DNNOFF: Offloading DNN-Based Intelligent IoT Applications in Mobile Edge Computing

Xing Chen, Member, IEEE, Ming Li, Hao Zhong, Member, IEEE, Yun Ma, Member, IEEE, and Ching-Hsien Hsu, Senior Member, IEEE

Abstract—A deep neural network (DNN) has become increasingly popular in industrial Internet of Things scenarios. Due to high demands on computational capability, it is hard for DNN-based applications to directly run on intelligent end devices with limited resources. Computation offloading technology offers a feasible solution by offloading some computation-intensive tasks to the cloud or edges. Supporting such capability is not easy due to two aspects: Adaptability: offloading should dynamically occur among computation nodes. Effectiveness: it needs to be determined which parts are worth offloading. This article proposes a novel approach, called DNNOFF. For a given DNN-based application, DNNOFF first rewrites the source code to implement a special program structure supporting on-demand offloading and, at runtime, automatically determines the offloading scheme. We evaluated DNNOFF on a real-world intelligent application, with three DNN models. Our results show that, compared with other approaches, DNNOFF saves response time by 12.4–66.6% on average.

Index Terms—Computation offloading, deep neural networks (DNNs), intelligent Internet of Things (IoT) application, mobile edge computing (MEC), software adaptation.

I. INTRODUCTION

Recent years have witnessed the remarkable improvements of a deep neural network (DNN). As the core machine learning technique [1], the DNN has been applied in industrial Internet of things (IoT) scenarios such as computer vision [2] and self-driving cars [3]. Meanwhile, more and more trained deep learning models have been deployed on intelligent end devices, such as wearable devices [4], vehicles [5], and unmanned aerial vehicles [6]. In this article, we call such trained models as DNN-based intelligent IoT applications.

Due to limited resources about computation and storage, complex DNN-based applications cannot be directly run on intelligent end devices. One feasible solution is to offload all or part of computational tasks to the cloud with sufficient resources [7], [8]. More specifically, DNNs are divided by the granularity of neural network layers [9]. Thus, some computation-intensive neural network layers can be offloaded to the cloud for execution, while other simpler neural network layers are processed locally.

However, the network communication between end devices and the cloud is likely to cause significant execution delay, and it seriously affects the user experience. To address this delay problem, mobile edge computing (MEC) has been introduced [10]. The mobile edges provide computing capabilities in close proximity to end devices and enable the execution of highly demanding applications in end devices while offering significantly lower latencies. Although MEC provides new opportunities to offload DNN-based applications among end devices, the cloud, and nearby edges, the prior approaches do not consider how to offload them in the new environment. On the one hand, as the environment is constantly changing, the offloading scheme of the DNN model shall be flexible for the need of adaptation. On the other hand, an offloading scheme shall make tradeoffs between the reduced execution time and the network delay, when it determines which layers will be offloaded and where to offload them, based on the changes of environment.

To fully release the potential of offloading, an offloading mechanism shall support on-demand changes for DNN-based applications and shall enable the execution of some parts of the DNN model on different computing nodes (including end devices, cloud, and edge servers). Afterward, there needs to be an efficient estimation model, which can determine which of its layers shall be offloaded. In summary, our main research questions are: 1) How to design a mechanism to support the automatic offloading of DNN-based applications in the MEC environment? 2) How to build an estimation model to determine the optimal offloading schemes? After the above questions are carefully handled, the problem of offloading can be reduced to a traditional optimization problem [11].

1551-3203 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
To address the aforementioned questions, we present a novel approach called DNNOff, which supports offloading DNN-based applications in the MEC environment. This article makes the following major contributions:

1) an offloading mechanism that enables DNN-based applications to be offloaded automatically and dynamically in the MEC environment. To achieve this, DNNOff translates a DNN-based application to a target program that is easier to offload;

2) an effective model to predict the latency of offloading schemes. DNNOff first extracts the structure and parameters of the DNN model and then uses a random forest regression model to predict the execution cost of each layer. Based on the prediction model, DNNOff can determine which parts shall be moved to MEC servers;

3) an evaluation on a real-world DNN-based application with AlexNet, VGG, and ResNet models. Our results show that DNNOff reduces the response time by 12.4–66.6% for complex DNN-based applications.

