# EDGE DEVICES INFERENCE PERFORMANCE COMPARISON

### A PREPRINT

### Rafał Tobiasz\*

Bulletprove sp. z o. o. Ignacego Mościckiego 1 24-110 Puławy, Poland rafal.tobiasz@bulletprove.com

### Piotr Graszka

Bulletprove sp. z o. o. Ignacego Mościckiego 1 24-110 Puławy, Poland piotr.graszka@bulletprove.com

# Grzegorz Wilczyński\*

Bulletprove sp. z o. o. Ignacego Mościckiego 1 24-110 Puławy, Poland grzegorz.wilczynski@bulletprove.com

### Nikodem Czechowski

Bulletprove sp. z o. o. Ignacego Mościckiego 1 24-110 Puławy, Poland nikodem.czechowski@bulletprove.com

### Sebastian Łuczak

Bulletprove sp. z o. o. Ignacego Mościckiego 1 24-110 Puławy, Poland sebastian.luczak@bulletprove.com

June 22, 2023

# **ABSTRACT**

In this work, we investigate the inference time of the MobileNet family, EfficientNet V1 and V2 family, VGG models, Resnet family, and InceptionV3 on four edge platforms. Specifically NVIDIA Jetson Nano, Intel Neural Stick, Google Coral USB Dongle, and Google Coral PCIe. Our main contribution is a thorough analysis of the aforementioned models in multiple settings, especially as a function of input size, the presence of the classification head, its size, and the scale of the model. Since throughout the industry, those architectures are mainly utilized as feature extractors we put our main focus on analyzing them as such. We show that Google platforms offer the fastest average inference time, especially for newer models like MobileNet or EfficientNet family, while Intel Neural Stick is the most universal accelerator allowing to run most architectures. These results should provide guidance for engineers in the early stages of AI edge systems development. All of them are accessible at https://bulletprove.com/research/edge\_inference\_results.csv.

**Keywords** Edge device · Deep learning · Computer vision

### 1 Introduction

A variety of applications exploit machine learning models, be it on a cloud [Brown et al., 2020, OpenAI, 2022, Rombach et al., 2021], personal computers, mobile devices [Bazarevsky et al., 2019, Team, 2017], or edge devices [Zhang et al., 2017]. Especially for the latter we can observe intensified development of devices and suitable algorithms, since progressively more IoT applications use AI solutions.

A significant fraction of all those applications is in the computer vision domain, such as classification [Liu et al., 2022, 2021], object detection [Redmon and Farhadi, 2018, Liu et al., 2015], image segmentation [He et al., 2017]. Those algorithms require vast amount of computational power to perform inference since the input data is high resolution.

<sup>\*</sup>Equal contribution.

Also, other reasons encourage the development of edge devices:

- network load sending high-resolution data from a vast amount of IoT devices to the computational unit may result in unwanted and unpredicted time delays [Yu et al., 2018, Mao et al., 2017],
- computational unit load analyzing high-resolution data using current state-of-the-art models may result in a cost-inefficient system [Amodei and Hernandez, 2018],
- safety sending raw data to the cloud may get targeted by hackers [Neshenko et al., 2019][Franceschi-Bicchierai, 2017a,b] or could lower the trust of a user who, e.g., for a face detection system, does not want his photos on an undisclosed server. Instead, it is better to use a feature extractor on an edge device and send only those anonymous features to the cloud.

To address those problems, many companies are targeting specialized inference chips, which result in a vast amount of different edge accelerators. Moreover, currently, various architectures are well-scalable and can extract features correctly. However, picking "the best" algorithm or platform is not possible. It depends on the application. Therefore, an engineer wanting to choose a platform and a model to start with, faces the time-consuming task of testing different variants.

Few papers address model profiling on different AI accelerators, like in work done by Reza et al. [2021], although only for a few architectures (often with default parameters). Excellent research in this field is done by Reddi et al. [2019] which defines the proper way of different profiling methods. It allows users to perform their tests and share their results on a moderated platform.

However, none of those papers present any data regarding feature extractors. Authors analyze models only as classifiers. Whereas, those pre-trained models are mostly used as feature extractors. Therefore, analyzing them as classifiers may be slightly confusing in this setting. Profiling those algorithms only as feature extractors provides clear information on how long the first component of the final algorithm will last.

In this work, we present an extensive comparison of the most popular models available in TensorFlow [Google, 2023a] Keras Applications: MobileNet family [Howard et al., 2017, Sandler et al., 2018, Howard et al., 2019], EfficientNet (V1 and V2) family [Tan and Le, 2019, 2021], ResNet (V1 and V2) family [He et al., 2015, 2016], VGG family [Simonyan and Zisserman, 2014], and InceptionV3 [Szegedy et al., 2015] on multiple platforms: NVIDIA Jetson Nano, Google Coral USB, Google Coral PCI, and Intel Neural Stick. Those algorithms were also analyzed in different settings, e.g., input sizes, scaling parameters, and more. Despite focusing on profiling models as feature extractors we also examined them with a classic ImageNet head (1000 classes) and a more real-life scenario (5 output neurons).

The motivation behind this work is to create an in-depth comparison of the performance of different models on multiple edge devices so that it could make the work of fellow ML engineers more time- and cost-efficient.

# 2 Edge AI Accelerators

AI on Edge is focusing on running artificial intelligence models on Edge devices [Deng et al., 2019]. It favors low power consumption and small physical size at the cost of performance. An Edge AI accelerator is hardware specialized in processing AI workloads at the edge. Computation is local, close to data collection, which can be beneficial in preserving data privacy or in offline scenarios. Moreover, it reduces latency and communication costs when compared to Cloud AI.

### 2.1 Google Coral Accelerator

Google Coral Accelerator expands the user's system with an application-specific integrated circuit (ASIC) called Edge TPU, designed to deploy high-quality AI at the edge. Coral can perform 4 trillion operations per second using 2 watts of power. Edge TPU supports only 8-bit quantized Tensorflow Lite models compiled using a dedicated tool [Google, 2020a]. Potential use cases cover predictive maintenance, voice recognition, anomaly detection, machine vision, robotics, and more [Google, 2023b]. In this paper, we tested two IO Interface versions of Google Coral: PCIe and USB.

