

Machine Learning at the Network Edge: A Survey

M.G. Sarwar Murshed[§] Christopher Murphy^{‡‡} Daqing Hou[§]
 Nazar Khan[¶] Ganesh Ananthanarayanan[†] Faraz Hussain[§]

[§]Clarkson University, Potsdam, NY, USA
 {murshem,dhou,fhussain}@clarkson.edu

^{‡‡}SRC, Inc., North Syracuse, NY, USA
 cmurphy@srcinc.com

[†]Microsoft Research, Redmond, WA, USA
 GA@microsoft.com

[¶]Punjab University College of Information Technology, Lahore, Pakistan
 nazarkhan@pucit.edu.pk

Abstract

Devices comprising the Internet of Things, such as sensors and small cameras, usually have small memories and limited computational power. The proliferation of such resource-constrained devices in recent years has led to the generation of large quantities of data. These data-producing devices are appealing targets for machine learning applications but struggle to run machine learning algorithms due to their limited computing capability. They typically offload input data to external computing systems (such as cloud servers) for further processing. The results of the machine learning computations are communicated back to the resource-scarce devices, but this worsens latency, leads to increased communication costs, and adds to privacy concerns.

Therefore, efforts have been made to place additional computing devices at the edge of the network, i.e. close to the IoT devices where the data is generated. Deploying machine learning systems on such edge devices alleviates the above issues by allowing computations to be performed close to the data sources. This survey describes major research efforts where machine learning has been deployed at the edge of computer networks.

Keywords— edge device, low-power, machine learning, single board computer, edge computing, cloud, fog, resource-constrained, IoT, deep learning, resource-scarce, embedded systems.

1 Introduction

Due to the explosive growth of wireless communication technology, the number of Internet of Things (IoT) de-

vices has increased dramatically in recent years. It has been estimated that by 2020, more than 25 billion devices will have been connected to the Internet [69] and the potential economic impact of the IoT will be \$3.9 trillion to \$11.1 trillion annually by 2025 [70]. IoT devices typically have limited computing power and small memories. Examples of such resource-constrained IoT devices include sensors, microphones, smart fridges, and smart lights. IoT devices and sensors continuously generate large amounts of data, which is of critical importance to many modern technological applications such as autonomous vehicles. This data must be fed to a machine learning system to extract information and make decisions.

Unfortunately, limitations in the computational capabilities of resource-scarce devices inhibit the implementation of ML algorithms on them. So, the data is offloaded to remote computational infrastructure, most commonly cloud servers, where computations are performed. Transferring raw data to cloud servers increases communication costs, causes delayed system response, and makes any private data vulnerable to compromise. To address these issues, it is natural to consider processing data closer to its sources and transmitting only the necessary data to remote servers for further processing [23].

Edge computing refers to computations being performed as close to data sources as possible, instead of on far-off, remote locations [93, 17]. This is achieved by adding an *edge computing device* (or simply, *edge device*) close to the resource-constrained devices where data is generated. Edge devices possess both computational and communication capabilities. For example, an embedded device such as an Nvidia Jetson TX2 could serve as an edge device if it takes data from a camera on a robotic

arm and performs data processing tasks that are used to determine the next movement of the arm. Computations too intense for edge devices are sent over to more powerful remote servers. Performing computations at the network edge (see [93, 95]) has several advantages:

- The volume of data needed to be transferred to a central computing location is reduced because some of it is processed by edge devices.
- The physical proximity of edge devices to the data sources makes it possible to achieve lower latency which improves real-time data processing performance.
- For the case of data that still must be processed remotely, edge devices can be used to discard personally identifiable information (PII) prior to data transfer, thus enhancing user privacy and security.
- Decentralization can make systems more robust by providing transient services during a network failure or cyber attack.
- Edge computing increases scalability by expanding compute capacity through a combination of edge and IoT devices.

Edge computing is quickly emerging as an important paradigm for IoT based systems; attempts are being to use devices at the network edge to do as many computations as possible, instead of only using the cloud for processing data. For example, Floyer studied data management and processing costs (see Figure 1) of a remote wind-farm using a cloud-only system versus a combined edge-cloud system [36]. The wind-farm consisted of several data producing sensors and devices such as video surveillance cameras, security sensors, access sensors for all employees, and sensors on wind-turbines. The edge-cloud system turned out to be 36% less expensive and the volume of data required to be transferred was observed to be 96% less, compared to the cloud-only system.

Figure 2 shows the growth of edge computing search over the years. Major technology firms, the defense industry, and the open source community have all been at the forefront in investments in edge technology. For example, researchers have developed a new architecture called *Agile Condor* which uses machine learning algorithms to perform real-time computer vision tasks (e.g. video, image processing, and pattern recognition) [50]. This architecture can be used for automatic target recognition (ATR) at the network edge, near the data sources. As an example from the technology world, Microsoft introduced HoloLens 2¹, a holographic computer, in early 2019. The HoloLens is built onto a headset for an augmented reality experience. It is a versatile and powerful edge device that can work offline and also connect to the cloud. Microsoft aims to design standard computing, data analysis, medical imaging, and gaming-at-the-edge tools using the HoloLens. The Linux Foundation recently launched the

¹<https://www.microsoft.com/en-us/hololens/>

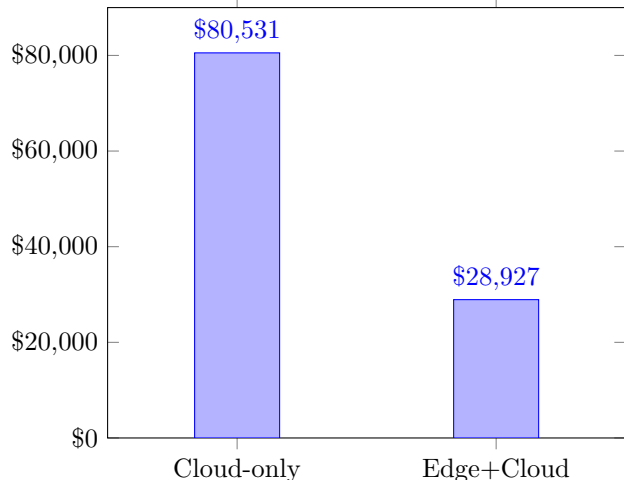


Figure 1: Cost comparison of cloud-only management and processing system of a remote wind-farm with an edge-cloud combined system.

*LF Edge*² project to facilitate applications at the edge and establish a common open source framework that is independent of OS and hardware. EdgeX Foundry³ is another Linux Foundation project that involves developing a framework for industrial IoT edge applications [37]. GE⁴, IBM [2], Cisco [82] and Dell⁵ have all committed to investing in edge computing and VMware is also developing a framework for boosting enterprise IoT efforts at the edge [13]. In this last few years, edge computing is being increasingly used for the *deployment of machine learning based intelligent systems in resource-constrained environments* [111], which is the motivation for this survey.

A note on terminology. A related term, *fog computing*, describes an architecture where the ‘cloud is extended’ to be closer to the IoT end-devices, thereby improving latency and security by performing computations near the network edge [49]. So fog and edge computing are related, the main difference being about where the data is processed: in edge computing, data is processed directly on the devices to which the sensors are attached (or on gateway devices physically very close to the sensors); in fog computing, data is processed further away from the edge, on devices connected using a LAN [51].

This survey focuses on machine learning systems deployed on edge devices, and also covers efforts made to train ML models on edge devices. We begin by summarizing the machine learning algorithms that have been used in edge computing (§2, §3), discuss ML applications that use edge devices (§4), and then list the ML frameworks frequently used to build such systems (§5). We then de-

²<https://www.lfedge.org/>

³<https://www.edgexfoundry.org/>

⁴<https://www.ge.com/digital/iiot-platform/predix-edge>

⁵<https://www.dell EMC.com/en-us/service-providers/edge-computing.htm>

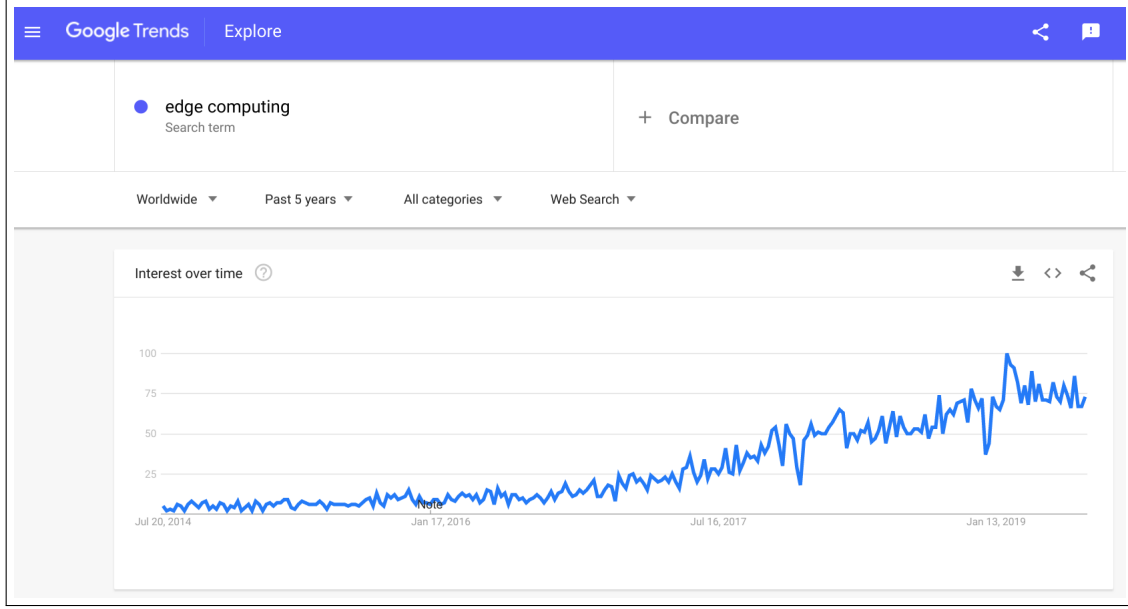


Figure 2: The growth of ‘edge computing’ as a search term 2014-19 (Google Trends).

scribe the hardware and software required for fast deployment of ML algorithms on edge devices (§6.1, §6.2). Before concluding, we discuss the challenges hindering more widespread adoption of this technology and present our thoughts about possible future research directions in the area (§7).

Paper Collection Methodology

We conducted exact keywords searches on Google Scholar, Microsoft Academic, DBLP, IEEE, ACM digital library and arXiv to collect papers related to machine learning and edge computing, resulting in 124 papers. The following search terms were used:

- (machine | deep) learning + edge computing
- (machine | deep) learning + (resource-scarce devices | IoT)
- modified learning algorithms + edge computing
- (SVM | k-means | decision trees | convolutional neural networks | recurrent neural networks) + edge computing
- resource-efficient deep neural networks + edge computing

The following questions were used to determine whether to include a paper in the survey:

1. Was edge computing used to improve ML efficiency in a networked ML system?
2. Were ML algorithms specialized for resource-scarce devices designed or applied in a new setting?

3. Were new results describing the deployment of machine learning systems on the edge reported?

If any of the questions above could be answered in the affirmative, the paper was included. After careful analysis regarding the three questions mentioned above, we considered 88 out of the 124 papers collected for this survey.

2 Machine learning at the edge

We now discuss machine learning algorithms that have been used in resource-constrained settings, where some computations were done at the edge of the network. Note that deep learning systems deployed on edge devices are discussed separately (§3).

2.1 SVM, K-means, linear regression

With the increasing amount of information being generated at the network edge, the demand for machine learning models that can be deployed at the edge has also increased.

Wang et al. introduced a technique that helps *train machine learning models at the edge* of the network without the help of external computation systems such as cloud servers [110]. They focus on algorithms that use gradient-based approaches for training, including SVMs, K-means, linear regression and convolutional neural networks (CNN). Their technique minimizes the loss function of a learning model by using the computational power of edge devices. Multiple edge devices are used to perform ML tasks and one edge device is used to aggregate the results generated by the other (edge) devices.

The aggregator has three main functions: collect processed data from edge devices, make necessary changes to that data, and send processed data back to the edge devices for further processing. The algorithm repeatedly performs two steps:

- Each edge device processes input data and generates a (local) model parameter. This is called *local update*. To minimize the loss function, each edge device uses gradient-descent to adjust the model parameter. Finally, it sends this model parameter to the aggregator.
- The aggregator collects local model parameters from edge devices, aggregates these parameters by taking a weighted average, and sends back a single updated parameter to the edge devices for running another iteration to generate the next local update.

This continues until the ML model reaches an acceptable accuracy. The frequent aggregation process quickly provides an accurate result but increases resource consumption and communication cost. The algorithm determines the optimum frequency of the aggregation phase to ensure an accurate result using the available resources.

To demonstrate the effectiveness of their technique, Wang et al. used three Raspberry Pi 3 devices and a laptop computer as their edge devices. All these devices were connected via Wi-Fi and an additional laptop computer was used as the aggregator. They used three datasets (MNIST, Facebook metrics, and User Knowledge Modeling) to evaluate several models (including smooth SVM, K-means, linear regression, and deep convolutional neural networks) and reported that their algorithm performs close to the optimum with various ML models, making it very effective for resource-constrained settings.

2.2 k-NN

Real-world applications increasingly demand accurate and real-time prediction results for a variety of areas such as autonomous cars, factories, and robots. Powerful computing devices can provide real-time prediction results, but in many situations, it is nearly impossible to deploy them without significantly affecting performance (e.g. on a robot). Instead, ML models must be adapted in a way that makes them suitable for deployment on (small) edge devices having limited computational power and storage capacity. In this section, we discuss a modified version of k-NN that helps to implement the k-NN algorithm in resource-constrained settings.

Gupta et al. introduced a new technique, **ProtoNN**, to train an ML model on a small edge device and perform real-time prediction tasks accurately [40]. ProtoNN is a k-NN based algorithm but has lower storage requirements. The following issues arise when trying to implement a traditional k-NN on a resource-scarce edge device:

1. Selecting an appropriate training data size: k-NN generates prediction results using the entire training

dataset. Resource-scarce devices are unable to store the entire dataset when the training data is large.

2. Prediction time: k-NN calculates the distance of a given sample (which is to be classified) with each training example of the training data. Due to small computational power, resource-constrained devices are unable to calculate all the distances required to predict a sample in real-time.
3. Selecting an appropriate distance metric for better accuracy: k-NN does not explicitly suggest which distance metric a developer should use to get a better prediction result. Standard metrics like Euclidean distance and Hamming distance are not task-specific and sometimes generate poor results [40].

