A Survey on Edge Intelligence

Dianlei Xu, Tong Li, Student Member, IEEE, Yong Li, Senior Member, IEEE, Xiang Su, Member, IEEE, Sasu Tarkoma, Senior Member, IEEE, and Pan Hui, Fellow, IEEE

Abstract—Edge intelligence refers to a set of connected systems and devices for data collection, caching, processing, and analysis in locations close to where data is captured based on artificial intelligence. The aim of edge intelligence is to enhance the quality and speed of data processing and protect the privacy and security of the data. Although recently emerged, spanning the period from 2011 to now, this field of research has shown explosive growth over the past five years. In this paper, we present a thorough and comprehensive survey on the literature surrounding edge intelligence. We first identify four fundamental components of edge intelligence, namely edge caching, edge training, edge inference, and edge offloading, based on theoretical and practical results pertaining to proposed and deployed systems. We then aim for a systematic classification of the state of the solutions by examining research results and observations for each of the four components and present a taxonomy that includes practical problems, adopted techniques, and application goals. For each category, we elaborate, compare and analyse the literature from the perspectives of adopted techniques, objectives, performance, advantages and drawbacks, etc. This survey article provides a comprehensive introduction to edge intelligence and its application areas. In addition, we summarise the development of the emerging research field and the current state-of-the-art and discuss the important open issues and possible theoretical and technical solutions.

Index Terms—Artificial intelligence, edge computing, edge caching, model training, inference, offloading

I. INTRODUCTION

With the breakthrough of Artificial Intelligence (AI), we are witnessing a booming increase in AI-based applications and services. AI technology, e.g., machine learning (ML) and deep learning (DL), achieves state-of-the-art performance in various fields, ranging from facial recognition [1], [2], natural language processing [3], [4], computer vision [5], [6], traffic prediction [7], [8], and anomaly detection [9], [10]. Benefiting from the services provided by these intelligent applications and services, our lifestyles have been dramatically changed. For example, smart homes have greatly improved the quality of our lives and we are increasingly reliant on intelligent techniques.

However, existing intelligent applications are computation-intensive, which has strict requirements on resources, e.g., CPU, GPU, memory, and network, and makes it impossible to be available anytime and anywhere for users. Although current end devices are increasingly powerful, it is still not sufficient to support some deep models. For example, most voice assistants, e.g., Apple Siri and Google Microsoft’s Cortana, are based on cloud computing and they would not function if the network is down. Moreover, existing intelligent applications generally adopt centralised paradigms of data management, which require users to upload their data to central data-centre, e.g., the cloud data-centre. However, there is giant volume of data which has been generated or collected by billions of mobile users and Internet of Things (IoT) devices distributed at the network edge. According to Cisco’s forecast, there will be 850 ZB of data generated by mobile users and IoT devices by 2021 [11]. Uploading such volume of data to the cloud consumes significant bandwidth resources, which would also result in unacceptable latency for users. On the other hand, users are increasingly concerned about privacy issues. For example, the European Union has promulgated General Data Protection Regulation (GDPR) to protect users’ private information [12]. If mobile users upload their personal data to the cloud for a specific intelligent application, they would face the risk of privacy issues, i.e., the personal data might be extracted by malicious hackers or companies for illegal purposes.

Edge computing [13]–[17] has recently been proposed as an extension of cloud computing to push cloud services to the proximity of end users. Edge computing is a virtual computing platform which provides computing, storage, and networking services between end devices and the cloud server, which are usually located at the edge of networks. The device that provides services for end devices is referred to as edge servers, which could be IoT gateways, routers, and micro data centres in mobile network base stations, on vehicles, and amongst other places. End devices such as mobile phones, IoT devices, and embedded devices that requests services from edge servers are called edge devices. The main advantages of the edge computing paradigm could be summarised into three aspects. (i) Ultra-low latency: computation usually takes place in the proximity of the source data, which saves substantial amounts of time on data transmission. Edge servers provides nearly real-time responses to end devices. (ii) Saving energy for end devices: since end devices could offload computing tasks to edge servers, the energy consumption on end devices would significantly shrink. Consequently, the battery life of end devices would be extended. (iii) Scalability: cloud computing is still available if there are no enough resource on edge devices or edge servers. In such a case, the cloud server would help to perform tasks. In addition, end devices with idle resources could communicate amongst themselves to collaboratively finish a task. The capability of the edge computing paradigm is flexible to accommodate different application scenarios.

Since the advantages of edge computing could make up for the drawbacks of AI based applications, it is a promising solution to combine edge computing and AI together. This new paradigm of intelligence is called edge intelligence [18].
Fig. 1. The comparison of traditional intelligence and edge intelligence from the perspective of implementation. In traditional intelligence, all data must be uploaded to a central cloud server, whilst in edge intelligence, intelligent application tasks are partially or entirely done at the edge with locally-generated data.

Edge intelligence refers to a set of connected systems and devices for data collection, caching, processing, and analysis in a spot close to where data is captured, with the purpose of enhancing the quality and speed of data processing and to protect the privacy and security of data. Compared with traditional cloud-based intelligent applications that require end devices to upload generated or collected data to the remote cloud, edge intelligence processes and analyses data locally, which effectively protects users’ privacy, reduces response time, and saves on bandwidth resources [21], [22]. Moreover, users could also customise intelligent applications by training ML/DL models with self-generated data [23], [24].

In effect, there are already lots of works which have proved the feasibility of edge intelligence by applying an edge intelligence paradigm to practical applications. Yi et al. implement a face recognition application across a smartphone and edge server [25]. Results show that the latency is reduced from 900ms to 169ms, compared with cloud based paradigm. Ha et al. use a cloudlet to help a wearable cognitive assistance execute recognition tasks, which saves energy consumption by 30%-40% [26]. Some researchers pay attention to the performance of AI in the context of edge computing. Lane et al. successfully implement a constrained DL model on smartphones for activity recognition [27]. The demo achieves a better performance than shallow models, which demonstrates that ordinary smart devices are qualified for simple DL models. Similar verification is also done on wearable devices [28] and embedded devices [29]. The most famous edge intelligence application is Google G-board, which uses federated learning [30] to collaboratively train the typing prediction model on smartphones. Each user uses their own typing records to train G-board. Hence, the trained G-board could be used immediately, powering experiences personalised by the way users use this application.

The aim of our survey is to provide a comprehensive introduction to the development and the state-of-the-art of edge intelligence. As far as we know, there exist few recent efforts in this direction, but they are very different from ours. Table I summarizes the comparison among these works. Specifically, Yang et al. provide a survey on federated learning, in which they mainly focus on the architecture and applications of federated learning [31]. The authors divide literature of federated learning into three classifications: horizontal federated learning, vertical federated learning, and federated transfer learning. Federated learning is also involved in our survey. For federated learning part in our survey, we focus on the challenges and solutions of how to realize it at edge. The focus of [32] is how to realize the training and inference of DL models on a single mobile device. They briefly introduce some challenges and existing solutions from the perspective of training and inference. By contrast, we provide a more comprehensive and deeper review on solutions from the perspective of model design, model compression, and model acceleration. We also survey how to realize model training and inference with collaboration of edge devices and edge servers, even the assistance from the cloud server, in addition to solo training and inference at edge. Mohammadi et al. review works on how to use DL method for IoT big data analytic [33]. They briefly mention the relationship between cloud-based and edge-based DL algorithms for IoT data analytic and summarize challenges of realization. The survey [34] mainly focus on how to use DL techniques to deal with the problems in wireless environment, including edge computing problems. They also introduce the hardware and software foundation of edge intelligence.
To our best knowledge, [35] is the most relevant article to our survey. The focus of [35] is the inter-availability between edge computing and DL. Hence, the work involves two parts: DL for edge computing, and edge computing for DL. The former part focuses on some optimisation problems at edge with DL approaches, whilst the latter part focuses on applying DL in the context of edge computing (i.e., techniques to perform DL at edge). By contrast, we pay more attention to the implementation of AI based applications and services (including ML and DL) with the assistance of edge resources. More specifically, not only the implementation of AI based applications and services (including both training and inference), but also the management of edge data and the required computing power are involved in our work. We provide more comprehensive and detailed classification and comparison on existing works of this research area from multi-dimensions. Moreover, statistically, there are only 40 coincident surveyed papers between [35] and our work.

In our survey, we focus on how to realise systematic edge intelligence. There are three key components in AI: data, model/algorithm, and computation. The complete process of implementing AI applications should involve data collection and management, model training, and model inference. Computation is essential throughout the whole process. Hence the scope of our survey involves how to cache data to fuel intelligent applications (i.e., edge caching), how to train intelligent applications at the edge (i.e., edge training), how to infer intelligent applications at the edge (edge inference), and how to provide sufficient computing power for intelligent applications at the edge (edge offloading). Our contributions are summarized as following:

- We survey recent research achievements on edge intelligence and identify four key components: edge caching, edge training, edge inference, and edge offloading. For each component, we outline a systematical and comprehensive classification from a multi-dimensional view, e.g., practical challenges, solutions, optimisation goals, etc.
- We present thorough discussion and analysis on relevant papers in the field of edge intelligence from multiple views, e.g., applicable scenarios, methodology, performance, etc. and summarise their advantages and shortcomings.
- We discuss and summarise open issues and challenges in the implementation of edge intelligence, and outline five important future research directions and development trends, i.e., data scarcity, data consistency, adaptability of model/algorithms, privacy and security, and incentive mechanisms.

The rest of this article is organized as follow. Section II overviews the research on edge intelligence, with considerations of the essential elements of edge intelligence, as well as the development situation of this research field. We present detailed introduction, discussion, and analysis on the development and recent advances of edge caching, edge training, edge inference, and edge offloading in Section III to Section VI, respectively. Finally, we discuss the open issue and possible solutions for future research in Section VII, and conclude the paper in Section VIII.

### II. Overview

As an emerging research area, edge intelligence has received wide interest in the past few years. Such development benefits from the contribution of edge computing and artificial intelligence techniques, whose combination makes intelligent applications much closer to users and less dependent on the centralised cloud. For convenience, we show the comparison between traditional centralised intelligence with edge intelligence from the perspective of implementation in Fig. [1]. Traditional centralised intelligence is shown in Fig. [1(a)], where all edge devices first upload data to the central server for intelligent tasks, e.g., model training or inference. The central server/data-centre is usually, but not necessarily, located in the cloud. The central server/data-centre is usually in remote cloud. After the processing on the central server, results, e.g., recognition or prediction results, are transmitted back to edge devices. Fig. [1(b)] demonstrates the implementation of edge intelligence, where a task, e.g., recognition and prediction is either done by edge servers and peer devices, or with the edge-cloud cooperation paradigm. A very small amount, or none of the data is uploaded to the cloud. For example, in area (1) and (2), cloudlet, i.e. BS and IoT gateway could run complete intelligent models/algorithms to provide services for

---

1 model and algorithm are interchangeable in this article
edge devices. In area (3), a model is divided into several parts with different functions, which are performed by several edge devices. These edge devices work together to finish the task. Area (4) presents the edge-cloud cooperation paradigm, where edge servers first pre-processes the raw data received from vehicles, e.g., extracts features, and then uploads intermediate result to a cloud server for further processing.

It is known that three most important elements for an intelligent application are: data, model, and computation. Suppose that an intelligent application is a ‘human’, model would be the ‘body’, and computation is the ‘heart’ which powers the ‘body’. Data is then the ‘book’. The ‘human’ improves their abilities by learning knowledge extracted from the ‘book’. After learning, the ‘human’ starts to work with the learned knowledge. Correspondingly, the complete deployment of most intelligent applications (unsupervised learning based application is not included) includes three components: data collection and management (preparing the ‘book’), training (learning), and inference (working). Computation is a hidden component that is essential for the other three components. Combined with an edge environment, these three obvious components turn into edge cache (data collection and storage at edge), edge training (training at edge), and edge inference (inference at edge), respectively. Note that edge devices and edge servers are not powerful. Computation at edge usually is done via offloading. Hence, the hidden component turn into edge offloading (computation at edge). Our classification is organised around these four components, each of which features multidimensional analysis and discussion. The global outline of our proposed classification is shown in Fig. 2. For each component, we identify key problems in practical implementation and further break down these problems into multiple specific issues to outline a tiered classification. Next, we present an overview of these modules shown as Fig. 2.

A. Edge Caching

In edge intelligence, edge caching refers to a distributed data system, which collects and stores the data generated by edge devices and surrounding environments, and the data received from the Internet to support intelligent applications for users at the edge. Fig. 3 shows the illustration of edge caching. Data is distributed at the edge. For example, mobile users’ information generated by themselves is stored in their smartphones. Edge devices such as monitoring devices and sensors record the environmental information. Such data is stored at reasonable places and used for processing and analysis by intelligent algorithms to provide services for end users. For example, the video captured by cameras could be cached on vehicles for aided driving [36]. BS caches the data that users recently accessed from the Internet to characterise users’ interests for better a recommendation service [37]. To implement edge caching, we answer three questions: (i) what to cache, (ii) where to cache, and (iii) how to cache. The structure of this section is organised as the bottom module in Fig. 2.

For the first problem, what to cache, we know that caching is based on the redundancy of requests. In edge caching, the collected data is inputted into intelligent applications and results are sent back to where data is cached. Hence, there are two kinds of redundancy: data redundancy and computation redundancy. Data redundancy, also named communication redundancy, means that the inputs of an intelligent application may be the same or partially the same. For example, in continuous mobile vision analysis, there are large amounts of similar pixels between consecutive frames. Some resource-constrained edge devices need to upload collected videos to edge servers or the cloud for further processing. With cache, edge devices only needs to upload different pixels or frames. For the repeated part, edge devices could reuse the results to avoid unnecessary computation. Ref. [38–41] have
investigated the pattern of data redundancy. Caching based on such redundancy could effectively reduce computation and accelerate the inference. Computation redundancy means that the requested computing tasks of intelligent applications may be the same. For example, an edge server provides image recognition services for edge devices. The recognition tasks from the same context may be the same, e.g., the same tasks of flower recognition from different users of the same area. Edge servers could directly send the recognition results achieved previously back to users. Such kind of caching could significantly decrease computation and execution time. Some practical applications based on computation redundancy are developed in [42]–[44].

For the second problem, where to cache, existing works mainly focus on three places to deploy caches: macro BSs, micro BSs, and edge devices. The work in [45], [46] have discussed the advantages of caching at macro BSs from the perspective of covering range and hit probability. Some researchers also focus on the cached content at macro BSs. According to statistics, two kinds of content are considered: popular files [47]–[51] and intelligent models [52]–[56]. In edge intelligence, macro BSs usually work as edge servers, which provide intelligent services with cached data. In addition, some works [14], [57]–[59] consider how to improve the performance of caching with the cooperation among macro BSs. Compared with macro BSs, micro BSs provide smaller coverage but higher quality of experience [60]–[66]. Existing efforts on this area mainly focus on two problems: how to deliver the cached content, and what to cache. For the aspect of delivery, research mainly focuses on two directions: delivery from single BS [67], and delivery from multiple BSs based on the cooperation amongst them [68]–[72]. Considering the small coverage of micro BSs and the mobility of mobile users, research on handover and users’ mobility for better delivery service [73]–[76] is also carried out. In addition, the optimal content to cache, i.e., data redundancy based content [77]–[84] and computation redundancy based content [42], [85]–[91] is thoroughly investigated. Edge devices are usually of limited resources and high mobility, compared with macro BSs and micro BSs. Therefore, only few efforts pay attention to the problem of caching on a single edge device. For example, [36], [36], [92]–[94] studies the problem of what to cache based on communication and computation redundancy in some specific applications, e.g., computer vision. Most researchers adopt collaborative caching amongst edge devices, especially in the network with dense users. They usually formulate the caching problem into an optimisation problem on the content replacement [95]–[104], association policy [73], [105]–[107], [107]–[110], and incentive mechanisms [111], [112].

Since the storage capacity of macro BSs, micro BSs, and edge devices is limited, the content replacement must be considered. Works on this problem focus on designing replacement policies to maximise the service quality, such as popularity based schemes [113], [114], and ML based schemes [113], [115].

B. Edge Training

Edge training refers to a distributed learning procedure that learns the optimal values for all the weights and bias, or the hidden patterns based on the training set cached at the edge. For example, Google develops an intelligent input application, named G-board, which learns user’s input habits with the user’s input history and provides more precise prediction on the user’s next input [30]. The architecture of edge training is shown as Fig. 3. Different from traditional centralised training procedures on powerful servers or computing clusters, edge training usually occurs on edge servers or edge devices, which are usually not as powerful as centralised servers or computing clusters. Hence, in addition to the problem of training set (caching), four key problems should be considered for edge training: (i) how to train (the training architecture), (ii) how to make the training faster (acceleration), (iii) how to optimise the training procedure (optimisation), and (iv) how to estimate the uncertainty of the model output (uncertainty estimates).

The structure of this section is organised as the left module in Fig. 3. For the first problem, researchers design two training architectures: solo training [27]–[29], [116] and collaborative training [30], [117]–[120]. Solo training means training tasks are performed on a single device, without assistance from others, whilst collaborative training means that multiple devices cooperate to train a common model/algorithm. Since solo training has a higher requirement on the hardware, which is usually unavailable, most existing literature focuses on collaborative training architectures.

Different from centralised training paradigms, in which powerful CPUs and GPUs could guarantee a good result with a limited training time, edge training is much slower. Some researchers pay attention to the acceleration of edge training. Corresponding to training architecture, works on training acceleration are divided into two categories: acceleration for solo training [116], [121]–[124], and collaborative training [120], [125].

Solo training is a closed system, in which only iterative computation on single devices is needed to get the optimal parameters or patterns. In contrast, collaborative training is based on the cooperation of multiple devices, which requires periodic communication for updating. Update frequency and
update cost are two factors which affect the performance of communication efficiency and training result. Researchers on this area mainly focus on how to maintain the performance of the model/architecture with lower update frequency \[52\], \[126\], \[126\]–\[131\], and update cost \[52\], \[125\], \[132\]–\[135\]. In addition, the public nature of collaborative training is vulnerable to malicious users. There is also some literature which focuses on the privacy \[127\], \[136\]–\[144\] and security \[149\]–\[151\], \[151\]–\[154\] issues.

In DL training, the output results may be erroneously interpreted as model confidence. Estimating uncertainty is easy on traditional intelligence, whilst it is resource-consuming for edge training. Some literature \[155\], \[156\] pays attention to this problem and proposes various kinds of solutions to reduce computation and energy consumption.

We also summarised some typical applications of edge training \[30\], \[52\], \[126\], \[127\], \[157\]–\[162\], \[162\]–\[166\] that adopt the above-mentioned solutions and approaches.

C. Edge Inference

Edge inference is the stage where a trained model/architecture is used to infer the testing instance by a forward pass to compute the output on edge devices and servers. For example, developers have designed a face verification application based DL, and employ on-device inference \[167\], \[168\], which achieves high accuracy and low computation cost. The architecture of edge inference is shown as Fig. 5. Most existing AI models are designed to be implemented on devices which have powerful CPUs and GPUs, this is not applicable in an edge environment. Hence, the critical problems of employing edge inference are: (i) how to make models applicable for their deployment on edge devices or servers (design new models, or compress existing models), and (ii) how to accelerate edge inference to provide real-time responses. The structure of this section is organised as the right module in Fig. 5.

For the problem of how to make models applicable for the edge environment, researchers mainly focus on two research directions: design new models/architectures that have less requirements on the hardware, naturally suitable for edge environments, and compress existing models to reduce unnecessary operation during inference. For the first direction, there are two ways to design new models: let machines themselves design optimal models, i.e., architecture search \[169\], \[170\], \[170\]–\[173\]; and human-invented architectures with the application of depth-wise separable convolution \[174\]–\[176\] and group convolution \[177\], \[178\]. We also summarise some typical applications based on these architectures, including face recognition \[167\], \[168\], \[179\], human activity recognition (HAR) \[180\]–\[188\], vehicle driving \[189\]–\[192\], and audio sensing \[193\], \[194\]. For the second direction, i.e., model compression, researchers focus on compressing existing models to obtain thinner and smaller models, which are more computation- and energy-efficient with negligible or even no loss on accuracy. There are five commonly used approaches on model compression: low-rank approximation \[195\]–\[200\], knowledge distillation \[201\]–\[209\], compact layer design \[210\]–\[218\], network pruning \[219\]–\[228\], and parameter quantisation \[196\], \[229\]–\[244\]. In addition, we also summarise some typical applications \[245\]–\[250\] that are based on model compression.