The rest of this article is organized as follows. Section II reviews the related work. Section III presents our approach, and Section IV evaluates it on a real-world application. Section V discusses some issues about applicability. Section VI introduces industrial applications. Finally, Section VII concludes this article.

II. RELATED WORK

Mobile devices are generally limited to storage space, battery life, and computing power [12]. To improve the performance of mobile applications, computation offloading has become the most widely used technology. MCC improves the performance of applications by sending computing-intensive components from end devices to the cloud. These applications are partitioned at different granularities, such as method, thread, and class. For example, MAUI [13] supports offloading at the granularity of methods. It allows annotating which parts of a program can be offloaded to the cloud and makes offloading decisions at runtime. CloneCloud [14] is a thread-based computation offloading framework, and it modifies virtual machines to support seamless offloading of threads to the cloud. DPartner [15] can offload classes, and it uses a proxy mechanism to access class instances. Furthermore, it calculates the coupling of classes and divides them into two sets. The two sets are deployed on the end device and the cloud server, respectively. However, MCC has inherent limitations, namely, long latency between end devices and remote clouds. Hence, MEC has been proposed, in which the service of cloud is increasingly moving toward nearby edges [16]. AndroidOff [17] supports mobile applications with the offloading capability at the granularity of objects for MEC.

It provides the mechanism to offload an object-oriented application and determine which parts shall be offloaded. However, the proposed works above cannot apply to DNN-based applications.

Computation offloading for DNN-based applications is further advanced in recent years. Neurosurgeon [9] showed that large amounts of data produced by DNN models should be uploaded to the cloud via wireless network, leading to high latency and energy consumption. For the sake of better performance and energy efficiency of modern DNN-based applications, Neurosurgeon designed a light weight scheduler to partition DNN-based applications automatically between end devices and the cloud at the granularity of neural network layers. Edgent [18] is a framework that automatically and intelligently selects the best partition point of a DNN model to satisfy the requirement on the execution latency. Compared with Neurosurgeon, Edgent can offload computation-intensive DNN layers to the remote server at a low transmission overhead, namely, nearest computation node. Liu et al. [19] proposed an image recognition framework based on the DNN in the MEC environment and realized the food image recognition system by employing an edge-computing-based service infrastructure. It allows the system to overcome some inherent limitations of the traditional MCC paradigm, such as high latency and energy consumption. Zhou et al. [20] proposed a robust mobile crowd sensing framework in the MEC environment. It can reduce the service delay with edge-computing-based local processing. The above approaches assume that end devices use a single remote server for computation offloading and cannot make efficient use of dispersed and changing computing resources in the MEC environment.

III. APPROACH

Fig. 1 presents the overview of DNNOff. For the nodes, we use rectangles to denote its components and circles to denote its internal data. For the edges, red ones denote data flows, and blue ones denote requests. DNNOff has three main components, namely, extraction, offloading mechanism, and estimation model. First, the extraction component extracts the structure and parameters of a DNN model (see Section III-A). Second, the offloading mechanism translates a DNN model to a target program that enables offloading (see Section III-B) and deploys it on end devices and remote servers where offloading may occur. Finally, the estimation model component deployed on the end device synthesizes an optimized offloading scheme to execute different parts of the target program on proper locations, based on the DNN network structure information and the surrounding MEC environment (see Section III-C). Moreover, the estimation model will update the offloading decision when the surrounding...
MEC environment changes. In Fig. 1, a DNN-based application and its MEC environment are presented on the right.