To address these issues, ProtoNN excludes unnecessary training data, resulting in a smaller training dataset, which is projected to a low dimension matrix and jointly learned across all data points to get an acceptable accuracy. The main ideas behind ProtoNN are:

- To reduce the size of the input data, this algorithm converts high dimensional data into fewer dimensions. This conversion operation reduces overall model accuracy. To compensate for this, the algorithm uses a sparse projection matrix that is jointly learned to convert the entire dataset into fewer dimensions. This joint learning technique helps to provide good accuracy in the projected space.
- To represent the entire dataset, this algorithm learns: a) a small number of prototypes from the training data, and b) the label of each prototype. Two techniques are used to determine prototypes: random sampling and k-means clustering. In the former, training data points are randomly selected in the transformed space and these points are assigned as the prototypes. This approach is used for multi-label problems. For binary and multi-class problems, a k-means clustering algorithm is run on all data points in the transformed space to determine the cluster centers of data points, which are chosen as prototypes. For example, a dataset with five kinds of classes can be represented using just five data points (the five cluster centers). This reduction decreases the prediction time because fewer distances need to be computed to classify a new sample.
- The algorithm learns prototypes & their labels jointly with the sparse projection matrix.

On multi-label datasets, ProtoNN's compression process reduces the accuracy by 1% compared to popular methods such as RBF-SVM and 1-hidden layer NN, while providing 100 times the compression. ProtoNN can handle general supervised learning in a dataset with millions of examples and can run with just 16 kB of memory.

Gupta et al. used 14 datasets⁶ in their experiments. They deployed ProtoNN on an Arduino Uno (8 bit, 16 MHz Atmega328P microcontroller, 2kB of SRAM and 32kB of flash memory) to evaluate its performance and reported almost the same classification accuracy as the state-of-the-art.

They compared their result with k-NN, Stochastic Neighborhood Compression (SNC), Binary Neighbor Compression (BNC), Gradient Boosting Decision Tree (GBDT), 1-hidden layer NN and RBF-SVM for this test. They reported that ProtoNN attained an accuracy that was within 1–2% of the best uncompressed baseline methods with 1–2 orders of magnitude reduction in model size. In another experiment, they compared the performance of ProtoNN with other methods: BudgetRF, Decision Jungle, LDKL, Tree Pruning, GBDT, Budget Prune, SNC and NeuralNet Pruning. ProtoNN was 5% more accurate on most datasets in severely resource-constrained settings, where model sizes are required to be less than 2 kB. For character recognition, ProtoNN was 0.5% more accurate than RBF-SVM, one of the best character recognition method, while using an approximately 400 times smaller model size. For large multi-label and multi-class datasets like ALOI, MediaMill, and Eurlex, ProtoNN was within 1% the accuracy of RBF-SVM, but used a model size 50 times smaller, with approximately 50 times fewer floating-point computations per prediction than RBF-SVM.

2.3 Tree-based ML algorithms

Tree-based machine learning algorithms are used for classification, regression and ranking problems. Classification and regression are very common in resource-constrained IoT settings. For example, they are used with IoT sensors to classify defective products being produced on a production line. In this section, we discuss a modified version of a tree-based ML algorithm designed to run on a resource-scarce device.

Even though the time-complexity of tree-based algorithms is logarithmic with respect to the size of the training data, their space complexity is linear, so they are not easily amenable to implementation on resource-scarce devices. Aggressively pruning or learning shallow trees are ways to shrink the model but lead to poor prediction results.

Kumar et al. introduced a novel tree-based algorithm, **Bonsai**, with improved space complexity [56]. They achieved acceptable prediction accuracy by testing this algorithm in a resource-constrained environment. The following techniques make Bonsai suitable for resource-scarce devices:

- To reduce the model size, **Bonsai learns a single, shallow, sparse tree**. Unfortunately, this learning process decreases the overall accuracy. To retain accuracy Bonsai makes each node (all leaf and inter-

nal nodes) more powerful by allowing it to predict a non-linear score. Each node learns matrices \mathbf{W} and \mathbf{V} so that each node predicts the result vector $\tanh(\mathbf{V}_k^T \cdot \mathbf{x}) \circ \mathbf{W}_k^T \cdot \mathbf{x}$ where \circ represents the element wise Hadamard product and \mathbf{x} represents the prediction point. The overall predicted vector for a point \mathbf{x} of the Bonsai model is generated by adding all individual predicted vectors with the nodes lying along the path traversed by point \mathbf{x} . The difference between Bonsai and the other tree-based algorithms is that Bonsai uses both leaf and internal nodes to generate a prediction result, while other algorithms use only leaf nodes. This property allows Bonsai to accurately learn non-linear decision boundaries using just a few nodes of a shallow tree.

- The parameters \mathbf{W} and \mathbf{V} are dependent on the dimension of the dataset. High dimensional input data increases the overall model size. To address this issue, a learned sparse projection matrix is used to input feature vector \mathbf{x} reducing the vector into a low dimensional space. This algorithm uses fixed point arithmetic, which reduces computational overheads, when deployed on resource scarce devices.
- All node parameters are learned jointly with optimally allocated memory budget for each node. Bonsai allows parameter sharing during the training phase which helps to reduce the model size and maximize prediction accuracy.

Bonsai was implemented⁷ and tested with a number of binary and multi-class datasets. When deployed on an Arduino Uno board, Bonsai required only 70 bytes and 500 bytes of writable memory for binary and multi-class classification respectively. Table 1 shows the overall accuracy of the Bonsai model on different datasets.

3 Deep learning at the edge

Most deep learning (DL) algorithms need large amounts of input data and huge computational power to generate results. IoT devices are a good source of large data but their limited computational power makes them unsuitable for training and inference of DL models. So, edge devices are placed near the IoT (end) devices and used for deploying DL models that operate on IoT-generated data. DL models have been adapted and deployed on edge devices in the following ways:

- Divide a DL network into two parts and deploy one part on edge devices and the other part on cloud servers (§3.1, §3.2). Such techniques have been shown to provide the same accuracy as cloud-based DL models while reducing the overall involvement of cloud servers [81].
- Modify the architecture of deep learning models so

⁶CIFAR, Character Recognition, Eye, MNIST, USPS, Ward, Letter-26, MNIST-10, USPS-10, CURET-61, ALOI, MediaMill, Delicious, and EURLEX.

⁷<https://github.com/Microsoft/EdgeML>

Table 1: Prediction accuracy attained by deploying Bonsai on an Arduino [56].

Dataset	Model size	Accuracy(%)
RTWhale-2	2KB	61.74
	16KB	61.74
Chars4K-2	2KB	74.28
	16KB	76.67
Eye-2	2KB	88.26
	16KB	90.31
WARD-2	2KB	95.85
	16KB	96.89
CIFAR10-2	2KB	73.02
	16KB	76.64
USPS-2	2KB	94.42
	16KB	95.72
MNIST-2	2KB	94.38
	16KB	96.47
Chars4K-62	101 KB	58.59
CURet-61	115 KB	95.23
MNIST-10	84 KB	97.01

that they can run on an edge device exclusively [57]. Unfortunately, this decreases the accuracy of DL models, but special techniques to improve accuracy have recently been designed (§3.3, §3.4).

3.1 Mapping deep learning models on the cloud and edge

Li et al. developed a technique to train a DL model using cloud servers and then divide the trained model into two parts [64]. They refer to the layers near the output of a DL network as higher layers and those near the input as lower layers. The lower layers are deployed on edge devices and the higher layers on the cloud. **The main challenge is to divide a deep learning network into two parts because different networks have different computational overhead. Also, the size and type of intermediate data in different DL networks are different.**

To deal with these challenges they designed a **scheduling algorithm which calculates the input data size and computational overhead of all the tasks of a DL layer. Then, the algorithm checks the service capability and network bandwidth of all edge devices.** If all edge devices are capable of handling this input data then tasks are deployed to all edge devices. If an edge device is unable to accommodate the input data then **the algorithm reduces the task number and finds an appropriate input size and computational overhead for deploying tasks in all edge devices.** If the scheduling algorithm fails to find a task with an appropriate input data size, then nothing gets deployed on the edge devices and all the data is transferred to the cloud server for computation. Li et al. used the Caffe framework and tested ten CNN tasks with different networks, claiming their online algorithm

can deploy more tasks to edge devices than two popular online scheduling algorithms, viz. FIFO and LBF.

Convolutional neural networks are one of the most popular machine learning models which have been used to process the data coming from multimedia IoT applications such as scene tagging, image classification, face recognition, and gender classification. Osia et al. introduced a new CNN architecture for analyzing multimedia data from different IoT sensors [81]. Currently, most multimedia IoT applications use cloud-based CNN models to analyze the data. However, cloud-based processing methods may compromise user privacy and consume a lot of Internet bandwidth. Osia et al. designed an edge-to-cloud CNN model, which enhances user privacy by running shallow portions of CNN processing tasks on edge devices. The key is that the feature extraction module, which runs on the edge. The rest of the CNN tasks are performed on the cloud using modules which they call analyzers. The general structure of their edge-to-cloud CNN model is shown in Figure 3 (page 7).

The main challenge of such a cloud-edge CNN architecture is to identify an intermediate layer to design a good feature-extractor module that only **removes sensitive information from user data without compromising the accuracy of the CNN model.** Osia et al. used the gender classification model developed by Rothe et al. [91] for identifying the best intermediate layer to discard personal information from input data. This classification model is a pre-trained model with 16 layers. They chose the 5th layer as the intermediate layer and then used a Siamese network architecture for fine-tuning the CNN model. The fine-tuning by the Siamese architecture helps the intermediate layer identify sensitive user information from the input. After discarding sensitive data, principal component analysis was used to reduce the dimensionality of the intermediate features for compressing the data. The data is then transferred to the cloud for further processing. They reported the accuracy of their hybrid edge-to-cloud model for gender classification to be 93%, which is similar to the gender classification model proposed by Rothe et al., but provides more privacy to the user's data.

Most DL models discussed above use static configurations to execute inference, such as dividing a model using a particular layer and running one part on an edge device, another part on cloud servers. However, edge devices may have different computational capabilities and may therefore need dynamic configurations to deploy a DL model on a device to achieve better accuracy. Ogden and Guo designed a novel mobile deep inference platform called **MODI**, which provides multiple deep learning models and *dynamically selects the best model at run-time* [80]. Their platform has the following major components for **dynamically configuring a deep learning model:**

- *inference profilers*: calculate resource requirements for each inference task
- *decision stubs*: work with decision engines to collect information from inference profilers and determine

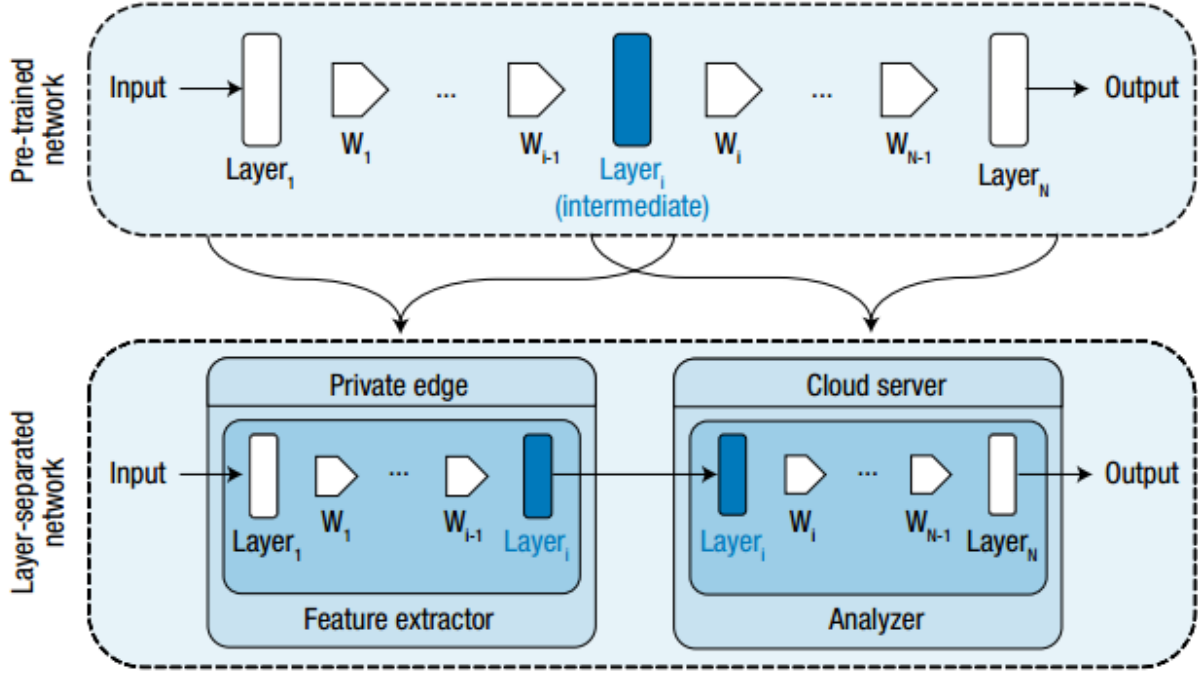


Figure 3: The CNN layer-separation mechanism of Osia et al. [81] (reproduced with permission from the authors). The top part of this figure shows a pre-trained network. In order to divide this network into two parts, an intermediate layer is chosen from this pre-trained network. The layers before the intermediate layer are deployed on the edge device and the rest on the cloud server. The edge discards private information from the input data and uploads intermediate data to the cloud server for further processing.

where to perform inference tasks

- *decision engines*: generate model distribution plans
- *inference engines*: execute a diverse set of deep learning models across servers
- *centralized managers*: collect model usage and inference statistics across the system and select which models are to be deployed on mobile devices and servers.

These components run jointly on end devices and servers to provide the most suitable model for mobile devices based on device resources and installed applications.

3.2 Distributed DNN architectures

Distributed deep neural network architectures map DNN sections across the computing hierarchy (i.e. on the edge or cloud) to facilitate local and fast inference on edge devices wherever possible. The inference is done as early as possible in the network.

3.2.1 DDNNs

Teerapittayanon et al. introduced distributed deep neural networks (DDNN⁸), where sections of a deep neural network are mapped across distributed computing hierarchies [103]. Their algorithm runs simultaneously on cloud servers, edge devices, and resource-constrained end-devices (such as surveillance cameras).

