EDGE INTELLIGENCE: A BLEND OF AI WITH EDGE COMPUTING

A Survey on the technologies and applications of AI in Edge computing

Lavish Thomas, Department of Computing, Letterkenny Institute of Technology, Ireland.

Abstract— A survey on the current state of Edge Intelligence is carried out in order to compare and contrast the different techniques in play, to discuss different application areas and to identify existing challenges in the field. Initially, the paper gives an overview of the field "Artificial Intelligence" and the Deep neural network which is the core of the Edge Intelligence. Then, some popular Deep Neural Networks are explained. Furthermore, the concept of Edge Intelligence is explained in detail by presenting various performance indicators which can be used to measure the EI systems. Afterwards, various architecture and techniques used for Training and Inferencing are presented of which a comparison report is generated based on the presented key characteristics of the EI systems. Finally, various application areas of EI and the challenges identified in this field is presented.

***Index Terms*— Edge Computing, Artificial Intelligence, Industry 4.0, IoT, Internet of Things, Smart Devices**

I. INTRODUCTION

THIS paper is presented in order to provide a layout of the current landscape of Edge Computing and how Artificial Intelligence is spread across various sectors. Initially, the paper will present an introduction to the concept of 'Artificial Intelligence', a timeline of its evolution and its current state. Furthermore, it will introduce the area of Edge Computing with architecture and examples. Afterwards, the central topic of the paper will be introduced: Edge Intelligence. Edge intelligence is the idea of moving the Artificial Intelligence models near to the source of the data in order to reduce the bandwidth and latency required for decision-making. Furthermore, various performance criteria for an Edge Intelligence system is presented. Consequently, the paper will compare and contrast the techniques used based on the EI performance criteria such as accuracy, latency, communication, privacy, energy and memory. Afterwards, applications of the Edge Intelligence systems are discussed. Finally, challenges which are faced in the field of Edge Intelligence currently and foreseeable challenges as the technology moves forward in this field will be debated.

II. ARTIFICIAL INTELLIGENCE

"The science and engineering of making intelligent machines" – John McCarthy [1]

This is the modern definition of "Artificial intelligence" by John McCarthy, who coined the term. Throughout its course, several significant scientists have given various definitions for the AI. According to Russel et al [2], the AI can be defined in two dimensions, thought-process and the behaviour for the same. The relevance of the first dimension is that some machines can simulate human behaviour, especially in physical activity, for e.g. moving a log of wood. But it can be just as well programmed in conventional robotics, not necessarily a thought process like a human. The second dimension deals with the practicality of the solution. Unless a beneficial action can be derived from an agent in spite of thinking like human holds no value.

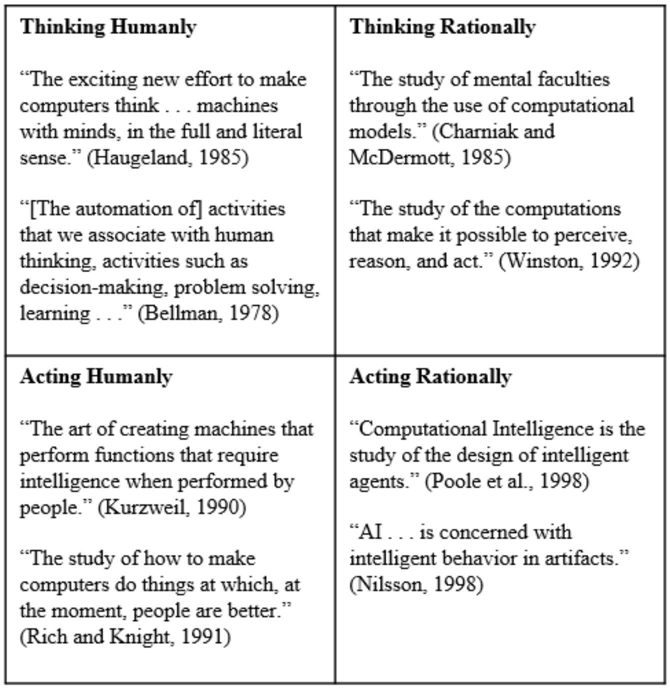
Another aspect which is needed to be considered while evaluating the AI is whether the thought process and behaviour was humane or rational. Human behaviour cannot be completely explained with logical reasoning all the time, but rationality is defined as what is the right action with the provided knowledge. Often used to achieve ideal performance. Table I [2] gives some popular definition and its categorical position based on the previously discussed dimensions.

The various categories have been tried and blend together to achieve the unified goal of the AI paradigm. Firstly, Acting Humanly is the concept stating that a program which can simulate human behaviour is considered as AI [2]. In other

This paper is presented as per the curriculum requirement for the module Artificial Intelligence under the guidance of Dr. Shagufta Henna, Lecturer, LYIT.

Lavish Thomas is the author of this paper, currently pursing Masters of Science in Letterkenny institute of Technology, Ireland, mailto:L0150445@student.lyit.ie

words, which pass the Turning test. Thinking humanly or cognitive modelling states that, for a machine to be considered Intelligent, it should be able to think like human [2]. This is a difficult task to exact as the cognitive science itself is not matured enough to give a clear explanation of how humans think. Thinking rationally is basically representing all knowledge and its relations in precise notations and in theory could solve any solvable problem represented in logical notation [2]. Finally, Acting Rationally is a notion in which an agent can perform autonomously by perceiving its environment, continuously learning, adapting and overcoming the challenges.



Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think . . . machines with minds, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole et al., 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Table. I Categorization of popular definitions of AI-based on based on whether the action and thought process was humane or rational [2].