A. Extracting Structure for the DNN Model

Fig. 2 shows an example of the DNN model. A DNN model consists of layers. In Fig. 2, layers are represented as squares in different colors. In particular, the yellow one represents a convolution layer, which translates an image to a feature map with learned filters. The blue one represents an activation layer, which is a nonlinear function. The function accepts a feature map and generates an output with the same dimension. The purple one represents a pooling layer, which can be defined as a general pooling, an average pooling, or a max pooling. The green one represents a fully connected layer, which calculates the weighted sum of the inputs by learned weights. The top of square is the parameters of layer. For example, “channel:3” denotes that the corresponding value of the parameter “channel” of square is the parameters of layer. For example, “channel:3” denotes that the corresponding value of the parameter “channel” is “3.” The black arrow represents the data flow. DNNOff first extracts the structure of a DNN model, and the structure includes the parameters of each layer and the data flow between layers. Its definitions are as follows.

Definition 1 (DNN model structure): A DNN model structure is a directed graph \( G_D = (L, R) \) representing data transmissions between layers of a DNN \( D \), where \( L = \{l_1, l_2, \ldots, l_n\} \) is the set of layers of \( D \) and \( R \) is the set of data flow edges. Each edge \( r_{ij} \in R \) represents a data flow from \( l_i \) to \( l_j \).

Definition 2 (DNN layer information): A layer consists of type and parameters as \( l_i = \langle \text{type}, \text{feature} \rangle \), where type is the type of the layer and feature denotes the set of features of the layer.

In general, the DNN-based application stores its trained model in the configuration file, such as prototxt of Caffe.\(^1\) Our approach takes this file as the input and gets the DNN model graph \( G_D = (L, R) \) through code analysis.

B. Offloading Mechanism for the DNN Model

First, we translate an original application to a target program, and the translated target program follows the pipe-and-filter style. In this style, DNN layers are modeled as filters that receive and send data, and data flows between two layers are modeled as pipes that transmit the intermediate results. Second, we propose a “Pipe” engine to determine which neural network layer shall be offloaded.

\(^1\)[Online]. Available: https://caffe2.ai/
Algorithm 1: Pipe.

**Input:** \( m \) — the label of the “Pipe” function

**Output:** \( \text{CurrentL} \) — the label of the layer to be executed

Declare:
- \( \text{config} \) — the offloading scheme that records execution locations of each layer;
- \( \text{EndL} \) — the label of the last layer that is executed at Local;
- \( l_i \) — the result of the \( i \)th layer

1: \[ \text{if } m < \text{CurrentL} \text{ then} \]
2: \[ \text{return } \text{CurrentL} \]
3: \[ \text{end if} \]
4: \[ \text{if } m == \text{CurrentL} \text{ and } \text{config}[m] == \text{Local} \text{ then} \]
5: \[ \text{return } \text{CurrentL} \]
6: \[ \text{end if} \]
7: \[ \text{if } m == \text{CurrentL} \text{ and } \text{config}[m] != \text{Local} \text{ then} \]
8: \[ k \leftarrow \text{calculate the label of the next layer that shall be executed and the value of } \text{EndL} \]  
9: \[ \text{if } \text{EndL} == \text{Local} \text{ then} \]
10: \[ \text{return } \text{CurrentL} \]
11: \[ \text{end if} \]
12: \[ \text{end if} \]
13: \[ l_{k-1} \leftarrow \text{remote}(l_{\text{CurrentL}-1}, \text{CurrentL}, k-1) \]
14: \[ \text{CurrentL} \leftarrow k \]
15: \[ \text{return } \text{CurrentL} \]
16: \[ \text{end if} \]

Step 2 (Adding a “Pipe(i)” function before each layer \( l_i \)):
This function determines whether the layer \( l_i \) shall be offloaded (see Section III-B3 for details).