# 2.2 Nvidia Jetson Nano

Nvidia Jetson Nano is a small computer equipped with: 128-core NVIDIA Maxwell GPU, Quad-core ARM Cortex-A57 MPCore processor and 4GB of 64-bit LPDDR4 of memory. It is the only standalone edge device used in this comparison. The platform has an AI performance of 472 giga-floating point operations per second (GFLOPS) and uses

5 to 10 watts of power [NVIDIA, 2023a]. Jetson utilizes NVIDIA JetPack SDK, which provides useful features like CUDA, cuDNN and TensorRT. TensorRT optimizes inference of deep learning frameworks. For example, it allows for the use of FP16 precision for inference [NVIDIA, 2023b]. Potential use cases include predictive maintenance, voice recognition, anomaly detection, machine vision, robotics, patient monitoring, traffic management, and many more [NVIDIA, 2023c].

### 2.3 Intel Neural Compute Stick 2

Neural Compute Stick 2 (NCS2) is a plug-and-play AI accelerator. It contains Intel Movidius Myriad X Vision Processing Unit (VPU), which includes 16 SHAVE cores and a dedicated DNN hardware accelerator. Intel Distribution of the OpenVINO toolkit is used to convert and optimize models for this platform. Potential use cases cover anomaly detection, machine vision, and more [Intel, 2019].

# 3 Methodology

# 3.1 Compared models

TensorFlow is one of the most popular deep learning frameworks. Alongside many tools for creating, training, and profiling neural networks it also provides a set of already trained popular image classification networks. Those architectures are broadly used in the industry. We picked multiple algorithm families based on multiple motives.

### 3.1.1 MobileNet

This family of models consists of MobileNetV1, MobileNetV2, MobileNetV3-Small, and MobileNetV3-Large. They were particularly designed to run efficiently on mobile devices, hence the name.

MobileNetV1 [Howard et al., 2017] is based on a streamlined architecture that uses depthwise separable convolutions. The authors introduced two hyperparameters that efficiently tradeoff between latency and accuracy:

- Width multiplier  $\alpha$  number of layer's input channels M (where M is the default number of channels) becomes  $\alpha * M$ , and the number of output channels N becomes  $\alpha * N$ . It takes place for every layer.
- Resolution multiplier  $\rho$  is implicitly set by setting the input resolution.

Its fast inference is an effect of putting nearly all of the computation into dense 1x1 convolutions. It is crucial because they are implemented with highly optimized general matrix multiply (GEMM) functions and do not need any initial reordering in memory.

MobileNetV2 [Sandler et al., 2018] is based on an inverted residual structure where the shortcut connections are between the thin bottleneck layers. Lightweight depthwise convolutions are here the source of non-linearity since they were removed in the narrow layers. The two prior introduced hyperparameters remain the same. This network improved the state-of-the-art for a wide range of performance points at that time for ImageNet [Deng et al., 2009].

MobileNetV3 [Howard et al., 2019] models leverage complementary search techniques complemented by the NetAdapt algorithm. The authors also introduced the swish activation function and rebuilt MobileNetV2's bottleneck by adding squeeze and excitation in the residual layer. MobileNetV3 family outperformed, at that time, the current state-of-the-art models on ImageNet taking into account top-1 accuracy and latency.

### 3.1.2 EfficientNet (V1 and V2)

In their work [Tan and Le, 2019], the authors improved model scaling and identified that balancing network depth, width, and resolution leads to better performance. They focused on optimizing FLOPS rather than latency since they did not target specific hardware. By using NAS (neural architecture search) they came up with a new scalable baseline network and called this family of models EfficientNets (now EfficientNets V1). There are eight models available in Tensorflow applications from B0 to B7.

In the following paper, the authors once again used a combination of NAS and scaling to optimize training speed and parameter efficiency. It resulted in the EfficientNetV2 [Tan and Le, 2021] family. It achieved the state-of-the-art top-1 accuracy on ImageNet, outperforming even the famous ViT [Dosovitskiy et al., 2020]. This family consists of seven models, from B0 to B3 and S, M, and L.

### 3.1.3 InceptionV3

As the Inception family of models had a crucial role in the development of convolutional neural networks for vision tasks, we decided to profile InceptionV3 [Szegedy et al., 2015]. Here authors aimed to utilize added computation more efficiently by factorizing convolutions and adding aggressive regularization. InceptionV3 achieved, at the time, the state-of-the-art top-1 error on ImageNet.

#### 3.1.4 VGG

As with the former model, we decided to profile this [Simonyan and Zisserman, 2014] family of networks based on historical reasons. The authors' main contribution was to increase the depth of the convolutional network using small 3x3 kernels. This allowed them to build 16-19 (VGG16, VGG19) layers architectures that earned first and second place at the ImageNet Challenge 2014 in the localization and classification tasks respectively.

### 3.1.5 ResNet (V1 and V2)

In their work [He et al., 2015], the authors introduced residual connections that allowed training substantially deeper networks than their predecessors. On ImageNet they evaluated ResNet152 - networks 8x deeper than VGGs and yet having lower complexity - achieving state-of-the-art. We profiled ResNet50, ResNet101, and ResNet152 (we refer to this family as ResNetV1).

In the following work [He et al., 2016], the authors took on the propagation formulations behind the residual building blocks, trying to make the forward and backward signals flow easier. They rethought the residual blocks, moreover removed ReLU from the "easiest" path after the addition operation. For the ResNetV2 family, we profiled analogous models to its predecessor. We profile ResNet families because of the same reasons as InceptionV3 and VGGs.

# 3.2 Model parameters

The only measurement performed in this work is the inference time. As we wanted to profile the influence of different model hyperparameters, we created extensive sets of architectures for every model family. Each neural network is built with different parameters.