A. History of AI

The term "Artificial Intelligence" was coined by John McCarthy in 1956 so as to distinguish the topic of 'intelligence of computer' from the popular topic of 'Cybernetics' [3]. Cybernetics was considering intelligence is always based on signals, controls and programmed logic between them. But Artificial intelligence is the study of logic and rational reasoning behind them. In the initial stages of AI, the enthusiasm was high the funding of the projects had increased, but the AI projects were not rewarding as expected by the investors especially with the military projects. The best example was the machine translation projects which performed pathetically when science papers were translated to and forth from the Russian Language. This led to financial withdrawals and high criticism as the expectation had gone too high. But, AI was limited by the technological limitation of the computational field. This period in the 1970s is called AI winter when most of the projects were halted or withdrawn.

The second wave of AI systems was born with the introduction of expert systems in the early '80s. These systems were designed to capture human reasoning skills into two basic structures, knowledge base in which "facts" are stored and "inferencing rules" for validating decisions, deriving new observations and to arrive at a reasonable conclusion. But after a few years of rapid advancements, the technology hit a saturation point because the knowledge available was too much and rules to infer them was complex than anticipated. This led to the decline in the interest in AI and eventually brought in the "second AI winter" from the late '80s till the early '90s.

The artificial intelligence once again started getting traction as the 21st-century technology empowered the AI computing necessities with cloud computing and satisfied the AI data appetite with data generated from IoT and wearable technologies. This new age of AI boom is considered as AI spring by several AI enthusiasts. Machine learning is the subset of AI field which got most of the attention with excellent application success, especially in the field of Data analytics. Deep-Learning which is a further subset of Machine Learning is the latest rampant where previously unapplied techniques are being experimented. But, the concept of AI is ubiquitous and possibly will spread into all fields where mankind has their hands-on.

B. Background on Artificial Intelligence

Even though artificial intelligence has not been fully realized as desired by the scientific community, with the current state itself AI holds a lot of advantages over the conventional software programs. Currently, most of the applications of AI lies in the data analytics field where an abundance of data is collected from various sources, cleansed and feed to machine learning algorithms to derive valuable insights to businesses. But most of these applications are still human interfered and ML is merely used as a supporting tool in the process. Nevertheless, the field of Artificial Intelligence is much vast than data analytics and can be applied to various fields. The power of Artificial Intelligence has been proved by developing gaming machines which could beat the human champions of Chess, Go etc. But then again, these are definite environment-based applications, the real challenges lie in creating the agents which can strive through non-definite solutions space.

C. Deep Learning

Even though there are several Machine Learning Techniques which are efficient and is used in the previous generation infrastructures. The new GPU based computing architecture has paved the way for Deep Neural Network-based Machine Learning methods. The DNN leverages the concept of Artificial Neural Network. Neural Networks are essentially an interconnected network of Artificial Neurons which are modelled after the Human Neurons Cells and their connecting behaviour. Neuron cells are connected to each other through with weighted connections with multiple inputs and single output as illustrated in Fig. 1.

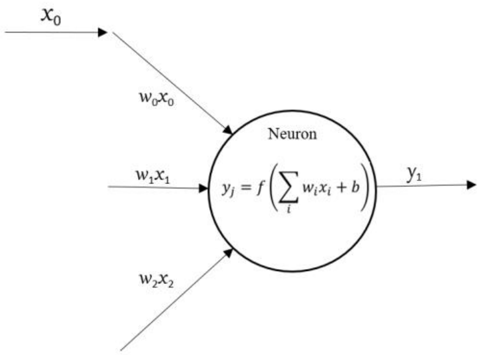


Fig. 1 Structure of Neuron [4]

Most implementations of the Neural Network consist of multiple layers of neurons interconnections through which the data is propagated to infer insights. The DL model usually follows a 3-layer approach, as illustrated in Fig. 2.

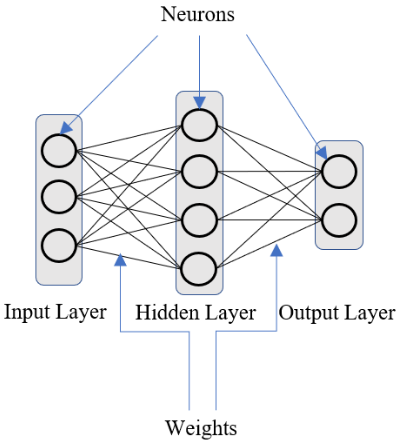


Fig. 2 The layers in a Deep Learning Model [4]

The layers work on weighted connections, each connection is given a random weight at first and adjusted same based on the corrections required on the output with respect to training data set output. The Input layer processes the input data and creates the outputs required based on the assigned weights in each connection. There can be more than one hidden layer (middle layers) based on applications. Advantage of the DNN is that it can process high-level features because of the abstract layers than other Machine Learning Models.

Some of the most used DNN's are [5]:

1) Multilayer Perceptron:

It is one of the simplest DNN models. The architecture is simple, where neurons in each layer are fully connected to the neurons in the adjacent layer. Due to this, the model is prone to overfitting of the data and redundancies in the data representation leading to inefficient memory usage [6].

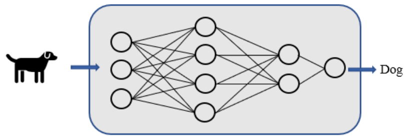


Fig. 3 Multi-Layer Neural Network [6]

2) Convolution Neural Network:

Using convolution operations, it can extract simple features. It is best suited for computer vision programs. It is also very helpful in deriving the hierarchal relations between the data [7].

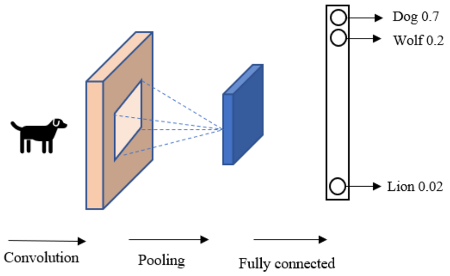


Fig. 4 Convolution Neural Network [7]

3) Recurrent Neural Network:

For sequential data feeding, this type of DNNs is used in order to resolve the issue of time-series [8]. Internal memory is used in each neuron cells in order to retain the previous state of the cell. These are used in natural language processing and language translation services.