The general structure of a distributed deep neural network is shown in Figure 4 (page 8). In such a model, a big portion of the raw data generated by sensors is processed on edge devices and then sent to the cloud for further processing. Teerapittayanon reported that using this DDNN reduced their data communication cost by a factor of more than 20 for classification on a multi-view multi-camera dataset [103]. They faced the following challenges when designing the DDNN model:

- mapping the DNN on to a small device: designing a suitable DNN architecture to fit in the small memory of the end devices is the primary task for reducing communication cost between computational devices and keeping the same accuracy as a cloud-

⁸<https://github.com/kunglab/ddnn>

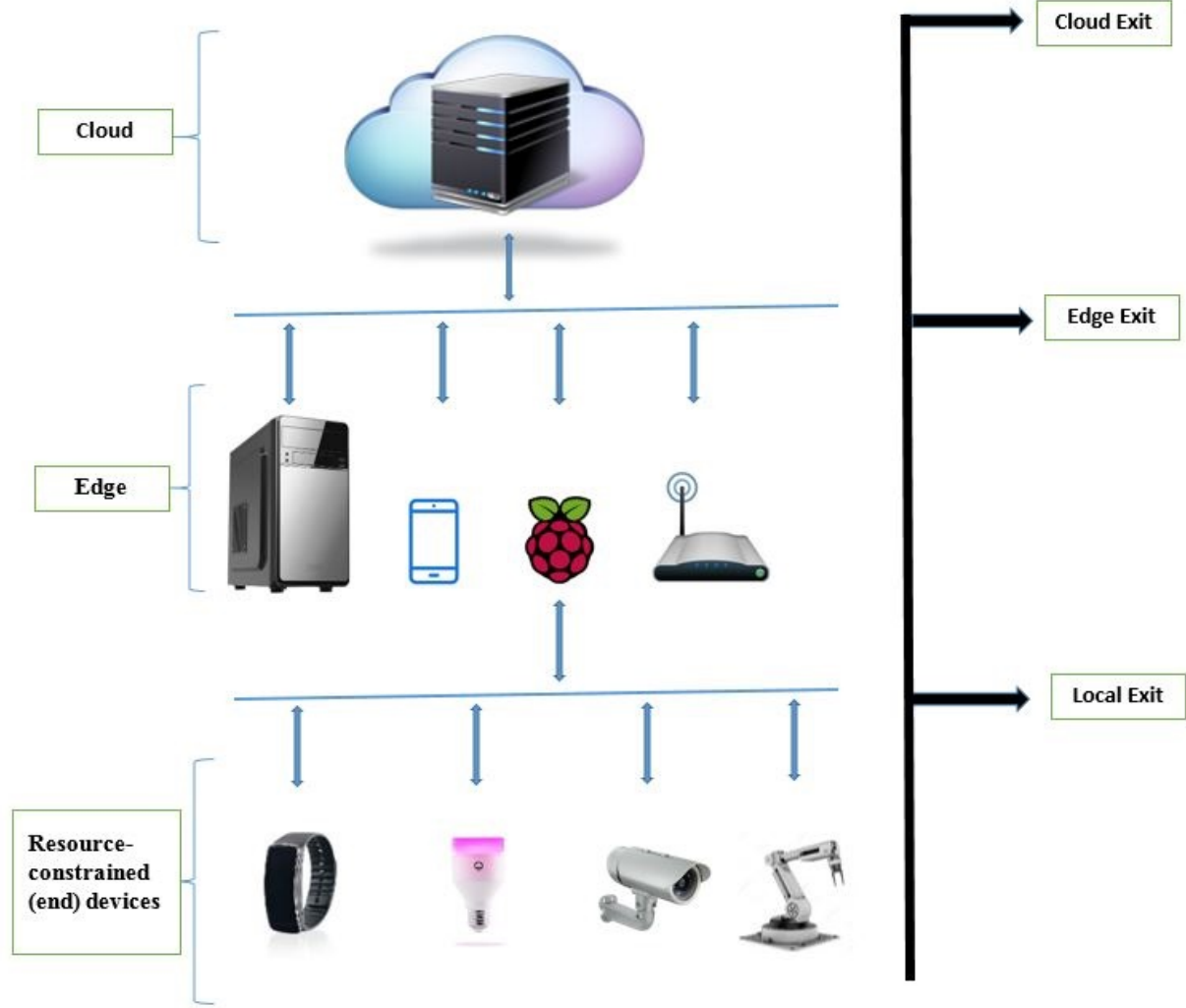


Figure 4: Structure of a distributed deep neural network (DDNN). Resource-constrained devices send summary information to a local aggregator which serves as a layer of the DDNN. **The DDNN is jointly trained with all resource constrained end-devices and exit points**, so the network is capable of automatically collecting and combining input data from different end-devices. If the information collected from a particular end-devices is sufficient to classify a sample then classification is done locally (i.e. on the end-device itself). Otherwise, the information is sent to the edge devices for further processing. If edge devices can complete the classification, they send the result back to the end-devices. Otherwise, edge devices send the information to the cloud, where the classification process is completed, and the results returned to the end-devices.

based model

- **data aggregation:** aggregating the data produced by sensors attached to end-devices
- **learning jointly:** multiple models may that may be running on the cloud-edge network must be learned jointly in order to make coordinated decisions
- **creating an early exit point:** typically a neural network model has one input and one output layer, but DDNNs need multiple output layers to support fast inference at local devices.

The techniques they used to address these issues are summarized below:

BNNs DDNNs use binary neural networks [47, 74] in order to deploy trained models on resource-scarce devices.

DDNN aggregation The data from end devices must be aggregated to perform classification or regression tasks. This process is called DDNN aggregation. The DDNN is jointly trained with all end devices and exit points, so the network is capable of automatically collecting and combining inputs from different end-devices before each exit point. This automatic aggregation process helps avoid the need for additional aggregator devices or manually combining output from end devices. There are different types of aggregation methods, such as max pooling, average pooling⁹, and concatenation¹⁰.

Max pooling aggregates input vectors by taking the maximum value of each component. The expression $v_j = \max_{1 \leq i \leq n} v_{ij}$ represents how aggregation is done using max pooling; v_j represents the value of the j -th component of the output vector, n represents the number of inputs and v_{ij} represents the j -th component of the input vector.

Training the network Since a DDNN is distributed on different devices (including edge devices and the cloud), the most powerful device can be used for training the network. Teerapittayanon et al. used a cloud server for training the distributed deep neural networks. The training of DDNNs is difficult because of multiple exit points. To address this issue, the network was trained jointly by combining losses from each exit point during back-propagation. The joint training technique is described in their work on deep learning inference using early exiting [104].

Inference Distributed neural networks use local aggregators to perform inference on new data. A local aggregator collects data from end-devices and summarizes the information. If the summary informa-

tion is sufficient to classify a new sample accurately then classification happens in this stage. If summary information is insufficient, then end devices send all data to the next stage where the edge devices are located. After processing the data, if edge devices can classify the sample correctly then it classifies that sample at the edge, otherwise edge devices forward summary information to the cloud and then the cloud performs the final classification.

The success of a distributed neural network model depends on keeping inter-device communication costs as low as possible. As mentioned earlier, they achieved a 20 times reduction in communication costs on a multi-camera multi-view dataset.

3.2.2 MoDNN

Mao et al. proposed a local distributed mobile computing system (MoDNN) to deploy DNNs on resource-constrained devices [71]. MoDNN uses a pre-trained DNN model and scans each layer of a DNN model to identify layer types. If a convolutional layer is detected, the layer input is partitioned by a method called Biased One-Dimensional Partition (BODP). BODP helps reduce computing cost by reducing the input size of convolutional layers. If a fully connected layer is detected, the layer input is assigned to different work nodes (mobile devices) to achieve the minimum total execution time. They used the ImageNet VGG-16 dataset and a pre-trained DNN model for their experiment. MoDNN was implemented on the LG Nexus 5 (with 2 GB memory, a 2.28 GHz processor, running Android 4.4.2). They reported having successfully accelerated DNN computations by 2.17-4.28 times with 2 to 4 mobile devices.

3.3 Depthwise separable convolutions

It is challenging to deploy CNNs on edge devices, owing to their limited memory and compute power. *Depthwise Separable Convolutions* provide a lightweight CNN architecture which reduces the computational cost of standard CNNs by a $(\frac{1}{N} + \frac{1}{D_k^2})$ factor, where N is the number of output channels and D_k represents both the height and width of the square kernel used in a convolutional neural network. Depthwise separable convolutions were introduced by Sifre and Mallat [96]. For the rest of this section, we closely follow the terminology used by Howard et al. [44, §3.1].

A standard CNN model uses each convolutional layer to generate a new set of outputs by filtering and combining the input. In contrast, depthwise separable convolutions divide each convolutional layer into two separate layers which serve the same purpose as a single convolutional layer. This separation greatly reduces the computational complexity and model size of this algorithm. The two separate convolutional layers are:

- **depthwise convolution:** this layer applies a single-channel filter to each multi-channel input.

⁹In average pooling, the input vector is aggregated by taking the average of each component.

¹⁰In the concatenation method, all information of each component is stored in the input vector. This aggregation method is used in the cloud layer where full information is needed to extract higher-level features.

- pointwise convolution: this is a 1×1 convolutional which combines the output of the depthwise convolution.

Typically, the computational cost of a standard convolutional layer of a CNN is given by:

$$CC_{Standard} = D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F \quad (1)$$

where $D_K \times D_K$ is the dimension of the kernel, M , N are the number of the input and output channels respectively, and $D_F \times D_F \times M$ is the dimensionality of the feature map F . Depthwise separable convolutions split this cost into two parts by using depthwise and pointwise convolutional techniques. The total convolutional cost of depthwise separable convolutions is equal to the combined cost of depthwise and pointwise convolutions:

$$CC_{DSC} = D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F$$

Depthwise and pointwise convolutions help to reduce computational complexity as represented by the following equation:

$$\frac{CC_{DSC}}{CC_{Standard}} = \frac{1}{N} + \frac{1}{D_K^2} \quad (2)$$

Equation 2 shows that using depthwise separable convolutions reduces the computational cost by factor of $(\frac{1}{N} + \frac{1}{D_K^2})$ over standard CNNs. This cost reduction technique makes them faster and also more efficient in terms of power consumption, which is ideal for edge devices.

3.3.1 MobileNets

Based on depthwise separable convolutions, Howard et al. presented a new neural network architecture called MobileNets [44]. MobileNets strives to minimize the latency of smaller scale networks so that computer vision applications can run on mobile (edge) devices. They compared the performance of the MobileNets model to the traditional CNN model using ImageNet dataset, and reported that the MobileNets uses only 4.2 million parameters, when a traditional CNN uses 29.3 million parameters. This parameter reduction process reduced the accuracy by 1% but the number of multiplications and additions dropped by more than 8 times, significantly improving overall performance.

3.3.2 ShuffleNet

Zhang et al. have designed a new architecture called ShuffleNet [117], using depthwise separable convolutions. To reduce computational cost, they introduce two new operations, pointwise group convolutional and channel shuffle. A grouped convolution is simply several convolutions where input channels are grouped in different groups and convolution is performed independently for each group of channels. This strategy greatly reduces the computational cost. For example, if a convolutional layer has

4 input channels and 8 output channels. Then, the total computational cost is $D_K \cdot D_K \cdot 4 \cdot 8 \cdot D_F \cdot D_F$ (Equation 1). While with two groups, each taking 2 input channels and 4 output channels, the computational cost is $(D_K \cdot D_K \cdot 2 \cdot 4 \cdot D_F \cdot D_F) \cdot 2$ which is half as many operations. In this model, finding the appropriate number of groups is crucial. The authors reached the best results with 8 groups on the ImageNet dataset. Finally, a channel shuffle operation is added in this model to mix the output channels of the group convolution. The authors claim this new architecture is 3.1% more accurate than MobileNets.

3.3.3 L-CNN

Nikouei et al. introduced a lightweight convolutional neural network called L-CNN [78] that is inspired by depthwise separable convolutions. L-CNN can run on edge devices and is able to detect pedestrians in a real-time human surveillance system. They used a Raspberry Pi 3 Model B (with an ARMv7 1.2 GHz processor and 1 GB RAM) to run this algorithm and reported that it was 64% faster than MobileNets in detecting human objects in a resource-scarce edge environment.

3.4 Other NN architectures

3.4.1 SqueezeNet

SqueezeNet [48] is a parameter efficient neural network used in resource-constrained settings. This small CNN-like architecture has 50 times fewer parameters than AlexNet while preserving AlexNet-level accuracy on the ImageNet dataset. By using deep compression with 6-bit quantization, this model can be compressed to 0.47 MB, which is 510 times smaller than 32-bit AlexNet. The authors used two techniques to reduce the model size: a) by using 1×1 filters instead of 3×3 filters b) by decreasing the number of input channels. These techniques decrease the number of parameters but affect the accuracy. To compensate, the authors downsample later in the network to have larger activation maps which lead to higher accuracy. The authors report that SqueezeNet exceeds the top-1 and top-5 accuracy of AlexNet while using 50 times smaller size model.

Pradeep et al. deployed a CNN model on an embedded FPGA platform [88]. They use low bit floating-point representation¹¹ to reduce the computational resources required for running a CNN model on FPGA. Their architecture was tested with the SqueezeNet model on the ImageNet dataset. The authors reported having achieved 51% top-1 accuracy on a DE-10 board at 100 MHz that only consumed 2W power.

3.4.2 FastRNN & FastGRNN

Recurrent Neural Networks (RNNs) are powerful neural networks used for processing sequential data, but suffer

¹¹8-bit for storage and 12-bit during computations.

from inaccurate training and inefficient prediction. Techniques to address these issues, such as unitary RNNs and gated RNNs, increase the model size as they add extra parameters. To shrink the model size Kusupati et al. designed two new techniques, FastRNN and FastGRNN [57]. In FastRNN, an additional residual connection is used that has only two additional parameters to stabilize the training by generating well-conditioned gradients. FastRNN used the following formulas to update any hidden state (h_t):

$$\begin{aligned}\bar{h}_t &= \sigma(W \cdot x_t + U \cdot h_{t-1} + b), \\ h_t &= \alpha \bar{h}_t + \beta h_{t-1}\end{aligned}$$

where $\alpha \ll 1$ and $\beta \approx 1 - \alpha$ are trainable weights that are parameterized by the sigmoid function, σ . FastRNN controls the extent of its hidden layers by limiting α and β . FastRNN has two additional parameters compared to the traditional RNN, but it requires very little computation which does not affect the performance of this algorithm.

After analyzing the performance of FastRNN Kusupati et al. found that the expressive power of this model might be limited for some datasets. So, they designed another architecture called FastGRNN by converting the residual connection to a gate while reusing the RNN matrices. FastGRNN has 2-4 times fewer parameters than other leading gated RNN models such as LSTM, GRU, but provides the same or sometimes better accuracy than gated RNN models [57].