Step 3 (Adding two if statements to check each layer \( l_i \)):
The first statement is “\( \text{if CurrentL} == i \)”, where \( i \) represents the \( i \)th layer. It checks whether the layer \( l_i \) is to be executed currently.
The second statement is “\( \text{if CurrentL} - 1 == \text{EndL} \)” If the layer \( l_{\text{EndL}} \) has been executed, its result shall be returned, and the layers after \( l_{\text{EndL}} \) are skipped.

3) Computation Offloading at Runtime: At runtime, the “Pipe” functions connect each layer that can be executed locally or remotely, according to the offloading scheme. Algorithm 1 shows how the “Pipe” function works. Pipe(m) denotes the pipe between the layers \( l_m-1 \) and \( l_m \). config is the offloading scheme that records execution locations of each layer, and \( \text{CurrentL} \) denotes the label of the layer to be executed.

When \( m < \text{CurrentL} \) it indicates that the layer \( l_m \) has been executed and does not need to be repeated (lines 1–3). Therefore, the layer \( l_m \) will be skipped.

When \( m == \text{CurrentL} \) and \( \text{config}[m] == \text{Local} \) (Local is a keyword, representing the local node), it means that the layer \( l_m \) is to be locally executed (lines 4–6). Therefore, \( l_m \) is executed and the value of \( \text{CurrentL} \) is added by 1.

When \( m == \text{CurrentL} \) and \( \text{config}[m] != \text{Local} \), it means that the layer \( l_m \) is to be remotely executed (lines 7–15). Then, we calculate the label of the next layer that shall be executed at local (line 8), and if \( k \) does not exist, we assign “\( \text{EndL}+1 \)” to it (lines 9–11). Finally, we run the program \( DP(l_{\text{CurrentL}}, \text{CurrentL}, (k-1)) \) on the remote node according to \( \text{config}[m] \) and assign its result to \( l_{k-1} \) (line 12).

C. Estimation Model for the Offloading Scheme

1) Predicting Cost With Random Forest Regression: The execution time of each layer is an essential factor in the estimation model. If the layer \( l_i \) is executed on the node \( n_k \), we define its execution cost as follows.

Definition 3 (Execution cost): \( Cost_{l_i}^{n_k} = \text{time} \) 
\( \text{datasize} \) denotes the execution time from setting input data to generating output data, which depends on the performance of the execution node, while \( \text{datasize} \) denotes the amount of data transmission, which is a fixed value obtained by the extraction component.

With the number of layers and the diversity of computing nodes, it is difficult to get execution time of each layer on each computing node. Thus, we used the random forest regression to build prediction models for different layer types and computing nodes, which is to predict \( Cost_{l_i}^{n_k} \text{time} \). The RF regression model is proposed by Breiman [22] and is proved to carry out the nonlinear relation between the attributes. It is a nonlinear model-building tool, which is widely used in classification [23] and prediction [24].
Definition of the prediction model:

\[ Y = \text{predict}(X) \]  
\[ X_{\text{conv}} = (\text{channel}, k_{\text{size}}, k_{\text{number}}, \text{stride}, \text{padding}) \]  
\[ X_{\text{pooling}} = (\text{channel}, k_{\text{size}}, \text{stride}) \]  
\[ X_{\text{relu}} = (\text{in}_{\text{number}}, \text{out}_{\text{number}}) \]  
\[ X_{\text{fc}} = (\text{in}_{\text{number}}, \text{out}_{\text{number}}). \]  
\[ (1) \]  
\[ (2) \]

We use the dataset of history data to train the prediction model, which is collected from DNN applications, including Alexnet [25], VGG16, VGG19 [26], ResNet-50, and ResNet-152 [27]. The RF regression prediction model is represented as Equation (1). The input \( (X) \) depends on the type of layers as Equation (2), and the layer types include convolution layer, pooling layer, activation layer, and fully connected layer.

2) Contributory Factor: In this subsection, we introduce a context model that describes the environment (e.g., computation nodes) and the factors that affect the offloading decision.