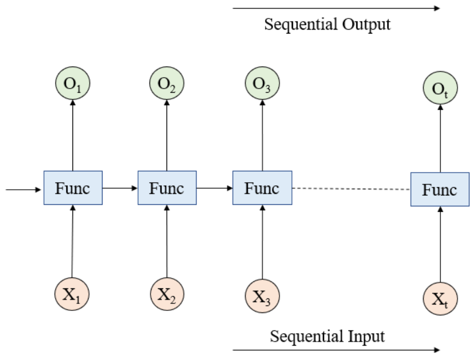


Fig. 5 Recurrent Neural Network [8]

III. Edge Computing

The cloud architecture was a breakthrough in the world of connected devices and data services as it provided lots of advantages over the limited computing resources in the local devices. Hence, the industries and information technology world enthusiastically adopted the cloud, and the spread of the cloud services was tremendous. But as time progressed, the issues in the cloud technology started to pop up, in which the prominent ones were high latency and high bandwidth consumption. These issues demanded some innovative techniques for mitigating these constraints. After a period of experimentation, by shifting the computing focus into the nodes in the network systems, a new concept called Edge Computing [9] was born. New territory for computing got generated between the data generating devices and cloud servers as explained in Fig. 6.

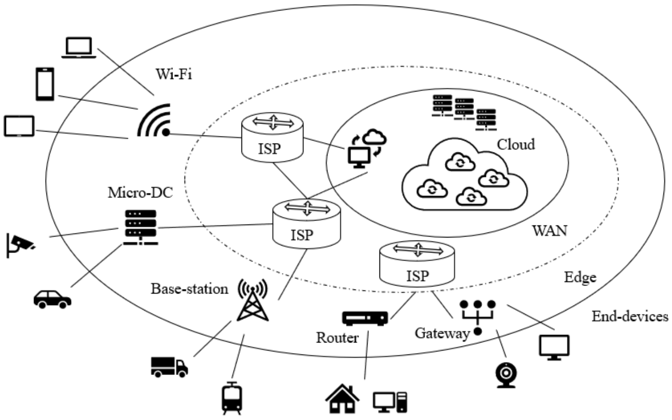


Fig. 6 Edge Computing space in the data network [9]

The Edge computing can be seen as intermediate processing hub in between the devices where the data is generated and the cloud where the data is stored and processed. In another version, the edge computing can be considered as an extended distributed computing system where the processing is carried out near to the local nodes generating the data.

IV. Edge Intelligence: A Blend of AI with Edge Computing

Thanks to the new age digital world, where an abundance of data is getting generated every second. According to Cisco [10], the global *amount* of data generated by the network edge will reach 850 Zettabytes by 2021. But data-centre capacity will have comparatively very fewer data, that is 20.6 Zettabytes. This contrast between source and capacity will inevitably lead to loss of data. This is where Edge Analytics and AI systems play a significant role. By deploying Intelligence to edge nodes in the network, there are two benefits: the reduction of network bandwidth needed to transport the data and reduction of latency between the data generation and reverse propagation of the insight (action required based on the decision). This new Intelligence deployment paradigm is called as Edge Intelligence where Artificial Intelligence programs come down from the clouds to the edge nodes in order provide more contextual meaning thus by reducing the network bandwidth usage and latency in communication.

Initially, in this chapter the Edge Intelligence level and performance indicators for an EI systems will be presented as [8] per Zhou et al. Furthermore, based on the level and performance, four aspects of EI will be evaluated: 1) Training Architecture, 2) Training Techniques, 3) Inferencing Architecture, 4) Inferencing Techniques. At the end of each subsection, a comparison table and analysis are provided to consolidate the learning advantages of each method.

A. *Edge Intelligence Levels*

Based on the availability of the data, modelling of the DNN and the inferencing, the levels of the intelligence is classified into 6 Levels by Zhou et al [8] as illustrated in Fig. 6.

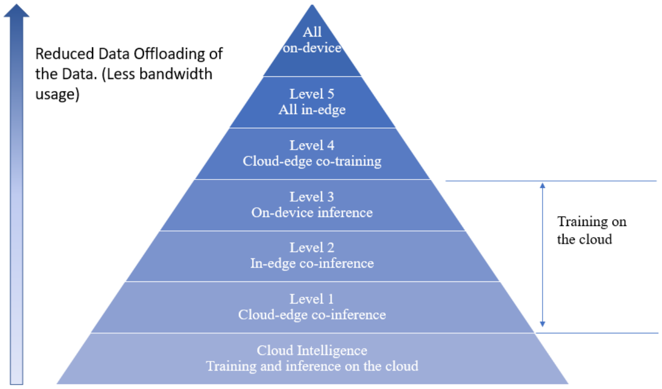

Reduced Data Offloading of the Data. (Less bandwidth usage)
All on-device
Level 5 All in-edge
Level 4 Cloud-edge co-training
Level 3 On-device inference
Level 2 In-edge co-inference
Level 1 Cloud-edge co-inference
Cloud Intelligence Training and inference on the cloud
Training on the cloud

Fig. 7 Levels of Edge Intelligence [8]

1) Cloud Intelligence:

In this conventional cloud-based AI system architecture, the DNN is trained in the cloud server and the information from the devices across the network is sent to the cloud for training and inferencing.

2) Level 1 - Cloud-Edge Co-Inference and Cloud Training:

Cloud Edge cooperation is achieved by a limited offloading of the data into the cloud from the end device. The cloud server act as a central data repository and training centre for the DNN model but inferencing is done on the edge nodes by creating a synergy between the cloud and edge programs.

3) Level 2 - In-Edge Co-Inference and Cloud Training:

Here also the cloud server acts as a central data repository and training centre for the DNN model but the inferencing is mainly done on the Edge Network.

4) Level 3 - On-Device Inference and Cloud Training

Still, the cloud server act as a central data repository and training centre for the DNN model but the inferencing is completely done on the end device. Replica of the DNN model on the device which is used for inferencing is updated regularly as changes happened to the model in the central server based on the new data sets and feedbacks.