It is possible to fit FastGRNN in 1-6 kilobytes which makes this algorithm suitable for IoT devices, such as Arduino Uno. The developers of this model reported 18–42 times faster prediction results compared to other leading RNN models by deploying FastGRNN on an Arduino MKR1000. FastRNN and FastGRNN, both of which are open-source¹², have been benchmarked using the following:

- utterance detection (e.g. "Hey Cortana") using the Wakework-2 dataset
- utterance detection with background noise and silence using the Google-12 dataset
- human activity recognition using HAR-2 and DSA-19 datasets
- language modeling using the Penn Treebank (PTB) dataset
- star rating prediction using the Yelp review dataset and image classification using the MNIST dataset.

3.5 Training DNN models on edge devices

In this section, we discuss DNN models in which *both training and inference is done on edge devices*. These models are especially important in situations where there

is a need to handle new problems arising after deployment of a trained model on an edge device. Also, edge training is necessary to provide personalized support for smart device users. For example, a pre-trained speech recognition model would not work equally well for people from different parts of the world. In that case, re-training a DNN model using a mobile phone would improve accuracy. However, due to resource limitation, training a deep convolutional neural network on an edge device is daunting. Usually, deep learning applications deployed on the network edge use cloud servers for training a DL model and then run inference on edge devices.

3.5.1 Re-training of pruned networks

Chandakkar et al. designed a new architecture to re-train a pruned network on an edge device (such as a smartphone) [19]. This model runs the following steps in cyclic order to re-train a DNN model:

- A complete DNN is trained for an epoch (when an entire dataset is passed forward and backward through the DNN) on the original data.
- Layer-wise magnitude-based weight pruning is performed with a user-defined threshold value. This greatly reduces the computational complexity by removing connections in a DNN model and makes it suitable to run on a resource-scarce device. Unfortunately, any pruning process reduces the accuracy of a model. To overcome this issue, this approach finds the indices of most important weights for an important feature and excludes these elements from being pruned.
- Finally, the pruned DNN network is used while training the next epoch.

3.5.2 Privacy-preserving learning

Mao et al. presented a privacy-aware DNN training architecture that uses differential privacy to protect user data during the training [72]. This deep learning scheme is capable of training a model using multiple mobile devices and the cloud server collaboratively, with minimal additional cost. First, this algorithm selects one convolutional layer to partition the neural network into two parts. One part runs on edge devices another part runs on the server (AWS cloud server). The first part takes raw data with sensitive user information as input and uses a differentially private activation algorithm to generate the volume of activations as output. These output activations contain Gaussian noise, which prevents external cloud servers from reproducing original input information using reversing activations. These noisy output activations are then transmitted to cloud servers for further processes. The servers take output activations as input and run the subsequent training process. This model was evaluated using a Nexus 6P phone and AWS based servers. The authors report that they have achieved good accuracy by deploying this model on Labeled Faces in the Wild dataset (LFW).

¹²<https://github.com/Microsoft/EdgeML/>

Table 2: eSGD gradient drop ratio data reported by Tao and Li [99].

Drop Ratio	Total iterations	Accuracy
25%	200000	95.31
50%	200000	91.22
87.5%	200000	88.46
75%	200000	83.85
75%	150000	83.76
75%	100000	81.13

3.5.3 eSGD

Exchanging model parameters and other data between edge devices and cloud servers is mandatory for training an edge-cloud-based DL model. However, as the size of the training model increases, more data needs to be exchanged between edge devices and servers. The high network communication cost is a bottleneck for a training model. To reduce communication costs and keep model accuracy high, Tao and Li introduced a new method called Edge Stochastic Gradient Descent (eSGD) [99]. In this approach, all edge devices run training tasks separately with independent data and the gradient values generated by the edge devices are sent to the cloud servers.

The server collects all gradients from edge devices and performs gradient synchronization by taking their average. After that, it updates the parameters by using this average value. These updated parameters are sent back to the edge devices for the next training step. This process is called parameter synchronization.

The authors noticed that **only a small fraction of the gradients need to be updated after each mini-batch**. This finding helps the authors to reduce communication cost by taking only the required gradients and sending them to the server. Unfortunately, **this gradient selection technique decreases model accuracy**. So, eSGD uses two mechanisms to maintain satisfactory training accuracy:

- **‘Important’ updating:** After each mini-batch only a small fraction of the gradient coordinates need to be updated. eSGD determines these important gradients and transfers them to the server for updating the parameters. This process significantly reduces communication cost. Random weight selection is used to select important gradients.
- **Momentum residual accumulation:** This mechanism is applied for tracking and accumulating out-of-date residual gradients, which helps to avoid low convergence rate caused by the previous important updating method.

eSGD is capable of reducing the gradient size of a CNN model by up to 90%. Unfortunately, high gradient shrinking leads to bad accuracy. Tao and Li used MNIST in their experiments and reported 91.22% accuracy with a 50% gradient drop (Table 2).

4 ML systems at the edge

The previous two sections covered techniques that have been developed for machine learning inferencing (and, in some cases, also training) on the network edge. This section focuses on the actual *applications* of edge-based ML and DL methods for deploying intelligent systems.

4.1 Real-time video analytics

Real-time video analytics systems are an integral part of a wide range of applications, e.g. self-driving cars, traffic safety & planning, surveillance, and augmented reality [9]. Until recently, video analytics systems using ML algorithms could only process about 3 fps whereas most real-time video cameras stream data at 30 fps [52]. Edge computing with IoT cameras has been used to address this problem and provide improved real-time video analytics services.

Ananthanarayanan et al. developed a video analytics system called Rocket [9] that produces high-accuracy outputs with low resource costs. Rocket collects video from different cameras and uses vision processing modules for decoding it. Each module uses predefined interfaces and application-level optimizations to process video data. A resource manager is used to execute data processing tasks on different resource-constrained edge devices and cloud servers. A traffic analytics system based on the Rocket software stack has been deployed in Bellevue, WA to track cars, pedestrians, and bikes. After processing the data in real-time, it raises an alert if anomalous traffic patterns are detected. Rocket has been shown to be effective in a variety of applications [10], which include *a)* a smart crosswalk for pedestrians in a wheelchair, *b)* a connected kitchen to pre-make certain food to reduce customer wait-times, *c)* traffic dashboard for raising an alarm in abnormal traffic volumes, and *d)* retail intelligence for product placement.

As the volume of video data produced by IoT and other cameras has increased sharply, searching for relevant video in large video datasets has become more time consuming and expensive. Hsieh et al. introduced a low-cost, low-latency video querying technique called Focus [45]. This system generates indexes of multiple object classes in the video and uses them to search a relevant video. They used GT-CNN, a low-cost CNN architecture with fewer convolutional layers for recognizing objects and indexing the object classes of each video stream.

Xu et al. used an SVM classifier with the Histogram of the Oriented Gradient (HOG) feature extraction algorithm on edge devices to develop a real-time human surveillance system [78, 114]. They used the COCO image set archive to train the SVM classifier with around 20K images and a Raspberry Pi 3 (model B with an ARMv7 1.2 GHz processor, 1 GB of RAM) to run the HOG and SVM and were able to successfully distinguish between human and nonhuman objects in real-time.

Qi and Liu used an embedded GPU to reach real-time video processing speed (30fps) by using a quantized DL

model [89]. They used Nvidia’s TensorRT framework to quantize the CNN model parameter to 16-bit float type. They deployed their model on an embedded GPU (Nvidia Jetson TX2) and showed real-time videos analysis with different video resolutions: 1080p and 720p. They claimed that approaches based on model quantization and pruning have the capability of improving the deployment of deep learning models on the IoT edge.

Wang et al. introduced a bandwidth-efficient video analytics architecture based on edge computing that enables real-time video analytics on small autonomous drones [109].

Kar et al. presented a CNN-based video analytics system to count vehicles on a road and estimate traffic conditions without the help of surveillance cameras [54]. They consider a vehicle as an edge device and deploy their model on a dashboard camera on-board the vehicle. The deployed model uses an object detection framework called YOLO [90] to detect other vehicles on the road. They used 8000 car images to train the model and then deployed the trained model to identify a moving vehicle and achieved an accuracy of 90%.

Ali et al. designed an edge-based video analytics system using deep learning to recognize an object in a large-scale IoT video stream [3]. There are four different stages in this video analytics system: *a)* frame loading/decoding and motion detection, *b)* preprocessing, *c)* object detection and decomposition, and *d)* object recognition. The first three actions are performed on the edge infrastructure and the fourth one in the cloud. To improve accuracy, this model uses a filter that finds important frames from the video stream and forwards them to the cloud for recognizing objects. Their edge-cloud based model was 71% more efficient than the cloud-based model in terms of throughput on an object recognition task.

4.2 Image recognition

Image recognition refers to the process of extracting meaningful information from a given image, e.g. to identify objects in that image. Deep learning techniques such as CNNs can be used to detect people, places, handwriting, etc. in an image. The prevalence of IoT cameras and mobile devices has increased the importance of improved image recognition techniques.

Until recently, data would almost always be transferred to the cloud where the images captured by IOT or mobile phones would be processed. Researchers have increasingly begun to use edge computing techniques to process images close to where they are captured. For example, there are currently more than 2.5 billion social media users in the world and millions of photographs and videos are posted daily on social media [97]. Mobile phones and other devices can capture high-resolution video which, uploading which may require high bandwidth. By processing this data on the edge, the photos and videos are adjusted to a suitable resolution before being uploaded to the Internet

[95]. Caffe2Go¹³ is a lightweight framework that allows deploying DL systems on a mobile device and helps to reduce the size of the input layer of a DL model.

PII information in videos and images can be removed at the edge before they are uploaded to an external server, thereby enhancing user privacy.

IoT cameras and edge devices have been used with DL algorithms to understand animal behavior [32], which helps in the study of changes in animal habitats.

Liu et al. developed a food image recognition model for automatic dietary assessment [66]. In this model, machine learning algorithms were deployed on edge devices and cloud servers collaboratively. The main function of the edge device is to identify a blurry image taken by the user. Generally, the percentage of edge pixels for a blurry image is lower than a clear image. They used this feature and some other texture features (e.g., contrast, correlation) as the input of a two-step K-means clustering algorithms to identify a blurry image. If the algorithm finds a blurry and low quality image, then the user gets a real-time notification to retake the picture. Otherwise it segments the original image with different filters to generate a clear image. Due to the completion of these notification and segmentation processes in real-time on an edge device, the overall run time for the algorithms decreases greatly. After processing the food image on the edge device, a clear image is sent to the cloud server for further processing. The communication between the edge device and the cloud servers is maintained by Apache HttpClient services. In this experiment two publicly available datasets have been used for recognizing food images, UEC-256/UEC-100 and Food-101. They reported their food recognizing system is 5% more accurate than the existing approach (FoodCam(ft)) [55] using the same dataset (UEC-100).

Drolia et al. designed a prefetching and caching technique to reduce image recognition latency for mobile applications [30]. They used a Markov model to predict which parts of the trained classifiers a user might use in the future to recognize a new image. Based on this prediction, this model caches parts of the trained classifiers and generates smaller image recognition models. It also changes feature extraction parameters based on network conditions and computing capability of the mobile device. Finally, it uses smaller recognition models that are cached on the device and adjusted feature parameters to recognize an image.

Tuli et al. introduced a deep learning-based real-time object detection system using IoT, fog, and cloud computing [106]. They used the YOLOv3 architecture [90] by training on the COCO dataset to evaluate their system. The authors have developed an open-source fog-cloud deployment system called EdgeLens¹⁴ and demonstrated the capability of this system by deploying object detection YOLO software on multiple Raspberry Pi de-

¹³<https://code.fb.com/android/delivering-real-time-ai-in-the-palm-of-your-hand/>
¹⁴<https://github.com/Cloudslab/EdgeLens>

vices and cloud VMs.

4.3 Automatic Speech Recognition

There is immense community interest in developing an offline speech recognition system that supports a digital voice-assistant without the help of the cloud. Limited-vocabulary speech recognition, also known as keyword spotting is one method to achieve offline speech recognition [112]. A typical keyword spotting system has two components:

- a feature extractor: extract necessary features from the human voice
- a neural network-based classifier: takes voice features as input and generates a probability for each keyword as output

DNN-based keyword spotting systems are not easily deployable on resource-constrained devices. Lin et al. designed a highly efficient DNN called EdgeSpeechNets to deploy DL models on mobile phones or other consumer devices for human voice recognition [65]. Their model achieved higher (approx. 97%) accuracy than state-of-the-art DNNs with a memory footprint of about 1MB using Google Speech Commands dataset. They used the Motorola Moto E phone with a 1.4 GHz Cortex-A53 mobile processor as the edge device. EdgeSpeechNets used 36 times fewer mathematical operations resulting in 10 times lower prediction latency, a 7.8 times smaller network size, and a 16 times smaller memory footprint than state-of-the-art DNNs.

Chen et al. introduced a small-footprint keyword spotting technique based on DNNs called Deep KWS, which is suitable for mobile edge devices [21]. Deep KWS has three components: a feature extractor, a deep neural network, and a posterior handling module. The feature extractor uses the KWS algorithm to analyze the input audio and generate a feature vector. This feature vector is used as input to the DNN to generate frame-level posterior probabilities scores. Finally, the posterior handling module uses these scores to generate the final output score of every audio frame to recognize the audio. Their model achieved 45% relative improvement with respect to the Hidden Markov Model-based system. Chen et al. also developed a query-by-example keyword spotting technique using a long short-term memory (LSTM) network instead of standard DNN [22]. This model also has low computational cost and small memory footprint, and can easily be executed on a resource-constrained edge device.

4.4 Privacy and security

The widespread use of IoT devices and their ability to generate information has boosted personal data production, but poor security protections on these devices has simultaneously increased the potential for misuse of user data.

Osia et al. designed a hybrid architecture that works

with edge and cloud servers collaboratively to protect user privacy. All personal data is collected and processed on personal edge devices to remove sensitive information [81]. Only data that is free of sensitive information is sent to the cloud server for further processing.

Das et al. introduced a distributed privacy infrastructure for IoT that notifies users about nearby cameras, what data is collected about them and how this data is being used [27]. This framework can denatures user faces if desired and uses five components to ensure user privacy:

- Internet of Things Resource Registry (IRR): Stores and advertises the privacy-related information, which helps the user to understand the policy.
- IoT Assistant (IoTA): IoTA is an android application which needs to be installed on the user phone. This application captures the resource published by IRR and informs the user about nearby cameras, collected data and what it does with the collected data. After being informed about the data capturing system, the user can configure their privacy preferences (e.g. opt-in/opt-out of any service using sensitive data).
- Policy Enforcement Point (PEP): Ensures user-defined privacy settings are deployed on cameras and maintains a database for storing user's setting for future use.
- Face Trainer: Recognizes human faces using OpenFace (a DNN based face recognition library).
- Privacy Mediator: Denatures human faces from live video feeds.