The context architecture consists of an end device (ED), several nearby edges (NE), and a remote cloud (RC). We use a graph to present this network \( GC = (N, E) \), where \( N \) denotes a set of compute nodes, including end device and remote servers, and \( E \) represents a set of communication links among nodes \( n_i \in N \).

Each edge \((n_i, n_j) \in E\) is associated with a data transmission rate \( v_{n_i,n_j} \) and a round-trip time \( rtt_{n_i,n_j} \) between \( n_i \) and \( n_j \). A typical offloading scenario is as follows: The data are generated on the end device (the only \( n_i \)), and the layers can be offloaded to nearby edges (some nodes of \( N_{\text{NE}} \)) or the remote cloud (\( n_{\text{RC}} \)).

Table I shows our factors for estimating an offloading scheme. Among them, \( n_k \in N \), \( v_{n_k,n_p} \), and \( rtt_{n_k,n_p} \) are defined before. We next introduce \( DEP = \{\text{dep}(l_1), \text{dep}(l_2), \ldots, \text{dep}(l_n)\} \), where \( DEP \) denotes the offloading scheme. Each \( l_i \in L \) is executed on a computation node \( \text{dep}(l_i) \in N \). Let \( T_r(l_i) \) denote the execution time of \( l_i \) and let \( T_d(l_k, l_m) \) denote the data transmission time between layers \( l_k \) and \( l_m \). The response time of application can be represented by \( \text{Time} \), which is equal to the moment after the execution of the last layer \( t_{n} \). In addition, an objective function is constructed to calculate \( \text{Time} \) and estimate the offloading scheme.

3) Objective Function: Our objective function makes predictions, based on contributory factors. In particular, based on the factors in Section III-C2, we construct the objective function as shown in Equation (3). Here, we consider that a scheme is optimal, if its objective value is the smallest. As Table I shows that \( t_{i} \) is the moment after the execution of layer \( l_i \), the total response time is obtained when the last layer \( l_n \) is executed

\[ \text{Time} = t_n \]  
\[ (3) \]

The description of Equation (4) is expounded as follows: First, the moment before the execution of current layer \( l_i \) is calculated as the moment after the execution of previous layer \( (t_{p,i}) \) with the addition of the transmission time between two layers \((T_d(p_{i}, l_i))\). Second, according to the characteristic of the DNN, the current layer can only be executed when all branches from previous layers have already been executed. Hence, \( t_i \) includes the execution time of layer \( l_i \), and the maximum sum of the time at the moment after the execution of each parent layer with the addition of the transmission time between two layers. Among them, the execution time of layer \( l_i \) is represented as Equation (5) \((\text{Cost}_{\text{dep}(l_i)}^{l_i})(\text{Time})\) is mentioned in Definition 3) and the transmission time with previous layer is represented as Equation (6).
As a result, given an offloading scheme, the calculation of response time is shown in Algorithm 2. According to line 11 of Algorithm 2, we first initialize the value of $t_0$. Then, we use the “currentTime” function to calculate $t_n$ recursively according to line 12. The calculation principle of the “currentTime” function corresponds to Equation (4).

### IV. EVALUATION

We implemented DNNOff and conducted evaluations to explore the following research questions.

- **(RQ1)** To what degree does DNNOff improve performance of DNN-based applications (see Section IV-A)?
- **(RQ2)** How does DNNOff perform in cost prediction of each neural network layer (see Section IV-B)?
- **(RQ3)** How much extra overhead does DNNOff introduce (see Section IV-C)?

For RQ1, our results show that DNNOff saved 12.4–66.6% response time compared with other approaches. For RQ2, DNNOff achieved high accuracy for predicting execution time in different layers types and computing nodes. For RQ3, the overhead of our offloading mechanism is acceptable.