5) Level 4 - Cloud-Edge Co-Training & Inference:

In this Level, the communication between the cloud and edge network becomes very crucial. The training and inferencing the responsibilities are shared between the cloud server and edge servers.

6) Level 5 - All In-Edge:

The dependency on the cloud is completely removed at this level. The complete cycle of data collection, training, model generation and inferencing is taken up by Edge Network.

7) All on-device:

This represents the purely distributed AI systems where all the responsibility which was carried out by Edge network in level 5 is offloaded to the end devices in the network. This creates a high contextual modelling (prone to overfitting) but reduces the latency to a greater extent. The model may lack the ability to infer new data as it lacks diversity in the training data.

B. *Performance Indicators*

The Edge computing systems vary in terms of their utility, purpose and capability of the network infrastructure. On a ground-breaking study of Edge Intelligence by Zhou et al. [8], the performance of Edge Intelligent systems can be measured along 6 axes. They are

1) *Latency:*

The time required from the generation of a request to the serving of the same is defined as latency. In the context of Edge Intelligence, this time includes the pre-processing, data transmission, training process, model inference, and post-processing[8].

2) *Accuracy:*

Accuracy of a Neural Network is base for any artificial intelligence systems, the same applies to the Edge Intelligence systems also. Since the Edge intelligence has a high degree of dependency on the communications, the speed in which the input and output have an effect on the performance as well. If the speed is low, the feedback may be not fast enough and if the speed is too high the modelling of DNN model may not be able to process is leading to loss of data points in time [8].

3) *Energy:*

Moving the processing towards the end devices raises another concern of power consumption. In most use cases, the end devices are supplied with power from the battery, and the battery life is limited. Hence, the model should be optimized in terms of processing power which can, in turn, lower the energy consumption for running model [8].

4) *Privacy:*

One of the popular use cases of Edge Intelligence resides in the area of personalized services and customized application features. This brings in the question of privacy and related security. The anonymization of data and its decoding is a great challenge for the harmony of the models in the Edge Computing systems and inherited to the Edge-based Artificial Intelligence systems [8].

5) *Communication overload:*

The communication bandwidth is always limited and efficient use of the same is a very important aspect of any application running in the network-based architecture. This applies to the Edge applications also and is considered as one of the performance indicators for the Edge Intelligent systems [8].

6) *Memory footprint:*

For any software applications, memory is a significant performance parameter. In the case DNN applications, conventionally there is high availability of dedicated GPU's and memory available in the cloud environment, which is convenient and usually the applications concentrate on perfecting the accuracy in trade for higher memory footprint. But when it comes to the Edge, the memory becomes a constraint and techniques of DNN have to be adapted to reduce the memory footprint.[8]

C. *Modelling Architectures*

One of the most important processes in any AI systems is modelling. The accuracy of the DNN is solely determined by how well the modelling was done and data used for the same. Due to the high amount of data generated by the end-devices and IoT networks, the modelling architecture for the Edge Intelligence plays a major role. Based on the various Edge Intelligence levels described for implementation of training and inferencing, the architectures for training the DNN can be broadly classified into three types [11].

1) *Centralized:*

Cloud Intelligence, Level 1, Level 2, Level 3 follow this architecture for the DNN training. The information from the devices across the systems is sent to the cloud as illustrated in Fig. 7. Further, based on this data, the DNN model is trained. In this architecture, the DNN models are trained in the cloud server [12].

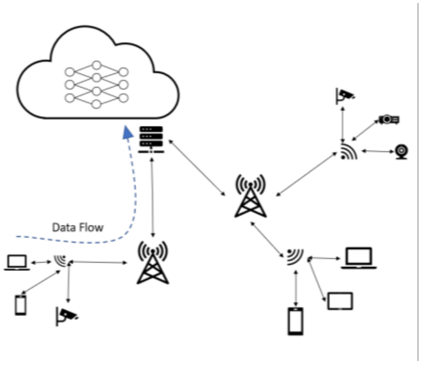


Fig. 8 Centralized architecture for EI training [12]

2) *Decentralized:*

The data is not collected at any central repository in the cloud rather the DNN models are trained locally using the data available in the device as outlined in Fig. 8. Even though the maturity if the local models will not be high individually, it saves bandwidth substantially. The omniscient model is derived by sharing the local DNN data and communicating the updates in the same regularly [11]. Another advantage is that the requirement for a cloud-based datacentre is eliminated. This architecture is used in the Level 5 systems of EI.

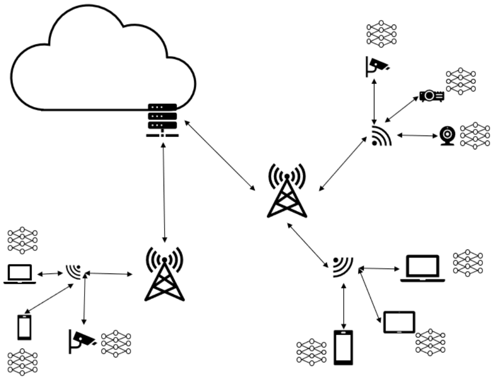


Fig. 9 Decentralized architecture for EI training [11]

3) Hybrid:

This architecture is an amalgamation of the previously discussed two architecture, where synergy is created between cloud models and edge servers bypassing regular updates between them about their own DNN models [11]. This architecture is used in order to realize the Level 4 and Level 5 EI intelligence systems. As illustrated in Fig. 10 the edge server does the heavy-lifting in contrast with the cloud-based centralized architecture.