Similar to edge training, edge devices and servers are not as powerful as centralised servers or computing clusters. Hence, edge inference is much slower. Some literature focuses on solving this problem by accelerating edge inference. There
are two commonly used acceleration approaches: hardware acceleration and software acceleration. Literature on hardware acceleration [93], [251]–[260], [260]–[274] mainly focuses on the parallel computing which is available as hardware on devices, e.g., CPU, GPU, and DSP. Literature on software acceleration [36], [92], [93], [275]–[282] focus on optimising resource management, pipeline design, and compilers, based on compressed models.

D. Edge offloading

As a necessary component of edge intelligence, edge offloading refers to a distributed computing paradigm, which provides computing service for edge caching, edge training, and edge inference. If a single edge device does not have enough resource for a specific edge intelligence application, it could offload application tasks to edge servers or other edge devices. The architecture of edge offloading is shown as Fig. 6. Edge offloading layer transparently provides computing services for the other three components of edge intelligence. In edge offloading, Offloading strategy is of utmost importance, which should give full play to the available resources in edge environment. The structure of this section is organised as the top module in Fig. 2.

Available computing resources are distributed in cloud servers, edge servers, and edge devices. Correspondingly, existing literature mainly focuses on four strategies: device-to-cloud (D2C) offloading, device-to-edge server (D2E) offloading, device-to-device (D2D) offloading, and hybrid offloading. Works on the D2C offloading strategy [283]–[298] prefer to leave pre-processing tasks on edge devices and offload the rest of the tasks to a cloud server, which could significantly reduce the amount of uploaded data and latency. Works on D2E offloading strategy [19], [299]–[304] also adopt such operation, which could further reduce latency and the dependency on cellular network. Most works on D2D offloading strategy [305]–[313] focus on smart home scenarios, where IoT devices, smartwatches and smartphones collaboratively perform training/inference tasks. Hybrid offloading schemes [314]–[316] have the strongest ability of adaptiveness, which makes the most of all the available resources.

We also summarise some typical applications that are based on these offloading strategies, including intelligent transportation [317], smart industry [318], smart city [319], and healthcare [320]–[321].

E. Summary

In our survey, we identify four key components of edge intelligence: edge caching, edge training, edge inference, and edge offloading. Although recently emerged, edge intelligence shows an explosive developing trend. Huge amount of researches have been carried out to investigate and realise edge intelligence over the past five years. We count the publication volume of edge intelligence, as shown in Fig. 7.

We see that this research area started from 2011, then grew at a slow pace before reaching 2014. Most strikingly, after 2014, there is a rapid rise in the publication volume of edge training, edge inference, and edge offloading. Meanwhile, the publication volume of edge caching is gradually winding down. Overall, the publication volume of edge intelligence is booming, which demonstrates a research field replete with activity. Such prosperity of this research filed owes to the following three reasons.

First, it is the booming development of intelligent techniques, e.g., deep learning and machine learning techniques that provides a theoretical foundation for the implementation of edge intelligence [322]–[324]. Intelligent techniques achieve state-of-the-art performance on various fields, ranging from voice recognition, behaviour prediction, to automatic piloting. Benefiting from these achievements, our life has been dramatically changed. People hope to enjoy smart services anywhere and at any time. Meanwhile, most existing intelligent services are based on cloud computing, which brings inconvenience for users. For example, more and more people are using voice assistant on smartphone, e.g., MI AI and Apple Siri. However, such applications can not work without networks.
Second, the increasing big data distributed at the edge, which fuels the performance of edge intelligence [325], [327]. We have entered the era of IoT, where a giant amount of IoT devices collect sensory data from surrounding environment day and night and provide various kinds services for users. Uploading such giant amount of data to cloud data centres would consume significant bandwidth resources. Meanwhile, more and more people are concerned about privacy and security issues behind the uploaded data. Pushing intelligent frontiers is a promising solution to solve these problems and unleash the potential of big data at the edge.

Third, the maturing of edge computing systems [14], [16] and peoples’ increasing demand on smart life [328], [329] facilitate the implementation of edge intelligence. Over the past few years, the theories of edge computing have moved towards application, and various kinds of applications have been developed to improve our life, e.g., augmented reality [330], [331], [332]. At the same time, with the wide spreading of 5G networks, more and more IoT devices are implemented to construct smart cities. People are increasingly reliant on the convenient service provided from a smart life. Large efforts from both academia and industry are enacted to realise these demands.

III. EDGE CACHING

Initially, the concept of caching comes from computer systems. The cache was designed to fill the throughput gap between the main memory and registers [333] by exploring correlations of memory access patterns. Later, the caching idea was introduced in networks to fill the throughput gap between core networks and access networks. Nowadays, the cache is deployed in edge devices, like various base stations and end devices. By leveraging the spatiotemporal redundancy of communication and computation tasks, caching at the edge can significantly reduce transmission and computation latency and improve users’ QoE [334], [335].

From existing studies, the critical issues of caching technologies in edge networks fall into three aspects: the cached content, caching places, and caching strategies. Next, we discuss and analyse relevant literature in edge caching in terms of the above three perspectives. The related subjects include the preliminary of caching, cache deployment, and cache replacement.

A. Preliminary of Caching

The critical idea of edge caching technologies is to explore the spatiotemporal redundancy of users’ requests. The redundancy factor largely determines the feasibility of caching techniques. Generally, there are two categories, i.e., communication redundancy, and computation redundancy.

1) Communication Redundancy: Communication redundancy is caused by the repetitive access of popular multimedia files, such as audio, video, and webpages. Content with high popularity tends to be requested by mobile users many times. Thus, the network needs to transmit the content to these users over and over again. In this case, caching popular content at edge devices can eliminate enormous duplicated transmissions from core networks.

To better understand communication redundancy, many existing studies investigate content request patterns of mobile users. For example, Crane et al. [38] analyse 5 million videos on YouTube and regards the number of daily views as the popularity of videos. They then discover four temporal popularity patterns of videos and demonstrate the temporal communication redundancy of popular videos from the angle of content providers’. Meanwhile, Adar et al. [39] analyse five weeks of webpage interaction logs collected from over 612,000 users. They show that temporal revisitation also exists at the individual level. In [40], Traverso et al. characterise the popularity profiles of YouTube videos by using the access logs of 60,000 end-users. They propose a realistic arrival process model, i.e., the Shot Noise Model (SNM), to model the temporal revisitation of online videos. Moreover, Dernbach et al. [41] exhibit the existence of regional movie preferences, i.e., the spatial communication redundancy of content by analysing the MovieLens dataset, which contains 6,000 user viewing logs. Consequently, the above studies apply large-scale and real-world datasets demonstrating communication redundancy from both the temporal and spatial dimensions. These studies support the feasibility of edge caching ideas and pave the way for employing edge caching technologies in real-world scenarios.

2) Computation Redundancy: Computation redundancy is caused by commonly used high computational complexity applications or AI models. In the wave of AI, we are now surrounded by various intelligent edge devices such as smartphones, smart watches, and smart brands. These intelligent edge devices provide diverse applications to augment users’ understanding of their surroundings [337], [338]. For example, speech-recognition based AI assistants, e.g., Siri and Cortana, and song identification enabled music applications, have been widely used in peoples’ daily lives. Such AI-based applications are of high computational complexity and cause high power consumption of the device [339], [340].

Meanwhile, some researchers have discovered the computation redundancy in AI-based applications. For example, in a park, nearby visitors may use their smartphones to recognise the same flowers and then search the information accordingly. In this case, there are a lot of unnecessary computations across devices. Therefore, if we offload such a painting recognition task to edges and cache the computation results, redundant computations can be further eliminated [42]. In [42], Guo et al. crawl street views by using the Google Streetview API [43] and builds an ‘outdoor input image set’. They then find that around 83% of the images exhibit redundancy, which leads to a large number of potential unnecessary computations for image reorganisation application. Also, they analyse NAVVIS [44], an indoor view dataset, and observed that nearly 64% of indoor images exhibit redundancy. To our knowledge, this work is the earliest one using real-world datasets to demonstrate the existence of computation redundancy.

In the elimination of computation redundancy, an important step is to capture and quantify the similarity of users’ requests. As shown in Fig. 8 in the case of communication redundancy,
Fig. 8. Illustration of image recognition (IR) cache, as well as the comparison with traditional web cache. The request/input for IR is an image. The system run the trained model to recognize the image, which will be labelled with an identifier. Then, the recognized image’s identifier is used to find relevant content within cache. If there is a cache hit, content is returned. Otherwise, IR modelling is performed and the result is placed in the cache.

A unique identifier can identify users’ request content, e.g., Universal Resource Identifier, URI. However, for computation redundancy, we first need to obtain the features of users’ requests and then find the best match of the computation results according to the extracted features. It is notable that in computation redundancy, the cached content is computation results instead of requested files.

B. Cache Deployment

As shown in Fig. 9 in edge networks, there are three main places to deploy cache units, i.e., macro base stations, small base stations, and end devices. Caching at different places have different characteristics, and we now provide a detailed discussion.

1) Caching at Macro Base Stations: The main purpose of deploying cache units in macro base stations (MBSs) is to relax the burden of backhaul [45] by exploring the communication redundancy and cache machine learning models by reducing the computation redundancy. Caching popular files and models in MBSs, content could be directly fetched from MBSs instead of core networks. In this way, redundant transmission and computation are eliminated. Compared with other cache places in the edge networks, MBSs have the most extensive coverage range and the most massive cache spaces. The typical coverage radius of an MBS is nearly 500 meters [46]. Due to its broad coverage range and vast cache spaces, MBS can serve more users. Thus, caching at MBSs can explore more of both communication redundancy and computation redundancy, and obtain a better cache hit probability. Also, since MBSs are deployed by operators, the topology structure of MBSs is stable and does not change over time. The comparison of different cache places is summarised in Table II.

Some researchers first studied the most fundamental problem, i.e., what files should be cached at MBSs to improve end-users’ QoE. One straightforward idea is to cache the most popular files. However, in [47], Blaszczyzyn et al. find that always caching the most popular contents maybe not the optimal strategy. Specifically, they assume the distribution of MBSs follows a Poisson Poisson point process [48], and both users’ requests and the popularity of content follow Zipf’s law [49]. By maximising users’ cache hit ratio, they derived the optimal scheme, in which less popular contents are also cached. Furthermore, Ahlehagh and Dey [50] take user preference profiles into account, and the videos of user-preferred video categories have a high priority for caching. Alternatively, Chatzieleftheriou et al. [51] explore the effect of content recommendations on the MBSs caching system. They discover that caching the recommended content can improve the cache hit ratio. Still, the recommendations of service providers may distort user preferences.

Apart from popular files, caching machine learning models at MBSs is another promising field. In 2016, Google proposed a novel machine learning framework, i.e., federated learning with the objective of training a global model, e.g., deep neural networks, across multiple devices [52]. In detail, the multiple devices train local models on local data samples. And they only exchange parameters, e.g., the weights of a deep neural network, between these local models and finally converge to a global model. Note that data is only kept in local devices and not exchanged across devices. Thus, federated learning can ensure data privacy and data secrecy and attract widespread attention from both industry [53] and academia [54]. Because MBSs can serve more mobile devices and have more powerful computation units, they are usually acted as the central server to orchestrate the different steps of the federated learning algorithm by caching the global model.
and collecting parameters from multiple devices [55], [56].

To fully explore the communication and computation redundancy, MBSs are allowed to cooperatively cache content, including files, computation results, and AI-models. In other words, a mobile user can be neighbouring base stations [57]. In [58], Peng et al. consider a collaborative caching network where all base stations are managed by a controller. They discover that contents with the highest popularity should be cached first in the case of long latency through backhaul network. Otherwise, caches should keep content diversity, i.e., to cache as much different content as possible. Also, Tuyen et al. [14] propose a collaborative computing framework across multiple MBSs. The computation workload at different base stations usually exhibits spatial diversity [59]. Therefore, offloading computation tasks to nearby idle MBSs and cache the computation results or trained-AI models cooperatively can facilitate the performance of mobile networks.

2) Caching at Small Base Stations: Small base stations (SBSs) are a set of low-powered mobile access points that have a range of 20 metres to 200 metres, e.g., microcell, picocell, and femtocell [60]. By deploying small base stations on hot spots, mobile users will have a better quality of experience, such as high end-rate, due to the benefit from spatial spectrum reuse [61], [341]. Therefore, densely deploying small base stations [62], [63] is a promising approach in future mobile networks and also brings huge potential to reduce both communication and computation redundancy by caching at SBSs [64]–[66].

In [342], Bacstug et al. use a stochastic geometry method to model the caching network of small cells, where users’ most popular files or computation results are stored on SBSs. They theoretically demonstrate that employing storage units in SBSs indeed brings gains in terms of the average delivery rate. Unlike [342] which considers a centralised control method, Chen et al. [343] propose a distributed caching strategy where each SBS only considers local users’ request pattern instead of global popularity. Each SBS maintains a content list by sampling requested files.

Compared with MBS, SBS has less redundancy potential due to its smaller coverage range, the number of served users, and even cache spaces. Thus, various technologies are adopted to explore caching benefits fully. A commonly used technique is the cooperative transmission. In SBS networks, the coverage area of SBSs is usually overlapped with each other, especially for the dense deployment scenario. In other words, a mobile user is able to receive content from multiple SBSs, making cooperative caching of multiple SBSs possible. There are a lot of transmission technologies applied in SBS networks, such as multicast, beamforming, cooperative multi-point (CoMP), and so on.

In [344], Liao et al. investigate caching enabled SBS networks. By exploring the potential of multicast transmission from SBSs to mobile users, their approach could reduce the backhaul cost in SBS networks. Similarly, Poularakis et al. [345] also delve into multicast opportunities in cache-enabled SBS networks. Different from [344], they assume that both MBSs and SBSs are able to use multicast. Each SBS can create multicast transmissions to end-users, while each MBS can provide multicast of popular content to SBSs within the coverage area. In terms of simulation results, the serving cost reduction up to 88% compared to unicast transmissions.

In SBS networks, since a mobile user has the potential to receive signals from multiple SBSs, there are two association cases, as shown in Fig. 10. In the first case, the mobile user is only associated with one SBS. Alternatively, the mobile user is associated with multiple SBSs. By using the beamforming [346] and CoMP technology [347], the associated SBSs can jointly send content for downlink transmission.

In [67], Pantisano et al. consider the first association case and presents a cache-aware user association approach. They adjust users’ association to improve the local cache hit ratio based on whether the associated base station contains files and AI models required by the user. The problem is formulated as a one-to-many matching game and they propose a distributed algorithm based on the deferred acceptance scheme to solve it.

Alternatively, some scholars focus on the second association case and explore the power of cooperative transmission in cache-enabled SBS networks [68]–[72]. In [68], Shanmugam et al. study the problem of content deployment in SBS networks. They assume that mobile users could communicate with multiple cache units and formulate the optimisation problem to minimise the average downloading delay. Liu et al. [69] explore the potential of energy efficiency of cache-enabled SBS networks where all SBSs use CoMP to transmit cached content cooperatively. By maximising energy efficiency, the optimal transmit power of SBSs are worked out. In [70] and [71], Ao et al. study a distributed caching strategy in SBS networks where CoMP technology is applied. In the CoMP enabled networks, caching strategy can bring two different gains. On the one hand, diverse contents could be cached in nearby BSs to maximise the cache hit. On the other hand, caching the same content in nearby BSs can let corresponding BSs transmit concurrently and bring multiplexing gain. By trading off both gains, they devised a near-optimal strategy to maximise the system throughput. Moreover, they find when content is of skewed popularity distribution, caching multiple copies of popular files and AI models yields larger caching gains. Further, Chen et al. [72] consider a similar system.
Unlike [70] and [71], where all nearby SBSs can employ CoMP, Chen et al. group SBSs into multiple disjointed clusters and only the SBSs in the same cluster are able to transmit content cooperatively. To trade off the parallel transmission and joint transmission, they divide the cache space into two parts. One is in charge of caching content with less popularity to improve content diversity, while the other is used to cache contents with the highest popularity. Then, they optimize the problem of space assignment.

Since the coverage range of each SBS is too small, mobile users will go through multiple SBSs within a short time, as shown in Fig. 11. This frequent handover behaviour will cause the degradation of caching performance. In [73], Krishnan et al. investigate the retransmission in cache-enabled SBS networks. Because of the frequent handover between different SBSs, sometimes, when none of the SBSs in the user’s vicinity has cached the requested file or AI models, the file transmission will be interrupted. Nevertheless, the retransmission will be triggered when the requested file or AI models are cached at vicinity SBSs. By using stochastic geometry to analyse the cache hit probability, Krishnan et al. find that SBSs should cache content diversely for mobile users. In [74], Guan et al. assume that users’ preferences for content and mobility patterns are known prior, and users’ preferences remain constant over a short period. They then formulate an optimisation problem with the objective of maximising the utility of caching and devise a heuristic caching strategy. In [75] and [76], the same caching system model is investigated where mobile users migrate between multiple SBSs. Due to the limited transmission time, users may not be able to download the complete requested files or parameters of AI models from the associated SBS, and the requests can be redirected to MBS. In [75], Poularakis et al. use random walks to model user movements and formulated an optimisation problem based on the Markov chain aiming to maximize the probability of serving by SBSs. They further propose two caching strategies, i.e., a centralised solution for the small-scale system and a distributed solution for the large-scale system. Unlike [75], Ozfatura et al. propose a distributed greedy algorithm to minimise the amount of data downloaded from MBS [76]. Requests with deadlines below a given threshold are responded by SBSs while other requests are served by MBS.

In SBS caching systems, predicting users’ requests to improve the cache hit ratio is also a commonly used method, which falls into the field of artificial intelligence applications. By applying AI technology to historical users’ request logs, we can profile user preference or content popularity patterns. Next, we can predict users’ requests or content popularity, respectively [77]. In [78], Kader et al. design a big data platform and collects mobile traffic data from a telecom operator in Turkey. They then used collaborative filtering, a common machine learning method, to estimate content popularity. The simulation results demonstrate that the caching benefits are further explored with the help of content popularity prediction. Similar to [78], in [79], Pantisano et al. also apply collaborative filtering to predict content popularity. They then devised a user-SBS association scheme based on estimated popularity and the current cache composition to minimise the backhaul bandwidth allocation. In [80], Bastug et al. focus on individual content request probability instead of global content request probability. They propose to use the Bayesian learning method to predict personal preferences and then incorporate this crucial information into the caching strategy. If we lack the historical data of user request logs, how can we predict content popularity? In [81], Bastug et al. investigate this open issue and proposes a transfer learning-based caching procedure. Specifically, they exploit contextual information, e.g., social networks, and referred to it as a source domain. Then the prior information in the source domain is incorporated in the target domain to estimate content popularity.

Also, SBSs are used to cache the data from end devices, like smartphones and IoT. In recent times, IoT devices are widely distributed in homes, streets, and even whole cities to allow users to monitor the ambient environment [82], [83]. By collecting and analysing the big data on IoT devices, a smarter physical world can be built. Considering the demand of real-time data analysis, caching and processing the data at the edge is a common and promising method. In [83], Quevedo et al. introduce a caching system for IoT data and prove that the caching system could reduce the energy consumption of IoT sensors. In [84], Sharma et al. propose a collaborative edge and cloud data processing framework for IoT networks where SBSs are in charge of caching IoT data, extracting useful features and uploading features to cloud part.

Meanwhile, since SBSs are often deployed at hot points, the requested computation tasks from served users will exhibit spatiotemporal locality. Therefore, by caching computation results at SBSs, the redundant computation tasks can be eliminated. Drolia et al. [85] propose a caching strategy, Cachier, to cache the recognition results on edge servers by release repetitive recognition computation. Specifically, Cachier first extracts features of the requested recognition task, and then tries to match a similar object from the cache. If there is a cache hit, the corresponding computation results would be sent back to the mobile device. Otherwise, the request would be sent to the cloud. To identify similar recognition tasks, they used a Locality Sensitive Hashing (LSH) algorithm [86] to determine the best match. Furthermore, to overcome the unbalanced and time-varying distribution of users’ requested tasks, Guo et al. [42] design an Adaptive LSH-Homogenized kNN
joint algorithm which outperforms LSH in terms of evaluation results. Drolia et al. further introduce a proactive caching strategy into their system by predicting the requirement of users and proactively caching parts of models on SBSs server for pre-processing to further reduce the latency [87]. Such a strategy is also used in [88] to deal with unstructured data at SBSs.