To mitigate security risks to IoT networks, Pajouh et al. proposed a model to identify suspicious behaviors such as a user-to-root attack or a remote-to-local attack within IoT networks [83]. They used the NSL-KDD dataset [100] with the k-NN and Naive Bayes algorithm to classify normal and suspicious behaviors. This experiment was conducted using a personal computer as an edge device without the help of cloud servers, which showed that an edge device can detect suspicious activity in IoT networks.

4.5 Fraud detection

With the increase in data being generated by IoT and smart devices, incidents of data fraud and theft are also increasing. Machine learning is being used to prevent data falsification and to authenticate data validity. Ghoneim et al. developed a new medical image forgery detection framework that can identify altered or corrupted medical images [38]. They used a multi-resolution regression filter on a noise map generated from a noisy medical image and then used SVM and other classifiers to identify corrupted images. The first part of their algorithm (that creates a noisy map from a medical image) is done on an edge computing device, and the rest is done on a cloud server. This distributed approach decreases the time for data processing as well as bandwidth consumption. They used the

SVM algorithm and CASIA 1 and CASIA 2 datasets in their experiment and reported 98.1% accuracy on CASIA 1 and 98.4% accuracy on CASIA 2 dataset. Also, the maximum bandwidth consumption during the experiment of their proposed system without edge computing is 322 bits per second and with edge computing is 281 bits per second.

4.6 Creating training datasets

Assembling and labeling a large training dataset and producing accurate classifiers are typically accomplished by the collaboration of human and ML algorithms [34]. Unfortunately, raw data collected from various data sources often contains a huge amount of noise. Human experts with specialized image identification skills or programming skills are needed to create an accurate classifier in domains such as military, environmental or medical research. Discarding noise from the data as early as possible increases the speed of this training process. Early discard refers to the removal of irrelevant data in the initial phase of processing pipelines. This could be on the operating system layer, application layers, or transport layers across the Internet. This early discard helps label data quickly and reduces human involvement during data assembling hence improve the efficiency of the system. This early discard approach may apply to both live data sources (e.g. video cameras) and archival data sources (datasets dispersed over the Internet).

Feng et al. designed an architecture using edge computing that produces a labeled training dataset which can be used to train a machine learning model [34]. They used edge computing to discard irrelevant data at the early stage using three targets (deer, the Taj Mahal, and fire hydrants) that do not have a public dataset for training. Their approach, named Eureka, uses HoG, SVM, MobileNets and R-CNN at the edge of the network to build a labeled training dataset for the three targets. They were able to reduce the human labeling effort by two orders of magnitude compared to a brute-force approach.

4.7 Autonomous vehicles

An autonomous vehicle on average generates more than 50 GB of data every minute¹⁵. This data must be processed in real-time to generate driving decisions. The bandwidth of an autonomous vehicle is not large enough for transferring this enormous amount of data to remote servers. Therefore, edge computing is becoming an integral part of autonomous driving systems.

Navarro et al. designed a pedestrian detection method for autonomous vehicles [76]. A LIDAR sensor gathers data to detect pedestrians and features are extracted from this data. These features include stereoscopic information, the movement of the object, and the appearance of a pedestrian (local features like Histogram of Oriented

Gradients). Using the features obtained from raw LIDAR data, an n -dimensional feature vector is generated to represent an object on the road. This feature vector is used as input for the machine learning model to detect pedestrians. They use a Nuvo-1300S/DIO computer to run the machine learning model inside an autonomous vehicle and report 96.8% accuracy in identifying pedestrians.

Hochstetler et al. have shown that it is possible to process real-time video and detect objects using deep learning on a Raspberry Pi combined with an Intel Movidius Neural Compute Stick [43]. They reported that their embedded system can independently process feeds from multiple sensors in an autonomous vehicle.

4.8 Healthcare monitoring

Deep learning models are increasingly being used to analyze medical data and images [108]. Anguita et al. designed a new algorithm, Multiclass Hardware Friendly Support Vector Machine (MC-HF-SVM), for building models geared towards edge-devices, focusing on healthcare applications such as human activity recognition and monitoring [11]. Their algorithm uses fixed-point arithmetic to reduce computational cost. Signals from the sensors (accelerometer, gyroscope) are used as input of the MC-HF-SVM algorithm. Noise reduction filters are applied on the sensor signal and the resulting noise-free signal is sampled into fixed-width sliding windows to generate a feature vector, which is fed to the MC-HF-SVM algorithm. Their algorithm requires less processing time, memory and power, with only a 0.3% reduction in accuracy compared to the standard floating-point multi-class SVM.

4.9 Smart homes and cities

Homes equipped with intelligent systems built using numerous resource-constrained devices are increasingly being designed [41]. An important goal in the design of such smart homes is ensuring the safety of children and the elderly. Hsu et al. developed a fall detection system that generates an alert message when an object falls [46]. Their approach has three steps: a) a skeleton extraction performed followed by ML prediction model to detect falls, b) a Raspberry Pi 2 is used as an edge computing device for primary data processing and to reduce the size of videos or images, c) falls are detected using machine learning on the cloud and users are notified in appropriate cases.

Tang et al. designed a hierarchical distributed computing architecture for a smart city to analyze big data at the edge of the network, where millions of sensors are connected [98]. They used SVMs in a smart pipeline monitoring system to detect threatening events on the pipeline. Their architecture has multiple layers:

- Sensing Networks: this layer contains numerous low-cost sensors, which generate a massive amount of data.

¹⁵<https://datafloq.com/read/self-driving-cars-create-2-petabytes-data-annually/>
172

- Edge computing layer: this level consists of low-power computing devices (see §6). The main function of this layer is to quickly detect hazardous events to avoid potential damages. For example, if a house in the smart city experiences a leakage or a fire in the gas line, this layer will detect the threat and quickly shut down the gas supply of that home/area without any help from cloud computing. Tang et al. used support vector machine in edge devices to detect potential threat patterns.
- Data center layer: This layer performs complex calculations which are not possible on edge computing devices. For example, relationship modeling for detecting a massive threat event by analyzing big data generated by the lower layer.

Chang et al. designed an edge-based energy management framework for smart homes [20] that improves the use of renewable energy to meet the requirements of IoT applications. Since sunlight is the main source of renewable energy, they used a numerical weather prediction (NWP) model [86] to predict weather impacting solar energy generation. After obtaining forecast information from NWP models, an energy scheduling module generates a cost-effective energy utilization schedule for the smart home. Raspberry Pi 3 B was used to run their framework at the location of its users, helping to protect privacy-sensitive data.

Park et al. designed a new fault detection system for edge computing (LiRed) using LSTM recurrent neural networks [84]. A Raspberry Pi 3 Model B with 1 GB memory and 16 GB flash storage was used as an edge device to run LiRed for detecting faults in a smart factory. Evaluated in terms of F1 and F2 scores, their classification technique performs better than SVM and RF for detecting machine faults.

4.10 Safety and security

To improve pedestrian security, Miraftabzadeh et al. introduced a new embedded algorithm pipeline using artificial neural networks with edge computing to identify a person in real time [75]. Their technique maps the facial image of a pedestrian to a high-dimensional vector by using a facial privacy-aware parameterized function. This vector helps to identify and track individuals without determining their true identity. They embedded a ResNet (concurrent residual neural network) model with every camera to extract a facial feature vector by real-time CCTV video analysis. ResNet is trained with a vectorized-l2-loss function for face recognition and a multivariate kernel density estimation matching algorithm is applied for identity identification and security verification. They reported achieving a 2.6% higher accuracy over other state-of-the-art approaches.

Liu et al. developed an edge-based phone application software to detect attacks in ride-sharing services [67]. Their model has three parts: a) audio capture & analy-

sis, b) driving behavior detection, and c) video capture & analysis. In the audio capture and analysis step, an android mobile application runs a speech recognition model to capture and analyze the audio during a ride. If it detects certain keywords (e.g. ‘help’, ‘rescue me’) with abnormal high-pitched sounds, it captures and uploads video to the cloud server in real-time, where the video is analyzed using a trained CNN model. If any abnormal movement or dangerous objects are detected, the cloud server shares the video with law enforcement agencies.

Dautov et al. introduced an intelligent surveillance system using IoT, cloud computing, and edge computing [28]. This system processes sensitive data at the edge of the network to enhance data security and reduces the amount of data transferred through the network.

Xu et al. developed a smart surveillance system for tracking vehicles in real-time [115]. Their system helps to identify suspicious vehicles after traffic accidents. They processed camera streams at the edge of the network and stored space-time trajectories of the vehicles instead of storing raw video data which reduces the size of the stored data. Vehicle information can be found by querying these space-time trajectories.

4.11 Document classification

Ding and Salem designed a novel architecture for automatic document classification (D-SCAML) which improves data protection in edge environments [29]. D-SCAML uses natural language processing and machine learning techniques, such as decision trees and Naive Bayes, to predict the likely nature of raw data and enforces security procedures on that data. Then this data is processed, analyzed and classified using edge and cloud servers collaboratively.

5 ML Frameworks

This section describes common frameworks that have been used to deploy machine learning models on edge devices. Table 3 (on page 17) summarizes common ML frameworks, along with their core language, interface, and applications.

TensorFlow Lite TensorFlow is a popular machine learning framework, developed by Google. TensorFlow Lite¹⁶ is a lightweight implementation of TensorFlow for edge devices and embedded systems. TensorFlow Lite has been used for classification and regression on mobile devices. It supports DL without the help of a cloud server and has some neural network APIs to support hardware acceleration¹⁷.

TensorFlow Lite can be run on multiple CPUs and GPUs and is therefore well-suited for distributed ML algorithms. The main programming languages for

¹⁶<https://www.tensorflow.org/lite>

¹⁷https://www.tensorflow.org/lite/performance/gpu_advanced

Table 3: Machine learning frameworks that have been used on edge devices.

Framework	Core language	Interface	Part running on the edge	Example applications
TensorFlow Lite (Google)	C++ Java	C/C++ Java	TensorFlow Lite NN API	computer vision [109], speech recognition [42, 1]
Caffe2 Caffe2Go (Facebook)	C++	Android iOS	NNPack	image analysis, video analysis [53]
Apache MXNet	C++ Python R	Linux MacOS Windows	Full Model	object detection, recognition [78]
Core ML2 (Apple)	Python	iOS	CoreML	image analysis [16] NLP [105]
ML Kit (Google)	C++ Java	Android iOS	Full Model	image recognition, text recognition, bar-code scanning [26]
AI2GO	C, Python Java, Swift	Linux macOs	Full Model	object detection, classification [5]
DeepThings	C/C++	Linux	Full Model	object detection [119]
DeepIoT	Python	Ubilinux	Full Model	human activity recognition, user identification [116]
DeepCham	C++ Java	Linux Android	Full Model	object recognition [62]
SparseSep	-	Linux Android	Full Model	mobile object recognition, audio classification [15]
Edgent	-	Ubuntu	Major part of the DNN	image recognition [63]

this framework are Java, Swift, Objective-C, C, and Python. A performance evaluation study of TensorFlow Lite by Zhang et al. showed that it occupied only 84MB memory and took 0.26 seconds to execute an inference task using MobileNets on a Nexus 6p mobile device [118].

Caffe2 and Caffe2Go Caffe2¹⁸ is a fast and flexible deep learning framework developed by Facebook. Caffe2Go is a lightweight and modular framework built on top of Caffe2. Both frameworks provide a straightforward way to implement deep learning models on mobile devices and can be used to analyze images in real time¹⁹. Caffe2Go²⁰ can run on the Android and iOS platforms with the same code. It implements debugging tools by abstracting the neu-

ral network math. It uses fewer convolutional layers than traditional neural networks and optimizes the width of each layer to reduce model size.

Apache MXNet Apache MXNet²¹ is a lean, scalable, open-source framework for training deep neural networks and deploying them on resource-scarce edge devices. MXNet supports distributed ecosystems and public cloud interaction to accelerate DNN training and deployment. It comes with tools which help to tracking, debugging, saving checkpoints, and modifying hyperparameters of DNN models.

CoreML3 CoreML3²² is an iOS-based ML framework developed by Apple for building ML models and integrating them with Apple mobile applications. It allows an application developer to create an ML model to perform regression and image classification. This

¹⁸Caffe2 is now being merged with Pytorch.

¹⁹<https://code.fb.com/android/delivering-real-time-ai-in-the-palm-of-your-hand/>

²⁰<https://caffe2.ai/blog/>

²¹<https://mxnet.apache.org/>

²²<https://developer.apple.com/machine-learning/core-ml/>

framework allows ML to run on edge devices without a dedicated server. A trained DNN model is translated into CoreML format and this translated model can be deployed using CoreML APIs to make an image classifier inside a mobile phone.

ML Kit ML Kit²³ is a mobile SDK framework introduced by Google. It uses Google’s cloud vision APIs, mobile vision APIs, and TensorFlow Lite to perform tasks like text recognition, image labeling, and smart reply. Curukoglu et al. tested ML Kit APIs for image recognition, bar-code scanning, and text recognition on an Android device and reported that these APIs recognize different types of test objects such as tea cup, water glass, remote controller, and computer mouse successfully [26].

AI2GO Introduced by Xnor²⁴, AI2GO helps to tune deep learning models for popular use cases on resource-scarce devices. More than 100 custom ML models have been built with this framework for on-device AI inferencing. These custom models have the ability to detect objects, classify foods and many other AI applications [5]. This platform targets specialized hardware which includes the Raspberry Pi, Ambarella S5L, Linux and macOS based laptops, and Toradex Apalis iMX6. Allan conducted tests to benchmark the AI2GO platform on a Raspberry Pi and reported this platform to be 2 times faster than TensorFlow Lite in machine learning inferencing using a MobileNets v1 SSD 0.75 depth model [5].

DeepThings DeepThings is a framework for adapting CNN-based inference applications on resource-constrained devices [119]. It provides a low memory footprint of convolutional layers by using Fused Tile Partitioning (FTP). FTP divides the CNN model into multiple parts and generates partitioning parameters. These partitioning parameters with model weights are then distributed to edge devices. When all edge devices complete their computational tasks, a gateway device collects the processed data and generates results. The authors deployed YOLOv2 using DeepThings on Raspberry Pi 3 devices to demonstrated the deployment capability of this framework on IoT devices.