#### A. RQ1 Improvement Over the State of the Art

1) **Experimental Settings:**

   a) **Network environment:** The network context consists of four computation nodes; one end device and three remote servers. We simulate four locations, which are named community, traffic road, parking lot, and store. Table II lists the connections among our computation nodes. The column and the row of a cell denote the round-trip time and the data transmission rate between computation nodes. We utilize the network simulation tool DummyNet\(^2\) to control the available bandwidth. A smaller $rtt$ and a higher $v$ denotes a better signal strength.

   b) **Devices:** We take three desktop computers to emulate the Elastic Compute Service (ECS) and edge servers $E1$ and $E2$. The ECS is equipped with a 3.6-GHz 16-core CPU and 16-GB RAM, server $E1$ is equipped with a 2.5-GHz eight-core CPU and 8-GB RAM, and server $E2$ is equipped with a 3.0-GHz eight-core CPU and 8-GB RAM. We further use a smartphone to act as the end device, and the end device is equipped with a 2.2-GHz CPU and 4-GB RAM.

   c) **Application:** We use a real-world DNN-based image recognition application in the evaluation. It is written in Python and powered by the Caffe2 deep learning framework.


<table>
<thead>
<tr>
<th>Community</th>
<th>Traffic Road</th>
<th>Parking Lot</th>
<th>Store</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E1$</td>
<td>$RTT = 30$ ms $V = 1$ Mb/s</td>
<td>$RTT = 30$ ms $V = 1$ Mb/s</td>
<td>$RTT = 30$ ms $V = 800$ Kbs</td>
<td>$RTT = 30$ ms $V = 800$ Kbs</td>
</tr>
<tr>
<td>$E2$</td>
<td>$RTT = 30$ ms $V = 1$ Mb/s</td>
<td>$RTT = 60$ ms $V = 700$ Kbs</td>
<td>$RTT = 30$ ms $V = 200$ Kbs</td>
<td>$RTT = 30$ ms $V = 200$ Kbs</td>
</tr>
<tr>
<td>Cloud</td>
<td>$RTT = 150$ ms $V = 200$ Kbs</td>
<td>$RTT = 150$ ms $V = 200$ Kbs</td>
<td>$RTT = 150$ ms $V = 200$ Kbs</td>
<td>$RTT = 150$ ms $V = 200$ Kbs</td>
</tr>
</tbody>
</table>

We mainly concern with three models, which are the core of the DNN-based application, including AlexNet, VGG16, and ResNet-50. The most complex model is ResNet-50, while AlexNet is the simplest one. The inference latency and recognition accuracy are increasing as the model is more complex.

**d) Compared approaches:** In our evaluation, we compared DNNOff with four other approaches.

The original application is executed on end device, without any offloading.

**Neurosurgeon** [9] selected the best DNN partition point and sent the remaining DNN layers from end device to the cloud.

**Edgent** [18] is similar to Neurosurgeon [9], but offloads computation-intensive DNN layers to the remote server at a low transmission overhead, namely, nearest computation node.

For the ideal plan, it has to get execution time of each layer on each computing node and choose the fastest one after executing all the schemes in reality. The ideal plan is infeasible in practice since it needs to get the execution time at different levels in advance and try all the possibilities. We introduce the ideal plan to illustrate how close DNNOff is to the ideal one.

**e) Measurement:** To show the effectiveness of DNNOff, we define the following metrics.

1) **Total response time:** We use the total response time as the metric for performance. To make a fair comparison, we pick ten different images from the video in each location and calculate their averages for comparison. Here, the start time is recorded when the image is input, and the end time is recorded when the recognition result is output. It includes local inference, data transmission, and remote inference. The less response time indicates better results.

2) **Local inference:** This is the time about inference process on the end device. The inference on remote servers is usually more efficient than end device.

3) **Remote inference:** This is the time about inference process on the remote servers.

4) **Data transmission:** This is the time to transmit the feature vectors result by the partitioned layers of DNN model, and it is often slow under poor network connection.