Moreover, in some task-fickle scenarios, multiple different kinds of tasks, e.g., voice recognition and object recognition, are offloaded from devices to SBSs. By pre-caching multiple kinds of deep learning models at SBSs for different kinds of tasks, we can reduce the computation time and further improve users’ QoE. Taylor et al. propose an adaptive model selection scheme to select the best model for users [89]. They use a supervised learning method to train a predictor model offline and then deploy it on an edge server. When a request arrives, the predictor will select an optimal model for the task. In [90], Zhao et al. propose a system, Zoo, to compose different models to provide a satisfactory service for users. Ogden et al. propose a deep inference platform, MODI, to determine what model to cache and what model to use for specific tasks [91]. There is a decision engine inside MODI, which aggregates previous results to decide what new models are required to cache.

3) Caching at Devices: Caching at devices exploits the available storage space of end equipment, like mobile phones and IoT devices. These devices can leverage the communication and computation redundancy locally. Furthermore, they can fetch the requested content or computation results from other devices in proximity through device-to-device (D2D) communication [348], [349].

First of all, end devices can explore the communication and computation redundancy locally. For instance, in some static continuous computer vision applications, such as monitoring, the captured consecutive images are similar to some extent. Therefore, the results of previous images could be reused for the latter inference. In [36], [92], [93], they cache the results of previous frames to reduce redundant computation and latency. In some mobile continuous computer vision applications, such as driving assistance, the system is required to provide high trackability. The system needs to recognise, locate, and label the tracked object, e.g., road signs, on the screen in real-time. The recognised object would repeatedly appear in multiple images for a period. Chen et al. develop an active cache based continuous object recognition system, called Glimpse, to achieve real-time trackability [94]. The structure of Glimpse is shown as Fig. 12. Glimpse caches frames locally and only uploads trigger frames to the cloud server. Trigger frames refers to the frames, for which the recognition from the server is different from current local tracking. The cloud server sends back the recognised object, its labels, bounding boxes, and features, which would be cached locally on devices. Then the devices would track the object with the labels, bounding boxes, and features locally on captured frames. A similar approach is also adopted in CNNCache [36].

On the other hand, compared with MBSs and SBSs, devices have very limited cache spaces and coverage range, due to the cost constraint of end devices and their low transmission power. Although these limitations seem to only take the small caching benefits for local devices, the situation will be changed when the networks are with dense users. The benefits of caching will be amplified with the number of users increases. In [350], Chen et al. study the difference between caching at SBSs and devices where content is cached according to a joint probability distribution. By applying stochastic geometry, they derive the closed-form expression of hit probability and request density. Although the cache hit probability of device caching is always lower than SBS caching due to the small cache spaces, the request density of device caching is much higher than SBS caching. This is because, in device caching, more concurrent links are allowed, compared with the case of SBS caching, especially in the dense user scenario. Similarly, in [351], Gregori et al. investigate caching at both devices and SBSs as well. However, they do not compare these two different scenarios and only design joint transmission and caching policies for them to minimise energy consumption separately.

In the device caching system, the number of coexisting D2D links affect the device caching performance dramatically. The D2D links are the fundamental requirement for end-devices sharing files, models, and computation results and further reducing communication and computation redundancy amongst devices. First of all, the establishment of a D2D connection depends on content placement. In other words, when a device discovers that the requested content is placed in a nearby device within the D2D transmission range, the D2D link can be built, and the content will be transmitted directly. Therefore, there are a lot of scholars trying to maximise caching performance by optimising the content placement [95]–[99]. In [98], Malak et al. model senders and receivers as members of Poisson Point Process and compute the probability of delivery in D2D networks. Considering the low transmission noise case, they find that the optimal content allocation could
be approximately achieved by Benford’s law when the path loss exponent equals 4. A similar system model is applied in [99] but with a different performance metric. In [99], Peng et al. analyse the outage probability of D2D transmission in cache-enabled D2D networks. They then obtain the optimal caching strategy by using a gradient descent method. In [95], Chen et al. aim to maximise successful offloading probability. Different from [93], [99] which did not consider the time divided transmission, Chen et al. divide time into multiple slots, and each transmitter independently chooses a time slot to transmit files. Employing the gradient descent method, they design a distributed caching policy. Unlike the above studies where each end-user applies the same caching strategy, Giatsoglou et al. [96] divide the 2K most popular contents into two groups of the same size. K is the cache capacity of each user. Then randomly allocate these two groups to users, i.e., some users cache group A, whilst others cache group B.

Apart from content placement, association policy is also important to the establishment of D2D links. In [105], Golrezaei et al. optimise the collaboration distance for D2D communications with distributed caching, where the collaboration distance is the maximum allowable distance for D2D communication. They assume each user employs a simple randomised caching policy. In [106], Naderializadeh et al. propose a greedy association method, i.e., greedy closest-source policy. In this association policy, starting from the first user, each user chooses the closest user with the desired file forming a D2D pair. They assume that each file is randomly cached in devices and derive a lower bound on the cached content reuse.

Generally, end devices are controlled by end consumers who are able to decide whether to cache and share content or not do so. Therefore, the incentive mechanism is introduced in D2D networks to encourage users to exploit the storage space of their equipment and share cached content, e.g., files, AI models and computation results, with other users. In [111], Chen et al. propose an incentive mechanism where the base station rewards the users who shares content with others via D2D communications. Since the base station will determine the reward to minimise its total cost while users would like to maximise their reward by choosing the caching policy, Chen et al. model this conflict as a Stackelberg game and proposed an iterative gradient algorithm to obtain the Stackelberg Equilibrium. In [112], Taghizadeh et al. consider a similar case where content providers pay the download cost to encourage users to download and share content. However, they do not model the conflict and merely design the caching strategy to minimise content provisioning costs.

Since end devices are bound up with users and affected by user attributes to some extent, some researchers focus on exploring the knowledge of user attributes, like social ties and interests, to assist device caching as shown in Fig. 13. In [108], Bastug et al. propose to let influential users cache content, such that these users could disseminate the cached contents to others through their social ties. The influential users are determined by their social networks. First, a social graph is built based on past action history of users’ encounters and file requests. Then, the influence of users is measured in terms of the centrality metric [109]. Apart from social ties, Bai et al. [110] consider users’ interests as well. They use the hypergraph to model the relationships among social ties, common interests, and spectrum resources and design an optimal caching strategy to maximise the cache hit ratio.

C. Cache Replacement

In practice, the request distribution of content varies with time, and new content is constantly being created. Hence, it is critical to update caches at intervals. The cache update process generally takes place when new content is delivered and needs to be cached, but all cache units are occupied. Hence, some cached old content needs to be replaced. Therefore, the cache update process is called cache replacement as well.

Several conventional cache replacement strategies have been proposed, such as first-in first-out (FIFO), least frequently used (LFU), least recently used (LRU), and their variants [561]. FIFO evicts the content in terms of cached time without any regard to how often or how many times it was accessed before. LFU keeps the most frequently requested content, while LRU keeps the most recently accessed content. However, these replacement strategies merely consider content request features in a short time window and may not obtain the global optimal solutions.

Another popular method is to replace the content based on its popularity. In [113], Blasco et al. divide time into periods and within each period there is a cache replacement phase. During each cache replacement phase, the content of the lowest popularity is discarded. Apart from historical popularity, Baustug et al. [114] take the future content popularity into consideration as well. They propose a proactive popularity caching (PropCaching) method to estimate content popularity and then determine which content should be evicted.

Mathematically, the cache replacement problem could be formulated as a Markov Decision Process (MDP) [114], [115]. The MDP model can be represented into a tuple \((S, A, R(s, a))\). S refers to the set of possible states for
| Ref. | Cache places | Performance metrics | Mathematical tools | Control methods | Transmission Cooperativity |
|------|--------------|---------------------|--------------------|-----------------|----------------------------|
| 47   | MBSs         | Hit probability     | Stochastic geometry| Centralised     | Non-cooperative            |
| 51   | MBSs         | Cache hit ratio     | Optimisation       | Centralised     | Non-cooperative            |
| 52-53| MBSs         | Energy efficiency   | Stochastic geometry| Centralised     | Non-cooperative            |
| 50   | MBSs         | The number of concurrent videos | Optimisation | Distributed | Non-cooperative |
| 58   | MBSs         | The average download delay | Optimisation | Centralised | Cooperative |
| 55   | MBSs         | Storage space       | Optimisation       | Centralised     | Cooperative               |
| 55   | MBSs         | Aggregate operational cost | Optimisation | Centralised | Cooperative               |
| 357  | MBSs         | The aggregated caching and download cost | Optimisation | Centralised | Cooperative               |
| 341  | SBSs         | Outage probability  | Contract theory    | Centralised     | Non-cooperative            |
| 342  | SBSs         | Content delivery rate | Contract theory | Centralised | Non-cooperative            |
| 343  | SBSs         | The cache service probability | Stochastic geometry | Distributed | Non-cooperative |
| 358  | SBSs         | The profit of NSP and VPs | Stochastic geometry | Centralised | Non-cooperative |
| 359  | SBSs         | The number of requests served by SBSs | Optimisation | Centralised | Non-cooperative            |
| 344  | SBSs         | Backhaul cost       | Optimisation       | Centralised     | Non-cooperative            |
| 345  | SBSs         | Servicing cost      | Optimisation       | Centralised     | Non-cooperative            |
| 67   | SBSs         | Caching utility     | Matching theory    | Distributed     | Non-cooperative            |
| 68   | SBSs         | Downloading time of files | Optimisation | Centralised | Cooperative |
| 69   | SBSs         | Energy efficiency   | Optimisation       | Centralised     | Cooperative               |
| 70   | SBSs         | System throughput   | Optimisation       | Distributed     | Cooperative               |
| 72   | SBSs         | Cache hit probability and energy efficiency | Stochastic geometry | Centralised | Cooperative |
| 73   | SBSs         | Cache hit probability | Stochastic geometry | Centralised | Non-cooperative            |
| 74   | SBSs         | Maxmise caching utility | Optimisation | Centralised | Non-cooperative            |
| 75   | SBSs         | The probability of response from MBS | Optimisation | Centralised & Distributed | Non-cooperative |
| 76   | SBSs         | The amount of data downloaded from MBS | Optimisation | Distributed | Non-cooperative |
| 78   | SBSs         | Backhaul load       | Machine learning   | Centralised     | Non-cooperative            |
| 79   | SBSs         | Backhaul bandwidth allocation | Machine learning | Distributed | Non-cooperative |
| 80   | SBSs         | System throughput   | Machine learning   | Centralised     | Non-cooperative            |
| 81   | SBSs         | Backhaul offloading gains | Machine learning | Centralised | Non-cooperative |
| 350  | Devices      | Density of cache-served requests | Stochastic Geometry | Distributed | Non-cooperative |
| 351  | Devices      | Energy consumption  | Optimisation       | Distributed     | Non-cooperative            |
| 98   | Devices      | The probability of successful content delivery | Stochastic Geometry | Distributed | Non-cooperative |
| 99   | Devices      | Outage probability  | Optimisation       | Distributed     | Non-cooperative            |
| 95   | Devices      | Offloading probability | Optimisation | Distributed | Non-cooperative |
| 96   | Devices      | Offloading gain     | Stochastic Geometry| Centralised     | Non-cooperative            |
| 105  | Devices      | Number of D2D links | Optimisation       | Distributed     | Non-cooperative            |
| 106  | Devices      | Spectral reuse      | Optimisation       | Distributed     | Non-cooperative            |
| 107  | Devices      | System throughput   | Optimisation       | Centralised     | Non-cooperative            |
| 108  | Devices      | Network throughput  | Optimisation       | Centralised     | Cooperative               |
| 109  | Devices      | Coverage probability | Stochastic Geometry| Centralised | Non-cooperative |
| 107  | Devices      | Service success probability | Stochastic Geometry | Distributed | Non-cooperative |
| 110  | Devices      | Coverage probability | Stochastic Geometry| Distributed | Non-cooperative |
| 111  | Devices      | Caching reward      | Game theory        | Distributed     | Non-cooperative            |
| 112  | Devices      | Content provision costs | Optimisation | Distributed | Non-cooperative |
| 108  | Devices      | Backhaul costs      | Graph Theory       | Centralised     | Non-cooperative            |
| 110  | Devices      | Cache hit ratio     | Graph Theory       | Centralised     | Non-cooperative            |
caches. \( A \) is the set of eviction actions. \( R(s, a) \) is the reward function that determines the reward when cache performs action \( a \) in the state \( s \). The reward is usually modelled as the cache hit or the changes in transmission cost. In [114], Bacstuug et al. obtain the cache replacement actions based on Q-learning. In [115], the method is upgraded. Wang et al. apply deep reinforcement learning to solve it.

### IV. Edge Training

The standard learning approach requires centralising training data on one machine, whilst edge training relies on distributed training data on edge devices and edge servers, which is more secure and robust for data processing. The main idea of edge training is to perform learning tasks where the data is generated or collected with edge computing resources. It is not necessary to send users’ personal data to a central server, which effectively solves the privacy problem and saves network bandwidth.

Training data could be solved through edge caching. We discuss how to train an AI model in edge environments in this section. Since the computing capacity on edge devices and edge servers is not as powerful as central servers, the training style changes correspondingly in the edge environment. The major change is the distributed training architecture, which must take the data allocation, computing capability, and network into full consideration. New challenges and problems, e.g., training efficiency, communication efficiency, privacy and security issues, and uncertainty estimates, come along with the new architecture. Next, we discuss these problems in more detail.

#### A. Training architecture

Training architecture depends on the computing capacity of edge devices and edge servers. If one edge device/server is powerful enough, it could adopt the same training architecture as a centralised server, i.e., training on a single device. Otherwise, cooperation with other devices is necessary. Hence, there are two kinds of training architectures: solo training, i.e., perform training tasks on a single edge device/server, and collaborative training, i.e., few devices and servers work collaboratively to perform training tasks.

1) Solo training: Early researchers mainly focus on verifying the feasibility of directly training deep learning models on mobile platforms. Chen et al. find that the size of neural network and the memory resource are two key factors that affect training efficiency [116]. For a specific device, training efficiency could be improved significantly by optimising the model. Subsequently, Lane et al. successfully implement a constrained deep learning model on smartphones for activity recognition and audio sensing [27]. The demonstration achieves a better performance than shallow models, which demonstrates that ordinary smart devices are qualified for simple deep learning models. Similar verification is also done on wearable devices [28] and embedded devices [29].

2) Collaborative training: The most common collaborative training architecture is the master-slave architecture. Federated learning [30] is a typical example, in which a server employs multiple devices and allocates training tasks for them. Li et al. develop a mobile object recognition framework, named DeepCham, which collaboratively trains adaptation models [117]. The DeepCham framework consists of one master, i.e., edge server and multiple workers, i.e., mobile devices. There is a training instance generation pipeline on workers that recognises objects in a particular mobile visual domain. The master trains the model using the training instance generated by workers. Huang et al. consider a more complex framework with additional scheduling from the cloud [118]. Workers with training instances first uploads a profile of the training instance and requests to the cloud server. Then, the cloud server appoints an available edge server to perform the model training.

Peer-to-peer is another collaborative training architecture, in which participants are equal. Valerio et al. adopt such training architecture for data analysis [119]. Specifically, participants first perform partial analytic tasks separately with their own data. Then, participants exchange partial models and refine them accordingly. The authors use an activity recognition model and a pattern recognition model to verify the proposed architecture and find that the trained model could achieve similar performance with the model trained by a centralised server. Similar training architecture is also used in [120] to enable knowledge transferring amongst edge devices.

#### B. Training Acceleration

Training a model, especially deep neural networks, is often too computationally intensive, which may result in low training efficiency on edge devices, due to their limited computing capability. Hence, some researchers focus on how to accelerate the training at edge. Table IV summarizes existing literature on training acceleration.

Chen et al. find that the size of a neural network is an important factor that affects the training time [116]. Some efforts [121], [122] investigate transfer learning to speed up the training. In transfer learning, learned features on previous models could be used by other models, which could significantly reduce the training time. Valery et al. propose to transfer features learned by the trained model to local models, which would be re-trained with the local training instances [121]. Meanwhile, they exploit the shared memory of the edge devices to enable the collaboration between CPU and GPU. This approach could reduce the required memory and increase computing capacity. Subsequently, the authors further accelerate the training procedure by compressing the model by replacing float-point with 8-bit fixed point [122].

In some specific scenarios, interactive machine learning (iML) [362], [363] could accelerate the training. iML engages users in generating classifiers. Users iteratively supply information to the learning system and observe the output to improve the subsequent iterations. Hence, model updates are more fast and focused. For example, Amazon often asks users targeted questions about their preferences for products. Their
preferences are promptly incorporated into a learning system for recommendation services. Some efforts [123], [124] adopt such approach in model training on edge devices. Shahmehammadi et al. apply iML on human activity recognition, and find that only few training instances are enough to achieve a satisfactory recognition accuracy [124]. Based on such theory, Flutura et al. develop DrinkWatch to recognise drink activities based on sensors on smartwatch [364].

In a collaborative training paradigm, edge devices are enabled to learn from each other to increase learning efficiency. Xing et al. propose a framework, called RecycleML, which uses cross modal transfer to speed up the training of neural networks on mobile platforms across different sensing modalities in the scenario that the labelled data is insufficient [120]. They design an hourglass model for knowledge transfer for multiple edge devices, as shown in Fig. 14. The bottom part denotes lower layers of multiple specific models, e.g., AudioNet, IMUNet, and VideoNet. The middle part represents the common layers of these specific models. These models project their data into the common layer for knowledge transfer. The upper part represents the task-specific layers of different models, which are trained in a targeted fashion. Experiments show that the framework achieves 50x speedup for the training.

C. Training optimisation

Training optimisation refers to optimising the training process to achieve some given objectives, e.g., energy consumption, accuracy, privacy-preservation, security-preservation, etc. Since solo training is similar to training on a centralised server to a large extent, existing work mainly focuses on collaborative training. Federated learning is the most typical collaborative training architecture, and almost all literature on collaborative training is relevant to this topic.

Federated learning is a kind of distributed learning [119], [365], [366], which allows training sets and models to be located in different, non-centralised positions, and learning can occur independent of time and places. This training architecture is first proposed by Google, which allows smartphones to collaboratively learn a shared model with their local training data, instead of uploading all data to a central cloud server [30]. The learning process of federated learning is shown as Fig. 15. There is an untrained shared model On the central server, which will be allocated training participants for training. Training participants, i.e., edge devices train the model with the local data. After local learning, changes of the model are summarised as a small focused update, which will be sent to the central server through encrypted communication. The central server averages received changes from all mobile devices and updates the shared model with the averaged result. Then, mobile devices download the update for their local model and repeats the procedure to continuously improve the shared model. In this learning procedure, only the encrypted changes are uploaded to the cloud and the training data of each mobile user remains on mobile devices. Transfer learning and edge computing are combined to learn a smarter model for mobile users. In addition, since learning occurs locally, federated learning could effectively protect user privacy, when compared with a traditional centralised learning approach.

Typical edge devices in federated learning are smartphones with unreliable and slow network connections. Moreover, due to the unknown mobility, these devices may be intermittently available for working. Hence, the communication efficiency between smartphones and the central server is of the utmost importance to the training. Specifically, there are factors affecting the communication efficiency: communication frequency...
and communication cost. In addition, the update from edge devices is vulnerable to malicious users. Hence privacy and security issues should also be considered. We discuss these problems in detail next. Table IV summarises literature on training optimisation.

1) Communication Frequency: In federated learning, communication between edge devices and the cloud server is the most important operation, which uploads the updates from edge devices to the central server and downloads the aggregated update from the shared model to local models. Due to the possible unreliable network condition of edge devices, minimising the number of update rounds, i.e., communication frequency between edge devices and cloud server is necessary. Jakub et al. are the first to deploy federated learning framework and propose the setting for federated optimisation [126]. In [52], the authors characterise the training data as massively distributed (data points are stored across massive edge devices), non-IID (training set on devices may be drawn from different distributions), and unbalanced (different devices have different number of training samples). In each round, each device sends an encrypted update to the central server. Then they propose a federated stochastic variance reduced gradient (FSVRG) algorithm to optimise the federated learning. They find that the central shared model could be trained with a small number of communication rounds.