DeepIoT DeepIoT is a framework that shrinks a neural network into smaller dense matrices but keeps the performance of the algorithm almost the same [116]. This framework finds the minimum number of filters and dimensions required by each layer and reduces the redundancy of that layer. Developed by Yao et al., DeepIoT can compress a deep neural network by more than 90%, shorten execution time by more than 71%, and decrease energy consumption by 72.2% to 95.7%.

DeepCham Li et al. introduced a framework called DeepCham which allows developers to deploy DL models on mobile environments with the help of edge computing devices [62]. DeepCham is developed for recognizing objects captured by mobile cameras, specifically targeting Android devices.

SparseSep SparseSep, developed by Bhattacharya and Lane, is a framework for optimizing large-scale DL models for resource-constrained devices such as wearable hardware [15]. It runs large scale DNNs and CNNs on devices that have ARM Cortex processors with very little impact on inference accuracy. It can also run on the NVidia Tegra K1 and the Qualcomm Snapdragon processors and was reported to run inference 13.3 times faster with 11.3 times less memory than conventional neural networks.

DeepX Lane et al. presented a software accelerator for low-power deep learning inference called *DeepX*, which allows developers to easily deploy DL models on mobile and wearable devices [60]. DeepX dramatically lowers resource overhead by decomposing a large deep model network into unit-blocks. These unit-blocks are generated using two resource control algorithms, namely Runtime Layer Compression and Deep Architecture Decomposition, and executed by heterogeneous processors (e.g. GPUs, LPUs) of mobile phones.

Edgent Li et al. developed Edgent²⁵, a framework for deploying deep neural networks on small devices [63]. This framework adaptively partitions DNN computations between a small mobile device (e.g. Raspberry Pi) and the edge computing devices (e.g. laptops), and uses early-exit at an intermediate DNN layer to accelerate DNN inference.

6 Hardware and Software

This section describes the software and low-power hardware that have been used for deploying machine learning systems at the network edge.

6.1 Hardware

Table 4 on page 19 lists embedded computing boards on which ML models have been successfully deployed at the network edge. Below, we describe the context in which they have been used.

Raspberry Pi

The Raspberry Pi, a single-board computer developed by the Raspberry Pi Foundation, is one of the most common devices used for edge computing. It has been used to run ML inference without any extra hardware. The Raspberry Pi 3 Model B has a Quad

²³<https://firebase.google.com/products/ml-kit>

²⁴<https://www.xnor.ai/>

²⁵This is unrelated to Apache Edgent, an SDK for developing streaming analytics at the edge that we discuss later (§5).

Table 4: Computing devices that have been used for machine learning at the edge.

Device	GPU	CPU	RAM	Flash memory	Power consumption	Example applications
Raspberry Pi	400MHz VideoCore IV	Quad Cortex A53 @ 1.2GHz	1 GB SDRAM	32 GB	2.5 Amp	video analysis [78, 114]
Coral Dev Board (Edge TPU)	GC7000 Lite Graphics + Edge TPU coprocessor	Quad Cortex-A53, Cortex-M4F	1 GB LPDDR4	8 GB LPDDR4	5V DC	image processing [18]
SparkFun Edge	-	32-bit ARM Cortex-M4F 48MHz (with 96MHz burst mode) processor	384KB	1MB	6uA/MHz	speech recognition [31]
Jetson TX1	Nvidia Maxwell 256 CUDA cores	Quad ARM A57/2 MB L2	4 GB 64 bit LPDDR4 25.6 GB/s	16 GB eMMC, SDIO, SATA	10-W	video, image analysis [68] [61] robotics [33]
Jetson TX2	Nvidia Pascal 256 CUDA cores	HMP Dual Denver 2/2 MB L2 + Quad ARM A57/2 MB L2	8 GB 128 bit LPDDR4 59.7 GB/s	32 GB eMMC, SDIO, SATA	7.5-W	video, image analysis [68], [92] robotics [24]
Intel Movidius Neural Compute Stick	High Performance VPU	Myriad 2 Vision Processing Unit	1 GB	4 GB	2 trillion 16-bit operations per second within 500 mW	classification [73] computer vision [14, 43]
ARM ML	-	ARM ML processor	1 GB	-	4 TOPs/W (Tera Operations)	image, voice recognition [107]
RISC-V GAP8	-	nona-core 32-bit RISC-V microprocessor @250 MHz	16 MiB SDRAM	-	1 GOPs/mW	image, audio processing [35]
OpenMV Cam	-	ARM 32-bit Cortex-M7	512KB	2 MB	200mA @ 3.3V	image processing [6]
BeagleBone AI	-	Cortex-A15 Sitara AM5729 SoC with 4 EVEs	1 GB	16 GB	-	computer vision [25]
EMC3531	-	ARM Cortex-M3 NXP Coolflux DSP	-	-	-	audio, video analysis ³⁸

Cortex A53 @ 1.2GHz CPU, 400MHz VideoCore IV GPU, 1GB SDRAM. Xu et al. used Raspberry Pi 3 as edge devices to develop a real-time human surveillance system [78, 114]. Their system is able to distinguish between human and nonhuman objects in real-time.

It has a micro-SD card slot to support flash memory up to 32 GB. Xnor.ai has developed a new AI platform to run deep learning models efficiently on edge devices such as embedded CPUs (e.g. Raspberry Pi), phones, IoT devices, and drones without using a GPU or TPU [58, 113].

Edge TPU and Coral Dev Board

The edge tensor processing unit (TPU²⁶) is an ASIC chip designed by Google for ML inference on edge devices. It accelerates ML inference and can execute convolutional neural networks (CNN). It is capable of running computer vision algorithms, such as MobileNets V1/V2, MobileNets SSD V1/V2, and Inception V1-4. The TPU can also run TensorFlow Lite and NN APIs. If a user selects the power efficient mode, the Edge TPU can execute state-of-the-art mobile vision models at 100+ fps²⁷. The Coral Dev Board uses the Edge TPU as a co-processor to run machine learning applications.

The Coral Dev board has two parts, a baseboard and a system-on-module (SOM). The baseboard has a 40-pin GPIO header to integrate with various sensors or IoT devices and the SOM has a Cortex-A53 processor with an additional Cortex-M4 core, 1GB of RAM and 8GB of flash memory that helps to run Linux OS on Edge TPU.

The Coral USB accelerator²⁸ is a device that helps to run ML inference on small devices like the Raspberry Pi. The accelerator is a co-processor for an existing system, which can connect to any Linux system using a USB-C port. Allan deployed machine learning inference on different edge devices, including the Coral Dev Board and conducted tests to benchmark the inference performance [4, 8]. He reported that the Coral Dev Board performed 10 time faster than Movidius NCS, 3.5 time faster than Nvidia Jetson Nano (TF-TRT) and 31 time faster than Raspberry Pi for MobileNetV2 SSD model.

SparkFun Edge

SparkFun Edge is a real-time audio analysis device, which runs machine learning inference to detect a keyword, for example, "yes" and responds accordingly²⁹. Developed by Google, Ambiq, and SparkFun collaboratively, it's used for voice and gesture recognition at the edge³⁰ without the help of remote

services. It has a 32-bit ARM Cortex-M4F 48MHz processor with 96MHz burst mode, extremely low-power usage, 384KB SRAM, 1MB flash memory and a dedicated BLE 5 Bluetooth processor.

It also has two built-in microphones, a 3-axis accelerometer, a camera connector, and other input/output connectors. This device can run for 10 days with a CR2032 coin cell battery. Ambiq Apollo3 is a Software Development Kit is available for building AI applications with the SparkFun Edge.

Nvidia Jetson

The Nvidia Jetson is an embedded computing board that can process complex data in real-time. The Jetson AGX Xavier can operate with a 30W power supply and perform like a GPU workstation for edge AI applications.

Jetson TX1 and TX2 are embedded AI computing devices powered by Nvidia Jetson. These two small, but powerful, computers are ideal for implementing an intelligent system on edge devices such as smart security cameras, drones, robots, and portable medical devices.

JetPack is an SDK for building AI applications with the Jetson. This SDK includes TensorRT, cuDNN, Nvidia DIGITS Workflow, ISP Support, Camera imaging, Video CODEC, Nvidia VisionWorks, OpenCV, Nvidia CUDA, and CUDA Library tools for supporting ML. It is also compatible with the Robot Operating System (ROS³¹).

Intel Movidius

Intel Movidius³² is a vision processing unit which can accelerate deep neural network inferences in resource-constrained devices such as intelligent security cameras or drones. This chip can run custom vision, imaging, and deep neural network workloads on edge devices without a connection to the network or any cloud backend.

This chip can be deployed on a robot placed in rescue operations in disaster-affected areas. The rescue robot can make some life-saving decisions without human help. It can run real-time deep neural networks by performing 100 gigaflops within a 1W power envelope. Movidius Myriad 2 is the second generation vision processing unit (VPU) and Myriad X VPU is the most advanced VPU from Movidius to provide artificial intelligence solutions from drones and robotics to smart cameras and virtual reality³³. Intel also provides a Myriad Development Kit (MDK) which includes all necessary tools and APIs to implement ML on the chip.

ARM ML

The ARM ML processor [12] allows developers to

²⁶<https://cloud.google.com/edge-tpu/>

²⁷<https://coral.withgoogle.com/docs/edgetpu/faq/>

²⁸<https://coral.withgoogle.com/products/accelerator/>

²⁹<https://learn.sparkfun.com/tutorials/sparkfun-edge-hookup-guide/all>

³⁰<https://www.sparkfun.com/products/15170>

³¹<https://www.ros.org>

³²<https://www.movidius.com/>

³³<https://www.movidius.com/myriadx>

accelerate the performance of ML algorithms and deploys inference on edge devices. The ecosystem consists of the following:

- ARM NN³⁴, an inference engine that provides a translation layer that bridges the gap between existing Neural Network frameworks and ARM ML processor, and
- ARM Compute Library³⁵, an open source library containing functions optimized for ARM processors.

The ARM ML processor can run high-level neural network frameworks like TensorFlow Lite, Caffe, and ONNX. The ARM NN SDK has all the necessary tools to run neural networks on edge devices. This processor is designed for mobile phones, AR/VR, robotics, and medical instruments. Lai and Suda designed a set of efficient neural network kernels called CMSIS-NN³⁶ to maximize the performance of a neural network using limited memory and compute resources of an ARM Cortex-M processor [59].

RISC-V

The RISC-V [85] is an open instruction set architecture (ISA) and the GAP8 is a RISC-V architecture microprocessor that has been used for edge computing [101]. It has 9 cores capable of running 10 GOPS at the order of tens of mW. This 250 MHz processor is designed to accelerate CNNs for the edge computing and IoT market. Greenwaves has developed a tool, TF2GAP8, that automatically translates TensorFlow CNN applications to GAP8 source [102].

OpenMV Cam

OpenMV Cam³⁷ is a small, low-powered camera board. This board is built using an ARM Cortex-M7 processor to execute machine vision algorithms at 30 FPS. This processor can run at 216 MHz and has 512KB of RAM, 2 MB of flash, and 10 I/O pins. The main applications of this device are face detection, eye tracking, QR code detection/decoding, frame differencing, AprilTag tracking, and line detection [6].

BeagleBone AI

BeagleBone AI is a high-end board for developers building machine-learning and computer-vision applications [25]. This device is powered by an SoC – TI AM5729 dual core Cortex-A15 processor featuring 4 programmable real-time units, a dual core C66x digital-signal-processor, and 4 embedded-vision-engines core supported through a TIDL (Texas Instruments Deep Learning) machine learning API. It can perform image classification,

object detection, and semantic segmentation using TIDL.

ECM3531

ECM3531³⁸ is an high-efficiency ASIC based on the ARM Cortex-M3 and NXP Coolflux DSP processors for machine learning applications. The name of the processor of this ASIC is called Tensai, which can run TensorFlow or Caffe framework. This processor offer 30-fold power reduction in a specific CNN-based image classification.

SmartEdge Agile & Brainium

The SmartEdge Agile³⁹ device along with the accompanying Brainium software help build artificial intelligence models and deploy them on resource-constrained devices. This SmartEdge Agile device sits at the edge environment and uses Brainium’s zero-coding platform to deploy a trained intelligent model on the edge [7].

6.2 Software

SeeDot SeeDot is a programming language, developed by Microsoft researchers, to express machine learning inference algorithms and to control them at a mathematical-level [39]. Typically, most learning models are expressed in floating-point arithmetic, which are often expensive. Most resource-constrained devices do not support floating-point operations. To overcome this, SeeDot generates fixed-point code with only integer operations, which can be executed with just a few kilobytes of RAM. This helps SeeDot run a CNN on a resource-scarce microcontroller with no floating-point support. SeeDot-generated code for ML classification that uses Bonsai (§2.3) and ProtoNN (§2.2) is 2.4 to 11.9 times faster than floating-point microcontroller-based code. Also, SeeDot-generated code was 5.2-9.8 times faster than code generated by high-level synthesis tools on an FPGA-based implementation.

AWS IoT Greengrass AWS IoT Greengrass⁴⁰ is software which helps an edge device to run serverless AWS Lambda functions. It can be used to run ML inference on an edge device, filter device data, sync data and only transmit important information back to the cloud.

Azure IoT Edge The Azure IoT Edge⁴¹ is a platform that can be used to offload a large amount of work from the cloud to the edge devices. It has been used to deploy machine learning models on edge devices and cloud servers. Such workload migration reduces

³⁴<https://mlplatform.org/>

³⁵<https://www.arm.com/why-arm/technologies/compute-library>

³⁶https://github.com/ARM-software/CMSIS_5

³⁷<https://openmv.io/collections/cams/products/openmv-cam-m7>

³⁸<https://etacompute.com/products/>

³⁹<https://www.avnet.com/wps/portal/us/solutions/iot/building-blocks/smartedge-agile/>

⁴⁰<https://aws.amazon.com/greengrass/>

⁴¹<https://azure.microsoft.com/en-gb/services/iot-edge/>

data communication latency and operates reliably even in offline periods.

Zephyr OS Zephyr⁴² is a real-time operating system with a small-footprint kernel, specially designed for resource-constrained devices. This OS supports multiple architectures, such as Intel x86, ARM Cortex-M, RISC-V 32, NIOS II, ARC, and Tensilica Xtensa. Zephyr OS is a collaborative project, which is hosted by the Linux Foundation under the Apache 2.0 license.

7 Challenges and Future Directions

In order to fully utilize the benefits offered by edge computing, researchers have to address a number of issues that significantly inhibit the emergence of edge-based machine learning applications.