### 3.2.1 Input size

Nowadays, AI applications need to analyze images of various sizes, ranging from  $224^2$  to  $1024^2$  or larger. This trend is also noticeable on edge devices. Taking into account restricted computational resources an engineer has to be able to pick a suitable architecture that will allow using the model efficiently, e.g. in real-time.

# 3.2.2 Preprocessing

Every analyzed model has some specific input preprocessing, e.g., reducing the mean and dividing by standard deviation. It is important to know how such simple data alteration will affect the model's inference time.

# 3.2.3 Classification heads

In this work, our key focus is on analyzing models as feature extractors (no classification heads). Currently, those architectures are used in this fashion in the industry hence it is crucial to know how computing features will affect the inference time. Furthermore, we profiled architectures with classic ImageNet head (1000 classes) to make our results comparable to others. Last but not least, we analyzed smaller classification heads (5 classes) which represent more "real-life" case scenarios for image classification.

#### 3.2.4 Width multiplier $\alpha$

The aforementioned parameter is specific only to the MobileNet family which allows efficient scaling of the models.

# 3.2.5 Precision

Only applicable to models on Jetson Nano, possible types are FLOAT32 and FLOAT16 (this board does not support INT8).

# 3.3 Benchmark prerequisites

Each platform imposes different model preparation techniques as well as environment requirements.

### 3.3.1 Google Coral

Each model has to be quantized, converted to a TFLite format, and compiled with the edgetpu-compiler.

The Coral USB dongle was tested on Ubuntu 20.04.5 LTS with Intel Core i5-1135G7 2.4GHz, 32GB of RAM, a USB 3.0 port, and Edge TPU runtime for Linux that operates at the maximum clock frequency [Google, 2020a].

The Coral PCIe was tested on Radxa CM3 IO Board with Ubuntu 20.04.4 LTS, 4GB of RAM, and, analogous to PC, Edge TPU runtime.

The coral benchmarking script is based on [Google, 2020b], which is also the reason for using *timeit* [Foundation, 2023] package.

### 3.3.2 Inter Neural Stick

Models were not quantized, as Neural Stick does not support quantization, only converted to ONNX format with *tf2onnx* package [ONNX, 2021].

The Intel Neural Stick was tested on Ubuntu 20.04.4 LTS with AMD Ryzen 5 1600, 128GB of RAM, and a USB 3.0 port. Moreover, it needed OpenVino-dev with additional libraries (like TensorFlow, and ONNX) [Intel, 2022a], an OpenVino toolkit [Intel, 2022b], and set up variables with *setupvars.sh* from the prior downloaded toolkit.

The Neural Stick benchmarking script is based on the *hello\_classification.py* script from the OpenVino toolkit [Intel, 2022b]. Additionally, models were compiled with the latency hint.

# 3.3.3 Nvidia Jetson Nano

Each model is saved to Protocol Buffers format, and then, is converted to TensorRT [NVIDIA, 2023d] architecture using TensorFlow's experimental converter. To make experiments quicker we compiled TensorRT models on RTX3090 and then built them on the target device.

# 3.4 Benchmarking process

Inferences in our tests are synchronous and blocking, similar to the single-stream scenario described by Reddi [Reddi et al., 2019], i.e., the batch size is equal to 1, but our metric is the whole latency time. The Benchmarking script is written in Python programming language, and inference time is measured with the *timeit* package. Models are analyzed in a grid search fashion over all possible parameters:  $input\_size$ , preprocessing,  $classification\_head$ ,  $\alpha$ , and precision. The process runs as follows:

- model creation from TensorFlow's application module,
- outcome compilation to meet platform-specific requirements,
- an input image creation of size [input\_size, input\_size, 3] with platform-specific data type,
- warmup inferences with a compiled model on a selected platform (10 for every run),
- proper inferences along with time measurement.

At the end of every grid search run, the script creates a CSV file with

- · model parameters,
- minimum inference time,
- maximum inference time,
- mean inference time.

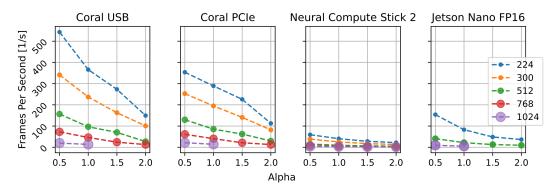


Figure 1: Scaling up MobileNetV2 alpha parameter on four platforms. Missing data points are the result of compilation errors.

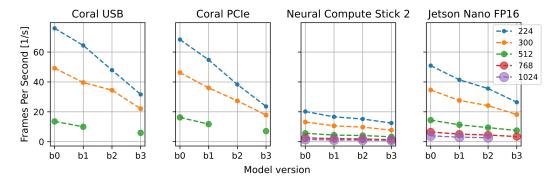


Figure 2: Scaling EfficientNetV2 on four platforms. Missing data points on EfficientNetV2B2 are the result of compilation errors.

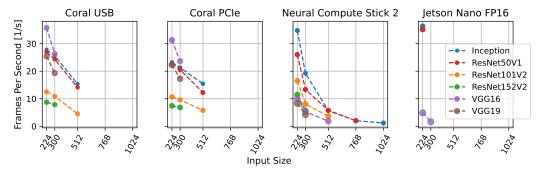


Figure 3: Model input size vs frames vs platform on larger model families. Missing data points for specific networks are mainly the result of exceeding the platform's model size limits.

- the standard deviation of inference time,
- median inference time for Jetson Nano (skewness of inference time distribution).

The default number of proper inferences is 1024. In special cases, it might vary, e.g., for the MobileNetV3 family on the Neural Stick it was 128 caused by the duration of the inference.

### 4 Results

Our experiment covered 3095 unique test cases, focusing predominantly on MobileNet and EfficientNet model families. For visualizations and Table 1 results, we used models which include input data preprocessing, and have no classification heads. Those are the results that we wanted to focus on, since, in this work, we first analyze architectures for being time-efficient feature extractors. Those picked results provide good insight into the rest of the collected data.