McMahan et al. propose a federated averaging algorithm (FedAvg) to optimise federated learning in the same scenario with [52], [126] and further evaluate the framework with five models and four datasets to proof the robustness of the framework [127]. Although FedAvg could reduce the number of communication rounds for certain datasets, Zhao et al. find that using this algorithm to train CNN models with highly skewed non-IID dataset would result in the significant reduction of the accuracy [128]. They find the accuracy reduction results from the weight divergence, which refers to the difference of learned weights between two training processes with the same weight initialisation. Earth mover’s distance (EMD) between the distribution over classes on each mobile device and the distribution of population are used to quantify the weight divergence. They then propose to extract a subset of data, which is shared by all edge devices to increase the accuracy.

Strategies that reduce the number of updates should be on the premise of not compromising the accuracy of the shared model. Wang et al. propose a control algorithm to determine the optimal number of global aggregations to maximise the efficiency of local resources [129]. They first analyse the convergence bound of SGD based federated learning. Then, they propose an algorithm to adjust the aggregation frequency in real-time to minimise the resource consumption on edge devices, with the joint consideration of data distribution, model characteristics, and system dynamics.

Above-mentioned works adopt a synchronous updating method, where in each updating round, updates from edge devices are first uploaded to the central server, and then aggregated to update the shared model. Then the central server allocates aggregated updates to each edge device. Some researchers think that it is difficult to synchronise the process. On one hand, edge devices have significantly heterogeneous computing resources, and the local model are trained asynchronously on each edge device. On the other hand, the connection between edge devices and the central server is not stable. Edge devices may be intermittently available, or response with a long latency due to the poor connection. Wang et al. propose an asynchronous updating algorithm, called CO-OP through introducing an age filter [130]. The shared model and downloaded model by each edge device would be labelled with ages. For each edge device, if the training is finished, it would upload its update to the central server. Only when the update is neither obsolete nor too frequent, it will be aggregated to the shared model. However, most works adopt synchronous approaches in federated learning, due to its effectiveness [127], [131].

2) Communication cost: In addition to communication frequency, communication cost is another factor that affects the communication efficiency between edge devices and the central server. Reducing the communication cost could significantly save bandwidth and improve communication efficiency. Konevcny et al. propose and proof that the communication cost could be lessened through structured and sketched updates [52], [132]. The structured update means learning an update from a restricted space that could be parametrised with few variables through using low rank and random mask structure, while sketched update refers to compressing the update of the full model through quantisation, random rotations and subsampling.

Lin et al. find most of the gradient update between edge devices and the central server is redundant in SGD based federated learning [133]. Compressing the gradient could solve the redundancy problem and reduce the update size. However, compression methods, such as gradient quantisation and gradi-

| Ref. | Model | Approach | Learning method | Object | Performance |
|------|-------|----------|----------------|--------|-------------|
| 116  | DNN   | Hardware acceleration | Transfer learning | Review training factors | N/A |
| 121  | CNN   | Hardware acceleration | Transfer learning | Alleviate memory constraint | Faster than Caffe-OpenCL trained |
| 122  | CNN   | Hardware acceleration parameter quantisation | Transfer learning | Alleviate memory constraint | Faster than Caffe-OpenCL trained |
| 124  | RF, ET, NB, LR, SVM | Human annotation | Incremental learning | Investigate iML for HAR | 93.3% accuracy |
| 126  | Naive Bayes | Human annotation | Incremental learning | Reduce limitations in learning | 6-8 hours to train a model |
| 120  | CNN   | Software acceleration | Transfer learning | Reduce required labelled data | 50× faster |
| 125  | Statistical model | Software acceleration | Federated learning | Address statistical challenges | Outperform global, local manners |
ent sparsification would lead to the decreased accuracy. They propose a deep gradient compression (DGC) method to avoid the loss of accuracy, which use momentum correction and local gradient clipping on top of the gradient sparsification. Hardy et al. also try to compress the gradient and propose a compression algorithm, called AdaComp [134]. The basic idea of AdaComp is compute staleness on each parameter and remove a large part of update conflicts.

Smith et al. propose to combine multi-task learning and federated learning together, which train multiple relative models simultaneously [125]. It is quite cost-effective for a single model, during the training. They develop an optimisation algorithm, named MOCHA, for federated setting, which allows personalisation through learning separate but related models for each participant via multi-task learning. They also prove the theoretical convergence of this algorithm. However, this algorithm is inapplicable for non-convex problems.

Different from the client-to-server federated learning communication in [52], [133], [134], Caldas et al. propose to compress the update from the perspective of server-to-client exchange and propose Federated Dropout to reduce the update size [135]. In client-to-server paradigm, edge devices download the full model from the server, while in a server-to-client paradigm, each edge device only downloads a sub-model, which is a subset of the global shared model. This approach both reduces the update size and the computation on edge devices.

3) Privacy and security issues: After receiving updates from edge devices, the central server needs to aggregate these updates and construct an update for the shared global model. Currently, most deep learning models rely on variants of stochastic gradient descent (SGD) for optimisation. FedAvg, proposed in [127], is a simple but effective algorithm to aggregate SGD from each edge device through weighted averaging. Generally, the update from each edge device contains significantly less information of the users’ local data. However, it is still possible to learn the individual information of a user from the update [136]. If the updates from users are inspected by malicious hackers, participant edge users’ privacy would be threatened. Bonawitz et al. propose Secure Aggregation to aggregate the updates from all edge devices, which makes the participant updates un-inspectable by the central server [137]. Specifically, each edge device uploads a masked update, i.e., parameter vector to the server, and then the server accumulates a sum of the masked update vectors. As long as there is enough edge devices, the masks would be counteracted. Then, the server would be able to unmask the aggregated update. During the aggregation, all individual updates are non-inspectable. The server can only access the aggregated unmasked update, which effectively protect participants’ privacy. Liu et al. introduce homomorphic encryption to federated learning for privacy protection [138]. Homomorphic encryption [139] is an encryption approach that allows computation on ciphertexts and generates an encrypted result, which, after decryption, is the same with the result achieved through direct computation on the plain text. The central server could directly aggregate the encrypted updates from participants.

Geyer et al. propose an algorithm to hide the contribution of participants at the clients’ based on differential privacy [140]. Similar to differential privacy-preserving traditional approaches [141], [142], the authors add a carefully calibrated amount of noise to the updates from edge devices in federated learning. The approach ensures that attackers could not find whether an edge device participated during the training. A similar differential privacy mechanisms are also adopted in federated learning based recurrent language model and federated reinforcement learning in [143] and [144].

In federated learning, the participants could observe intermediate model states and contribute arbitrary updates to the global shared model. All aforementioned research assumes that the participants in federated learning are un-malicious, which provides a real training set and uploads the update based on the training set. However, if some of the participants are malicious, who uploads erroneous updates to the central server, the training process fails. In some cases, the attack would result in large economic losses. For example, in a backdoor attacked face recognition based authentication system, attackers could mislead systems to identify them as a person who can access a building through impersonation. According to their attack patterns, attacks could be classified into two categories: data-poisoning and model-poisoning attacks. Data-poisoning means compromising the behaviour and performance of the model through changing the training set, e.g., accuracy, whilst model-poisoning only change the model’s behaviour on specific inputs, without impacting the performance on other inputs.

The work in [145] tests the impact of a data-poisoning attack on SVM through injecting specially crafted training data, and find that the SVM’s test error increases with the attack. Steinhardt et al. construct the approximate upper bound of the attack loss on SVM and provides a solution to eliminate the impact of the attack [146]. In particular, they first remove outliers residing outside a feasible bound, and then minimise the margin-based loss on the rest data.

Fung et al. evaluate the impact of sybil-based data-poisoning attack on federated learning and propose a defense scheme, FoolsGold, to solve the problem [147]. A sybil-based attack [148] means that a participant edge device has a wrong training dataset, in which the data is the same with other participants whilst its label is wrong. For example, in digit recognition, the digit ‘1’ is labelled with ‘7’. They find that attackers may overpower other honest participants.
by poisoning the model with sufficient sybils. The proposed defense system, FoolGold, is based on contribution similarity. Since sybils share a common objective, their updates appear more similar than honest participants. FoolGold eliminates the impact of sybil-based attacks through reducing the learning rate of participants that repeatedly upload the same updates.

Blanchard et al. evaluate the Byzantine resilience of SGD in federated learning \[149\]. Byzantine refers to arbitrary failures in federated learning, such as erroneous data and software bugs. They find that linear gradient aggregation has no tolerance for even one Byzantine failure. Then they propose a Krum algorithm for aggregation with the tolerance of \( f \) Byzantines out of \( n \) participants. Specifically, the central server computes pairwise distances amongst all updates from edge devices, and takes the sum of \( n - f - 2 \) closest distance for all updates. The update with the minimum sum would be used to update the global shared model. However, all updates from edge devices are inspectable during computation, which may result in the risk of privacy disclosure. Chen et al. propose to use the geometric median of gradients as the update in federated learning \[150\]. This approach could tolerate \( q \) Byzantine failures up to \( 2q(1 + \varepsilon) \leq m \), in which \( q \) is the number of Byzantine failures, \( m \) refers to the headcount of participants, and \( \varepsilon \) is a small constant. This approach groups all participants into mini-batches. However, Yin et al. find that the approach fails if there is one Byzantine in each mini-batch \[151\]. They then propose a coordinate-wise median based approach to deal with the problem.

In fact, data-poisoning based attacks on federated learning is low in efficiency in the condition of small numbers of malicious participants. Because there are usually thousands of edge devices participating in the training in federated learning. The arbitrary update would be offset by averaging aggregation. In contrast, model-poisoning based attacks are more effective. Attackers directly poison the global shared model, instead of the updates from thousands of participants. Attackers introduce hidden backdoor functionality in the global shared model. Then, attackers use key, i.e., input with attacker-chosen features to trigger the backdoor. The model-poisoning based attack is shown as Fig. 17. Works on model-poisoning mainly focus on the problem of how backdoor functionality is injected in federated learning. Hence, we will focus on this direction as well.

Chen et al. evaluate the feasibility of conducting a backdoor in deep learning through adding few poisoning samples into the training set \[152\]. They find that only 5 poisoning samples out of 600,000 training samples are enough to create a backdoor. Bagdasaryan et al. propose a model replacement technique to open a backdoor to the global shared model \[153\]. As we aforementioned, the central server computes an update through averaging aggregation on updates from thousands of participants. The model replacement method scales up the weights of the ‘backdoored’ update to ensure that the backdoor survives the averaging aggregation. This is a single-round attack. Hence, such attack usually occurs during the last round update of federated learning. Different from \[153\], Bhagoji et al. propose to poison the shared model even when it is far from convergence, i.e., the last round update \[154\]. To prevent that, the malicious update is offset by updates from other participants, they propose a explicit boosting mechanism to negate the aggregation effect. They evaluate the attack technique against some famous attack-tolerant algorithms, i.e., Krum algorithm \[149\] and coordinate-wise median algorithm \[151\], and find that the attack is still effective.

D. Uncertainty Estimates

Standard deep learning method for classification and regression could not capture model uncertainty. For example, in model for classification, obtained results may be erroneously interpreted as model confidence. Such problems exist as well in edge intelligence. Efficient and accurate assessment of the deep learning output is of crucial importance, since the erroneous output may lead to undesirable economy loss or safety consequence in practical applications.

In principle, the uncertainty could be estimated through extensive tests. \[155\] propose a theoretical framework that casts dropout training in DNNs as approximate Bayesian inference in deep Gaussian processes. The framework could be used to model uncertainty with dropout neural networks through extracting information from models. However, this process is computation intensive, which is not applicable on mobile devices. This approach is based on sampling, which requires sufficient output samples for estimation. Hence, the main challenge to estimate uncertainty on mobile devices is the computational overhead. Based on the theory proposed in \[155\], Yao et al. propose RDeepSense, which integrates scoring rules as training criterion that measures the quality of the uncertainty estimation to reduce energy and time consumption \[156\]. RDeepSense requires to re-train the model to estimate uncertainty. The authors further propose ApDeepSense, which replaces the sampling operations with layer-wise distribution approximations following closed-form representations \[367\].

E. Applications

Bonawitz et al. develop a scalable product system for federated learning on mobile devices, based on TensorFlow \[157\]. In this system, each updating round consists of three phases: client selection, configuration, and reporting, as shown
in Fig. 18. In the client selection phase, eligible edge devices, e.g., devices with sufficient energy and computing resources, periodically send messages to the server to report the liveness. The server selects a subset among them according to a given objective. In the configuration phase, the server sends a shared model to each selected edge device. In the reporting phase, each edge device reports the update to the server, which would be aggregated to update the shared model. This protocol presents a framework of federated learning, which could adopt multiple strategies and algorithms in each phase. For example, the client selection algorithm proposed in [158] could be used in the client selection phase. The communication strategy in [52], [126] could be used for updating, and the FedAvg algorithm in [127] is adopted as an aggregation approach.

Researchers from Google have been continuously working on improving the service of Gboard with federated learning. Gboard consists of two parts: text typing and a search engine. The text typing module is used to recognise users’ input, whilst the search engine provides user relevant suggestions according to their input. For example, when you type ‘let’s eat’, Gboard may display the information about nearby restaurants. Hard et al. train a RNN language model using a federated learning approach to improve the prediction accuracy of the next-word for Gboard [159]. They compare the training result with traditional training methods on a central server. Federated learning achieves comparable accuracy with the central training approach. Chen et al. use federated learning to train a character-level RNN to predict high-frequent words on Gboard [160]. The approach achieves 90.56% precision on a publicly-available corpus. McMahan et al. undertake the first step to apply federated learning to enhance the search engine of Gboard [30]. When users search with Gboard, information about the current context and whether the clicked suggestion would be stored locally. Federated learning processes this information to improve the recommendation model. Yang et al. further improve the recommendation accuracy by introducing an additional triggering model [161]. Similarly, there are some works [162], [162] focusing on emoji prediction on mobile keyboards.

Federated learning has great potential in the medical imaging domain, where patient information is highly sensitive. Sheller et al. train a brain tumour segmentation model with data from multi-institution by applying federated learning [163]. The encrypted model is first sent to data owners, i.e., institutions, then the data owners decode, train, encrypt and upload the model back to the central aggregator. Roy et al. further develop an architecture of federated learning that uses peer-to-peer communications to replace the central aggregator towards medical applications [164].

Samarakoon et al. apply federated learning in vehicular networks to jointly allocate power and resources for ultra reliable low latency communication [165]. Vehicles train and upload their local models to the roadside unit (RSU), and RSU feeds back the global model to vehicles. Vehicles could use the model to estimate the queue length in city. Based on the queue information, the traffic system could reduce the queue length and optimise the resource allocation.

Nguyen et al. develop DIoT, a self-learning system to detect infected IoT devices by malware botnet in smart home environments [166]. IoT devices connect to the Internet through a gateway. They design two models for IoT device identification and anomaly detection. These two models are trained through the federated learning approach.

V. EDGE INFERENCE

The exponential growth of network size and the associated increase in computing resources requirement have been become a clear trend. Edge inference, as an essential component of edge intelligence, is usually performed locally on edge devices, in which the performance, i.e., execution time, accuracy, energy efficiency, etc. would be bounded by technology scaling. Moreover, we see an increasingly widening gap between the computation requirement and the available computation capacity provided by the hardware architecture [166]. In this section, we discuss various frameworks and approaches that contribute to bridging the gap.
A. Model Design

Modern neural network models are becoming increasingly larger, deeper and slower, they also require more computation resources\cite{169},\cite{369},\cite{370}, which makes it quite difficult to directly run high performance models on edge devices with limited computing resources, e.g., mobile devices, IoT terminals and embedded devices. Guo \textit{emph} evaluate the performance of DNN on edge device and find inference on edge devices costs up to two orders of magnitude greater energy and response time than central server\cite{371}. Many recent works have focused on designing lightweight neural network models, which could be performed on edge devices with less requirements on the hardware. According to the approaches of model design, existing literature could be divided into two categories: architecture search, and human-invented architecture. The former is to let machine automatically design the optimal architecture, while the latter is to design architectures by human.

1) Architecture Search: Designing neural network architectures is quite time-consuming, which requires substantial effort of human experts. One possible research direction is to use AI to enable machine search for the optimal architecture automatically. In fact, some automatically searched architectures, e.g., NASNet\cite{169}, AmoebaNet\cite{170}, and Adanet\cite{171}, could achieve competitive even much better performance in classification and recognition. However, these architectures are extremely hardware-consuming. For example, it requires 3150 GPU days of evolution to search for the optimal architecture for CIFAR-10\cite{170}. Mingxing \textit{et al.} adopt reinforcement learning to design mobile CNNs, called MnasNet, which could balance accuracy and inference latency\cite{172}. Different from\cite{169}–\cite{171}, in which only few kinds of cells are stacked, MnasNet cuts down per-cell search space and allow cells to be different. There are more $5 \times 5$ depthwise convolutions in MnasNet, which makes MnasNet more resource-efficient compared with models that only adopt $3 \times 3$ kernels.

Recently, a new research breakthrough of differentiable architecture search (DARTS)\cite{173} could significantly reduce dependence on hardware. Only four GPU days are required to achieve the same performance as\cite{170}. DARTS is based on continuous relaxation of the architecture representation and uses gradient descent for architecture searching. DARTS could be used for both convolutional and recurrent architectures.

Architecture search is hot research area and has a wide application future. Most literature on this area is not specially for edge intelligence. Hence, we will not further discuss on this field. Readers interested in this field could refer to\cite{372},\cite{373}.

2) Human-invented Architecture: Although architecture search shows good ability in model design, its requirement on hardware holds most researchers back. Existing literature mainly focuses on human-invented architecture. Howard \textit{et al.} use depth-wise separable convolutions to construct a lightweight deep neural network, MobileNets, for mobile and embedded devices\cite{174}. Fig.\cite{19} shows the factorisation of a convolution filter into a depth-wise and a point-wise convolution filter. The drawback of depth-wise convolution is that it only filters input channels. Depth-wise separable convolution, which combines depth-wise convolution and $1 \times 1$ point-wise convolution could overcome this drawback. MobileNet uses $3 \times 3$ depth-wise separable convolutions, which only requires 8-9 times less computation than standard ones. Moreover, depth-
wise and point-wise convolutions could also be applied to implement keyword spotting (KWS) models \[175\] and depth estimation \[176\] on edge devices.

Group convolution is another way to reduce computation cost for model designing. Due to the costly dense $1 \times 1$ convolutions, some basic architectures, e.g., Xception \[374\] and ResNeXt \[375\] cannot be used on resource-constrained devices. Zhang et al. propose to reduce the computation complexity of $1 \times 1$ convolutions with pointwise group convolution \[177\]. However, there is a side effect brought on by group convolution, i.e., outputs of one channel are only derived from a small part of the input channels. The authors then propose to use a channel shuffle operation to enable information exchanging among channels, as shown in Fig 20.

Depth-wise convolution and group convolution are usually based on ‘sparsely-connected’ convolutions, which may hamper inter-group information exchange and degrades model performance. Qin et al. propose to solve the problem with merging and evolution operations \[178\]. In merging operation, features of the same location among different channels are merged to generate a new feature map. Evolution operation extracts the information of location from the new feature map and combines extracted information with the original network. Therefore, information is shared by all channels, so that the information loss problem of inter-groups is effectively solved.

3) Applications: A large number of models have been designed for various applications, including face recognition \[167\], \[168\], \[179\], human activity recognition (HAR) \[180\]–\[188\], vehicle driving \[189\]–\[192\], and audio sensing \[193\], \[194\]. We introduce such applications next.

Face verification is increasingly attracting interests in both academic and industrial areas, and it is widely used in device unlocking \[376\] and mobile payments \[377\]. Particularly, some applications, such as smartphone unlocking need to run locally with high accuracy and speed, which is challenging for traditional big CNN models due to constrained resources on mobile devices. Sheng et al. present a compact but efficient CNN model, MobileFaceNets, which uses less than 1 million parameters and achieves similar performance to the latest big models of hundreds MB size \[167\]. MobileFaceNets uses a global depth-wise convolution filter to replace the global average pooling filter and carefully design a class of face feature. Chi et al. further lighten the weight of MobileFaceNets and presents MobiFace \[168\]. They adopt the Residual Bottleneck block \[179\] with expansion layers. Fast downsampling is also used to quickly reduce the dimensions of layers over $14 \times 14$. These two adopted strategies could maximise the information embedded in feature vectors and keep low computation cost.