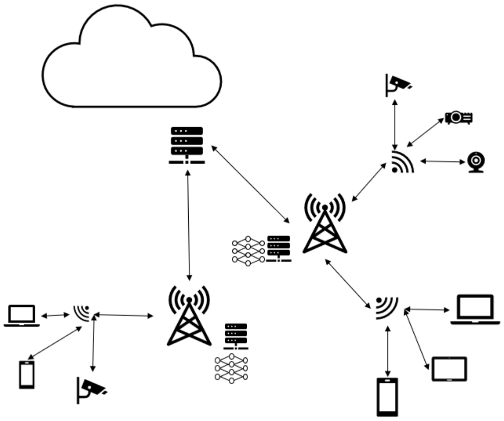


Fig. 10 Hybrid architecture for EI training [11]

The different architectures of modelling are used based upon application field and resource constraints in the deployment environment. Hybrid architecture holds a key advantage over others in terms of flexibility and performance, but the complexity of the system design and development is high. Table II presents a comparison of the training architectures in terms of the performance indicator for EI systems. As presented in the table, the accuracy is always high as we have more diversity and amount of data, that is, in the Centralized architecture. But the privacy is usually degraded in the centralized because the data has reached the central server through the network and anonymization is done after it reached the central data processing centre. The hybrid system has higher efficiency in terms of latency and communication as it has the advantage of proximity to the source of the data compared central server but better computing resource than the end device. In terms of energy and memory, the decentralized systems do not perform well as the end devices are usually battery powered and possess limited computing resources.

D. Technologies for Edge Intelligence Training:

Different techniques have been developed to enable the edge intelligence architectures in varying arenas of information technology. Few of them are

1) Federated Training [13]:

This technique designed to increase the privacy of the users by retaining the raw data in the end devices itself. The method aggregates the updates received from the locally-trained models in the end-devices and integrates to the shared model saved in Edge Server.

2) Aggregation Frequency Control [14]:

This is an improvement on the federated training by regulating the update frequency. This emphasis on reducing the bandwidth consumption by controlling the aggregation functions output size and the update strategy.

3) Gradient Compression[15] :

This technique is also aimed at reducing the communication overhead, using the concept of delta increments. The synergy between the local model and shared model is achieved by keeping track of the delta whenever an update is performed. The network traffic is reduced greatly by communicating only this delta to the shared DNN.

4) DNN Splitting [16]:

The DNN splitting is a simple concept of dividing the DNN into 2 sets of layers. Hence, when a data is received, partial processing is carried out in the end device and semi-processed data is sent to the edge server model for further processing. This increased the privacy of the user but takes a toll on the local computing resources and development time of the systems due to the complexity of the split DNN modelling strategy.

5) Knowledge Transfer Learning [17]:

This is one of the most innovative methods of training, where base-training is done based on a general dataset. Then this master model is passed on to the target devices where the more contextual training will be done on the target device dataset.

6) Gossip Training [18]:

This is based on decentralized edge intelligence training architecture where is no dependency on the central nodes. Each node trains its own model and broadcasts its knowledge to randomly selected peers. This process helps to reach convergence faster than the other techniques.

Table III summaries the techniques based on their performance indicators which were discussed in line with architectures. Various technique exhibit improvements in different combination of performance indicators. In federated learning, the focus is on decreasing the latency and increasing

Table. II Summary based on key performance indicators of training architectures

Architecture	Accuracy	Privacy	Latency	Communication Overload	Energy	Memory footprint
Centralized	Low	Low	High	High	Low	Low
Hybrid	Medium	Medium	Low	Medium	Medium	Medium
Decentralized	High	High	Low	Low	High	High

privacy. Aggregation Frequency Control improves on federated learning in terms of the efficiency of local resource and communication overload. Gradient Compression goes one step ahead compared to the previous two by implementing delta-change update strategy. DNN splitting method emphasis on the deployment flexibility of the EI system by dividing the DNN layers between the server and end-device based on the computing capabilities and latency. Knowledge transfer brings in completely a new concept in terms of AI by improving training time and reducing the exposure of user data. Gossip training presents a completely decentralized system which brings in high privacy, low latency and low memory footprint.

E. Inferencing Architectures

Model inferencing is the key aspect of the Edge intelligence systems as the predictions which are used for making decisions are derived in this stage of the AI system. Various architectures have been proposed and tested in various application areas. The significant architectures are:

1) Edge-based:

The inferencing in this architecture is Edge server-centric, where the device is responsible for collecting the data and pass on to the edge server. The prediction is one on the edge server based on the inputs provided by the device then, the prediction is passed on to the device [13]. This architecture reduces the variance required in DNN development as the device architecture plays a little role in inferencing.

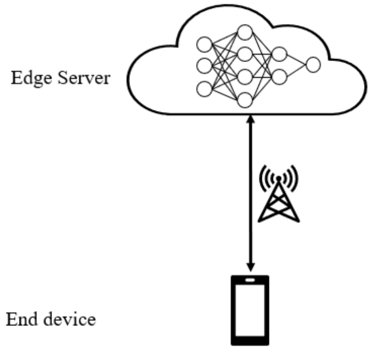


Fig. 11 Edge-based architecture for DNN Inferencing [13]

2) Device-based:

The complete inferencing process is carried out by the device itself, which reduces the latency between sensor and actuator significantly [13]. This also helps in decreasing the demand for network bandwidth. But, the demand for local computing resources such as GPU and memory are increased tremendously.

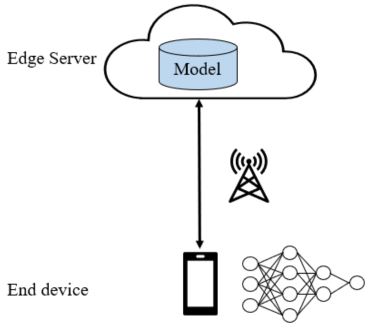


Fig. 12 Device-based architecture for DNN Inferencing [13]