All test results are available in a CSV format at https://bulletprove.com/research/edge\_inference\_results.csv.

#### 4.1 Performance

Figure 1 reveals the platform's performance on MobileNetV2. Coral devices achieve the highest frames-per-second performance. NCS2 was the only device to successfully run models with all input sizes, which makes this platform the most versatile.

Figure 2 shows that the Coral is generally faster than NCS2 and Jetson Nano in this model group, especially in the case of smaller input size. Apart from NCS2, platforms had difficulties running models with larger input sizes.

Figure 3 reflects the inference results on larger models. For models with input size 224 and outside the VGG family, the NCS2 is the fastest. Moreover, the device covers the broadest range of model configurations. With increasing model input size Coral slightly outperforms other platforms. Due to memory errors, only four inference test cases were completed successfully for Jetson Nano.

What also can be seen in figures 1, 2, 3 is the difference in the slope of curves, which indicates that inference time scales differently with the number of parameters across compared platforms.

As shown in Table 1, for almost every unique architecture and input size setting Coral (either USB Dongle or PCIe) is the best platform. The only exception occurs for the InceptionV3 and Resnet50 models however, it is crucial to mention that these architectures do not fit entirely on Google's platform. Nevertheless, the performance gap between Google's platforms and the others deepens with increasing the input size, which indicates better computation optimization for the prior platforms.

Results for Jetson Nano have, on average, 8.77 times bigger standard deviation of inference time than for other platforms. This value is based on data from Table 1 and only for MobileNetV2 and EfficientNetV2B0 however, this trend is correct for all measurements. It is a valuable insight for, e.g., designing a system with a strict maximum inference time constraint.

For MobileNetV2, architecture designed especially for mobile devices, we can observe a significant difference in FPS between Corals and other platforms. For an input size of 224 Coral USB was 4.42 times faster than Jetson and 9.08 than Neural Stick. In addition, Corals performed better for an input size of 512 than other devices for the smaller one - 1.16 times faster than Jetson Nano and 2.39 than Neural Stick.

#### 4.2 Limitations

On Google Coral, when the model size exceeds on-chip memory size limits, the model's data has to be fetched from the external memory, which results in additional latency. Exceeding the unspecified model size limit on Google Coral results in a compilation failure.

Coral does not support the hard-swish activation function, which is required to compile MobileNetV3 directly in a way that allows full TPU computation. Therefore, some operations are executed off-chip, increasing model latency.

For large enough models, Neural Stick does not behave like Google's platform (using on-chip and off-chip memory), i.e. throws  $NC\_OUT\_OF\_MEMORY$  error terminating running script. However, it is still able to work with more models than Coral.

Table 1: Inference time comparison for every model family representative per platform and input size. Bolded rows show the best result for a certain model and input size, whereas those without measurements indicate an inability of running that setting. Units: Input size is in pixels, FPS (frames per second) - 1/s, and the rest of the time variables are in ms.