Edge intelligence could be used to extract contextual information from sensor data and facilitate the research on Human Activity Recognition (HAR). HAR refers to the problem of recognising when, where, and what a person is doing \[378\], which could be potentially used in many applications, e.g., healthcare, fitness tracking, and activity monitoring \[379\], \[380\]. Table VI compares existing HAR technologies, regarding to their frameworks, models, ML methods, and objects. The challenges of HAR on edge platforms could summarised as follows.

- Commonly used classifiers for HAR, e.g., naive Bayes, SVM, DNN, are usually computation-intensive, especially when multiple sensors are involved.
- HAR requires to support near-real-time user experience in many applications.
- Very limited amount of labelled data is available for training HAR models.
- The data collected by on-device sensor includes noise and ambiguity.

Sourav et al. investigate how to deploy Restricted Boltzmann Machines (RBM)-based HAR models on smartwatch platforms, i.e., the Qualcomm Snapdragon 400 \[180\]. They first test the complexity of a model that a smartwatch can afford. Experiments show that although a simple RBM-based activity recognition algorithm could achieve satisfactory accuracy, the resource consumption on a smartwatch platform is unacceptably high. They further develop pipelines of feature representation and RBM layer activation functions. The
RBM model could effectively reduce energy consumption on smartwatches. Bandar et al. introduce time domain statistical features in CNN to improve the recognition accuracy [181]. In addition, to reduce the over-fitting problem of their model, they propose a data augmentation method, which applies a label-preserving transformation on raw data to create new data. The work is extended with extracting position features in [182]. Although deep learning could automatically extract features by exploring hidden correlations within and between data, pre-trained models sometimes cannot achieve the expected performance due to the diversities of devices and users, e.g., the heterogeneity of sensor types and user behaviour [381]. Prahalathan et al. propose to use on-device incremental learning to provide a better service for users [183]. Incremental learning [382] refers to a secondary training for a pre-trained model, which constrains newly learned filters to be linear combinations of existing ones. The re-trained model on mobile devices could provide personalised recognition for users.

Collecting fine-grained datasets for HAR training is challenging, due to a variety of available sensors, e.g., different sampling rated and data generation models. Valentin et al. propose to use RBM architecture to integrate sensor data from multiple sensors [184]. Each sensor input is processed by a single stacked restricted Boltzmann machine in RBM model. Afterwards, all outputted results are merged for activity recognition by another stacked restricted Boltzmann machine. Supervised machine learning is a most commonly utilised approach for activity recognition, which requires a large amount of labelled data. Manually labelling requires extremely large amounts of effort. Federico et al. propose a knowledge-driven automatic labelling method to deal with the data annotation problem [185]. GPS data and step count information are used to generate weak labels for the collected raw data. However, such an automatic annotation approach may create labelling errors, which impacts the quality of the collected data. There are three types of labelling errors, including inaccurate timestamps, mislabelling, and multi-action labels. Multi-action labels means that individuals perform multiple different actions during the same label. Xiao et al. solve the last two labelling errors through an ensemble of four stratified trained classifiers of different strategies, i.e., random forest, naive bayes, and decision tree [186].

The data collected by on-device sensors maybe noisy and it is hard to eliminate [381], [383]. For example, in movement tracking application on mobile devices, the travelled distance is computed with the sensory data, e.g., acceleration, speed, and time. However, the sensory data maybe noisy, which will result in estimation errors. Yao et al. develop DeepSense, which could directly extracts robust noise features of sensor data in a unified manner [187]. DeepSense combines CNN and RNN together to learn the noise model. In particular, the CNN in DeepSense learns the interaction among sensor modalities, while the RNN learn the temporal relationship among them based on the output of the CNN. The authors further propose QualityDeepSense with the consideration of the heterogeneous sensing quality [188]. QualityDeepSense hierarchically adds sensor-temporal attention modules into DeepSense to measures the quality of input sensory data. Based on the measurement, QualityDeepSense selects the input with more valuable information to provide better predictions.

Distracted driving is a key problem, as it potentially leads to traffic accidents [384]. Some researchers address this problem by implementing DL models on smartphones to detect distracted driving behaviour in real-time. Christopher et al. design DarNet, a deep learning based system to analyse driving behaviours and to detect distracted driving [189]. There are two modules in the system: data collection and analytic engine. There is a centralised controller in the data collection component, which collects two kinds of data, i.e., IMU data from drivers’ phones and images from IoT sensors. The analytic engine uses CNN to process image data, and RNN for sensor data, respectively. The outputs of these two models are combined through an ensemble-based learning approach to enable near real-time distracted driving activity detection. Fig. [27] presents the architecture of DarNet. In addition to CNN and RNN models, there are also other models could be used to detect unsafe driving behaviours, such as SVM [190], HMM [191], and decision tree [192].

Audio sensing has become an essential component for many applications, such as speech recognition [385], emotion detection [386], and smart homes [387]. However, directly running audio sensing models, even just the inference, would introduce a heavy burden on the hardware, such as digital signal processing (DSP) and battery. Nicholas et al. develop DeepEar, a DNN based audio sensing prototype for the smartphone platform [193], including four coupled DNNs of stacked RBMs that collectively perform sensing tasks. These four DNNs share the same bottom layers, and each of them is
responsible for a specific task, for example, emotion detection, and tone recognition. Experiments show that only 6% of the battery is enough to work through a day with the compromise of 3% accuracy drop. Petko et al. further improve the accuracy and reduces the energy consumption through applying multi-task learning and training shared deep layers [194]. The architecture of multi-task learning is shown as Fig. 22 in which the input and hidden layers are shared for audio analysis tasks. Each task has a distinct classifier. Moreover, the shared representation is more scalable than DeepEar, since there is no limitation in the integration of tasks.

B. Model Compression

Although neural networks are quite powerful in various promising applications, the increasing size of neural networks, both in depth and width, results in the considerable consumption of storage, memory and computing powers, which makes it challenging to run neural networks on edge devices. Moreover, statistic shows that the gaps between computational complexity and energy efficiency of deep neural networks and the hardware capacity are growing [388]. It has been proved that neural networks are typically over-parametrised, which makes deep learning models redundant [389]. To implement neural networks on powerless edge devices, large amounts of effort try to compress the models. Model compression aims to lighten the model, improve energy efficiency, and speed up the inference on resource-constraint edge devices, without lowering the accuracy. According to their approaches, we classify these works into five categories: low-rank approximation/matrix factorisation, knowledge distillation, compact layer, parameter quantising, and network pruning. Table VII summarises literature on model compression.

1) Low-rank Approximation: The main idea of low-rank approximation is to use the multiplication of low-rank convolutional kernels to replace kernels of high dimension. This is based on the fact that a matrix could be decomposed into the multiplication of multiple matrices of smaller size. For example, there is a weight matrix $W$ of $m \times k$ dimension. The matrix $W$ could be decomposed into two matrices, i.e., $X$ ($m \times d$) and $Y$ ($d \times k$), and $W = UV$. The computational complexity of matrix $W$ is $O(m \times k)$, while the complexity for the decomposed two matrices is $O(m \times d + d \times k)$. Obviously, the approach could effectively reduce the model size and computation, as long as $d$ is small enough.

Jaderberg et al. decompose the matrix of convolution layer $d \times d$ into the multiplication of two matrices $d \times 1$ and $1 \times d$ compress the CNNs [195]. The authors also propose two schemes to approximate the original filter. Fig. 23 presents the compression process. Fig. 23(a) shows a convolutional layer acting on a single-channel input. The convolutional layer consists of $N$ filters. For the first scheme, they use the linear combination of $M$ ($M < N$) filters to approximate the operation of $N$ filters. For the second scheme, they factorise each convolutional layer into a sequence of two regular convolutional layers but with rectangular filters. The approach achieves a 4.5x acceleration with 1% drop in accuracy. This work is a rank-1 approximation. Maji et al. apply this rank-1 approximation on compressing CNN models on IoT devices, which achieves 9x acceleration of the inference [197]. Denton et al. explore the approximation of rank-$k$ [196]. They use monochromatic and biclustering to approximate the original convolutional layer.

Kim et al. propose a whole network compression scheme with the consideration of entire convolutional and fully connected layers [198]. The scheme consists of three steps: rank selection, low-rank tensor decomposition, and fine-tuning. In particular, they first determine the rank of each layer through a global analytic solution of variational Bayesian matrix factorisation (VBMF). Then they apply Tucker decomposition to decompose the convolutional layer matrix into three components of dimension $1 \times 1$, $D \times D$ ($D$ is usually 3 or 5), and $1 \times 1$, which differs from SVD in [196]. The approach achieves a 4.26$\times$ reduction in energy consumption. We note that the component of spatial size $w \times h$ still requires a large amount of computation. Wang et al. propose a Block Term Decomposition (BTD) to further reduce the computation in operating the network, which is based on low-rank approximation and group sparsity [199]. They decompose the original weight matrix into the sum of few low multilinear rank weight matrices, which could approximately replace the original weight matrix. After fine-tuning, the compressed network achieves 5.91$\times$ acceleration on mobile devices with the network, and a less than 1% increase on the top-5 error.

Through optimising the parameter space of fully connected layers, weight factorisation could significantly reduce the memory requirement of DNN models and speed up the in-
| Ref. | Model         | Approach                  | Object                          | Performance              | Type       |
|------|---------------|---------------------------|---------------------------------|--------------------------|------------|
| 201  | NN            | Knowledge distillation    | Less resource requirement       | Faster                   | Lossless   |
| 202  | NN            | Knowledge distillation    | Compress model                  | 80% improvement          | Lossless   |
| 205  | NN            | Knowledge distillation    | Generate thiner model           | More accurate and smaller| Improved   |
| 204  | CNN           | Knowledge distillation    | Attention                       | Improve performance with shallow model | 1.1% top-1 better | Improved   |
| 205  | CNN           | Knowledge distillation    | Regularisation                  | Reduce storage           | 33.28× smaller | Improved   |
| 206  | CNN           | Knowledge distillation    | Less memory                     | 40× smaller              | Lossless   |
| 207  | GoogLeNet     | Knowledge distillation    | Less memory, acceleration       | 3× faster, 2.5× less memory | 0.4% drop  |
| 208  | CNN           | Knowledge distillation    | Reconstruct training set         | 6.4× smaller             | Lossy      |
| 209  | CNN           | Knowledge distillation    | Low-rank approximation          | Reduce runtime           | 4.5× faster | Lossy      |
| 210  | CNN           | Knowledge distillation    | Low-rank approximation          | Reduce computation       | 2× faster   | Lossy      |
| 211  | CNN           | Knowledge distillation    | Low-rank approximation          | Reduce computation       | 9× speedup  | Lossless   |
| 212  | CNN           | Knowledge distillation    | Low-rank approximation          | Reduce energy consumption| 4.26× energy reduction | Lossy      |
| 213  | CNN           | Knowledge distillation    | Low-rank approximation          | Group sparsity           | 5.91× faster | Improved   |
| 214  | CNN           | Knowledge distillation    | Kernel separation               | Use less resource        | 11.3× less memory | Lossless   |
| 215  | CNN           | Compact layer design      | Task decomposition              | Accelerates training     | 3.08% top-5 error | Lossless   |
| 216  | CNN           | Compact layer design      | Task decomposition              | Utilise storage to trade | 5.17× smaller | Improved   |
| 217  | CNN           | Compact layer design      | RNN                             | Train acceleration       | 28% relative improvement | Lossy      |
| 218  | YOLO          | Compact layer design      | Task decomposition              | Compress model           | 25MB total parameter | Lossy      |
| 219  | CNN           | Compressive sensing       | Training efficiency             | 6× faster                | Improved   |
| 220  | VGG-16        | Network pruning           | On-device customisation         | 1.24× faster             | 3% Lossy   |
| 221  | CNN           | Network pruning           | Network pruning                 | Reduce storage           | 13× fewer   | Lossless   |
| 222  | CNN           | Network pruning           | Network pruning                 | Higher energy efficiency | 20× faster  | Improved   |
| 223  | CNN           | Network pruning           | Network pruning                 | Reduce iterations        | 33× fewer   | Lossy      |
| 224  | CNN           | Network pruning           | Network pruning                 | Speed up inference       | 10× faster  | Lossy      |
| 225  | CNN           | Network pruning           | Global filter pruning           | Accelerate CNN           | 70% FLOPs reduction | Lossless   |
| 226  | CNN           | Network pruning           | Energy-efficiency               | 3.7× reduction on energy | Lossy      |
| 227  | RNN           | Network pruning           | Reduced model size              | 98.9% smaller, 94.3% faster | Lossless   |
| 228  | CNN           | Network pruning           | Network pruning                 | Reduces memory footprint  | 5× less computation | Lossless   |
| 229  | CNN           | Network pruning           | Network pruning                 | Maximise data reusability| 1.43× faster, 34× smaller | Lossless   |
| 230  | NN            | Hash function             | Parameter quantisation          | Compress required storage| 16 – 24× smaller | Lossy      |
| 231  | VGG-16        | Parameter quantisation    | Huffman coding                  | Compress model           | 49× smaller | Lossless   |
| 232  | CNN           | Parameter quantisation    | Network pruning                 | Compress model           | 20× smaller, 6× faster | Lossy      |
| 233  | DNN           | BinaryConnect             | Compress model                  | State-of-the-art         | Improved   |
| 234  | DNN           | Network Binarisation      | On-device customisation         | 32× smaller              | 95×7% energy saved | Lossy      |
| 235  | DNN           | Network Binarisation      | Compress model                  | Better than standard SGD | Improved   |
| 236  | DNN           | Parameter quantisation    | Binary Connect                  | Compress model speed up training | 2 – 3× faster, 5 – 10× smaller | Lossy      |
| 237  | HMM           | Parameter quantisation    | Compress model                  | Speed up training        | 10× speedup at most | Lossless   |
| 238  | LSTM          | Quantisation aware training| Compress model                  | Reduce size              | 4.16× smaller | Improved   |
| 239  | CNN           | Parameter quantisation    | Save energy                     | 4.45fps, 6.48 watts      | Lossy      |
| 240  | CNN           | Parameter quantisation    | Compress model                  | Reduce computation       | 1/10 memory shrinks | Improved   |
| 241  | CNN           | Parameter quantisation    | Posit number system             | Compress model           | 36.4% memory shrinks | Lossy      |
| 242  | MNN           | Network Binarisation      | Improve energy efficiency       | State-of-the-art         | Improved   |
| 243  | NN            | Network Binarisation      | Improve energy efficiency       | State-of-the-art         | Improved   |
Fig. 23. The decomposition and approximation of a CNN. (a) The original operation of a convolutional layer acting on a single-channel input. (b) The approximation of the first scheme. (c) The approximation of the second scheme.

Fig. 24. The application of attention mechanism in teacher-student paradigm transfer learning. (a) The left image is an input and the right image is the corresponding spatial attention map of a CNN model which shows which feature affects the classification decision. (b) Schematic representation of attention transfer. The attention map of the teacher model is used to supervise the training of the student model.

ference. However, the effect of the approach for CNN maybe not good, because there is a large amount of convolutional operations in CNN [29]. To solve the problem, Bhattacharya et al. propose a convolution kernel separation method, which optimises the convolution filters to significantly reduce convolution operations [200]. The authors verify the effectiveness of the proposed approach on various mobile platforms with popular models, e.g., audio classification and image recognition.

2) Knowledge Distillation: Knowledge distillation is based on transfer learning, which trains a neural network of smaller size with the distilled knowledge from a larger model. The large and complex model is called teacher model, whilst the compact model is referred as student model, which takes the benefit of transferring knowledge from the teacher network.

Bucilua et al. take the first step towards compressing models with knowledge distillation [390]. They first use a function learned by a high performing model to label pseudo data. Afterwards, the labelled pseudo data is utilised to train a compact but expressive model. The output of the compact model is compatible with the original high performing model. This work is limited to shallow models. The concept of knowledge distillation is first proposed in [202]. Hinton et al. first train a large and complex neural model, which is an ensemble of multiple models. This complex model is the teacher model. Then they design a small and simple student model to learn its knowledge. Specifically, they collect a transfer dataset as the input of the teacher model. The data could be unlabelled data or the original training set of the teacher model. The temperature in softmax is raised to a high value in the teacher model, e.g., 20. Since the soft target of the teacher model is the mean result of multiple components of the teacher model, the training instances are more informative. Therefore, the student model could be trained on much less data than the teacher model. The authors prove the effectiveness on MNIST and speech recognition tasks. Sau et al. propose to supervise the training of the student model with multiple teacher models, with the consideration that the distilled knowledge from a single teacher may be limited [205]. They also introduce a noise-based regulariser to improve the health in the performance of the student model.

Romero et al. propose FitNet, which extends [202] to create a deeper and lighter student model [203]. Deeper models could better characterise the essence of the data. Both the output of the teacher model and the intermediate representations are used as hints to speed up training of the student model, as well as improve its performance. Opposite to [203], Zagoruyko et al. prove that shallow neural networks could also significantly improve the performance of a student model by properly defining attention [204]. Attention is considered as a set of spatial maps that the network focuses the most on in the input to decide the output decision. These maps could be represented as convolutional layers in the network. In the teacher-student paradigm, the spatial attention maps are used to supervise the training of the student model, as shown in Fig. 24.

There are also some efforts focusing on how to design the student model. Crowley et al. propose to obtain the student model through replacing the convolutional layers of the teacher model with cheaper alternatives [206]. The new generated student model is then trained under the supervision of the teacher model. Li et al. design a framework, named DeepRebirth to merge the consecutive layers without weights, such as pooling and normalisation and convolutional layers vertically or horizontally to compress the model [207]. The newly generated student model learns parameters through layer-wise fine-tuning to minimise the accuracy loss. Fig. 25 presents the framework of DeepRebirth. After compression, GoogLeNet achieves 3x acceleration and 2.5x reduction in runtime memory.

The teacher model is pre-trained in most relevant works. Nevertheless, the teacher model and the student model could
be trained in parallel to save time. Zhou et al. propose a compression scheme, named Rocket Launching to exploit the simultaneous training of the teacher and student model [208]. During the training, the student model keeps acquiring knowledge learnt by the teacher model through the optimisation of the hint loss. The student model learns both the difference between its output and its target, and the possible path towards the final target learnt by the teacher model. Fig. 26 presents the structure of this framework.

When the teacher model is trained on a dataset concerning with privacy or safety, it is then difficult to train the student model. Lopes et al. propose an approach to distill the learned knowledge of the teacher model without accessing the original dataset, which only needs some extra metadata [209]. They first reconstruct the original dataset with the metadata of the teacher model. This step could find the images that best match these given by the network. Then they remove the noise of the image to approximate the activation records through gradients, which could partially reconstruct the original training set of the teacher model.

3) Compact layer design: In deep neural networks, if weights end up to be close to 0, the computation is wasted. A fundamental way to solve this problem is to design compact layers in neural networks, which could effectively reduce the consumption of resources, i.e., memories and computation power. Christian et al. propose to introduce sparsity and replace the fully connected layers in GoogLeNet [210]. Residual-Net replaces the fully connected layers with global average pooling to reduce the resource requirements [211]. Both GoogLeNet and Residual-Net achieve the best performance on multiple benchmarks.

Alex et al. propose a compact and lightweight CNN model, named YOLO Nano for image recognition [212]. YOLO Nano is a highly customised model with module-level macro- and micro-architecture. Fig. 27 shows the network architecture of YOLO Nano. There are three modules in YOLO Nano: expansion-projection (EP) macro-architecture, residual projection-expansion-projection (PEP) macro-architecture, and a fully-connected attention (FCA) module. PEP could reduce the architectural and computational complexity whilst preserving model expressiveness. FCA enables better utilisation of available network capacity.