Table. III Summary based on key performance indicators of training techniques

Technology	Accuracy	Privacy	Latency	Communication Overload	Energy	Memory footprint	Features
Federated Learning	Medium	High	High	Low	Medium	High	• Increased Privacy • Use of local computing resources • Reduced data exchange rate
Aggregation Frequency Control	NA	NA	Medium	Medium	High	Medium	• Increased efficiency in the use of local computing resources • Reduced data exchange rate
Gradient Compression	NA	NA	Low	Low	Medium	Low	• Delta change based update • Increased efficiency in model update strategy • Better
DNN Splitting	NA	High	Medium	High	High	Medium	• Better utilization of resources without compromise on accuracy • Flexibility in deployment • Higher complexity in development
Knowledge Transfer Learning	Medium	High	NA	Medium	NA	NA	• Brings in advantages of centralized learning in the base model by general training data set • Contextual precision created in the local training set
Gossip Training	Medium	Medium	High	Low	Medium	Low	• Asynchronous communication model creates better communication efficiency • Decentralization removes the dependency on an edge node server. • Privacy is preserved

Table. IV Summary based on key performance indicators of inferencing architectures

Architecture	Accuracy	Privacy	Latency	Communication Overload	Energy	Memory footprint
Device-based	Medium	High	Low	Low	High	High
Edge-Device	High	Medium	Medium	High	High	NA
Edge-Based	Medium	Medium	NA	High	Medium	Medium
Edge-Cloud	High	Low	High	High	Low	Low

3) Edge-Device:

The DNN is essentially is split into two sets of layers, in which the some of which will reside in the device and the rest in the edge server. When the data is collected the device execute its portion of the DNN layers then pass the semi-processed data to the edge server which can execute the data through remaining layers [13]. The edge server sends back the predictions to the device. This gives flexibility in terms of computational resources required at edge and device, but the complexity of the systems is increased.

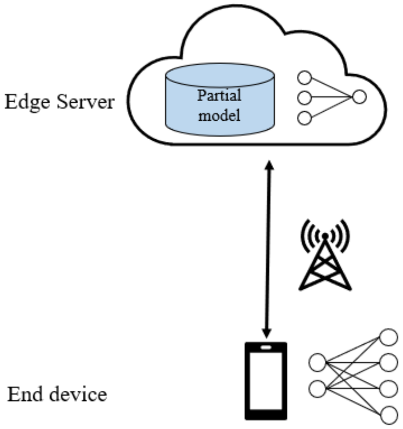


Fig. 13 Edge-Device architecture for DNN Inferencing [13]

4) Edge-Cloud

This is similar to the Edge-device model but the sharing of the DNN layers is done between the edge servers and cloud data Centre [13]. This reduces the resource demand in the device substantially but increases the latency.

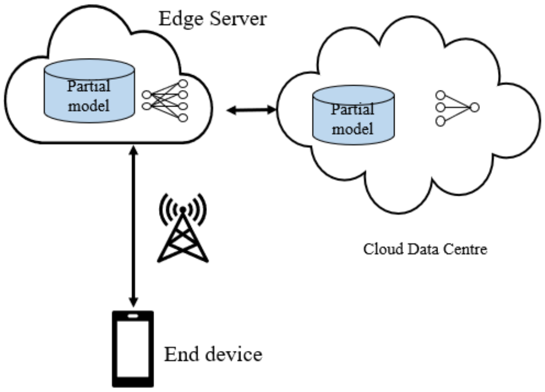


Fig. 14 Edge-Cloud based architecture for DNN Inferencing [13]

The various architecture is introduced to solve the variance in the systems in play while dealing with IoT environment. The computing power, communication bandwidth etc various a lot between the device and servers used for the ever-changing application arena. Table IV summaries the architecture based on the key performance indicators discussed previously. In terms of memory and energy requirements of the device, cloud-edge or edge-based architecture will be a good candidate. But, if the focus is on low latency and privacy, then device-based or edge-device based architecture will be beneficial.

F. Technologies for Edge Intelligence Inferencing

Based on the presented architectures, different technologies have been developed across the various use cases ranging from image processing to voice commands. Some of the techniques are:

1) Model Compression [19]:

The model compression is a way optimizing the representation of model which in turn reduces the processing time required for inferencing. Weight pruning is one of the popular methods used for compression which ranks the neurons based on contribution to the model and trim-off the one with lower ranks.

2) Model Partition[20]:

The idea is to split the DNN model into 2 sets of layers in which initial processing of the data will be done by the partition on the device and the rest on the partial model present on the edge server. The Neurosurgeon[20] is the most famous method in which regression methods are used find the optimal point at which the model has to be portioned in order to achieve the optimal performance based on the computing resource availability in the device and edge server.

3) Model Early-Exit [21]:

There are cases where a DNN with high accuracy can classify data using the first few layers itself. So, the DNN is designed in such a way that the exit points can be defined in intermediate layers, it will reduce the turnaround time for inferencing.

4) Edge Caching [22]:

This is an adaptation of the caching mechanism in most software programs where the recent results are stored temporarily in the edge server in order to avoid redundant processing of the same type of request.

5) Input Filtering [23]:

The method does the filtering of the data using conventional software programs which can be run at a local mobile device before sending data to the DNN network is called Input filtering. It gives advantages in two fronts: 1) DNN algorithms are expensive operations compared to logical programming. 2) the data to be sent across the network is reduced to a larger extent which saves the data.

The techniques listed here are only a subset of the techniques which are proposed in this solution space. Table V summaries the features as the improvement focus of each technique in terms of the EI performance indicators. Model partition focuses on mainly memory footprint by removing redundant weights in the model which in turns reduces the load on the DNN processing for inferencing. Model partition provides an improvement on privacy and latency by doing initial processing of the data using the first part of DNN layers present, this increases the privacy as the raw data stays on the end device. Model Early-Exit also have a similar goal but on a different scenario where the full potential of a deep network is not required, especially in classification problems. Even the Edge-caching provides a low latency by caching the previous results, this technique is apt in non-critical and generalized applications like recommendation systems. Input filtering is used mainly in application like video surveillance where the amount of data sent for inferencing is huge. Edge caching and Input Filtering reduces the latency to a greater extent

V. Applications And Trends

The use cases for Edge Intelligence is very broad as it is a concept which can be applied to any systems with a server, client and a network between them. The applications of EI is vast which varying from Human tasks like Natural Language processing to highly mechanised tasks like manufacturing. The growth of the IoT environment and advancements in the Edge Technology and High bandwidth wireless communications, EI has a lot scope to exhibit its capabilities over the previous generation AI programs. Following are some of the common applications of EI.