Platform   Model name   Input Size   FPS   Mean time   Std time   Min time   Max time	C III 1165.							
Coral PCle   MobileNetV2   512   96.51   10.36   0.08   10.05   10.57	Platform	Model name	Input Size	FPS	Mean time	Std time	Min time	Max time
Coral PCle   MobileNetV2   512   96.51   10.36   0.08   10.05   10.57	Coral USB	MobileNetV2	224	365.82	2.73	0.11	2.41	3.02
Coral PCIe   MobileNetV2   S12   S5.33   11.72   0.41   10.56   24.30     Jetson FP16   MobileNetV2   224   S2.75   12.08   1.91   5.20   17.58     Coral USB   EfficientNetV2B0   224   75.91   13.17   0.41   11.77   14.09     Coral PCIe   EfficientNetV2B0   224   68.39   14.62   0.15   12.88   15.03     Jetson FP16   EfficientNetV2B0   224   40.31   24.81   0.31   24.06   25.44     Jetson FP16   InceptionV3   224   36.35   27.51   4.49   6.75   47.51     Coral USB   VGG16   224   35.73   27.99   0.13   27.41   28.30     Jetson FP16   Resnet50   224   35.73   28.47   4.98   4.47   46.09     Neural Stick   InceptionV3   224   34.76   28.77   0.28   27.70   29.55     Coral PCIe   VGG16   224   31.31   31.94   0.16   31.04   34.92     Coral USB   InceptionV3   224   26.03   38.42   0.27   37.59   36.45     Coral USB   Resnet50   224   26.03   38.42   0.27   37.59   39.14     Coral PCIe   InceptionV3   224   22.71   44.04   0.14   43.09   44.68     Jetson FP16   MobileNetV2   512   21.91   45.64   8.35   6.82   84.63     Neural Stick   EfficientNetV2B0   512   21.91   45.64   8.35   6.82   84.63     Neural Stick   EfficientNetV2B0   512   16.26   61.50   0.21   58.74   62.63     Coral PCIe   InceptionV3   512   15.44   64.75   0.14   63.61   65.83     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral PCIe   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral PCIe   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral PCIe   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral PCIe   Resnet50   512   14.43   69.30   5.50   71.09     Coral PCIe   Resnet50   512   5.62   78.09   0.46   7	Coral PCIe	MobileNetV2	224	289.01	3.46	0.24	2.55	4.69
Jetson FP16	Coral USB	MobileNetV2	512	96.51	10.36	0.08	10.05	10.57
Coral VSB         EfficientNetV2B0         224         75.91         13.17         0.41         11.77         14.09           Coral PCIe         EfficientNetV2B0         224         68.39         14.62         0.15         12.88         15.03           Jetson FP16         EfficientNetV2B0         224         50.92         19.64         1.11         11.54         24.46           Neural Stick         MobileNetV2         224         40.31         24.81         0.31         24.06         25.44           Jetson FP16         InceptionV3         224         36.35         27.51         4.49         6.75         47.51           Jetson FP16         Resnet50         224         35.73         27.99         0.13         27.41         28.30           Jetson FP16         Resnet50         224         35.73         27.99         0.13         27.41         28.30           Jetson FP16         Resnet50         224         31.31         31.94         0.16         31.04         34.92           Coral USB         InceptionV3         224         27.78         35.99         0.23         35.09         36.45           Coral USB         Resnet50         224         26.96         37.09	Coral PCIe	MobileNetV2	512	85.33	11.72	0.41	10.56	24.30
Coral PCIe   EfficientNetV2B0   224   68.39   14.62   0.15   12.88   15.03     Jetson FP16   EfficientNetV2B0   224   50.92   19.64   1.11   11.54   24.46     Neural Stick   MobileNetV2   224   40.31   24.81   0.31   24.06   25.44     Jetson FP16   InceptionV3   224   36.35   27.51   4.49   6.75   47.51     Coral USB   VGG16   224   35.73   27.99   0.13   27.41   28.30     Jetson FP16   Resnet50   224   35.13   28.47   4.98   4.47   46.09     Neural Stick   InceptionV3   224   34.76   28.77   0.28   27.70   29.55     Coral PCIe   VGG16   224   31.31   31.94   0.16   31.04   34.92     Coral USB   InceptionV3   224   27.78   35.99   0.23   35.09   36.45     Coral USB   Resnet50   224   26.96   37.09   0.18   36.07   37.48     Neural Stick   Resnet50   224   26.96   37.09   0.18   36.07   37.48     Neural Stick   Resnet50   224   26.03   38.42   0.27   37.59   39.14     Coral PCIe   InceptionV3   224   23.23   43.05   0.14   41.92   43.99     Coral PCIe   Resnet50   224   22.71   44.04   0.14   43.09   44.68     Neural Stick   EfficientNetV2B0   224   20.20   49.50   0.38   48.18   50.26     Coral PCIe   InceptionV3   512   16.26   61.50   0.21   58.74   62.63     Coral USB   Resnet50   512   16.26   61.50   0.21   58.74   62.63     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   15.44   64.75   6.25   69.59   71.09     Coral USB   Resnet50   512   15.45   60.04   69.33   77.85   79.05   79.05   79.05   79.05   79.05   79.05   79.05   79.05   79.05   79.05   79.05   79.05   79.05	Jetson FP16	MobileNetV2			12.08			17.58
Jetson FP16	Coral USB	EfficientNetV2B0			13.17			14.09
Neural Stick   MobileNetV2   224   40.31   24.81   0.31   24.06   25.44     Jetson FP16   InceptionV3   224   36.35   27.51   4.49   6.75   47.51     Jetson FP16   Resnet50   224   35.73   27.99   0.13   27.41   28.30     Jetson FP16   Resnet50   224   35.13   28.47   4.98   4.47   46.09     Neural Stick   InceptionV3   224   34.76   28.77   0.28   27.70   29.55     Coral PCIe   VGG16   224   31.31   31.94   0.16   31.04   34.92     Coral USB   InceptionV3   224   27.78   35.99   0.23   35.09   36.45     Coral USB   Resnet50   224   26.96   37.09   0.18   36.07   37.48     Neural Stick   Resnet50   224   26.03   38.42   0.27   37.59   39.14     Coral PCIe   InceptionV3   224   22.71   44.04   0.14   43.09   44.68     Jetson FP16   MobileNetV2   512   21.91   45.64   8.35   6.82   84.63     Neural Stick   EfficientNetV2B0   224   20.20   49.50   0.38   48.18   50.26     Coral PCIe   EfficientNetV2B0   512   16.26   61.50   0.21   58.74   62.63     Coral USB   InceptionV3   512   15.44   64.75   0.14   63.61   65.83     Coral USB   EfficientNetV2B0   512   14.43   69.30   5.50   41.00   95.82     Coral USB   EfficientNetV2B0   512   13.56   73.76   2.53   70.05   83.77     Coral PCIe   Resnet50   512   14.13   70.75   0.22   69.59   71.09     Coral USB   EfficientNetV2B0   512   13.56   73.76   2.53   70.05   83.77     Coral PCIe   Resnet50   512   12.24   81.71   0.19   79.10   83.16     Neural Stick   VGG16   224   9.99   100.10   0.27   99.34   100.98     Neural Stick   MobileNetV2   512   5.68   176.08   0.62   173.58   177.87     Neural Stick   Resnet50   512   5.68   176.08   0.62   173.58   177.87     Neural Stick   VGG16   512   5.62   178.09   0.46   176.53   179.37     Detson FP16   VGG16   512   -	Coral PCIe	EfficientNetV2B0			14.62	0.15	12.88	15.03
Jetson FP16	Jetson FP16	EfficientNetV2B0	224	50.92	19.64	1.11	11.54	24.46
Coral USB         VGG16         224         35.73         27.99         0.13         27.41         28.30           Jetson FP16         Resnet50         224         35.13         28.47         4.98         4.47         46.09           Neural Stick         InceptionV3         224         34.76         28.77         0.28         27.70         29.55           Coral PCIe         VGG16         224         31.31         31.94         0.16         31.04         34.92           Coral USB         InceptionV3         224         27.78         35.99         0.23         35.09         36.45           Coral USB         Resnet50         224         26.96         37.09         0.18         36.07         37.48           Neural Stick         Resnet50         224         26.03         38.42         0.27         37.59         39.14           Coral PCIe         InceptionV3         224         23.23         43.05         0.14         41.92         43.99           Coral PCIe         Resnet50         224         22.71         44.04         0.14         43.09         44.68           Jetson FP16         MobileNetV2B0         251         16.26         61.50         0.21 <t< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></t<>								
