Making AI available for everyone at anywhere: A Survey about Edge Intelligence

Yusen Zhang\textsuperscript{1,2,*}, Bao Li\textsuperscript{1}, Yusong Tan\textsuperscript{1}

\textsuperscript{1} College of Computer, National University of Defence Technology, No.105, Deya Road, Changsha 410073, China
\textsuperscript{2} Center for Information Technology, Xi’an International Studies University, No. 6, South Wenyuan Road, Xi’an 710128, China

*zhangys04@nudt.edu.cn

Abstract. With the rapid development of the 5\textsuperscript{th} generation of wireless network communication technology, a high-speed data rate, broader bandwidth, and low-latency wireless network has already come to the truth. Meanwhile, Edge Computing which put data processing at the edge of the network has already been well developed, and it has changed the computing mode tremendously. And also, Artificial Intelligence (AI) has made a great breakthrough from Computer Vision, Unmanned Vehicle to Natural Language Processing. Nevertheless, with the increasing model size and model depth, the training data set and communication delay has become the bottleneck of AI popularization. As zillions of bytes of data are being generated at the network edge, and also lots of AI applications are being utilized by enormous customers lied at the edge, the combination of Edge Computing and AI is imperative. This paper has done a deep investigation into the new field about the confluence of Edge Computing and AI, aiming of discussing the concept, architecture, and research ideas on Edge Intelligence. To the end, this paper provides the road-map for future researching work.

Keywords: Edge Intelligence, Edge Computing, Confluence Architecture, AI.

1. Introduction

With the booming of communication technology, the 5\textsuperscript{th} generation of wireless networks has already reached maturity. Based on it, an extremely high-speed data rate, lower latency response, and broader bandwidth wireless network has been offered to enormous users and organizations. That’s to say 5G is on the way and it will accelerate the data speed up to 100 times faster than current technology. Hereafter, the Internet of Things (IoT) has already been playing an important role in today’s society, such as Video Surveillance, Unmanned Vehicle, Industry Internet of Things (IIoT), and so on. Due to the responsibility, all of these devices are collecting and processing data all the time, a huge scale of data is being generated continuously at the same time and plentiful information is on the way.
Because of the development of technology and the improvement of demands, more and more devices are being interconnected, the number of them is growing exponentially. In that case, the centralized Cloud Computing framework may not be suitable for most of the applications, such as Unmanned Aerial Vehicles, video surveillance, attributing to the requirements of real-time and privacy, making edge nodes to provide computing services is inevitable. Different from Cloud Computing, a big partition of data processing has been done at the edge side in Edge Computing, especially the privacy-sensitive and delay-sensitive data, that's because putting computational resources to the network edge could gain performance improvement, privacy protection, and bandwidth consumption reduction. From the white paper of Cisco, the number of end devices connected to the internet will grow to 75 billion in 2025, and the internet will become increasingly large and complex. In that case, the management and maintenance of the internet will be a sticky problem, and current technologies are no longer sufficient to meet this demand. Besides, under the prevalent network structure, security is also an issue that cannot be ignored.

Meanwhile, Artificial Intelligence (AI) is not a new term and was first proposed in 1956, which has already been well developed. At present time, plenty of AI-based applications has been deeply used and gained giant popularity. To enhance the user experience, AI models are becoming increasingly complex and large, and this has posed a huge challenge for model training and inference tasks. Although the 5th communication technology has already brought a high-speed and broad bandwidth wireless network, in some areas such as collision warning in unmanned vehicles, applications are still very sensitive to response time, even millisecond delays will bring huge consequences. As for the AI model inference, hardware platform architecture and operating mode will make a great influence on it. Coming to the AI model training, how to accelerate the training speed and improve the accuracy are emerging issues that need to be addressed.