Replacing a big convolution with multiple compact layers could effectively reduce the number of parameters and further reduce computations. Iandola et al. propose to compress CNN models with three strategies [213]. First, decomposing $3 \times 3$ convolution into $1 \times 1$ convolutions, since it has much fewer parameters. Second, cut down input channels in $3 \times 3$ convolutions. Third, downsample late to produce big feature maps. The larger feature maps could lead to higher classification accuracy. The first two strategies are used to decrease the quantity of parameters in CNN models and the last one is used to maximise the accuracy of the model. Based on three above mentioned strategies, the authors design SqueezeNet, which can achieve $50 \times$ reduction in the number of parameters, whilst remaining the same accuracy as the complete AlexNet. Similar approaches are also used in [214]. Shafiee et al. modify SqueezeNet for applications with fewer target classes and they propose SqueezeNet v1.1, which could be deployed on edge devices [215]. Yang et al. propose to decompose a recognition task into two simple sub-tasks: context recognition and target recognition, and further design a compact model, namely cDeepArch [216]. This approach uses storage resource
Shen et al. introduce Compressive Sensing (CS) to jointly modify the input layer and reduce nodes of each layer for CNN models [218]. CS [391] could be used to reduce the dimensionality of the original signal while preserving most of its information. The authors use CS to jointly reduce the dimensions of the input layer whilst extracting most features. The compressed input layer also enables the reduction of the number of parameters.

Besides the above-mentioned works about CNNs, Zhang et al. propose a dynamically hierarchy revolution (DirNet) to compress RNNs [217]. In particular, they mine dictionary atoms from original networks to adjust the compression rate with the consideration of different redundancy degrees amongst layers. They then adaptively change the sparsity across the hierarchical layers.

4) Network pruning: The main idea of network pruning is to delete unimportant parameters, since not all parameters are important in highly precise deep neural networks. Consequently, connections with less weights are removed, which converts a dense network into a sparse one, as shown in Fig. 28. There are some works which attempt to compress neural networks by network pruning.

The work [392] and [393] have taken the earliest steps towards network pruning. They prune neural networks to eliminate unimportant connections by using Hessian loss function. Experiment results prove the efficiency of pruning methods. Subsequent research focuses on how to prune the networks. Han et al. propose to prune networks based on a weight threshold [220]. Practically, they first train a model to learn the weights of each connection. The connections with lower weights than the threshold would then be removed. Afterwards, the network is retrained. The pruning approach is straightforward and simple. A similar approach is also used in [221]. In [221], the authors select and delete neurons of low performance, and then use a width multiplier to expand all layer sizes, which could allocate more resources to neurons of high performance. However, the assumption that connections with lower weights contribute less to the results may destroy the structure of the networks.

Identifying an appropriate threshold to prune neural networks usually takes iteratively trained networks, which consumes a lot of resources and time. Moreover, the threshold is shared by all the layers. Consequently, the pruned configuration maybe not the optimal, comparing with the case of identify thresholds for each layer. To break through these limitations, Manessi et al. propose a differentiability-based pruning method to jointly optimise the weights and thresholds for each layer [222]. Specifically, the authors propose a set of differentiable pruning functions and a new regulariser. Pruning could be performed during the back propagation phase, which could effectively reduce the training time.

Molchanov et al. propose a new criterion based on the Taylor expansion to identify unimportant neutrons in convolutional layers [223]. Specifically, they use the change of cost function to evaluate the result of pruning. They formulate pruning as an optimisation problem, trying to find a weight matrix that minimises the change in cost function. The formulation is approximately converted to its first-degree Taylor polynomial. The gradient and feature map's activation could be easily computed during back-propagation. Therefore, the approach could train the network and prune parameters simultaneously. You et al. propose a global filter pruning algorithm, named Gate Decorator, which transforms a CNN module through multiplying its output by the channel-wise scaling factors [224]. If the scaling factor is set to be 0, the corresponding filter would be removed. They also adopt the Taylor expansion to estimate the change of the loss function caused by the changing of the scaling factor. They rank all global filters based on the estimation and prune according to the rank. Compared with [223], [224] does not require special operations or structures.

In addition to minimum weight and cost functions, there are efforts trying to prune with the metric of energy consumption. Yang et al. propose an energy-aware pruning algorithm to prune CNNs with the goal of minimising the energy consumption [225]. The authors model the relationship between...
Fig. 29. Each channel is associated with a scaling factor $\gamma$ in convolutional layers. Then the network is trained to jointly learn weights and scaling factors. After that, the channels with small scaling factors (in orange colour) are pruned, which results in a compact model.

data sparsity and bit-width reduction through extrapolating the detailed value of consumed energy from hardware measurements. The pruning algorithm identifies the parts of a CNN that consumes the most energy and prunes the weights to maximise energy reduction.

Yao et al. propose to minimise the number of non-redundant hidden elements in each layer whilst retaining the accuracy in sensing applications and propose DeepIoT [226]. In DeepIoT, the authors compress neural networks through removing hidden elements. This regularisation approach is called dropout. Each hidden element is dropouted with a probability. The dropout probability is initialised with 0.5 for all hidden elements. DeepIoT develops a compressor neural network to learn the optimal dropout probabilities of all elements.

Liu et al. propose to identify important channels in CNN and remove unimportant channels to compress networks [227]. Specifically, they introduce a scaling factor $\gamma$ for each channel. The output $\hat{z}$ (also the input of the next layer) could be formulated as $\hat{z} = \gamma z + \beta$, where $z$ is the input of the current layer and $\beta$ is min-batch. Afterwards, they jointly train the network weight and scaling factors, with L1 regulation imposed on the latter. Following that, they prune the channels with the small scaling factor $\gamma$. Finally, the model is fine-tuned, which achieves a comparable performance with the full network. Fig. 29 presents this slimming process. However, the threshold of the scaling factor is not computed, which requires iterative evaluations to obtain a proper one.

Based on network pruning, the work in [228] investigates the data flow inside computing blocks and develops a data reuse scheme to alleviate the bandwidth burden in convolution layers. The data flow of a convolution layer is regular. If the common data could be reused, it is not necessary to load all data to a new computing block. The data reuse is used to parallelise computing threads and accelerate the inference of a CNN model.

5) Parameter quantisation: A very deep neural network usually involves many layers with millions of parameters, which consumes a large amount of storage and slows down the training procedure. However, highly precise parameters in neural networks are not always necessary in achieving high performance, especially when these highly precise parameters are redundant. It has been proved that only a small number of parameters are enough to reconstruct a complete network [389]. In [389], the authors find that the parameters within one layer could be predicted by 5% of parameters, which means we could compress the model by eliminating redundant parameters. There are some works exploiting parameter quantisation for model compression.

Gong et al. propose to use vector quantisation methods to reduce parameters in CNN [229]. Vector quantisation is often used in lossy data compression, which is based on block coding [171]. The main idea of vector quantisation is to divide a set of points into groups, which are represented by their central points. Hence, these points could be denoted with fewer coding bits, which is the basis of compression. In [229], the authors use k-means to cluster parameters and quantise these clusters. They find that this method could achieve 16–24× compression rate of the parameters with the scarification of no more than 1% of the top-5 accuracy. In addition to k-means, hash method has been utilised in parameter quantisation. In [230], Chen et al. propose to use hash functions to cluster connections into different hash buckets uniformly. Connections in the same hash bucket share the same weight. Han et al. combine parameter quantisation and pruning to further compress the neural network without compromising the accuracy [231]. Specifically, they first prune the neural network through recognising the important connections through all connections. Unimportant connections are ignored to minimise computation. Then, they quantise the parameters, to save the storage of parameters. After these two steps, the model will be retrained. These remaining connections and parameters could be properly adjusted. Finally, they use Huffman coding to further compress the model. Huffman coding is a prefix coding, which effectively reduces the required storage of data [394]. Fig. 30 presents the three-step compression.

For most CNNs, the fully connected layers consume most storage in neural network. Compressing parameters of fully connected layers could effectively reduce the model size. The convolutional layers consume most of the times during training and inference. Wu et al. design Q-CNN to quantise both fully connected layers and convolutional layers to jointly compress and accelerate the neural network [232]. Similar to [229],
the authors utilise k-means to optimally cluster parameters in fully connected and convolutional layers. Then, they quantise parameters by minimising the estimated error of response for each layer. They also propose a training scheme to suppress the accumulative error for the quantisation of multiple layers.

Enormous amount of floating point multiplications consumes significant times and computing resources in inference. There are two potential solutions to address this challenge. The first one is to replace floating point with fixed point, and the second one is to reduce the amount of floating point multiplications.

According to the evaluation of Xilinx, fixed point could achieve the same accuracy results as float [395]. Vanhoucke et al. evaluate the implementation of fixed point of an 8-bit integer on x86 platform [237]. Specifically, activation and the weights of intermediate layer are quantised into an 8-bit fixed point with the exception of biases that are encoded as 32-bit. The input layer remains floating point to accommodate possible large inputs. Through the quantisation, the total required memory shrinks $3-4 \times$. Results show that the quantised model could achieve a 10x speedup over an optimised baseline and a $4 \times$ speedup over an aggressively optimised float point baseline without affecting the accuracy. Similarly, Nasution et al. convert floating point to 8 and 16 bits to represent weights and outputs of layers, which lowers the storage to $4.16 \times [339]$. Peng et al. quantise an image classification CNN model into an 8-bit fixed-point at the cost of 1% accuracy drop [240]. Anwar et al. propose to use L2 error minimisation to quantise parameters [241]. They quantise each layer by one by one to induce sparsity and retrain the network with the quantised parameters. This approach is evaluated with MNIST and CIRAR-10 dataset. The results shows that the approach could reduce the required memory by 1/10.

In addition to fixed point, posit number could also be utilised to replace floating point numbers to compress neural networks. Posit number is a unique non-linear numerical system, which could represent all numbers in a dynamic range [396]. The posit number system represents numbers with fewer bits. Float point numbers could be converted into the posit number format to save storage. To learn more about the conversion, readers may refer to [397]. Langroudi et al. propose to use the posit number system to compress CNNs with non-uniform data [242]. The weights are converted into posit number format during the reading and writing operations in memory. During the training or inference, when computing operations are required, the number would be converted back to float point. Because this approach only converts the weight between two number systems, no quantisation occurs. The network does not require to be re-trained.

Network Binarisation is an extreme case of weight quantisation. Weight quantisation indicates that all weights are represented by two possible values (e.g., -1 or 1), which could overwhelmingly compress neural networks [233]. For example, the original network requires 32 bits to store one parameter, while in binary connect based network, only 1 bit is enough, which significantly reduces the model size. Another advantage of binary connect is that replacing multiply-accumulate operations by simple accumulations, which could drastically reduce computation in training. Courbariaux et al. extend the work [233] further and proposes Binary Neural Network (BNN), which completely changes the computing style of traditional neural networks [234]. Not only the weights, but also the input of each layer is binarised. Hence, during the training, all multiplication operations are replaced by accumulation operations, which drastically improves the power-efficiency. However, substantial experiments indicate that BNN could only achieve good performance on small scale datasets.

Rastegari et al. propose a XNOR-net to reduce storage and improve training efficiency, which is different with [234] in the binarisation method and network structure [235]. In Binary-Weight network, all weight values are approximately binarized, e.g., -1 or 1, which reduces the size of network by $32 \times$. Convolutions could be finished with only addition and subtraction, which is different with [234]. Hence, the training is speed up $2 \times$. With XNOR-net, in addition to weights, the input to convolutional layers are approximately binarised. Moreover, they further simplify the convolution with XNOR operations, which achieves a speed up of $58 \times$. The comparison amongst standard convolution, Binary-Weight and XNOR-net is presented as Table VIII.

Lin et al. propose to use binary connect to reduce multiplications in DNN [236]. In the forward pass, the authors stochastically binarise weights by binary connect. Afterwards,
they quantise the representations at each layer to replace the remaining multiply operations into bit-shifts. Their results show that there is no loss in accuracy in training and sometimes this approach surprisingly achieves even better performance than standard stochastic gradient descent training.

Soudry et al. prove that binary weights and activations could be used in Expectation Backpropagation (EBP) and achieves high performance [243]. This is based on a variational Bayesian approach. The authors test eight binary text classification tasks with EBP-trained multilayer neural networks (MNN). The results show that binary weights always achieve better performance than continuous weights. Esser et al. further develop a fully binary network with the same approach to EBP to improve the energy efficiency on neuromorphic chips [244]. They perform the experiment on the MNIST dataset, and the results show that the method achieves 99.42% accuracy at 108 µJ per image.

6) Applications: Some efforts try to use these compression techniques on practical applications and prototypes at the edge, including image analysis [245]–[247], compression service [250], and automotive [248], [249].

Mathur et al. develop a wearable camera, called DeepEye, that runs multiple cloud-scale deep learning models at edge provide real-time analysis on the captured images [245]. DeepEye enables the creation of five state-of-the-art image recognition models. After camera captures an image, the image pre-processing component deals with the image according to the adopted deep model. There is a model compression component inside the inference engine, which applies available compression techniques to reduce energy consumption and the running time. Finally, DeepEye use the optimised BLAS library to optimise the numeric operations on hardware.

To correctly identify prescription pills for patients based on their visual appearance, Zeng et al. develop MobileDeepPill, a pill image recognition system [246]. The pill image recognition model is based on ImageNet [398]. Fig. 31 presents the architecture of MobileDeepPill. In the training phase, the system first localises and splits the pill image in consumer and pill references. The system then enrich samples through running data augmentation module. Finally, the system imports CNNs as the teacher model to supervise the student model. In the inference phase, the system first processes the pill photo and extracts features to perform the student CNNs. As a last step, the system ranks the results according to their possibilities.

Wang et al. propose a fast image search framework to implement the content-based image retrieval (CBIR) service from cloud servers to edge devices [247]. Traditional CBIR services are based on the cloud, which suffers from high latency and privacy concerns. The authors propose to reduce the resource requirements of the model and to deploy it on edge devices. For the two components consuming most resources, i.e., object detection and feature extraction, the authors use low-rank approximation to compress these two parts. The compressed model achieves 6.1× speedup for inference.

Liu et al. develop an on-demand customised compression system, named AdaDeep [250]. Various kinds of compression approaches could be jointly used in AdaDeep to balance the performance and resource constraints. Specifically, the authors propose a reinforcement learning based optimiser to automatically select the combination of compression approaches to achieve appropriate trade-offs among multiple metrics such as accuracy, storage, and energy consumption.

With growing interests from the automotive industry, various large deep learning models with high accuracy have been implemented in smart vehicles with the assistance of compression techniques. Kim et al. develop a DL based object recognition system to recognise vehicles [248]. The vehicle recognition system is based on faster-RCNN. To deploy the system on vehicles, the authors apply network pruning and parameter quantisation to compress the network. Evaluations show that these two compression techniques reduce the network size to 16% and reduce runtime to 64%. Xu et al. propose an RNN based driving behaviour analysis system on vehicles [249]. The system uses the raw data collected by a variety of sensors on vehicles to predict the driving patterns. To deploy the system on automobiles, the authors apply parameter quantisation to reduce the energy consumption and model size. After compression, the system size is reduced to 44 KB and the power overhead is 7.7 mW.
C. Inference Acceleration

The computing capacities of edge devices have been increased and some embedded devices, such as NVIDIA Jetson TX2 [259] could directly perform CNN. However, it is still difficult for most edge devices to directly run large models. Model compression techniques reduce the required resources to create neural network models and facilitate the performance of these models on edge devices. Model acceleration techniques further speed up the performance of the compressed model on edge devices. The main idea of model acceleration in inference is to reduce the run-time of inference on edge devices and realise real-time responses for specific neural network based applications without changing the structure of the trained model. According to acceleration approaches, research works on inference acceleration could be divided into two categories: hardware acceleration and software acceleration. Hardware acceleration methods focus on parallelising inference tasks to available hardware, such as CPU, GPU, and DSP. Software acceleration method focuses on optimising resource management, pipeline design, and compiler.

1) Hardware Acceleration: Recently, mobile devices are becoming increasingly powerful. More and more mobile platforms are equipped with GPUs. Since mobile CPUs are not suitable for the computing of deep neural networks, the embedded GPU could be used to share the computing tasks and accelerate the inference. Table IX summaries existing literature on hardware acceleration.

Alzantot et al. evaluate the performance of CNNs and RNNs only on CPU, and compares against the execution in parallel on all available computing resources, e.g., CPU, GPU, DSP, etc. [251]. Results show that the parallel computing paradigm is much faster. Loukadakis et al. propose two parallel implementations of VGG-16 network on ODROID-XU4 board: OpenMP version and OpenCL version [252]. The former parallelises the inference within the CPU, whilst the latter one parallelises within the Mali GPU. These two approaches achieve 2.8× and 11.6× speedup, respectively. Oskouei et al. design a mobile GPU-based accelerator for using deep CNN on mobile platforms, which executes inference in parallel on both CPU and GPU [255]. The accelerator achieves 60× speedup. The authors further develop a GPU-based accelerated library for Android devices, called CNNdroid, which could achieve up to 60× speedup and 130× energy reduction Android platforms [254].

With the consideration that the memory on edge devices are usually not sufficient for neural networks, Tsung et al. propose to optimise the flow to accelerate inference [255]. They use a matrix multiplication function to improve the cache hit rate in memory, which indirectly speeds up the execution of the model.

Nvidia has developed a parallelisation framework, named Compute Unified Device Architecture (CUDA) for desktop GPUs to reduce the complexity of neural networks and improve inference speed. For example, in [400], CUDA significantly improves the execution efficiency of RNN on desktop GPUs. Some efforts implement the CUDA framework onto mobile platforms. Rizvi et al. propose an approach for image classification on embedded devices based on the CUDA framework [256]. The approach features the most common layers in CNN models, e.g., convolutions, max-pooling, batch-normalisation, and activation functions. General Purpose Computing GPU (GPGPU) is used to speed up the most computation-intensive operations in each layer. The approach is also used to implement an Italian license plate detection and recognition system on tablets [257]. They subsequently introduce matrix multiplication to reduce the computational complexity of convolution in a similar system to achieve real-time object classification on mobile devices [258]. They also apply the approach in a robotic controller system [259].

However, the experiments in [260] show that directly applying CUDA on mobile GPUs may be ineffective, or even deteriorates the performance. Cao et al. propose to accelerate RNN on mobile devices based on a parallelisation framework, called RenderScript [260]. RenderScript [261] is a component of the Android platform, which provides an API for hardware acceleration. RenderScript could automatically parallelise the data structure across available GPU cores. The proposed framework could reduce latency by 4×.

Motamedi et al. implement SqueezeNet on mobile and evaluates the performance on three different Android devices based on RenderScript [262]. Results show that it achieves 310.74× speedup on a Nexus 5. They further develop a general framework, called Cappuccino, for automatic synthesis of efficient inference on edge devices [263]. The structure of Cappuccino is shown as in Fig. 32. There are three inputs for the framework: basic information of the model, model file, and dataset. There are three kinds of parallelisation: kernel-level, filter bank-level, and output-level parallelisation. The thread workload allocation component allocates tasks by using these three kinds of parallelisation. They specially investigate the optimal degree of concurrency for each task, i.e., the number of threads in [264]. The data order optimisation component is
used to convert the data format. Cappuccino enables imprecise computing in exchange for high speed and energy efficiency. The inexact computing analyser component is used to analyse the effect of imprecise computing and determine the tradeoff amongst accuracy, speed and energy efficiency.

Huynh et al. propose Deepsense, a GPU-based CNN framework to run various CNN models in soft real-time on mobile devices with GPUs [265]. To minimise the latency, Deepsense applies various optimisation strategies, including branch divergence elimination and memory vectorisation. The structure of Deepsense is shown as in Fig. 33. The model converter first converts pre-trained models with different representations into a pre-defined format. Then, the model loader component loads the converted model into memory. When inference starts, the inference scheduler allocates tasks to the GPU sequentially. The executor takes inputted data and the model for executing. During the execution pipeline, CPU is only responsible for padding and intermediate memory allocation, whilst most computing tasks are done by the GPU. The authors further present a demo of the framework in [93] for continuous vision sensing applications on mobile devices.

The heterogeneous multi-core architecture, including CPU and GPU on mobile enables the application of neural networks. By reasonably mapping tasks to cores could improve energy efficiency and inference speed. Taylor et al. propose a machine learning based approach to map OpenCL kernels onto proper heterogeneous multi-cores to achieve given objectives, e.g., speedup, energy-efficiency or a tradeoff [266]. The framework first trains the mapping model with the optimisation setting for each objective, then it uses the learned model to schedule OpenCL kernels based on the information of the application.

Rallapalli et al. find that the memory of GPUs severely limits the operation of deep CNNs on mobile devices, and proposes to properly allocate part of computation of the fully-connected layers to the CPU [267]. The fully-connected layers are split into several parts, which are executed sequentially. Meanwhile, part of these tasks are loaded into the memory of the CPU for processing. They evaluate the method with an object detection model, YOLO [268] on Jetson TK1 board and achieve 60% speedup.