Neural Stick   InceptionV3   224   35.13   28.47   4.98   4.47   46.09	Jetson FP16	InceptionV3					6.75	47.51
Neural Stick   InceptionV3   224   34.76   28.77   0.28   27.70   29.55	Coral USB	VGG16	224		27.99			28.30
Coral PCIe         VGG16         224         31.31         31.94         0.16         31.04         34.92           Coral USB         InceptionV3         224         27.78         35.99         0.23         35.09         36.45           Coral USB         Resnet50         224         26.96         37.09         0.18         36.07         37.48           Neural Stick         Resnet50         224         26.03         38.42         0.27         37.59         39.14           Coral PCIe         InceptionV3         224         23.23         43.05         0.14         41.92         43.99           Coral PCIe         Resnet50         224         22.71         44.04         0.14         43.09         44.68           Jetson FP16         MobileNetV2         512         21.91         45.64         8.35         6.82         84.63           Neural Stick         EfficientNetV2B0         2512         16.26         61.50         0.21         58.74         62.63           Coral PCIe         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5	Jetson FP16	Resnet50		35.13				46.09
Coral USB         InceptionV3         224         27.78         35.99         0.23         35.09         36.45           Coral USB         Resnet50         224         26.96         37.09         0.18         36.07         37.48           Neural Stick         Resnet50         224         26.03         38.42         0.27         37.59         39.14           Coral PCIe         InceptionV3         224         23.23         43.05         0.14         41.92         43.99           Coral PCIe         Resnet50         224         22.71         44.04         0.14         43.09         44.68           Jetson FP16         MobileNetV2         512         21.91         45.64         8.35         6.82         84.63           Neural Stick         EfficientNetV2B0         224         20.20         49.50         0.38         48.18         50.26           Coral PCIe         InceptionV3         512         15.44         64.75         0.14         63.61         65.83           Coral USB         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.13         70.75         <	Neural Stick	InceptionV3						29.55
Coral USB         Resnet50         224         26.96         37.09         0.18         36.07         37.48           Neural Stick         Resnet50         224         26.03         38.42         0.27         37.59         39.14           Coral PCIe         InceptionV3         224         23.23         43.05         0.14         41.92         43.99           Coral PCIe         Resnet50         224         22.71         44.04         0.14         43.09         44.68           Jetson FP16         MobileNetV2         512         21.91         45.64         8.35         6.82         84.63           Neural Stick         EfficientNetV2B0         224         20.20         49.50         0.38         48.18         50.26           Coral PCIe         InceptionV3         512         16.26         61.50         0.21         58.74         62.63           Coral USB         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75	Coral PCIe	VGG16						34.92
Neural Stick         Resnet50         224         26.03         38.42         0.27         37.59         39.14           Coral PCIe         InceptionV3         224         23.23         43.05         0.14         41.92         43.99           Coral PCIe         Resnet50         224         22.71         44.04         0.14         43.09         44.68           Jetson FP16         MobileNetV2         512         21.91         45.64         8.35         6.82         84.63           Neural Stick         EfficientNetV2B0         224         20.20         49.50         0.38         48.18         50.26           Coral PCIe         EfficientNetV2B0         512         16.26         61.50         0.21         58.74         62.63           Coral USB         InceptionV3         512         15.44         64.75         0.14         63.61         65.83           Coral USB         InceptionV3         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral USB         EfficientNetV2B0         512         13.56         73.76		InceptionV3						
Coral PCIe         InceptionV3         224         23.23         43.05         0.14         41.92         43.99           Coral PCIe         Resnet50         224         22.71         44.04         0.14         43.09         44.68           Jetson FP16         MobileNetV2         512         21.91         45.64         8.35         6.82         84.63           Neural Stick         EfficientNetV2B0         224         20.20         49.50         0.38         48.18         50.26           Coral PCIe         EfficientNetV2B0         512         16.26         61.50         0.21         58.74         62.63           Coral PCIe         InceptionV3         512         15.44         64.75         0.14         63.61         65.83           Coral USB         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral PCIe         Resnet50         512         12.24         81.71								
Coral PCIe         Resnet50         224         22.71         44.04         0.14         43.09         44.68           Jetson FP16         MobileNetV2         512         21.91         45.64         8.35         6.82         84.63           Neural Stick         EfficientNetV2B0         224         20.20         49.50         0.38         48.18         50.26           Coral PCIe         EfficientNetV2B0         512         16.26         61.50         0.21         58.74         62.63           Coral PCIe         InceptionV3         512         15.44         64.75         0.14         63.61         65.83           Coral USB         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         MobileNetV2         512         9.35         106.94								
Jetson FP16   MobileNetV2   512   21.91   45.64   8.35   6.82   84.63     Neural Stick   EfficientNetV2B0   224   20.20   49.50   0.38   48.18   50.26     Coral PCIe   EfficientNetV2B0   512   16.26   61.50   0.21   58.74   62.63     Coral PCIe   InceptionV3   512   15.44   64.75   0.14   63.61   65.83     Coral USB   InceptionV3   512   15.35   65.15   1.54   62.04   69.33     Jetson FP16   EfficientNetV2B0   512   14.43   69.30   5.50   41.00   95.82     Coral USB   Resnet50   512   14.13   70.75   0.22   69.59   71.09     Coral USB   EfficientNetV2B0   512   13.56   73.76   2.53   70.05   83.77     Coral PCIe   Resnet50   512   12.24   81.71   0.19   79.10   83.16     Neural Stick   VGG16   224   9.99   100.10   0.27   99.34   100.98     Neural Stick   MobileNetV2   512   9.35   106.94   0.48   105.60   108.00     Neural Stick   InceptionV3   512   5.68   176.08   0.62   173.58   177.87     Neural Stick   EfficientNetV2B0   512   5.62   178.09   0.46   176.53   179.37     Jetson FP16   VGG16   512   -		InceptionV3						43.99
Neural Stick         EfficientNetV2B0         224         20.20         49.50         0.38         48.18         50.26           Coral PCIe         EfficientNetV2B0         512         16.26         61.50         0.21         58.74         62.63           Coral PCIe         InceptionV3         512         15.44         64.75         0.14         63.61         65.83           Coral USB         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral PCIe         Resnet50         512         13.56         73.76         2.53         70.05         83.77           Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         Resnet50         512         5.82         171.78         <		Resnet50						44.68