A. Natural Language Processing

According to Chen and Ran [11], natural language processing has become popular in recent times with personal entertainment and assistant applications. Some examples for NLP-based AI at Edge Nodes are the voice-based personal assistants like Apple Siri, Amazon Alexa, Cortana etc. Wake-word processing is the best example for EI based for hybrid-based architecture for inferencing. DNN splitting and Input filtering are used for this process, where the functions like noise cancellation, voice-filter extra are using the Input filtering then the inferencing of the command is split between the device processor and the cloud processor.

B. Video Surveillance

The video surveillance is one of the toughest yet one of the most crucial applications of vision processing. The challenge in using the cloud-based architecture for video surveillance is that the amount of data is huge to be transmitted. The challenge in using on-device inferencing is that usually, the complexity and computing requirements are usually higher than what the end device can handle. This is why Edge Intelligence is significant because using the Edge optimization techniques like Model Partitioning, DNN splitting etc [11]. the amount of data to be is reduced by having a semi-processing in the on-device and the processing in the central servers are reduced by pre-processing some of the DNN layers on the device itself. TensorFlow lite [24] is proposed for embedded devices with deep learning capabilities which can process images.

A wireless video surveillance system proposed by Zhang et al. [25] called Vigil, is a good example of Edge Intelligence based video processing. the systems concentrated on two optimization factors, reducing the frames to be sent to the cloud-based server and to balance the load on the server model when new devices are added. Another example is VideoEdge by Hung et al [26] in which a similar approach for optimization is implemented but the concentration of maintaining the high accurate prediction.

C. Industries

The Process industry had welcomed the cloud technologies with high enthusiasm, but the communication complexity, security concerns and latency associated with it, has made a

Table. V Summary based on key performance indicators of inferencing techniques

Technology	Accuracy	Privacy	Latency	Communication Overload	Energy	Memory footprint	Features
Model Compression	NA	Low	Low	Medium	Low	Low	• Weight pruning to reduce redundancy in the DNN network. • Quantization of the DNN to reduce data exchange
Model Partition	NA	Medium	Medium	High	Medium	High	• Splitting of DNN network to preserve privacy and decreased computing in the edge server • Flexibility in terms of the target deployment environment.
Model Early-Exit	Low	NA	Low	NA	Medium	NA	• Multi-use of DNN networks • Latency improvement is the key • Higher complexity in design and development
Edge Caching	NA	Medium	Low	Low	NA	High	• Easy to implement • Useful for high traffic applications • Reduced bandwidth usage and latency
Input Filtering	Medium	NA	Medium	Low	Low	Medium	• Reduced computing requirements on the DNN. • Reduced mundane tasks for DNN by implementing more efficient logical programs at the source itself.

space for Edge Computing, along with it comes the Edge Intelligence. There are various applications which are associated with Edge Intelligence in the industry like proximity decision-making services, energy management, adaptive route control, improvised batching process [27]. Sodhro [27]et al. has proposed an energy management system based on Edge intelligence such as 1) Forward Central Dynamic Available Approach (FCDAA) 2) Energy dissipation evaluation for industry-wide battery modelling 3) Data reliability model for Edge Intelligence-based IoT devices.

Logistics inside manufacturing plants is also a key consumer of Edge Intelligent systems [28] where the goal is to optimize the movements of the raw materials, semi-processed constituents and finished products inside the manufacturing plants. This is achieved by having elaborate IoT based devices which are capable of doing an update of inventory information. The Edge server AI applications are used to derive the requirements of materials at particular stages of assembly, based on the throughput, batch size and efficiency of the plant at that point of time.

VI. Challenges and Future Scope

Edge Intelligence is a comparatively new concept in the computing field; hence the technology has a lot of shortcomings which limit the implementation scope. These challenges are the areas in which the industry and academia have to concentrate in order to become an accepted norm in the pragmatic solutions arena. Starting with security concerns arousing due to the high network-driven computing to the functional use cases which lack the maturity to replace human dependency is a challenge which is encountered while advocating for the EI. This section discusses some of the areas where EI can improve in order to come in the mainstream of the technology world.

A. Security and privacy

As the cyber world has grown into people’s everyday life including medical data, personality assessment etc. This puts a lot of pressure on the security aspect of IoT and Cloud technologies, which is inherited to EI technologies [29]. The industry also considers the security aspect of EI with scepticism because for some companies the data itself is a revenue source and even a small chance of vulnerability is not accepted. Another major concern is over the privacy of the users of such EI systems. Even though in preliminary comparison of the cloud versus edge, the EI may seem to hold an upper hand with preserving the data locally than in the cloud. But on a closer examination, we can understand the maturity of the security measures in the cloud is at a higher level than the edge computing simply because of the timeline of introduction and acceptance into the market.

B. Latency

Currently, AI techniques are popular in the off-line

applications or non-critical applications because of the expensive training and inferencing operations. In Level 4 and above EI systems the latency created by communication is reduced significantly, but with current computing capabilities in the end-devices, the performance of the AI techniques has not reached the desired maturity [11].

C. Performance and Network

The expectation on the EI is enormous chiefly because it is expected to mitigate the issues in cloud and AI, but the ignored fact is that it also inherits the issues faced in the same. For example, the computation limitations in the Edge server is more compared to the cloud. For another instance, the network bandwidth between an end device and edge server may not be wide enough for the realization of edge computing. For the application of EI into the critical jobs the technology faces even a greater challenge; even with a slight delay in response or error can lead to catastrophic events [29].

D. Development complexity

The systems in the level of Cloud intelligence was simple in terms of architecture. But as the systems go up in the EI level, the implementation gets more complex as it faces multiple sub-challenges such as constrained memory, varying communication protocols, dynamic run-time environmental setups. As the complexity of the systems goes up there is a high chance of system scalability and maintenance issues [9].