The key point that pushing AI into the spotlight is big data, which is being generated by the edge devices. For today’s AI projects, the acquisition and process of training data set is the most crucial part of it, just like petroleum which is as important as to the automobile industry, and it could be named as the fuel of the AI model. In addition, intelligence applications also need to be widely deployed, so end users could enjoy the benefits of AI technologies. Furthermore, allocating and arranging resources such as computational and storage resources reasonably is the key step to make edge computing more powerful. Through the previous analysis, it is easy to figure out that AI has strong analytical and inferential capabilities, and Edge Computing could provide a large amount of training dataset and an excellent execution platform, which can decouple from the heterogeneous hardware. In that case, the fusion between Edge Computing and AI is bound to happen, which reproduces the Edge Intelligence (EI), as a natural process. Data produced at the network edge needs to be deeply excavated and fully unlocked its potential by AI, moreover, edge computing could provide abundant data and diverse running scenarios.

To indicate the benefits of the fusion between Edge Computing and AI, a comparison between different computing modes from 6 dimensions has been proposed as shown in Fig.1. It is easy to figure out from the radar chart that the conventional computing mode represented by Cloud Computing is not suitable for the current demand anymore, especially at privacy protection, response latency, availability, and scalability. Besides, edge computing and on-device computing each have their own merits, but they are all not perfect. The synergy between cloud, edge, and end devices will empower the whole system with more computation-intensive, delay-sensitive, and energy-constraint services.
Although EI is not a new concept and has been discussed several times, it is still at the theoretical stage, and in-depth research and practice are needed urgently. Besides, the EI is not a straightforward combination between Edge Computing and Artificial Intelligence, it requires a comprehensive understanding of the AI models and the edge computing features to design and build a power endurance computing system. As matters stand, literature reviews about EI are relatively rare, and all the published surveys only focus on discussing edge computing and AI separately. This is the motivation for proposing this paper with the tutorial of the EI computing paradigm and the comprehensive discussions about how to make AI available for everyone anywhere via edge computing. This paper is going to present the state of art EI frameworks and make a deep insight into this new interdisciplinary.

![Fig.1 Comparison of different computing mode](image)

2. Edge Intelligence Computing

As discussed previously, EI is derived from the combination of Edge Computing and AI, which could be explained in two dimensions that are AI in Edge and AI for Edge. AI in Edge means putting AI applications as close as possible to the end-users, by making full use of the superiority of edge computing architecture. Besides, it also includes distributing training tasks to the network edge to make the most of the training data generated by the terminal devices and speed up the model training progress. Coming to AI for Edge, it is going to supply more efficient management and maintenance services to edge computing, by exploring the capabilities of AI algorithms completely, especially for huge scale and complicated networks. In this paper, in-depth research and analysis work about EI computing, especially AI in Edge, will be proposed, which will cover the basic concepts, motivation, and challenges.

There is still no standard and unified definition for EI Computing, but it can be expressed as that the applications which combine edge computing with machine learning to complete the data collection, processing, and storage, and it aims to enhance the quality and speed of data processing, improving the privacy protection and connection security. This definition has adequately described the main elements of this concept, and also explained the motivation.

In the EI system, computational resources are deployed at the network edge, even some end devices connected to the network also have computing capability, so parts of model training or
inference jobs could be split into multiple sections and offloaded from the cloud center to the edge side. To align with the demand improvement, AI models have already got larger and deeper, and this brings great challenges to the model training and inference tasks. To tackle these problems, lots of enabling technologies such as weight pruning, model quantization, and model partition have been invented, and the main idea of these optimization methods is model compression, that’s to say completing the training or inference task quickly by using only part of the giant model, which gains speeding up by sacrificing the model accuracy and state space.

Simultaneously, Deep Learning models that are representative of AI are data-sensitive and data-hungry by nature, which means that the quality of the training dataset will impact the model accuracy directly. So the training dataset is the key component of AI, and it will become one of the main hazards which may hinder further advancements from being unleashed. In addition, how to make data more efficient is the most burning question needed to be addressed. According to the experiment, large amounts of high-quality training datasets will yield a high precision model, an increasing number of optimum proposals have been delivered such as high-speed data processing and data-flow acceleration, which will be conducive to generating enough labeled-data and furnishing to the AI applications. As Geoffrey Hinton reported in his presentation “The Next Generation of Neural Networks” in 2020, that the most important unsolved problem with artificial neural networks is how to do unsupervised learning as effectively as the brain, which is also talking about the feasible striving directions of treating the problems discussed previously.