Cost of training DL models. Since training a deep learning model on edge devices is difficult, most existing machine learning systems use the cloud for training. Some attempts have been made to train models on edge devices (such as by using model pruning and model quantization) but edge-trained models often have lower accuracy, and therefore designing power-efficient algorithms for training neural networks on edge devices is an active research area. Training deep neural networks on resource-constrained devices is difficult due to the memory and computational requirements of such deep learning models. There remains continued interest in developing new methods and frameworks that map sections of a deep learning model onto the distributed computing hierarchy and exploring the use of specialized hardware (such as ARM ML processors) to speed up deep learning training and inference.

Heterogeneous Data. The different types of IoT sensors available on the market often create a heterogeneous environment in edge-based intelligent systems. To deal with the diversity of data, ML algorithms need to learn using types of data that have different features like image, text, sound, and motion. Multimodal deep learning is used to learn features over multiple modalities (e.g., audio and video [77]). Even though these algorithms seem to be a potentially attractive, in practice, it is difficult *a)* to design appropriate layers for feature fusion with heterogeneous data and *b)* to deploy models on resource-scarce devices.

Challenges in Distributed ML. To handle the enormous amount of data produced by IoT sensors, researchers have designed an edge-based distributed learning algorithm (§3.2) over distributed computing hierarchies. The hierarchy consists of the cloud servers,

the edge devices, and end-devices like IoT sensors. Such algorithms provide acceptable accuracy with datasets that are naturally distributed, for example, fraud detection and market analysis. However, the influence of the heterogeneity of data on the accuracy of a distributed model is an open research issue [87].

Using unlabeled data and building new datasets. An important characteristic of deep learning algorithms is their ability to train using unlabeled input data. The availability of large amount of unlabeled data generated by edge and end devices is a very good source for building new datasets but advanced algorithms require to create new datasets with less noisy labels.

Augmentation of edge and other sensor data to enhance deep learning performance is another research opportunity. Data augmentation uses a small amount of data (transferred from sensor to edge devices or from edge devices to cloud servers) to generate new data. Augmentation helps ML models avoid overfitting issues by generating enough new training data [79]. However, the performance of data augmentation by edge devices or cloud servers needs to be evaluated before its use in a learning model, especially for small datasets.

Increasing Model Accuracy. It has been shown that edge-based deep learning techniques can be used in the health care sector (§4.8), but safety-critical areas such as this need the intelligent systems to have high accuracy before deploying in the health care sector. Another barrier that remains is the unavailability of annotated datasets.

Augmented Cognition. Researchers are now exploring how deep learning and edge computing can be used for augmenting human cognition to create adaptive human-machine collaboration by quickly giving expert guidance to the human for unfamiliar tasks and for amplifying the human’s memory capabilities [94]. Such techniques promise to transform the way humans with low cognitive abilities can perform both routine and complex tasks, but questions pertaining to security, privacy and ethics need to be addressed before such systems are deployed.

8 Conclusion

Edge-based machine learning is a fast-growing research area with numerous challenges and opportunities. Using edge devices for machine learning has been shown to improve not only the privacy and security of the user but also system response times.

This article provides a comprehensive overview of techniques and applications pertaining to the deployment of machine learning systems at the network edge. It highlights several new machine learning architectures that have been designed specifically for resource-scarce edge computing devices and reviews important applications of edge-based ML systems. Widely adopted frameworks es-

⁴²<https://www.zephyrproject.org/what-is-zephyr/>

sential for developing edge-based deep learning architectures as well as the resource-contained devices that have been used to deploy these models are described. Finally, the main challenges to deploying machine learning systems on the edge are listed and several directions for future work are briefly described.

References

- [1] Adafruit TensorFlow Lite Micro Speech. https://github.com/adafruit/Adafruit_TFLite_Micro_Speech. Accessed: 2019-07-11.
- [2] IBM scientists team with The Weather Company to bring edge computing to life. <https://www.ibm.com/blogs/research/2017/02/bringing-edge-computing-to-life/>, February 2017. Accessed: 2019-06-22.
- [3] M. Ali, A. Anjum, M. U. Yaseen, A. R. Zamani, D. Balouek-Thomert, O. Rana, and M. Parashar. Edge Enhanced Deep Learning System for Large-Scale Video Stream Analytics. In *2018 IEEE 2nd International Conference on Fog and Edge Computing (ICFEC)*, pages 1–10, May 2018.
- [4] Alasdair Allan. Benchmarking Edge Computing. <https://medium.com/@aallan/benchmarking-edge-computing-ce3f13942245>. Accessed: 2019-07-11.
- [5] Alasdair Allan. Benchmarking the Xnor AI2GO Platform on the Raspberry Pi. <https://blog.hackster.io/benchmarking-the-xnor-ai2go-platform-on-the-raspberry-pi-628a82af8aea>. Accessed: 2019-07-16.
- [6] Alasdair Allan. Deep Learning at the Edge on an Arm Cortex-Powered Camera Board. <https://blog.hackster.io/deep-learning-at-the-edge-on-an-arm-cortex-powered-camera-board-3ca16eb60ef7>. Accessed: 2019-07-10.
- [7] Alasdair Allan. Hands-On with the Smart-Edge Agile. <https://blog.hackster.io/hands-on-with-the-smartedge-agile-b7b7f02b5d4b>. Accessed: 2019-07-11.
- [8] Alasdair Allan. Measuring Machine Learning. <https://towardsdatascience.com/measuring-machine-learning-945a47bd3750>. Accessed: 2019-07-11.
- [9] G. Ananthanarayanan, P. Bahl, P. Bodík, K. Chintalapudi, M. Philipose, L. Ravindranath, and S. Sinha. Real-Time Video Analytics: The Killer App for Edge Computing. *Computer*, 50(10):58–67, 2017.
- [10] Ganesh Ananthanarayanan, Victor Bahl, Landon Cox, Alex Crown, Shadi Noghahi, and Yuanchao Shu. Video Analytics - Killer App for Edge Computing. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '19*, pages 695–696, New York, NY, USA, 2019. ACM.
- [11] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. "Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine". In *"Ambient Assisted Living and Home Care"*, pages 216–223. Springer, 2012.
- [12] ARM Limited. Machine Learning ARM ML Processor. <https://www.arm.com/products/silicon-ip-cpu/machine-learning/arm-ml-processor>. Accessed: 2019-06-11.
- [13] Asha Barbaschow. VMware looking towards IoT and the edge. <https://www.zdnet.com/article/vmware-looking-towards-iot-and-the-edge/>, 2018. Accessed: 2019-06-22.
- [14] B. Barry, C. Brick, F. Connor, D. Donohoe, D. Moloney, R. Richmond, M. O’Riordan, and V. Toma. Always-on Vision Processing Unit for Mobile Applications. *IEEE Micro*, 35(2):56–66, March 2015.
- [15] Sourav Bhattacharya and Nicholas D. Lane. Sparsification and Separation of Deep Learning Layers for Constrained Resource Inference on Wearables. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM, SenSys '16*, pages 176–189, New York, NY, USA, 2016. ACM.
- [16] Andrew A. Borkowski, Catherine P. Wilson, Steven A. Borkowski, Lauren A. Deland, and Stephen M. Mastorides. Using Apple Machine Learning Algorithms to Detect and Subclassify Non-Small Cell Lung Cancer. <https://arxiv.org/abs/1808.08230>, January 2019.
- [17] Brandon Butler. What is edge computing and how it’s changing the network. *Network World*, 2017. Accessed: July 21, 2019.
- [18] S. Cass. Taking AI to the edge: Google’s TPU now comes in a maker-friendly package. *IEEE Spectrum*, 56(5):16–17, May 2019.
- [19] P. S. Chandakkar, Y. Li, P. L. K. Ding, and B. Li. Strategies for Re-Training a Pruned Neural Network in an Edge Computing Paradigm. In *2017 IEEE International Conference on Edge Computing (EDGE)*, pages 244–247, June 2017.
- [20] X. Chang, W. Li, C. Xia, J. Ma, J. Cao, S. U. Khan, and A. Y. Zomaya. From Insight to Impact: Building a Sustainable Edge Computing Platform for Smart Homes. In *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 928–936, December 2018.
- [21] G. Chen, C. Parada, and G. Heigold. Small-footprint keyword spotting using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4087–4091, May 2014.

- [22] G. Chen, C. Parada, and T. N. Sainath. Query-by-example keyword spotting using long short-term memory networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5236–5240, April 2015.
- [23] Y. Chen, A. Wu, M. A. Bayoumi, and F. Koushanfar. Editorial Low-Power, Intelligent, and Secure Solutions for Realization of Internet of Things. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 3(1):1–4, March 2013.
- [24] Sandeep Chinchali, Apoorva Sharma, James Harrison, Amine Elhafsi, Daniel Kang, Evgenya Pergament, Eyal Cidon, Sachin Katti, and Marco Pavone. Network offloading policies for cloud robotics: a learning-based approach. <http://arxiv.org/abs/1902.05703>, 2019.
- [25] Christine Long. BeagleBone AI Makes a Sneak Preview. <https://beagleboard.org/blog/2019-05-16-beaglebone-ai-preview>, May 2019. Last Accessed: July 30, 2019.
- [26] N. Curukoglu and B. M. Ozyildirim. Deep Learning on Mobile Systems. In *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–4, October 2018.
- [27] A. Das, M. Degeling, X. Wang, J. Wang, N. Sadeh, and M. Satyanarayanan. Assisting Users in a World Full of Cameras: A Privacy-Aware Infrastructure for Computer Vision Applications. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1387–1396, July 2017.
- [28] Rustem Dautov, Salvatore Distefano, Dario Bruno, Francesco Longo, Giovanni Merlino, Antonio Puliafito, and Rajkumar Buyya. Metropolitan intelligent surveillance systems for urban areas by harnessing IoT and edge computing paradigms. *Softw., Pract. Exper.*, 48:1475–1492, 2018.
- [29] L. Ding and M. Ben Salem. A Novel Architecture for Automatic Document Classification for Effective Security in Edge Computing Environments. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 416–420, October 2018.
- [30] Utsav Drolia, Katherine Guo, and Priya Narasimhan. Precog: p refetching for image recognition applications at the edge. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, SEC ’17, pages 17:1–17:13, New York, NY, USA, 2017. ACM.
- [31] SparkFun Electronics. SparkFun Edge Hookup Guide. <https://learn.sparkfun.com/tutorials/sparkfun-edge-hookup-guide/all>, 2018. Accessed: 2019-06-08.
- [32] A. R. Elias, N. Golubovic, C. Krintz, and R. Wolski. Where’s the Bear? - Automating Wildlife Image Processing Using IoT and Edge Cloud Systems. In *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 247–258, April 2017.
- [33] E. Ezra Tsur, E. Madar, and N. Danan. Code Generation of Graph-Based Vision Processing for Multiple CUDA Cores SoC Jetson TX. In *2018 IEEE 12th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoc)*, pages 1–7, September 2018.
- [34] Z. Feng, S. George, J. Harkes, P. Pillai, R. Klatzky, and M. Satyanarayanan. Edge-Based Discovery of Training Data for Machine Learning. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 145–158, October 2018.
- [35] E. Flamand, D. Rossi, F. Conti, I. Loi, A. Pullini, F. Rotenberg, and L. Benini. GAP-8: A RISC-V SoC for AI at the Edge of the IoT. In *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pages 1–4, July 2018.
- [36] David Floyer. The Vital Role of Edge Computing in the Internet of Things. <https://wikibon.com/the-vital-role-of-edge-computing-in-the-internet-of-things>, 2015.
- [37] The Linux Foundation. The Open Platform for the IoT Edge. <https://www.edgexfoundry.org>. Accessed: 2019-06-19.
- [38] A. Ghoneim, G. Muhammad, S. U. Amin, and B. Gupta. Medical Image Forgery Detection for Smart Healthcare. *IEEE Communications Magazine*, 56(4):33–37, April 2018.
- [39] Sridhar Gopinath, Nikhil Ghanathe, Vivek Shashadri, and Rahul Sharma. Compiling KB-sized Machine Learning Models to Tiny IoT Devices. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, pages 79–95, New York, NY, USA, 2019. ACM.
- [40] Chirag Gupta, Arun Sai Suggala, Ankit Goyal, Harsha Vardhan Simhadri, Bhargavi Paranjape, Ashish Kumar, Saurabh Goyal, Raghavendra Udapa, Manik Varma, and Prateek Jain. ProtoNN: Compressed and Accurate kNN for Resource-scarce Devices. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1331–1340, International Convention Centre, Sydney, Australia, 08 2017. PMLR.
- [41] Richard Harper. *Inside the Smart House*. Springer-Verlag, Berlin, Heidelberg, 2003.
- [42] Evan Hennis, Mark Deoust, and Billy Lamberta. TensorFlow Lite Speech Command Recognition Android Demo. https://github.com/tensorflow/examples/tree/master/lite/examples/speech_commands/android. Last Accessed: July 21, 2019.
- [43] Jacob Hochstetler, Rahul Padidela, Qing Chen, Qiang Yang, and Songnian Fu. Embedded