In addition to commonly used hardware, i.e., CPUs, mobile GPUs, GPGPU, and DSP, field-programmable gate arrays (FPGAs) could also be used for acceleration. Different from CPUs and GPUs, which run software code, FPGA uses hardware level programming, which means that FPGA is much faster than CPU and GPU. Bettoni et al. implement an object recognition CNN model on FPGA via Tiling and Pipelining parallelisation [269]. Ma et al. exploit the data reuse and data movement in a convolution loop and proposes to use loop optimisation (including loop unrolling, tiling, and interchange) to accelerate the inference of CNN models in FPGA [270]. A similar approach is also adopted in [271].

Lots of literature focus on developing energy-efficiency DNNs. However, the diversity of DNNs makes them inflexible for hardware [272]. Hence, some researchers attempt to design special accelerating chips to flexibly use DNNs. Chen et al. develop an energy-efficient hardware accelerator, called Eyeriss [273]. Eyeriss uses two methods to accelerate the performance of DNNs. The first method is to exploit data reuse to minimise data movement, whilst the second method is to exploit data statistics to avoid unnecessary reads and computations, which improves energy efficiency. Subsequently, they change the structure of the accelerator and propose a new version, Eyeriss v2, to run compact and sparse DNNs [274]. Fig. 34 shows the comparison between Eyeriss and Eyeriss v2. Both of them consist of an array of processing elements (PE) and global buffers (GLB). The main difference is the structure. Eyeriss v2 is hierarchical, in which PEs and GLBs are grouped to reduce communication cost.

2) Software Acceleration: Different from hardware acceleration, which depends on the parallelisation of tasks on available hardware, software acceleration mainly focuses on optimising resource management, pipeline design, and compilers. Hardware acceleration methods speed up inference through increasing available computing powers, which usually does not affect the accuracy, whilst software acceleration methods maximise the performance of limited resources for speedup, which may lead to a drop in accuracy with some cases. For example, in [275], the authors sacrifice accuracy for real-time response. Table X summarises existing literature on software acceleration.

Georgiev et al. investigate the tradeoff between performance and energy consumption of an audio sensing model on edge devices [276]. Work items need to access different kinds of memory, i.e., global, shared, and private memory. Global memory has the maximum size but minimum speed, whilst private memory is fastest and smallest but exclusive to each work
item. Shared memory is between global and private memory. Typical audio sensing models have the maximum parameters, which surpasses the capacity of memory. They use memory access optimisation techniques to speed up the inference, including vectorisation, shared memory sliding window, and tiling.

Lane et al. propose DeepX to reduce the resource usage on mobile devices based on two resource management algorithms [277]. The first resource management algorithm is for runtime layer compression. The model compression method discussed in Section V-B could also be used to remove redundancy from original layers. Specifically, they use a SVD-based layer compression technique to simplify the model. The second algorithm is for architecture decomposition, which decomposes the model into blocks that could be performed in parallel. The workflow of DeepX is shown in Fig. 35. They further develop the DeepX toolkit (DXTK) [279]. A number of pre-trained and compressed deep neural network models are packaged in DXTK. Users could directly use DXTK for specific applications.

Yang et al. propose an adaptive software accelerator, Netadapt, which could dynamically speed up the model according to specific metrics [280]. They use empirical measurements on practical devices to evaluate the performance of the accelerator. Fig. 36 shows the structure of Netadapt. Netadapt adjusts the network according to the given budget, i.e., latency, energy, etc. During iteration, the framework generates many network proposals. Then, Netadapt evaluates these proposals according to direct empirical measurements, and selects one with maximum accuracy. The framework is similar to [250], which caches multiple model compression systems, and compresses the input model according to users’ demand.

Ma et al. introduce the concept of quality of service (QoS) in model acceleration and develop an accelerator, DeepRT [281]. The QoS of an accelerated model is defined as a tuple \( Q = (d, C) \), where \( d \) is a desired response time and \( C \) denotes model compression bound. There is a QoS manager component in DeepRT, which controls the system resources to support the QoS during the acceleration.

Liu et al. find that fast Fourier transform (FFT) could effectively speed up convolution operation [401]. Abtahi et al. apply FFT-based convolution ResNet-20 on NVIDIA Jetson.

### TABLE IX

| Ref. | Model   | Executor       | Strategy       | Object                  | Performance               |
|------|---------|----------------|----------------|-------------------------|---------------------------|
| [251]| CNN, RNN| CPU, GPU       | RenderScript   | Speed up inference      | 3× faster                 |
| [252]| VGG-16 | CPU, GPU       | SIMD           | Speed up inference      | 11.6× faster              |
| [253]| CNN    | GPU            | SIMD           | Speed up inference      | 60× faster                |
| [254]| CNN    | GPU            | SIMD           | Speed up inference      | 130× energy-saving        |
| [255]| DNN    | GPU            | Flow optimisation | Enable DNN on mobile device | 58× faster                 |
| [256]| CNN    | GPGPU          | CUDA           | Maxitimise throughput   | 104× energy-saving        |
| [257]| DNN    | GPU            | SIMD           | Real-time character detection | 250ms per time            |
| [258]| DNN    | GPU            | SIMD           | Real-time character detection | 3× faster                 |
| [260]| LSTM   | GPU            | RenderScript   | Rnn RNN on mobile platform | 4× reduction on latency |
| [262]| SqueezeNet | GPU        | RenderScript   | Acceleration, energy-efficiency | 249.47× energy-saving |
| [263]| CNN    | CPU, GPU, DSP  | RenderScript   | Optimise thread number  | 2.37× faster              |
| [265]| CNN    | CPU, GPU, DSP  | RenderScript   | Automatic speedup       | 272.03× faster at most    |
| [266]| YOLO   | GPU            | Tucker decompositon | Real-time response | VGG-F 361ms               |
| [267]| OpenCL | CPU, GPU       | Kernel mapping | Adaptive optimisation   | 1.2× faster                |
| [269]| YOLO   | CPU, GPU       | Memory optimisation | Enable CNN on mobile device | 0.42s for YOLO            |
| [270]| CNN    | FPGA           | Tiling, Pipelining | Enable CNN in FPGA | 15× faster                |
| [271]| CNN    | FPGA           | Loop optimisation | Memory and data movement | 3.2× faster              |
| [273]| DNN    | EyeLiss        | Data reuse     | Improve energy efficiency | 23% faster                |
| [274]| DNN    | EyeLiss v2     | Hierarchical mesh | Hardware processing efficiency | 9.05× energy-saving |

![Fig. 35. The workflow of DeepX. Layer compression could reduce the requirement on resource, whilst the architecture decomposition divides the model into multiple blocks that could be performed in parallel.](image)
TABLE X

| Ref. | Model | Strategy | Object | Performance | Accuracy |
|------|-------|----------|--------|-------------|----------|
| [275] | DNN  | memory access optimisation | Performance-energy tradeoff | 42ms, 83% accuracy | Lossy |
| [276] | DNN  | Resource management | Accelerate inference | 6.5× faster, 3 – 4× less energy | Lossless |
| [277] | DNN  | Compression, decomposition | Reduce resource use | 5.8× faster | 4.9% loss |
| [278] | DNN  | Compression, decomposition | Reduce resource use | 5.8× faster | 4.9% loss |
| [279] | DNN  | Caching | Reduce resource use | 5.8× faster | 4.9% Lossless |
| [280] | NN   | Caching | Adaptive speedup | 1.7× speedup | 4.9% Lossless |
| [281] | DNN  | QoS control | Improve QoS | N/A | N/A |
| [282] | CNN  | FFT-based convolution | Accelerate convolution | 10910× faster at most | N/A |
| [283] | CNN  | Cache mechanism | Accelerate inference | 20.2% faster | 3.51% drop |
| [284] | YOLO | Caching, pixel matching | Accelerate inference | 9.1× faster | 0.1% drop |
| [285] | YOLO | Caching, feature extraction | Real-time response | 36% faster | 3.8%-6.2% drop |

TABLE XI

| Ref. | Model | Strategy | Object | Performance | Accuracy |
|------|-------|----------|--------|-------------|----------|
| [276] | DNN  | Resource management | Accelerate inference | 6.5× faster, 3 – 4× less energy | Lossless |
| [277] | DNN  | Compression, decomposition | Reduce resource use | 5.8× faster | 4.9% loss |
| [278] | DNN  | Compression, decomposition | Reduce resource use | 5.8× faster | 4.9% loss |
| [279] | DNN  | Caching | Reduce resource use | 5.8× faster | 4.9% Lossless |
| [280] | NN   | Caching | Adaptive speedup | 1.7× speedup | 4.9% Lossless |
| [281] | DNN  | QoS control | Improve QoS | N/A | N/A |
| [282] | CNN  | FFT-based convolution | Accelerate convolution | 10910× faster at most | N/A |
| [283] | CNN  | Cache mechanism | Accelerate inference | 20.2% faster | 3.51% drop |
| [284] | YOLO | Caching, pixel matching | Accelerate inference | 9.1× faster | 0.1% drop |
| [285] | YOLO | Caching, feature extraction | Real-time response | 36% faster | 3.8%-6.2% drop |

VI. EDGE OFFLOADING

Computation is of utmost importance for supporting edge intelligence, which powers the other three components. Most edge devices and edge servers are not as powerful as central servers or computing clusters. Hence, there are two approaches to enable computation-intensive intelligent applications at the edge: reducing the computational complexity of applications and improving the computing power of edge devices and edge servers. The former approach has been discussed in previous sections. In this section, we focus on the latter approach.

Considering the hardware limitation of edge devices, computation offloading [16], [339], [402]–[404] offers promising approaches to increase computation capability. Literature of this area mainly focuses on designing an optimal offloading strategy to achieve a particular objective, such as latency minimisation, energy-efficiency, and privacy preservation. According to their realisation approaches, these strategies could be divided into five categories: device-to-cloud (D2C), device-to-edge (D2E), device-to-device (D2D), hybrid architecture, and caching.

A. D2C offloading strategy

It consumes considerable computing resources and energy to deal with streamed AI tasks, such as video analysis and continuous speech translation. Most applications, such as Apple Siri and Google Assistant, adopt pure cloud based offloading strategy, in which devices upload input data, e.g., speech or image to cloud server through cellular or WiFi networks. The inference through a giant neural model with high accuracy is done by powerful computers and the results are transmitted back through the same way. There are three main disadvantages in this procedure: (1) mobile devices are required to upload enormous volumes of data to the cloud, which has proved to be the bottleneck of the whole procedure [286]. Such a bottleneck increases users’ waiting time; (2) the execution depends on the Internet connectivity; once the device is offline, relative applications could not be used; (3) the uploaded data from mobile devices may contain private information of users, e.g., personal photos, which might be attacked by malicious hackers during the inference on cloud server [405]. There are some efforts trying to solve these problems, which will be discussed next. Table XI summarises existing literature on D2C offloading strategy.
There are usually many layers in a typical deep neural network, which processes the input data layer by layer. The size of intermediate data could be scaled down through layers. Li et al. propose a deep neural network layer schedule scheme for the edge environment, leveraging this characteristic of deep neural networks [283]. Fig. 37 shows the structure of neural network layer scheduling-based offloading scheme. Edge devices lacking computing resources, such as IoT devices, first upload the collected data to nearby edge server, which would process the original input data through few low network layers. The generated intermediate data would be uploaded to the cloud server for further processing and eventually output the classification results. The framework is also adopted in [284]. The authors use edge server to pre-process raw data and extract key features.

The model partitioning and layer scheduling could be designed from multiple perspectives, e.g., energy-efficiency, latency, and privacy. Eshratifar et al. propose a layer scheduling algorithm from the perspective of energy-efficiency in a similar offloading framework [285]. Kang et al. investigate this problem between edge and cloud side [286]. They propose to partition the computing tasks of DNN between local mobile devices and cloud server and design a system, called Neurosurgeon, to intelligently partition DNN based on the prediction of system dynamics. Osia et al. consider the layer scheduling from the perspective of privacy preservation [287]. They add a feature extractor module to identify private features from raw data, which will be sent to the cloud for further processing. Analogous approaches are also adopted in [288, 289].

In continuous computer vision analysis, video streams need to be uploaded to the cloud server, which requires a large amount of network resources and consumes battery energy. Ananthnarayanan et al. propose a geographically distributed architecture of clouds and edge servers for real-time video analysis [290]. Fixed (e.g., traffic light) and mobile cameras (e.g., event data recorder) upload video streams to available edge servers for pre-processing. The pre-processed data would be further transmitted to a central cloud server in a geographic location for inference. Similarly, Ali et al. leverage all available edge resources to pre-process data for large-scale video stream analytics [291]. Deep learning based video analytic applications contain four stages, including motion detection, frame enhancement, object detection based on shallow networks, and object detection based on deep networks. With the traditional cloud-based approach, these four stages are executed on a cloud server. The authors propose to execute the first two stages locally, which does not require much computation capacity. The output is then transmitted to edge servers for further processing (the third stage). The output is then uploaded to the cloud for final recognition.

Some efforts [406, 407] propose to upload only ‘interesting’ frames to the cloud, which significantly reduces the amount of uploaded data. However, detecting these ‘interesting’ frames also requires intensive computation. Naderiparizi et al. develop a framework, Glimpse, to select valid frames by performing coarse visual processing with low energy consumption [292]. Glimpse adopts gating sensors and redesigns the processing pipeline to save energy.

The easiest offloading strategy is to offload the inference task to the cloud when the network condition is good, otherwise perform model compression locally. For example, [293] only considers the network condition when offloading healthcare inference tasks. Cui et al. characterise the resource requirements of data processing applications on an edge gateway and cloud server [408]. Hanhirova et al. explore and characterises the performance of some CNNs, e.g., object detection and recognition, both on smartphones and cloud server [294]. They find that the latency and throughput are conflicting metrics, in turn, are difficult to be jointly optimised on both mobile devices and cloud server. Some efforts focus on designing an offloading scheme to decide when to use a neural network locally and when to offload the task to the cloud server, instead of always executing on cloud. Considering the capacities of local devices and the network conditions, Qi et al. design an adaptive decision scheme to dynamically perform tasks [295]. To enable the local execution, mobile devices adopt compressed models, which achieve lower accuracy than the complete model on the cloud server. If the network condition could not guarantee a real-time response, the inference task would be executed locally. [296] also proposes an offloading decision scheme for the same problem with extra consideration to energy consumption. Ran et al. subsequently extend the work with a measurement-driven mathematical framework for achieving a tradeoff between data compression, network condition, energy consumption, latency, and accuracy [297].

Georgiev et al. consider a collective offloading scheme for heterogeneous mobile processors and cloud for sensor based applications, which makes best possible use out of different kinds of computing resources on mobile devices, e.g., CPU, GPU, and DSP [298]. They designed a task scheduler running on low-power co-processor unit (LPU) to dynamically restructure and allocate tasks from applications across heterogeneous computing resources based on fluctuations in device
and network.

B. D2E offloading strategy

Three main disadvantages with the D2C offloading strategy have been discussed, i.e., latency, wireless network dependency, and privacy concern. Although various solutions have been proposed to alleviate these problems, they do not address these fundamental challenges. Users still need to wait for a long time. Congested wireless networks lead to failed inference. Moreover, the potential risk of private information leakage still exists. Hence, some works try to explore the potential of D2E offloading, which may effectively address these three problems. Edge server refers to the powerful servers (more powerful than ordinary edge devices) that is physically near mobile devices. For example, wearable devices could offload the inference tasks to their connected smartphones. Smartphones could offload computing tasks to cloudlets deployed at roadside. Table XI summarises the existing literature on D2E offloading strategy.

First, we review the works that offload inference tasks to specialised edge servers, e.g., cloudlets and surrogates [409], which refer to infrastructure deployed at edge of the network. There are two general problems that need to be considered in this scenario, including (1) which component of the model could be offloaded to the edge; and (2) which edge server should be selected to offload to. Ra et al. develop a framework, named Odessa for interactive perception applications, which enables parallel execution of the inference on local devices and edge server [299]. They propose a greedy algorithm to partition the model based on the interactive deadlines. The edge servers and edge devices in Odessa are assumed to be fixed, meaning they do not consider problem (2). Streiffer et al. appoint an edge server for mobile devices that requests video frame analytics [300]. They evaluate the impact of distance between edge server and mobile devices on latency and packet loss and find that offloading inference tasks to an edge server at a city-scale distance could achieve the similar performance with execution locally on each mobile devices.

Similar to D2C offloading, where the partitioned model layers could be simultaneously deployed on both cloud server and local edge device, the partitioned model layers could also be deployed on edge servers and edge devices. This strategy reduces the transmitted data, which further reduces latency and preserve privacy. Li et al. propose Edgent, a device-edge co-inference framework to realise this strategy [19]. The core idea of Edgen is to run computation-intensive layers on powerful edge servers and run the rest layers on device. They also adopt model compression techniques to reduce the model size and further reduce the latency. Similarly, Ko et al. propose a model partitioning approach with the consideration of energy efficiency [301]. Due to the difference of available resources between edge devices and edge servers, partitioning the network at a deeper layer would reduce the energy efficiency. Hence, they propose to partition the network at the end of the convolution layers. The output features through the layers on edge device would be compressed before transmitted to edge server to minimise the bandwidth usage.

Some efforts [410]–[412] attempt to encrypt the sensitive data locally before uploading. On cloud side, non-linear layers of a model are converted into linear layers, and then they use homomorphic encryption to execute inference over encrypted input data. This offloading paradigm could also be adopted on edge servers. However, the encryption operation is also computation-intensive. Tian et al. propose a private CNN inference framework, LEP-CNN, to offload most inference tasks to edge servers and to avoid privacy breaches [302]. The authors propose an online/offline encryption method to speed up the encryption, which trades offline computation and storage for online computation speedup. The execution of the inference over encrypted input data on edge server addresses privacy issues.

Mobility of devices introduces a challenge during the offloading, e.g., in autonomous driving scenarios. Mobile devices may lose the connection with edge server before the inference is done. Hence, selecting an appropriate edge server according to users’ mobility pattern is crucial. Zhang et al. use reinforcement learning to decide when and which edge server to offload to [303]. A deep Q-network (DQN) based approach is used to automatically learn the optimal offloading scheme from previous experiences. If one mobile device moves away
before the edge server finishes the execution of the task, the edge server must drop the task, which wastes the computing resources. Jeong et al. propose to move the execution state of the task back to the mobile device from the edge server before the mobile device moves away in the context of web apps [304]. The mobile device continues the execution of the task in this way.

Since the number of edge servers and computation capacity of edge servers are limited, edge devices may compete for resources on edge servers. Hence, proper task scheduling and resource management schemes should be proposed to provide better services at edge. Yi et al. propose a latency-aware video edge analytic (LAVEA) system to schedule the tasks from edge devices [308]. For a single edge server, they adopt Johnson’s rule [413] to partition the inference task into a two-stage job and prioritise all received offloading requests from edge devices. LAVEA also enables cooperation among different edge servers. They propose three inter-server task scheduling algorithms based on transmission time, scheduling time, and queue length, respectively.

C. D2D offloading strategy

Most neural networks could be executed on mobile devices after compression and achieve a compatible performance. For example, the width-halved GoogLeNet on unmanned aerial vehicles achieves 99% accuracy [414]. Some works consider a more static scenario, where edge devices, such as smart watches are linked to smartphones or home gateways. Wearable devices could offload their model inference tasks to connected powerful devices. There are two kinds of offloading paradigms in this scenario, including binary decision-based offloading and partial offloading. Binary decision offloading refers to executing the task either on a local device or through offloading. This paradigm is similar to D2C offloading. Partial offloading means dividing the inference task into multiple sub-tasks and offloading some of them to associated devices. In fact, although the associated devices are more powerful, the performance of the complete offloading is not necessarily better than partial offloading. Because complete offloading is required to transmit complete input data to the associated device, which increases the latency. Table XIII summarises the existing literature for D2D offloading strategy.

Xu et al. present CoINF, an offloading framework for wearable devices, which offloads partial inference tasks to associated smartphones [305]. CoINF partitions the model into two sub-models, in which the first sub-model could be executed on the wearable devices, while the second sub-model could be performed on smartphones. They find that the performance of partial offloading outperforms the complete offloading in some scenarios. They further develop a library and provide API for developers. Liu et al. also propose EdgeEye, an open source edge computing framework to provide real-time service of video analysis, which provides a task-specific API for developers [306]. Such APIs help developers focus on application logic. Similar methods are also adopted in [415].