Under the EI computing architecture, the greatest feature of the fusion is making full use of the decentralized resources, and edge nodes cooperating could provide much better services, even the modest computing resources could also be fully utilized. Besides, another natural characteristic of edge computing is the geographically distributed data sources, and it will be the abundant source of unlabeled data. A considerable amount of methodologies about distributed model training and inference have already been raised, such as parameter pruning and sharing, weights quantization, federated learning, and learning transfer. As a matter of fact, all these methods are trying to narrow the action space or compress the state space. Another key thing to remember is that the abundance of data, which should be made the most of its implicit features. At the present time, the focus has already been shifted to making deep learning more suitable for edge computing, and it seems that the production of training data has been overlooked. Although unsupervised learning and self-supervised learning are being fully studied these days, the carrier is still a conventional computing platform mostly, such as parallel computing and grid computing, and the final target is also sustaining the supervised learning. To cope with the evolving demands and AI techniques improvement, it is also very essential to pay a little more attention to the training dataset itself.

3. General Edge Intelligence Computing Architecture

General EI mainly contains two aspects, one is the EI model training, and the other is the model inference or the intelligence applications. With the help of edge computing, there are three types of model training architecture, Centralized, Decentralized, and Mixed (Cloud-Edge-Device) as discussed by Z. Zhou et al, and Fig.2 shows these architectures.
In Fig.1, the cloud represents the cloud data center and the terminal units are indicated by UAVs (Unmanned Aerial Vehicle), cars, and cell phones, which are the data sources. Besides, the EI models are expressed by star form networks, and the cell sites stand for the edge nodes.

In the centralized mode, the EI model training work is all implemented by the cloud data center, and the distributed end devices are only responsible for producing and gathering training data. The advantage of this mode is the fast convergence as a result of the computing power of the cloud datacenter, while the disadvantage is huge communication resource consumptions. This architecture is being widely used by lots of organizations, and it has already derived lots of mature prototype systems.

As for the decentralized mode, the EI model is trained respectively by edge nodes, and they use the local data to train their own model. Besides, edge nodes will communicate with each other and exchange local training improvement to get the global model. Under this architecture, the EI model can be trained by edge nodes separately without the cloud datacenter’s help. On account of this, the merits of this mode are flexible and small bandwidth requirement, but the shortcoming is also obvious, that is the uneven training process caused by random user access. And also, there are many mature applications derived from this training type such as Gossip Training.

Coming to the mixed mode, it is a converged architecture of centralized and decentralized mode. During the model training process, the edge nodes may treat it by decentralized improvements together or centralized training processes with the cloud datacenter dynamically, based on the network quality and data distribution. Compared with the Distributed Training Mode, the main differences are data distribution and communication consumption, especially the WAN occupation. Using the mixed training mode is easier to obtain more efficient AI models, and lots of fruitful technologies use this mode to train their models, such as the Federated Learning.

According to the model inference, there are also three main types of intelligence model inference architecture, centralized, local, and hybrid as represented in Fig.3.
The centralized mode is shown in Fig. Error! Reference source not found. (a), the AI model is based at the cloud datacenter, and the end devices only receive input data and directly send them back to the data center, all the model inference tasks have been done at the remote side. This mode is an alternative solution when local inference is not available or the accuracy is not suitable for use. As long as the powerful computing ability of cloud datacenter, it is able to hold a giant and complete AI model, the inference accuracy and capability is strongest. But the response latency and network consumption are the main disadvantages.

There are two types of local mode as shown in Fig. Error! Reference source not found. (b), for the left one, the AI model is on the edge server and the inference jobs are all done at the edge node locally. Similar to the centralized mode, terminal units are only in charge of input data collecting and sending them back to the edge server. The scalability of this mode is very well, and applications can be easily implemented on different mobile computing platforms by using this mode. For the right one, the AI model is transferred to the end devices and the intelligence inference tasks have been done by these devices separately. Limited by the computing capacity of the terminal units, such as GPU and memory, the inference performance relies on the local devices.