Coral PCIe         EfficientNetV2B0         512         16.26         61.50         0.21         58.74         62.63           Coral PCIe         InceptionV3         512         15.44         64.75         0.14         63.61         65.83           Coral USB         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral USB         EfficientNetV2B0         512         13.56         73.76         2.53         70.05         83.77           Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.62         178.09	Jetson FP16	MobileNetV2						
Coral PCIe         InceptionV3         512         15.44         64.75         0.14         63.61         65.83           Coral USB         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral USB         EfficientNetV2B0         512         13.56         73.76         2.53         70.05         83.77           Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         MobileNetV2         512         9.35         106.94         0.48         105.60         108.00           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.62         178.09								
Coral USB         InceptionV3         512         15.35         65.15         1.54         62.04         69.33           Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral USB         EfficientNetV2B0         512         13.56         73.76         2.53         70.05         83.77           Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         MobileNetV2         512         9.35         106.94         0.48         105.60         108.00           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Jetson FP16         VGG16         512         1.87         533.86         <								
Jetson FP16         EfficientNetV2B0         512         14.43         69.30         5.50         41.00         95.82           Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral USB         EfficientNetV2B0         512         13.56         73.76         2.53         70.05         83.77           Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         MobileNetV2         512         9.35         106.94         0.48         105.60         108.00           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         512         1.87         533.86								
Coral USB         Resnet50         512         14.13         70.75         0.22         69.59         71.09           Coral USB         EfficientNetV2B0         512         13.56         73.76         2.53         70.05         83.77           Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         MobileNetV2         512         9.35         106.94         0.48         105.60         108.00           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral VSB         VGG16         512         -         -         -		InceptionV3						
Coral USB         EfficientNetV2B0         512         13.56         73.76         2.53         70.05         83.77           Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         MobileNetV2         512         9.35         106.94         0.48         105.60         108.00           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -           Jetson FP16         VGG16         512         -         -         -         -								
Coral PCIe         Resnet50         512         12.24         81.71         0.19         79.10         83.16           Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         MobileNetV2         512         9.35         106.94         0.48         105.60         108.00           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         224         4.92         203.45         42.62         5.64         332.42           Neural Stick         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -           Jetson FP16         VGG16         512         -         -         -         -								
Neural Stick         VGG16         224         9.99         100.10         0.27         99.34         100.98           Neural Stick         MobileNetV2         512         9.35         106.94         0.48         105.60         108.00           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         224         4.92         203.45         42.62         5.64         332.42           Neural Stick         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -           Coral USB         VGG16         512         -         -         -         -         -         -           Jetson FP16         InceptionV3         512         -         -         -         -								
Neural Stick         MobileNetV2         512         9.35         106.94         0.48         105.60         108.00           Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         224         4.92         203.45         42.62         5.64         332.42           Neural Stick         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -           Coral USB         VGG16         512         -         -         -         -         -           Jetson FP16         VGG16         512         -         -         -         -         -           Jetson FP16         InceptionV3         512         -         -         -         -         -         - </td <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>								
Neural Stick         Resnet50         512         5.82         171.78         0.66         169.68         175.44           Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         224         4.92         203.45         42.62         5.64         332.42           Neural Stick         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -         -           Coral USB         VGG16         512         -								
Neural Stick         InceptionV3         512         5.68         176.08         0.62         173.58         177.87           Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         224         4.92         203.45         42.62         5.64         332.42           Neural Stick         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -           Coral USB         VGG16         512         -         -         -         -         -           Jetson FP16         VGG16         512         -         -         -         -         -         -           Jetson FP16         InceptionV3         512         -         -         -         -         -         -         -         -								
Neural Stick         EfficientNetV2B0         512         5.62         178.09         0.46         176.53         179.37           Jetson FP16         VGG16         224         4.92         203.45         42.62         5.64         332.42           Neural Stick         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -         -           Coral USB         VGG16         512         -								
Jetson FP16         VGG16         224         4.92         203.45         42.62         5.64         332.42           Neural Stick         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -         -           Coral USB         VGG16         512         -         -         -         -         -         -           Jetson FP16         VGG16         512         -         -         -         -         -         -         -           Jetson FP16         InceptionV3         512         -         -         -         -         -         -         -         -								
Neural Stick         VGG16         512         1.87         533.86         0.71         531.54         536.14           Coral PCIe         VGG16         512         -         -         -         -         -         -           Coral USB         VGG16         512         -         -         -         -         -         -           Jetson FP16         VGG16         512         -         -         -         -         -         -           Jetson FP16         InceptionV3         512         -         -         -         -         -         -								
Coral PCIe         VGG16         512         -								
Coral USB         VGG16         512         -				1.87	533.86	0.71	531.54	536.14
Jetson FP16         VGG16         512         -				-	-	-	-	-
Jetson FP16 InceptionV3 512				-	-	-	-	-
	Jetson FP16			-	-	-	-	-
Jetson FP16 Resnet50 512				-	-	-	-	-
	Jetson FP16	Resnet50	512		<u>-</u>			