If one edge device is not powerful enough to provide real-time response for model inference, a cluster of edge devices could cooperate and help each other to provide enough computation resources. For example, if a camera needs to perform image recognition task, it could partition the CNN model by layers, and transmit the partitioned tasks to other devices nearby. In this scenario, a cluster of edge devices could be organised as a virtual edge server, which could execute inference tasks from both inside and outside of the cluster. Hadidi et al. propose Musical Chair, an offloading framework that harvests available computing resources in an IoT network for cooperation [309]. In Musical Chair, the authors develop data parallelism and model parallelism scheme to speed up the inference. Data parallelism refers to duplicating devices that performs the same task whilst model parallelism is about performing different sub-tasks of a task on different devices. Fig. [38] shows the parallel structure for a layer-level task, which are further partitioned into two sub-tasks on two devices. These two devices adopt different input to double the system throughput.

![Fig. 38. The parallelism structure of Musical Chair. Task B is a layer-level task, which are further partitioned into two sub-tasks on two devices. These two devices adopt different input to double the system throughput.](image-url)
Coninck et al. develop DIANNE, a modular distributed framework, which treats neural network layers as directed graphs [311]. As shown in Fig. 39, each module provides forward and backward methods, corresponding to feedforward and back-propagation, respectively. Each module is a unit deployed on an edge device. There is a job scheduler component, which assigns learning jobs to devices with spare resources. Fukushima et al. propose the MicroDeep framework, which assigns neurons of CNN to wireless sensors for image recognition model training [312]. The structure of MicroDeep is similar to DIANNE. Each CNN unit is allocated to a wireless sensor, which executes the training of the unit parameters. In feedforward phase, sensors exchange their output data. Once a sensor receives the necessary input, it executes its unit and broadcasts its output for subsequent layer. If a sensor with an output layer unit obtains its input and ground-truth, it starts the back-propagation phase. They adopt a 2D coordinate based approach to approximately allocate a CNN unit to sensors.

Distributed solo learning enables edge devices or edge servers to train models with local data. Consequently, each model may become local experts that are good at predicting local phenomena. For example, RSUs use local trained models to predict local traffic condition. However, users are interested in the traffic condition of places they plan to visit, in addition to the local traffic condition. Bach et al. propose a routing strategy to forward the queries to devices that have the specific knowledge [313]. The strategy is similar to the routing strategy in TCP/IP networks. Each device maintains a routing table to guide the forwarding. The strategy achieve 95% accuracy in their experiments. However, latency is a big problem in such frameworks.

D. Hybrid offloading

The hybrid offloading architecture, also named osmotic computing [416], refers to the computing paradigm that is supported by the seamless collaboration between edge and cloud computing resources, along with the assistance of data transfer protocols. The hybrid computing architecture takes advantage of cloud, edge, and mobile devices in a holistic manner. There are some efforts focusing on distributing deep learning models in such environments. Table XIV presents a summary of these efforts.

Morshed et al. investigate ‘deep osmosis’ and analyses the challenges involved with the holistic distributed deep learning architecture, as well as the data and resource architecture [314]. Teerapittayanon et al. propose distributed deep neural networks (DDNNs) based on the holistic computing architecture, which maps sections of a DNN onto a distributed computing hierarchy [315]. All sections are jointly trained in the cloud to minimise communication and resource usage for edge devices. During inference, each edge device performs local computation and then all outputs are aggregated to output the final results.

There is always the risk that the physical nodes i.e., edge devices and edge servers may fail, which results in the failure of DNN units deployed on these physical nodes. Yousefpour et al. introduce ‘deepFogGuard’ to make the distributed DNN inference failure-resilient [316]. Similar to residual connections [211], which skips DNN layers to reduce the runtime, ‘deepFogGuard’ skips physical nodes to minimise the impact of failed DNN units. The authors also verify the resilience of ‘deepFogGuard’ on sensing and vision applications.

E. Applications

There exists some works applying the above mentioned offloading strategy to practical applications, such as intelligent transportation [317], smart industry [318], smart city [319], and healthcare [320,321]. Specifically, in [317], the authors design an edge-centric architecture for intelligent transportation, where roadside smart sensors and vehicles could work as edge servers to provide low latency deep learning based services. [318] proposes a deep learning based classification model to detect defective products in assembly lines, which leverages an edge server to provide computing resources. Tang et al. develop a hierarchical distributed framework to support data intensive analytics in smart cities [319]. They develop a pipeline monitoring system for anomaly detection. Edge devices and servers provide computing resources for the execution of these detection models. Liu et al. design an edge based food recognition system for dietary assessment, which

TABLE XIII

| Ref. | Model       | Problem                        | Object                          | Latency | Energy consumption |
|------|-------------|--------------------------------|---------------------------------|---------|--------------------|
| [305] | DNN         | Model partition                | Acceleration, save energy       | 23×     | 85.5% reduction    |
| [306] | DNN         | Open source framework          | Enable edge inference           | N/A     | N/A                |
| [307] | AlexNet, VGGNet | Incremental training        | Improve accuracy                 | 1.4 − 3.3× | 30%−70% saving    |
| [308] | DNN         | Task scheduling                | Minimise latency                 | 1.2 − 1.7× | N/A                |
| [309] | DNN         | Data and task parallelism      | Computing power                  | 50×     | 200× reduction     |
| [310] | N/A         | Execution management           | ML deployments at edge           | N/A     | N/A                |
| [311] | AlexNet, VGGNet | Data, model parallelism     | Modular architecture             | N/A     | N/A                |
| [312] | CNN         | Neuron assignment              | Enable training/inference        | N/A     | N/A                |
| [313] | Bayesian     | Knowledge retrieval            | Routing strategy                 | N/A     | N/A                |

Fig. 39. The illustration of a DIANNE module. Each module has references to its predecessors and successors for feedforward and back-propagation during training.
splits the recognition tasks between nearby edge devices and cloud server to solve the latency and energy consumption problem [320]. Muhammed et al. develop a ubiquitous healthcare framework, called UbeHealth, which makes full use of deep learning, big data, and computing resources [321]. They use big data to predict the network traffic, which in turn is used to assist the edge server to make the optimal offloading decision.

VII. FUTURE DIRECTIONS AND OPEN CHALLENGES

We present a thorough and comprehensive survey on the literature surrounding edge intelligence. The benefits of edge intelligence are obvious - it paves the way for the last mile of AI and to provide high-efficient intelligent services for people, which significantly lessens the dependency on central cloud servers, and can effectively protect data privacy. It is worth recapping that there are still some unsolved open challenges in realising edge intelligence. It is crucial to identify and analyze these challenges and seek for novel theoretical and technical solutions. In this view, we discuss some prominent challenges in edge intelligence along some possible solutions. These challenges include data scarcity at edge, data consistency on edge devices, bad adaptability of statically trained model, privacy and security issues, and Incentive mechanism.

A. Data scarcity at edge

For most machine learning algorithms, especially supervised machine learning, the high performance depends on sufficiently high-quality training instances. However, it often does not work in edge intelligence scenarios, where the collected data is sparse and unlabelled, e.g., in HAR and speech recognition applications. Different from traditional cloud based intelligent services, where the training instances are all gathered in a central database, edge devices use the self-generated data or the data captured from surrounding environments to generate models. High-quality training instances, e.g., good image features are lacking in such datasets. Most existing works ignore this challenge, assuming that the training instances are of good quality. Moreover, the training dataset is often unlabelled. Some works [120], [124] propose to use active learning to solve the problem of unlabelled training instances, which requires manual intervention for annotation. Such an approach could only be used in scenarios with few instances and classifications. Federated learning approaches leverage the decentralised characteristic of data to effectively solve the problem. However, federated learning is only suitable for collaboration training, instead of the solo training needed for personalised models.

We discuss several possible solutions for this problem as follows.

- Adopt shallow models, which could be trained with only a small dataset. Generally, the simpler the machine learning algorithm is, the better the algorithm will learn from the small datasets. A simple model, e.g., Naive Bayes, linear model, and decision tree, are enough to deal with the problem in some scenarios, compared with complicated models, e.g., neural network, since they are essentially trying to learn less. Hence, adopting an appropriate model should be taken into consideration when dealing with practical problems.

- Incremental learning based methods. Edge devices could re-train a commonly-used pre-trained model in an incremental fashion to accommodate their new data. In such a manner, only few training instances are required to generate a customised model.

- Transfer learning based methods, e.g., few-shot learning. Transfer learning uses the learned knowledge from other models to enhance the performance of a related model, typically avoiding the cold-start problem and reducing the amount of required training data. Hence, transfer learning could be a possible solution, when there is not enough target training data, and the source and target domains have some similarities.

- Data augmentation based methods. Data augmentation enables a model to be more robust by enriching data during the training phase [417]. For example, increasing the number of images without changing the semantic meaning of the labels through flipping, rotation, scaling, translation, cropping, etc. Through the training on augmented data, the network would be invariant to these deformations and have better performance to unseen data.

B. Data consistency on edge devices

Edge intelligence based applications, e.g., speech recognition, activity recognition, emotion recognition, etc., usually collect data from large amounts of sensors distributed at the edge network. Nevertheless, the collected data may not be consistent. Two factors contribute to this problem: different sensing environments, and sensor heterogeneity. The environment (e.g., street and library) and its conditions (e.g., raining, windy) add background noise to the collected sensor data, which may have an impact on the model accuracy. The heterogeneity of sensors (e.g., hardware and software) may also cause the unexpected variation in their collected data. For example, different sensors have different sensitivities, sampling rates, and sensing efficiencies. Even the sensor data collected from the same source may vary on different sensors. Consequently, the variation of the data would result in the variation on the model training, e.g., the parameters of features...
Such variation is still a challenge for existing sensing applications.

This problem could be solved easily if the model is trained in a centralised manner. The centralised large training set guarantees that the invariant features to the variations could be learned. However, this is not the scope of edge intelligence. Future efforts of this problem should focus on how to block the negative effect of the variation on model accuracy. To this end, two possible research directions maybe considered: data augmentation, and representation learning. Data augmentation could enrich the data during the model training process to enable the model to be more robust to noise. For example, adding various kinds of background noises to block the variation caused by the environments in speech recognition applications on mobile devices. Similarly, the noise caused by the hardware of sensors could also be added to deal with the inconsistency problem. Through the training of the augmented data, the models are more robust with these variations.

Data representation heavily affects the performance of models. Representation learning focuses on learning the representation of data to extract more effective features when building models [420], which could also be used to hide the differences between different hardware. For this problem, if we could make a ‘translation’ on the representations between two sensors which are working on the same data source, the performance of the model would be improved significantly. Hence, representation learning is a promising solution to diminish the impact of data inconsistency. Future efforts could be made on this direction, e.g., design more effective processing pipelines and data transformations.

C. Bad adaptability of statically trained model

In most edge intelligence based AI applications, the model is first trained on a central server, then deployed on edge devices. The trained model will not be retrained, once the training procedure is finished. These statically trained models cannot effectively deal with the unknown new data and tasks in unfamiliar environments, which results in low performance and bad user experience. On the other hands, for models trained with a decentralised learning manner, only the local experience is used. Consequently, such models may become experts only in their small local areas. When the serving area broadens, the quality of service decreases.

To cope with this problem, two possible solutions may be considered: lifelong machine learning and knowledge sharing. Lifelong machine learning (LML) [421] is an advanced learning paradigm, which enables continuous knowledge accumulation and self-learning on new tasks. Machines are taught to learn new knowledge by themselves based on previously learned knowledge, instead of being trained by humans. LML is slightly different from meta learning [422], which enables machines to automatically learn new models. Edge devices with a series of learned tasks could use LML to adapt to changing environments and to deal with unknown data. It is worth recapping that the LML is not primarily designed for edge devices, which means that the machines are expected to be computationally powerful. Accordingly, model design, model compression, and offloading strategies should be also considered if LML is applied.

Knowledge sharing [423] enables the knowledge communication between different edge servers. When there is a task submitted to an edge server that does not have enough knowledge to provide a good service, the server could send knowledge queries to other edge servers. Since the knowledge is allocated on different edge servers, the server with the required knowledge responds to the query and performs the task for users. A knowledge assessment method and knowledge query system are required in such a knowledge sharing paradigm.

D. Privacy and security issues

To realise edge intelligence, heterogeneous edge devices and edge servers are required to work collaboratively to provide computing powers. In this procedure, the locally cached data and computing tasks (either training or inference tasks) might be sent to unfamiliar devices for further processing. The data may contain users’ private information, e.g., photos and tokens, which leads to the risk of privacy leakage and attacks from malicious users. If the data is not encrypted, malicious users could easily obtain private information from the data. Some efforts [283], [285], [286], [405] propose to do some preliminary processing locally, which could hide private information and reduce the amount of transmitted data. However, it is still possible to extract private information from the processed data [136]. Moreover, malicious users could also attack and control a device that provides computing power through inserting a virus in the computing tasks. The key challenge is the lack of relevant privacy preserving and security protocols or mechanisms to protect users’ privacy and security from being attacked.

Credit system maybe a possible solution. This is similar with the credit system used in banks, which authenticates each user participated in the system and checks their credit information. Users with bad credit records would be deleted from the system. Consequently, all devices that provide computing powers are credible and all users are safe.

Encryption could be used to protect privacy, which is already applied in some works [52], [126]. However, the encrypted data need to be decrypted before the training or inference tasks are executed, which requires an increase in the amount of computation needed. To cope with the problem, future efforts could pay more attention to homomorphic encryption [139]. Homomorphic encryption refers to an encryption method that allows direct computation on ciphertexts and generate encrypted results. After decryption, the result is the same as the result achieved by computation on the unencrypted data. Hence, by applying homomorphic encryption, the training or inference task could be direct executed on encrypted data.

E. Incentive mechanism

Data collection and model training/inference are two utmost important steps for edge intelligence. For data collection, it is challenging to ensure the quality and usability of information
of the collected data. Data collectors consume their own resources, e.g., battery, bandwidth, and the time to sense and collect data. It is not realistic to assume that all data collectors are willing to contribute, let alone for preprocessing data cleaning, feature extraction and encryption, which further consumes more resources. For model training/inference in a collaborative manner, all participants are required to unselfishly work together for a given task. For example, the architecture proposed in [1] consists of one master and multi workers. Workers recognise objects in a particular mobile visual domain and provides training instances for masters through pipelines. Such architecture works in private scenarios, e.g., at home, where all devices are inherently motivated to collaboratively create a better intelligent model for their master, i.e., their owner. However, it would not work well in public scenarios, where the master initialises a task and allocates sub-tasks to unfamiliar participants. In this context, additional incentive issue arises, which not typically considered in smart environments where all devices are not under the ownership of a single master. Participants need to be incentivised to perform data collection and task execution.

Reasonable incentive mechanisms should be considered for future efforts. On one hand, participants have different missions, e.g., data collection, data processing, and data analysis, which have different resource consumptions. All participants hope to get as much reward as possible. On the other hand, the operator hopes to achieve the best model accuracy with as a low cost as possible. The challenges of designing the optimal incentive mechanism are how to quantify the workloads of different missions to match corresponding rewards and how to jointly optimise these two conflicting objectives. Future efforts could focus on addressing these challenges.

VIII. CONCLUSIONS

In this paper, we present a thorough and comprehensive survey on the literature surrounding edge intelligence. Specifically, we identify critical components of edge intelligence: edge caching, edge training, edge inference, and edge offloading. Based on this, we provide a systematic classification of literature by reviewing research achievements for each component and present a systematical taxonomy according to practical problems, adopted techniques, application goals, etc. We compare, discuss and analyse literature in the taxonomy from multi-dimensions, i.e., adopted techniques, objectives, performances, advantages and drawbacks, etc. Moreover, we also discuss important open issues and present possible theoretical research directions. Concerning the era of edge intelligence, We believe this is only the tip of iceberg. Along with the explosive development trend of IoT and AI, we expect that more and more research efforts would be carried out to completely realize edge intelligence in the following decades.

REFERENCES

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” 2015.
[2] Y. Sun, D. Liang, X. Wang, and X. Tang, “Deepid3: Face recognition with very deep neural networks,” arXiv preprint [arXiv:1502.00873] 2015.
[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” Journal of machine learning research, vol. 3, no. Feb, pp. 1137–1155, 2003.
[4] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in Proceedings of the 25th international conference on Machine learning. ACM, 2008, pp. 160–167.
[5] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?” in Advances in neural information processing systems, 2017, pp. 5574–5584.
[6] H. A. Aliaja, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, “Augmented reality meets deep learning for car instance segmentation in urban scenes,” in British machine vision conference, vol. 1, 2017, p. 2.
[7] W. Huang, G. Song, H. Hong, and K. Xie, “Deep architecture for traffic flow prediction: deep belief networks with multitask learning,” IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 5, pp. 2191–2201, 2014.
[8] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, “Traffic flow prediction with big data: a deep learning approach,” IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865–873, 2014.
[9] Z. Fang, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, and S. Chen, “Abnormal event detection in crowded scenes based on deep learning,” Multimedia Tools and Applications, vol. 75, no. 22, pp. 14 617–14 639, 2016.
[10] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, “Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds,” in 2016 Computing in Cardiology Conference (CinC). IEEE, 2016, pp. 621–624.
[11] “Cisco visual networking index: Global mobile data traffic forecast update (2017–2022),” http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, 2016.
[12] P. Voigt and A. Von dem Bussche, “The eu general data protection regulation (gdpr),” A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017.
[13] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, Oct 2016.
[14] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, “Collaborative mobile edge computing in 5g networks: New paradigms, scenarios, and challenges,” IEEE Communications Magazine, vol. 55, no. 4, pp. 54–61, 2017.
[15] P. Garcia Lopez, A. Montresor, D. Epema, A. Dutta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, and E. Riviere, “Edge-centric computing: Vision and challenges,” ACM SIGCOMM Computer Communication Review, vol. 45, no. 5, pp. 37–42, 2015.
[16] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and Y. Young, “Mobile edge computing key technology towards 5G,” ETSI white paper, vol. 11, no. 11, pp. 1–16, 2015.
[17] F. Bonomi, R. Militto, J. Zhu, and S. Addepalli, “Fog computing and its role in the internet of things,” in Proceedings of the first edition of the MCC workshop on Mobile cloud computing, 2012, pp. 13–16.
[18] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, “In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning,” IEEE Network, vol. 33, no. 5, pp. 156–165, 2019.
[19] E. Li, Z. Zhou, and X. Chen, “Edge intelligence: On-demand deep learning model co-inference with device-edge synergy,” in Proceedings of the 2018 Workshop on Mobile Edge Communications. ACM, 2018, pp. 31–36.
[20] Z. Wang, Y. Cui, and Z. Lai, “A first look at mobile intelligence: Architecture, experimentation and challenges,” IEEE Network, 2019.
[21] H. Khelifi, S. Luo, B. Noar, A. Sellami, H. Mounlla, S. H. Ahmed, and M. Guizani, “Bringing deep learning at the edge of information-centric internet of things,” IEEE Communications Letters, vol. 23, no. 1, pp. 52–55, 2018.
[22] N. D. Lane and P. Warden, “The (deep) learning transformation of mobile and embedded computing,” Computer, vol. 51, no. 5, pp. 12–16, 2018.
[23] F. Chen, Z. Dong, Z. Li, and X. He, “Federated meta-learning for recommendation,” arXiv preprint [arXiv:1802.07676] 2018.
[24] Y. Chen, J. Wang, C. Yu, W. Gao, and X. Qin, “Fedhealth: A federated transfer learning framework for wearable healthcare,” arXiv preprint [arXiv:1907.09173] 2019.
Ubiquitous Computing and Wearable Computers. ACM, 2018, pp. 1409–1413.

[420] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[421] Z. Chen and B. Liu, “Lifelong machine learning,” Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 10, no. 3, pp. 1–145, 2016.

[422] R. Vilalta and Y. Drissi, “A perspective view and survey of meta-learning,” Artificial intelligence review, vol. 18, no. 2, pp. 77–95, 2002.

[423] A. Soller, J. Wiebe, and A. Lesgold, “A machine learning approach to assessing knowledge sharing during collaborative learning activities,” in Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community. International Society of the Learning Sciences, 2002, pp. 128–137.