The hybrid mode is illustrated in Fig. Error! Reference source not found. (c), it is a cloud-edge-device architecture. Under this mode, the AI model will be divided into several parts and loaded on end-devices and edge nodes individually in the light of system configurations, such as computational resources, network bandwidth, and the workload. In that case, the AI model will be executed
hierarchically and synergistically. When data collected by end devices, they will be processed with the local model loaded at the devices. Then intermediate data will be transferred to the edge nodes and treated by the remaining parts of the model stored at the edge servers. In the end, the inference result will be calculated by the cloud datacenter and sent to the end devices. Sometimes, the final conclusion may be inferred by the edge nodes or the end devices, based on the distribution and accuracy of the model integrated on the device or the edge server. With the help of early-exit algorithms, this architecture has already been broadly applied in lots of scenarios, such as Smart Campus and Industry Internet of Things.

4. Continuous Learning in Edge Intelligence

As discussed before, lots of heuristic work has been put into practice and the EI has already ascended to the spotlight. Due to the natural advantage of Edge Computing, AI applications could obtain sufficient training data and computational resources which ensure the model training task could be done efficiently. Meanwhile, with the rapid development of IT technology, the needs of end-users have grown more and more diverse, and stricter and stricter. In other words, time-sensitive, privacy-sensitive, computation-intensive, and energy-intensive tasks have already come to be the main workload that edge computing needs to undertake. To capture all the demands, plenty of algorithms have been derived from the EI paradigms such as Model Compression, Input Filtering, and Edge Caching, and all of them are aiming to reduce resource consumption and improve performance, especially the response time and privacy protection. Although these enabling technologies have already been proved to have obvious effects, there are still some problems that need to be solved to maximize the overall performance of EI.

In addition, the Model Degradation in production is always a frustrating problem, and there is even a statement that “The moment you put a model in production, it starts degrading”. Through deep analysis, it was found that Model Degradation is caused by Concept Drifting, which is a highly demanding research job in AI. In a common scenario, the AI model training works are completed offline and separated from model inference, that’s to say the AI models are often obtained from the finite selected training dataset and fail to stay in sync with the run-time environment. As mentioned before, data is the most important component of an AI system, and the main reason which causes the problem is that the training data is already out of date. In order to give full play to the advantages of Edge Computing, continuously learning in EI has been proposed.

It is a useful feature of Edge Computing that is very close to the final tenant, acquiring an adequate fresh training dataset is very easy, and it could provide great conditions for the AI model training task. In order to fully use this property, architecture for continuously learning has been proposed as shown in Fig.4. Different from the previous mode, a training data collector has been added to the edge node to gather the first-hand training data set. Furthermore, for the purpose of privacy-protecting and reducing communication resource consumption, part of the model training task will be done at the edge server and only intermediate data will be sent back to the cloud datacenter. While the AI model inference is being carried out, model training work will be completed only between edge servers and cloud datacenter at the same time, contemplating the computation ability and energy constraint of end devices.
Fig.3 The architecture of continuous learning in EI computing.

The process of model inference and model training which has been done at the same time is shown in Fig.5. There are two types of data flow: inference data flow and model training data flow. When the end device has got the input data, the model inference module will do the prediction with the help of its local model first, if the accuracy passes the preset threshold, then it will give the inference result at once. If the accuracy is under the threshold, then the intermediate data will be submitted to the edge server and the inference task will be treated continuously under the model of the edge side. Like the previous procedure, the execution path is decided by the accuracy checking module. On condition that the result does not meet the prerequisites, model inference will be taken over by the cloud datacenter, and it will give the final result, based on its excellent computation power.