- Deep Learning for Vehicular Edge Computing. *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 341–343, 2018.
- [44] 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. <https://arxiv.org/abs/1704.04861>, 2017.
- [45] Kevin Hsieh, Ganesh Ananthanarayanan, Peter Bodik, Shivaram Venkataraman, Paramvir Bahl, Matthai Philipose, Phillip B. Gibbons, and Onur Mutlu. **Focus: Querying Large Video Datasets with Low Latency and Low Cost.** In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 269–286, Carlsbad, CA, 2018. USENIX Association.
- [46] C. C. . Hsu, M. Y. . Wang, H. C. H. Shen, R. H. . Chiang, and C. H. P. Wen. FallCare+: An IoT surveillance system for fall detection. In *2017 International Conference on Applied System Innovation (ICASI)*, pages 921–922, May 2017.
- [47] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in neural information processing systems*, pages 4107–4115, 2016.
- [48] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. 2016.
- [49] Michaela Iorga, Larry B. Feldman, Robert Barton, Michael Martin, Nedim S. Goren, and Charif Mahmoudi. Fog Computing Conceptual Model. <http://doi.org/10.6028/NIST.SP.500-325>, 2018. Special Publication (NIST SP) - 500-325.
- [50] D. Isereau, C. Capraro, E. Cote, M. Barnell, and C. Raymond. Utilizing high-performance embedded computing, agile condor, for intelligent processing: An artificial intelligence platform for remotely piloted aircraft. In *2017 Intelligent Systems Conference (IntelliSys)*, pages 1155–1159, Sep. 2017.
- [51] Kaya Ismail. Edge Computing vs. Fog Computing: What’s the Difference? <https://www.cmswire.com/information-management/edge-computing-vs-fog-computing-whats-the-difference/>, 2018. Accessed: 2019-06-5.
- [52] Daniel Kang, Peter Bailis, and Matei Zaharia. Challenges and Opportunities in DNN-Based Video Analytics: A Demonstration of the BlazeIt Video Query Engine. In *CIDR 2019, 9th Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 13-16, 2019, Online Proceedings*, 2019.
- [53] Duseok Kang, Euseok Kim, Inpyo Bae, Bernhard Egger, and Soonhoi Ha. C-GOOD: C-code Generation Framework for Optimized On-device Deep Learning. In *Proceedings of the International Conference on Computer-Aided Design, ICCAD ’18*, pages 105:1–105:8, New York, NY, USA, 2018. ACM.
- [54] Gorkem Kar, Shubham Jain, Marco Gruteser, Fan Bai, and Ramesh Govindan. Real-time Traffic Estimation at Vehicular Edge Nodes. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing, SEC ’17*, pages 3:1–3:13, New York, NY, USA, 2017. ACM.
- [55] Yoshiyuki Kawano and Keiji Yanai. Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools Appl.*, 74(14):5263–5287, July 2015.
- [56] Ashish Kumar, Saurabh Goyal, and Manik Varma. Resource-efficient Machine Learning in 2 KB RAM for the Internet of Things. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1935–1944, International Convention Centre, Sydney, Australia, 08 2017. PMLR.
- [57] Aditya Kusupati, Manish Singh, Kush Bhatia, Ashish Kumar, Prateek Jain, and Manik Varma. FastGRNN: A Fast, Accurate, Stable and Tiny Kilobyte Sized Gated Recurrent Neural Network. In *Advances in Neural Information Processing Systems 31*, pages 9017–9028. Curran Associates, Inc., 2018.
- [58] Gant Laborde. Perf Machine Learning on Rasp Pi. <https://medium.com/free-code-camp/perf-machine-learning-on-rasp-pi-51101d03dba2>, 2019. Accessed: 2019-07-11.
- [59] Liangzhen Lai and Naveen Suda. Enabling Deep Learning at the IoT Edge. In *Proceedings of the International Conference on Computer-Aided Design, ICCAD ’18*, pages 135:1–135:6, New York, NY, USA, 2018. ACM.
- [60] N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, L. Jiao, L. Qendro, and F. Kawsar. DeepX: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 1–12, April 2016.
- [61] S. Lee, K. Son, H. Kim, and J. Park. Car plate recognition based on CNN using embedded system with GPU. In *2017 10th International Conference on Human System Interactions (HSI)*, pages 239–241, July 2017.
- [62] D. Li, T. Salonidis, N. V. Desai, and M. C. Chuah. DeepCham: Collaborative Edge-Mediated Adaptive Deep Learning for Mobile Object Recognition. In *2016 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 64–76, October 2016.
- [63] En Li, Zhi Zhou, and Xu Chen. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In *Proceedings*

- of the 2018 Workshop on Mobile Edge Communications, MECOMM'18, pages 31–36, New York, NY, USA, 2018. ACM.
- [64] H. Li, K. Ota, and M. Dong. Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing. *IEEE Network*, 32(1):96–101, January 2018.
 - [65] Zhong Qiu Lin, Audrey G. Chung, and Alexander William Wong. EdgeSpeechNets: Highly Efficient Deep Neural Networks for Speech Recognition on the Edge. <https://arxiv.org/abs/1810.08559v2>, 2018.
 - [66] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, M. Yunsheng, S. Chen, and P. Hou. A New Deep Learning-Based Food Recognition System for Dietary Assessment on An Edge Computing Service Infrastructure. *IEEE Transactions on Services Computing*, 11(2):249–261, March 2018.
 - [67] L. Liu, X. Zhang, M. Qiao, and W. Shi. SafeShareRide: Edge-Based Attack Detection in Ridesharing Services. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 17–29, October 2018.
 - [68] Qiang Liu, Siqi Huang, and Tao Han. Fast and Accurate Object Analysis at the Edge for Mobile Augmented Reality: Demo. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing, SEC '17*, pages 33:1–33:2, New York, NY, USA, 2017. ACM.
 - [69] Mohammad Saeid Mahdavejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P. Sheth. Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks*, 4(3):161 – 175, 2018.
 - [70] James Manyika, Michael Chui, Peter Bisson, Jonathan Woetzel, Richard Dobbs, Jacques Bughin, and Dan Aharon. The Internet of Things: Mapping the Value Behind the Hype. Technical report, McKinsey and Company, 6 2015.
 - [71] J. Mao, X. Chen, K. W. Nixon, C. Krieger, and Y. Chen. MoDNN: Local distributed mobile computing system for Deep Neural Network. In *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, pages 1396–1401, March 2017.
 - [72] Y. Mao, S. Yi, Q. Li, J. Feng, F. Xu, and S. Zhong. Learning from Differentially Private Neural Activations with Edge Computing. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 90–102, October 2018.
 - [73] C. Marantos, N. Karavalakis, V. Leon, V. Tsoutsouras, K. Pekmestzi, and D. Soudris. Efficient support vector machines implementation on Intel/Movidius Myriad 2. In *2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pages 1–4, May 2018.
 - [74] Bradley McDanel, Surat Teerapittayanon, and H. T. Kung. Embedded Binarized Neural Networks. <https://arxiv.org/abs/1709.02260>, 2017.
 - [75] S. A. Miraftebadeh, P. Rad, K. R. Choo, and M. Jamshidi. A Privacy-Aware Architecture at the Edge for Autonomous Real-Time Identity Reidentification in Crowds. *IEEE Internet of Things Journal*, 5(4):2936–2946, August 2018.
 - [76] Pedro Navarro Lorente, Carlos Fernandez, Raul Borraz, and Diego Alonso. A Machine Learning Approach to Pedestrian Detection for Autonomous Vehicles Using High-Definition 3D Range Data. *Sensors*, 17:18, 12 2016.
 - [77] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*, pages 689–696, USA, 2011. Omnipress.
 - [78] Seyed Yahya Nikouei, Yu Chen, Sejun Song, Ronghua Xu, Baek-Young Choi, and Timothy R. Faughnan. Intelligent Surveillance as an Edge Network Service: from Harr-Cascade, SVM to a Lightweight CNN. <https://arxiv.org/abs/1805.00331v1>, 2018.
 - [79] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-garadi, and Uzoma Rita Alo. "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges". *Expert Systems with Applications*, 105:233 – 261, 2018.
 - [80] Samuel S. Ogden and Tian Guo. MODI: Mobile Deep Inference Made Efficient by Edge Computing. In *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*, Boston, MA, 2018. USENIX Association.
 - [81] S. A. Osia, A. S. Shamsabadi, A. Taheri, H. R. Rabiee, and H. Haddadi. Private and Scalable Personal Data Analytics Using Hybrid Edge-to-Cloud Deep Learning. *Computer*, 51(5):42–49, May 2018.
 - [82] Anand Oswal. Time to Get Serious About Edge Computing. <https://blogs.cisco.com/enterprise/time-to-get-serious-about-edge-computing>, 2018. Accessed: 2019-06-22.
 - [83] H. Haddad Pajouh, R. Javidan, R. Khayami, D. Ali, and K. R. Choo. A Two-layer Dimension Reduction and Two-tier Classification Model for Anomaly-Based Intrusion Detection in IoT Backbone Networks. *IEEE Transactions on Emerging Topics in Computing*, pages 1–1, 2018.
 - [84] Donghyun Park, Seulgi Kim, Yelin An, and Jae-Yoon Jung. LiReD: A Light-Weight Real-Time Fault Detection System for Edge Computing Using LSTM Recurrent Neural Networks. *Sensors*, 18(7), 2018.

- [85] David Patterson and Andrew Waterman. *The RISC-V Reader: An Open Architecture Atlas*. Strawberry Canyon LLC, 2017.
- [86] Richard Perez, Elke Lorenz, Sophie Pelland, Mark Beauharnois, Glenn Van Knowe, Karl Hemker, Detlev Heinemann, Jan Remund, Stefan C. Müller, Wolfgang Traunmüller, Gerald Steinmauer, David Pozo, Jose A. Ruiz-Arias, Vicente Lara-Fanego, Lourdes Ramirez-Santigosa, Martin Gaston-Romero, and Luis M. Pomares. Comparison of numerical weather prediction solar irradiance forecasts in the us, canada and europe. *Solar Energy*, 94:305 – 326, 2013.
- [87] Diego Peteiro-Barral and Bertha Guijarro-Berdiñas. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence*, 2(1):1–11, Mar 2013.
- [88] K. Pradeep, K. Kamalavasan, R. Natheesan, and A. Pasqual. EdgeNet: SqueezeNet like Convolution Neural Network on Embedded FPGA. In *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pages 81–84, December 2018.
- [89] Xuan Qi and Chen Liu. Enabling Deep Learning on IoT Edge: Approaches and Evaluation. *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 367–372, 2018.
- [90] Joseph Redmon and Ali Farhadi. YoloV3: An incremental improvement. <https://arxiv.org/abs/1804.02767>, 2018.
- [91] R. Rothe, R. Timofte, and L. V. Gool. DEX: Deep EXpectation of Apparent Age from a Single Image. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 252–257, December 2015.
- [92] K. Rungsuptaweekoon, V. Visoottiviseth, and R. Takano. Evaluating the power efficiency of deep learning inference on embedded GPU systems. In *2017 2nd International Conference on Information Technology (INCIT)*, pages 1–5, November 2017.
- [93] M. Satyanarayanan. The Emergence of Edge Computing. *Computer*, 50(1):30–39, January 2017.
- [94] M. Satyanarayanan and N. Davies. Augmenting Cognition Through Edge Computing. *Computer*, 52(7):37–46, July 2019.
- [95] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal*, 3:637–646, 2016.
- [96] Laurent Sifre and Stephane Mallat. Rigid-Motion Scattering for Texture Classification. <https://arxiv.org/abs/1403.1687>, 2014.
- [97] Flavio Souza, Diego Couto de Las Casas, Viniçius Flores Zambaldi, SunBum Youn, Meeyoung Cha, Daniele Quercia, and Virgilio A. F. Almeida. Dawn of the Selfie Era: The Whos, Wheres, and Hows of Selfies on Instagram. <https://arxiv.org/abs/1510.05700>, 2015.
- [98] B. Tang, Z. Chen, G. Heffernan, S. Pei, T. Wei, H. He, and Q. Yang. Incorporating Intelligence in Fog Computing for Big Data Analysis in Smart Cities. *IEEE Transactions on Industrial Informatics*, 13(5):2140–2150, October 2017.
- [99] Zeyi Tao and Qun Li. **eSGD: Communication Efficient Distributed Deep Learning on the Edge**. In *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*, Boston, MA, 2018. USENIX Association.
- [100] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the kdd cup 99 data set. In *Proceedings of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications, CISDA'09*, pages 53–58, Piscataway, NJ, USA, 2009. IEEE Press.
- [101] GreenWaves Technologies. GAP8 - GreenWaves. <https://en.wikichip.org/wiki/greenwaves/gap8>, 2018. Accessed: 2019-05-23.
- [102] GreenWaves Technologies. GAP8 Tensor-Flow to GAP8 Bridge Manual. <https://greenwaves-technologies.com/manuals/BUILD/TF2GAP8/html/index.html>, 2019. Accessed: 2019-05-23.
- [103] S. Teerapittayanon, B. McDanel, and H. T. Kung. Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 328–339, June 2017.
- [104] Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. BranchyNet: Fast inference via early exiting from deep neural networks. *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469, 2016.
- [105] Mohit Thakkar. "Custom Core ML Models Using Create ML", pages 95–138. Apress, Berkeley, CA, 2019.
- [106] Shreshth Tuli, Nipam Basumatary, and Rajkumar Buyya. Edgelens: Deep learning based object detection in integrated iot, fog and cloud computing environments. <https://arxiv.org/abs/1906.11056>, 2019.
- [107] C. Wang, L. Gong, Q. Yu, X. Li, Y. Xie, and X. Zhou. DLAU: A Scalable Deep Learning Accelerator Unit on FPGA. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 36(3):513–517, March 2017.
- [108] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren. Interactive Medical Image Segmentation Using Deep Learning With

- Image-Specific Fine Tuning. *IEEE Transactions on Medical Imaging*, 37(7):1562–1573, July 2018.
- [109] J. Wang, Z. Feng, Z. Chen, S. George, M. Bala, P. Pillai, S. Yang, and M. Satyanarayanan. Bandwidth-Efficient Live Video Analytics for Drones Via Edge Computing. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 159–173, October 2018.
- [110] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin S. Chan. When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning. *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 63–71, 2018.
- [111] Sally Ward-Foxton. AI at the Very, Very Edge. https://www.eetimes.com/document.asp?doc_id=1334918, 2019. Accessed: 2019-07-30.
- [112] Pete Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. <https://arxiv.org/abs/1804.03209>, 2018.
- [113] Matt Welsh. True AI on a Raspberry Pi, with no extra hardware. <https://medium.com/@mdwdotla/true-ai-on-a-raspberry-pi-with-no-extra-hardware-dcbbff12d068>, 2019. Accessed: 2019-07-11.
- [114] R. Xu, S. Y. Nikouei, Y. Chen, A. Polunchenko, S. Song, C. Deng, and T. R. Faughnan. Real-Time Human Objects Tracking for Smart Surveillance at the Edge. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2018.
- [115] Zhuangdi Xu, Harshit Gupta, and Umakishore Ramachandran. STTR: A System for Tracking All Vehicles All the Time At the Edge of the Network. In *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems*, DEBS '18, pages 124–135, New York, NY, USA, 2018. ACM.
- [116] Shuochao Yao, Yiran Zhao, Aston Zhang, Lu Su, and Tarek F. Abdelzaher. Compressing Deep Neural Network Structures for Sensing Systems with a Compressor-Critic Framework. <https://arxiv.org/abs/1706.01215>, 2017.
- [117] X. Zhang, X. Zhou, M. Lin, and J. Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, June 2018.
- [118] Xingzhou Zhang, Yifan Wang, and Weisong Shi. pCAMP: Performance Comparison of Machine Learning Packages on the Edges. In *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*, Boston, MA, 2018. USENIX Association.
- [119] Zhuoran Zhao, Kamyar Mirzazad Barijough, and Andreas Gerstlauer. DeepThings: Distributed Adaptive Deep Learning Inference on Resource-Constrained IoT Edge Clusters. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, PP:1–1, 10 2018.