E. Contextual Inferencing

The machine learning algorithms are as good as the data provided to the algorithms to train on. If the data is not good enough, the model also will not have accurate results in determining the actions. This is a great challenge the need to address in the new world of Artificial intelligence-based cyber world. One of the root causes for the same is that the in regular cloud-based Artificial intelligence systems, the data is aggregated without context. To accommodate the large volume of data, features are trimmed or merged. Another aspect is that when data is collected over various sectors, trying to fit a single model causes the algorithms to predict poorly because the features are not casing enough for the particular case [11].

VII. Conclusion

Artificial intelligence has gone through several upsurge and downturns throughout its course. Each time the AI bounced-back because of something innovative than before, and this time it is the Cloud and IoT which fertilized the growth of AI. The Edge Intelligence is the new milestone created by this amalgamation of IoT, Cloud and AI, which shall take the AI into previously unventured fields. Various architecture for training and inferencing has been discussed in this paper and have been evaluated based on the key factors, that is latency, energy, memory, accuracy, communication bandwidth and privacy. Furthermore, techniques in line with these architectures have been compared and contrasted based on their use cases. Additionally, the application of EI is discussed briefly, after which the challenges faced in the implementation of the Edge Intelligence is identified and presented.

VIII. References

[1] "Professor John McCarthy." [Online]. Available: http://jmc.stanford.edu/. [Accessed: 17-Oct-2019].

[2] S. J. Russell, P. Norvig, and E. Davis, *Artificial intelligence: a modern approach*, 3rd ed. Upper Saddle River: Prentice Hall, 2010.

[3] "Artificial intelligence," *ScienceDaily*. [Online]. Available: https://www.sciencedaily.com/terms/artificial_intelligence.htm. [Accessed: 10-Oct-2019].

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[5] S. Mittal, "A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform," *J. Syst. Archit.*, vol. 97, pp. 428–442, Aug. 2019.

[6] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," *Nat. Lang. Process.*, p. 45.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Inf. Process. Syst.*, vol. 25, Jan. 2012.

[8] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[9] F. Liu, G. Tang, Y. Li, Z. Cai, X. Zhang, and T. Zhou, "A Survey on Edge Computing Systems and Tools," *Proc. IEEE*, vol. 107, no. 8, pp. 1537–1562, Aug. 2019.

[10] "Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper - Cisco." [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html. [Accessed: 07-Oct-2019].

[11] J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug. 2019.

[12] S. B. Calo, M. Touna, D. C. Verma, and A. Cullen, "Edge computing architecture for applying AI to IoT," in *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, 2017, pp. 3012–3016.

[13] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *ArXiv160205629 Cs*, Feb. 2016.

[14] K. Hsieh *et al.*, "Gaia: Geo-Distributed Machine Learning Approaching LAN Speeds," p. 21.

[15] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," *ArXiv171201887 Cs Stat*, Dec. 2017.

[16] J. Zhao, R. Mortier, J. Crowcroft, and L. Wang, "Privacy-Preserving Machine Learning-Based Data Analytics on Edge Devices," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, New Orleans, LA, USA, 2018, pp. 341–346.

[17] J. Wang, J. Zhang, W. Bao, X. Zhu, B. Cao, and P. S. Yu, "Not Just Privacy: Improving Performance of Private Deep Learning in Mobile Cloud," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, London, United Kingdom, 2018, pp. 2407–2416.

[18] M. Blot, D. Picard, M. Cord, and N. Thome, "Gossip training for deep learning," *ArXiv161109726 Cs Stat*, Nov. 2016.

[19] Q. Qin *et al.*, *To Compress, or Not to Compress: Characterizing Deep Learning Model Compression for Embedded Inference*. 2018.

[20] Y. Kang *et al.*, "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge," in *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, New York, NY, USA, 2017, pp. 615–629.

[21] S. Teerapittayanon, B. McDanel, and H. T. Kung, "BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks," *ArXiv170901686 Cs*, Sep. 2017.

[22] U. Drolia, K. Guo, J. Tan, R. Gandhi, and P. Narasimhan, "Cachier: Edge-caching for recognition applications," p. 11.

[23] C. Zhang, Q. Cao, H. Jiang, W. Zhang, J. Li, and J. Yao, "FFS-VA: A Fast Filtering System for Large-scale Video Analytics," in *Proceedings of the 47th International Conference on Parallel Processing - ICPP 2018*, Eugene, OR, USA, 2018, pp. 1–10.

[24] "TensorFlow Lite." [Online]. Available: https://www.tensorflow.org/lite. [Accessed: 17-Oct-2019].

[25] T. Zhang, A. Chowdhery, P. (Victor) Bahl, K. Jamieson, and S. Banerjee, "The Design and Implementation of a Wireless Video Surveillance System," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking - MobiCom '15*, Paris, France, 2015, pp. 426–438.

[26] C.-C. Hung *et al.*, "VideoEdge: Processing Camera Streams using Hierarchical Clusters," in *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, Seattle, WA, USA, 2018, pp. 115–131.

[27] A. H. Sodhro, S. Pirbhulal, and V. H. C. de Albuquerque, "Artificial Intelligence-Driven Mechanism for Edge Computing-Based Industrial Applications," *IEEE Trans. Ind. Inform.*, vol. 15, no. 7, pp. 4235–4243, Jul. 2019.

[28] A. Luckow *et al.*, "Artificial Intelligence and Deep Learning Applications for Automotive Manufacturing," in *2018 IEEE International Conference on Big Data (Big*

Data), Seattle, WA, USA, 2018, pp. 3144–3152.

[29] G. Plastiras, M. Terzi, C. Kyrkou, and T. Theocharidcs, “Edge Intelligence: Challenges and Opportunities of Near-Sensor Machine Learning Applications,” in *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, Milan, 2018, pp. 1–7.