLO: A Report on Real-Time Object Detection Design," arXiv preprint arXiv:2211.15444 (2022), https://arxiv.org/abs/2211.15444.

[151] Ge et al., "YOLOX: Exceeding YOLO Series in 2021."

[152] Xu et al., "DAMO-YOLO: A Report on Real-Time Object Detection Design."

[153] Xu et al., "DAMO-YOLO: A Report on Real-Time Object Detection Design."

[154] Shangliang Xu et al., "PP-YOLOE: An evolved version of YOLO," arXiv preprint arXiv:2203.16250 (2022), https://arxiv.org/abs/2203.16250.

[155] Ge et al., "YOLOX: Exceeding YOLO Series in 2021."

[156] Wang, Bochkovskiy, and Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detector."

[157] Noah Shachtman, "Unlimited Talk, Only $679 Million: Inside the No-Bid Deal for Afghan Interpreters," *Wired*, May 2010, https://www.wired.com/2010/05/unlimited-talk-only-679-million-inside-the-no-bid-deal-for-afghan-interpreters/.

[158] Neil Shea, "Foreign Policy: Losing Afghanistan in Translation," NPR, August 2010, https://www.npr.org/2010/08/24/129396818/foreign-policy-losing-afghanistan-in-translation; Ron Synovitz, "Mistakes by Afghan Translators Endanger Lives, Hamper Antiterrorism Effort," *RFERL*, September 2008,

https://www.rferl.org/a/Mistakes_By_Translators_Hamper_Afghan_Antiterrorism_Campaign/1195783.html.

[159] Surangika Ranathunga et al., “Neural Machine Translation for Low-Resource Languages: A Survey,” arXiv preprint arXiv:2106.15115 (2021), https://arxiv.org/abs/2106.15115.

[160] Mitchell A. Gordon et al., “Data and Parameter Scaling Laws for Neural Machine Translation,” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, https://aclanthology.org/2021.emnlp-main.478/.

[161] For example, see Figure 7 in Angela Fan et al., “Beyond English-Centric Multilingual Machine Translation,” arXiv preprint arXiv:2010:11125 (2020), https://arxiv.org/abs/2010.11125.

[162] Julie Cattiau, “Offline translations are now a lot better thanks to on-device AI,” Google, June 2018, https://blog.google/products/translate/offline-translations-are-now-lot-better-thanks-device-ai/.

[163] Translation quality was determined by calculating the Bilingual Evaluation Understudy (BLEU) score for several pages of translated French-English text that were sourced from the European Parliament Proceedings Parallel Corpus, as well as from the French news website *Le Monde*. BLEU is a common metric to assess machine translation performance. Mean BLEU scores were 0.35 for online translation and 0.29 for offline translation. However, we should note that there are several shortcomings of this metric, including how it labels synonyms as incorrect translations, good and bad text variations can score the same, and the score is related to a specific test set and language pair (so they do not provide a clear measure of translation quality); “European Parliament Proceedings Parallel Corpus 1996-2011,” Statmt, accessed February 2023, https://www.statmt.org/europarl/.

[164] Increasing model size is of limited value if the data set is too small, as seen in: Jordan Hoffman et al., “Training Compute-Optimal Large Language Models,” arXiv preprint arXiv:2205.01917 (2022), https://arxiv.org/abs/2203.15556 and Gordon et al., “Data and Parameter Scaling Laws for Neural Machine Translation.”

[165] Angela Fan et al., “Beyond English-Centric Multilingual Machine Translation,” *Journal of Machine Learning Research*, March 2021, https://www.jmlr.org/papers/volume22/20-1307/20-1307.pdf.