Data collecting begins when the model inference task has been finished successfully and the result has already been output. At this moment, the training vector which contains the original data and label is ready, and it will be compressed by the tracker and sent to the edge server. Then the Training Data Collector module starts the data collection process, when the dataset grows to a certain size, the Data Preprocessing and Model Training module will join into the model training process. After that, the front layers of the AI model will be trained and the model updates will be submitted to the cloud datacenter. The remaining parts of the training task will be completed in the cloud datacenter, after that the Model Monitor will check the improvement of the training result. If the performance of the new model has significantly improved, such as the error rate, the Model Updater will update the AI model collaboratively, which has already been distributed on Cloud, Edge Nodes, and End Devices. Now all the processes and jobs have been done completely.
Fig. 4 The dataflow chart of the continuous learning via EI computing

5. Analysis and Future Working Focus

According to the discussion before, continuous learning in EI is able to inspire the AI applications’ potential. This is not only a new computing architecture and computing mode but also a novel ideology about training dataset generating without manual description. And the special characteristics of EI computing can be summarized as follows.

5.1. Powerful model generalization ability

Generalization is a key performance indicator of the AI model, which means the model is able to make great performance in different environments, and process the raw data which has never been met before. Under the framework of edge computing, the model training task is able to be done at the edge of the network, which has made the training process much closer to reality. Besides, edge servers have participated in the model training work and it could make better use of the advantage, that is the training dataset is collected from the real environment and could be updated in time. In the end, the
mechanism that model training work could be executed at the same time as the model inference task is in process, which has been proposed by this paper has also made contributions to it.

5.2. Platform independence

The other unique characteristic of EI computing is platform independence, which makes it be able to be widely deployed on various hardware platforms. First, EI computing has already supplied several architectures and they could be selected according to the execution situation. Then, there are plenty of enabling technologies such as Model Compression, Learning Transfer, which could make EI computing more suitable for various platforms. Finally, the cloud-edge-device collaborative working mechanism will also make EI computing more competitive. Based on these, AI-based applications could be installed on different hardware and run easily.

5.3. Complete privacy protection

It is an inevitable problem for AI that balances privacy and accuracy to get fabulous inference performance while training the model by using a training dataset transferred from the end devices. Edge Computing has already distributed computational resources at the edge of the network, which is mainly the edge servers, part of the model training work will be treated at the edge side. Besides, some of the terminal units also have computing ability, they can take part in the training task. Based on this, privacy-sensitive data will be processed between the edge servers and the terminal devices. Privacy policies have been fully considered and these legal subpoenas and extra-judicial surveillance have also been avoided successfully.

5.4. Continuous Learning

Traditionally, most AI models are highly dependent on the training dataset, especially the Deep learning-based applications are very sensitive to the updating and furnishing frequency of it, which is mutually exclusive to the data quality. It is a new ideology that generating training dataset during the model inference, which is an online data supplement for supervised learning, that is the main idea of continuous learning. Algorithms referring to continuous learning are designed to update and improve the AI models as the input data, run-time situation, and target change. With the help of it, AI model training and inference tasks will be stacked on top of each other, it will be the stepping stone of AI and make computing work more intelligence.

As discussed before, although the combination between Edge Computing and AI has already improved to a new level by the light of nature, there are still some challenges that may be future research directions. First, the research focus needs to be shifted to data acquisition and manipulation, which is still in its infancy. Then, EI computing is a definitely new concept and it has been described repeatedly from different perspectives, thus there is neither a complete definition nor a unified standard. Efforts about the EI standards, theories, and architectures should be strengthened and carried out in a short time. Next, security problems ought to be the key research issue as well, under the modality of the EI computing, AI model training work has already been distributed and offloaded to the edge side, vulnerable connections between edge nodes and end devices may become breach for attackers, and the abnormal behaviors at the edge side may forge or poison the training data that will result in worthless model updates and harming the global model. Finally, there is no available EI computing dedicated Test Performance Indicator, Test Case, and Benchmark now, and simulation tools and open platforms are also in shortage.
6. Conclusion Remarks

With the advances of 5G technology, communication has become more and more convenient, as a result, billions of sensors and devices can be interconnected directly. To keep up with the pace of this development, abundant computational resources have been deployed at the edge of the network. In the meantime, radically new applications based on AI have already become more and more popular, offloading AI computing workload from the cloud center to the edge side is a stringent need. As a consequence, EI computing has been delivered to fulfill the developing trend.