In the case of Jetson Nano, it takes significantly longer to prepare an inference model compared to other platforms. It turned out to be a bottleneck of our experiment. Similarly to Neural Stick, exceeding GPU's RAM results in an out-of-memory error.

### 5 Conclusions

This paper presents an extensive inference time performance comparison on Edge AI devices, specifically: NVIDIA Jetson Nano, Google Coral USB, Google Coral PCI, and Intel Neural Stick. For inference, we use variations of model families: MobileNet, EfficientNet, ResNet, VGG, and InceptionV3. Test configurations included mainly changes in the model's input size, classification head, type and scale. The experiments' results indicate that Google Coral is the platform that offers the fastest average inference time. Jetson Nano inference tests suggest that the platform is prone to latency spikes, which is undesirable in time-restricted use cases. The NCS2 platform is the most universal considering the model type and model size choice in the scope of this experiment. We hope this benchmark will help engineers in developing AI at the edge.

# 6 Acknowledgements

N.C. and P.G. conceived the experiment, R.T. and G.W. carried out the measurements and prepared the figures under N.C. supervision. G.W. and R.T. wrote the manuscript. S.L. supervised the project. All authors discussed the results and contributed to the final manuscript.

This research was undertaken as part of the "BulletProve - a system for supporting the training of long-distance shooters using machine learning methods" project and is jointly funded by the National Centre for Research and Development (Szybka Ścieżka 0552/21).

# References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *CoRR*, abs/1907.05047, 2019. URL http://arxiv.org/abs/1907.05047.

Apple ML Team. Apple face detection on mobile. https://machinelearning.apple.com/research/face-detection, 2017.

Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. Hello edge: Keyword spotting on microcontrollers. *CoRR*, abs/1711.07128, 2017. URL http://arxiv.org/abs/1711.07128.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. URL https://arxiv.org/abs/2201.03545.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. URL https://arxiv.org/abs/2103.14030.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. URL http://arxiv.org/abs/1804.02767.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015. URL http://arxiv.org/abs/1512.02325.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. URL http://arxiv.org/abs/1703.06870.

- Wei Yu, Fan Liang, Xiaofei He, William Grant Hatcher, Chao Lu, Jie Lin, and Xinyu Yang. A survey on the edge computing for the internet of things. *IEEE Access*, 6:6900–6919, 2018. doi:10.1109/ACCESS.2017.2778504.
- Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B. Letaief. A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys and Tutorials*, 19(4):2322–2358, 2017. doi:10.1109/COMST.2017.2745201.
- Dario Amodei and Danny Hernandez. Ai and compute. https://openai.com/blog/ai-and-compute/, 2018.
- Nataliia Neshenko, Elias Bou-Harb, Jorge Crichigno, Georges Kaddoum, and Nasir Ghani. Demystifying iot security: An exhaustive survey on iot vulnerabilities and a first empirical look on internet-scale iot exploitations. *IEEE Communications Surveys and Tutorials*, 21(3):2702–2733, 2019. doi:10.1109/COMST.2019.2910750.
- Lorenzo Franceschi-Bicchierai. How this internet of things stuffed animal can be remotely turned into a spy device. https://www.vice.com/en/article/qkm48b/how-this-internet-of-things-teddy-bear-can-be-remotely-turned 2017a.
- Lorenzo Franceschi-Bicchierai. Internet of things teddy bear leaked 2 million parent and kids message recordings. https://www.vice.com/en/article/pgwean/internet-of-things-teddy-bear-leaked-2-million-parent-and-kid 2017b.
- Sheikh Rufsan Reza, Yuzhong Yan, Xishuang Dong, and Lijun Qian. Inference performance comparison of convolutional neural networks on edge devices. In Sara Paiva, Sérgio Ivan Lopes, Rafik Zitouni, Nishu Gupta, Sérgio F. Lopes, and Takuro Yonezawa, editors, *Science and Technologies for Smart Cities*, pages 323–335, Cham, 2021. Springer International Publishing. ISBN 978-3-030-76063-2.
- Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, Carole-Jean Wu, Brian Anderson, Maximilien Breughe, Mark Charlebois, William Chou, Ramesh Chukka, Cody Coleman, Sam Davis, Pan Deng, Greg Diamos, Jared Duke, Dave Fick, J. Scott Gardner, Itay Hubara, Sachin Idgunji, Thomas B. Jablin, Jeff Jiao, Tom St. John, Pankaj Kanwar, David Lee, Jeffery Liao, Anton Lokhmotov, Francisco Massa, Peng Meng, Paulius Micikevicius, Colin Osborne, Gennady Pekhimenko, Arun Tejusve Raghunath Rajan, Dilip Sequeira, Ashish Sirasao, Fei Sun, Hanlin Tang, Michael Thomson, Frank Wei, Ephrem Wu, Lingjie Xu, Koichi Yamada, Bing Yu, George Yuan, Aaron Zhong, Peizhao Zhang, and Yuchen Zhou. Mlperf inference benchmark, 2019.
- Google. Tensorflow. https://www.tensorflow.org/, 2023a.
- Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. URL http://arxiv.org/abs/1704.04861.
- Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. URL http://arxiv.org/abs/1801.04381.
- Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *CoRR*, abs/1905.02244, 2019. URL http://arxiv.org/abs/1905.02244.
- Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. URL http://arxiv.org/abs/1905.11946.
- Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. *CoRR*, abs/2104.00298, 2021. URL https://arxiv.org/abs/2104.00298.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016. URL http://arxiv.org/abs/1603.05027.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv* preprint arXiv:1409.1556, 2014.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL http://arxiv.org/abs/1512.00567.
- Shuiguang Deng, Hailiang Zhao, Jianwei Yin, Schahram Dustdar, and Albert Y. Zomaya. Edge intelligence: the confluence of edge computing and artificial intelligence. *CoRR*, abs/1909.00560, 2019. URL http://arxiv.org/abs/1909.00560.
- Google. Google coral ai accelerator datasheet. https://coral.ai/docs/, 2020a.

Google. Google edge tpu. https://cloud.google.com/edge-tpu, 2023b.

NVIDIA. Jetson modules. https://developer.nvidia.com/embedded/jetson-modules, 2023a.

NVIDIA. Jetpack sdk. https://developer.nvidia.com/embedded/jetpack, 2023b.

NVIDIA. Embedded systems for product development. https://www.nvidia.com/en-us/autonomous-machines/embedded-sys 2023c.

Intel. Neural compute stick 2 product brief. https://www.intel.com/content/dam/support/us/en/documents/boardsandkit 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Google. Pycoral api repository. https://github.com/google-coral/pycoral, 2020b.

Python Soft. Foundation. Timeit package api. https://docs.python.org/3/library/timeit.html, 2023.

ONNX. tf2onnx repository. https://github.com/onnx/tensorflow-onnx, 2021.

Intel. Openvino-dev download page. https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/do 2022a.

Intel. Openvino toolkit packages. https://storage.openvinotoolkit.org/repositories/openvino/packages/2022.2/linu 2022b.

NVIDIA. Tensorrt. https://developer.nvidia.com/tensorrt, 2023d.