In this paper, with the motivation of specifically discussing EI computing, a survey with the tutorial of Edge Computing and AI, and a comprehensive literature review on the recent research efforts of EI have been proposed. Besides, continuous learning in EI is discussed in detail as a new potential direction. Finally, the analysis and future working focus on EI computing have been outlined.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No. 2018YFB1003602) and the National Natural Science Foundations of China (No U19A2060).

References

[1] Zhi Z, Xu C, En L, Zeng L, Ke L and Jushan Z 2019 Edge intelligence: paving the last mile of artificial intelligence with edge computing Proceedings of the IEEE JPROC 2918951

[2] Shuiguang D, Hailiang Z, Weijia F, Jianwei Y, Schahram D and Albert Y Z 2020 Edge intelligence: the confluence of edge computing and artificial intelligence IEEE Internet of Things Journal JIOT 2984887

[3] Weisong S, Jie C, Quan Z, Youhuizi L and Lanyu X 2016 Edge computing: vision and challenges IEEE Internet of Things Journal JIOT 2579198

[4] Muhammad A, Yong W, Kezhi W and Peiqiu H 2020 A review on computational intelligence techniques in cloud and edge computing IEEE Transactions on Emerging Topics in Computational Intelligence TETCI 3007905

[5] Xiaofei W, Yiwen H, Victor C M L, Dusit N, Xueqiang Y and Xu Ch 2020 Convergence of edge computing and deep learning: a comprehensive survey IEEE Communications Surveys & Tutorials COMST 2970550

[6] Dongzi C, Qinli Y, Jiaming L and Zhu Z 2020 Selective prototype-based learning on concept-drifting data streams J. Information Sciences 516 pp 20-32

[7] Parneeta D, Ajay K and Poonam C 2020 An approach for concept drifting streams: early dynamic weighted majority Procedia Computer Science 167 pp 2653-2661

[8] Tabassum S, Efrana J, Warafta R and Meher A C 2019 Big data mining in the presence of concept drifting International Conference on Advances in Science ICASERT 8934577

[9] Geoffrey H 2020 The next generation of neural networks International ACM SIGIR Conference 3397271 3402425

[10] Chunxiao J, Haijun Z, Yong R, Zhu H, Kwangcheng C and Lajos H 2017 Machine learning paradigms for next-generation wireless networks IEEE Wireless Communications MWC
[11] Nguyen C L, Dinh T Hoang, Shimin G, Dusit N, Ping W, Yingchang L and Dong I K 2019 Applications of deep reinforcement learning in communications and networking: a survey IEEE Communications Surveys & Tutorials COMST 2916583

[12] Dawei F, Xin G, Lin L, Yu P, Di S and Mohammad M H 2020 Edge intelligence based economic dispatch for virtual power plant in 5G internet of energy Computer Communications 151 pp 42-50

[13] Olivier D, Saïd M, Sidi M, Pierre M, Jérôme B and Frédéric L 2020 Edge computing and artificial intelligence for real-time poultry monitoring Procedia Computer Science 175 pp 534-541

[14] En L, Zhi Z and Xu C 2018 Edge intelligence on-demand deep learning model co-inference with device-edge synergy Workshop on Mobile Edge Communications 3229556 3229562

[15] Ke Z, Yongxu Z, Sabita M and Yan Z 2019 Edge intelligence and blockchain empowered 5G beyond for the industrial internet of things IEEE Network MNET 001 1800526

[16] Prakash P and Panduranga V K 2019 Design and development of advanced smart energy management system integrated with IoT framework in smart grid environment J. Journal of Energy Storage 100846

[17] Yin Z, Xiao M, Jing Z, M Shamim H, Ghulam M and Syed U A 2019 Edge intelligence in the cognitive internet of things: improving sensitivity and interactivity IEEE Network MNET 1800344

[18] Xin S, Giancarlo S, Vincenzo M, Antonio P, Christian E and Chang C 2019 An edge intelligence empowered recommender system enabling cultural heritage applications IEEE Transactions on Industrial Informatics 15 pp 4266-4275

[19] Zeyu H, Geming X, Zhaohang W and Sen Y 2020 Survey on edge computing security International Conference on Big Data ICBAIE 49996 00027