2) **Results:** The total response time consists of the inference time and the transmission time. Fig. 5 shows the time of compared approaches in the four locations. For each approach, the blue bar denotes the local inference time, the orange one denotes the data transmission time, while the gray one denotes the remote inference time.

Compared with the original application, DNNOff reduces the total response time by 30.4–66.6%. The result also shows that the more complex the model, the better the optimization of DNNOff. In general, the optimization of community is better than that of store, because community is closer to the better performing edge.
server, which can significantly reduce the reference time. In the traffic road, the ResNet-50 is optimized to 66.6% with DNNOff, since the data transfer volume between the layers in ResNet-50 is small and the location is connected to all remote servers, so that the offloading can alleviate the bottleneck of local inference time and, meanwhile, guarantee a lower data transmission time. It should be noted that the parking lot is only connected to the cloud server, so the performance improvement is not as obvious as that in other locations, but it can still reduce the time by 30.4–47.1%. Hence, DNNOff is still effective even if there are no edge servers.

Compared with Neurosurgeon, DNNOff reduces the total response time by 26.5–53.2%. The results show that DNNOff significantly outperforms Neurosurgeon in the traffic road with the VGG model. Because traffic road has the best network connection with remote servers, which provide more choices to DNNOff for offloading, while in the parking lot, DNNOff can keep the same performance as Neurosurgeon. Due to the poor network connection, multiple partitions will increase the data transmission time instead. In this case, DNNOff makes the same offloading scheme as Neurosurgeon does.

Compared with Edgent, DNNOff reduces the total response time by 12.4–39.3%. Although Edgent considers the use of nearest computation node, DNNOff can cut the DNN at multiple points and execute different parts over the end device, edges, and the cloud.

Compared with the ideal plan, DNNOff can achieve comparable performance in different cases, and the performance gap between them is about 5%.

In summary, DNNOff saved 12.4–66.6% of the total response time compared with other approaches. Meanwhile, the results show that DNNOff achieves optimal/near-optimal performance of offloading.

### B. RQ2 Accuracy for Cost Prediction of DNN Layers

#### 1) Experimental Settings:

**a) Model training:** We use the dataset of history data to train the random forest regression prediction model, which is collected from DNN-based applications running on different computing nodes. In total, we collected the layer information about convolution layers, pooling layers, activation layers, and fully connected layers of 425 items, 320 items, 582 items, and 96 items, respectively. Table III shows some convolution layer items, which is collected on the end device, as an example. Column “Channel” lists the number of channels of convolution kernel. “k size” and “k number” list the size and the number of their filters, respectively. Columns “Stride” and “Padding” list the stride and the padding with which the filters are being applied. The inputs \((X)\) include the channel, \(k\text{size}, k\text{number}, \text{stride}, \text{and padding}\). The output (time) is denoted as the predicted value of layer latency. Based on the dataset, we randomly split the data items into two categories: 70% for training the prediction model and 30% for testing the quality of our model.

**b) Measurement:** We regard root-mean-square error (RMSE) and \(R^2\) as the evaluation measures of the prediction model.

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (\text{observed}_t - \text{predicted}_t)^2 (7)}
\]

\[
R^2 = 1 - \frac{\sum (\text{observed}_t - \text{predicted}_t)^2}{\sum (\text{observed}_t - \text{mean}_t)^2}. (8)
\]

RMSE is the sample standard deviation of the differences between predicted and observed values. \(R^2\) is commonly used to evaluate the quality of regression models. They are calculated according to Equations (7) and (8).

#### 2) Results: Table IV shows the accuracy of the random forest regression prediction model. It illustrates the RMSE and
TABLE IV
RMSE AND R-SQUARED OF THE PREDICTION MODEL ON THE TEST SET

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>RMSE</th>
<th>R-Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolution Layer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td>0.289 ms</td>
<td>0.91</td>
</tr>
<tr>
<td>Edge1</td>
<td>0.155 ms</td>
<td>0.93</td>
</tr>
<tr>
<td>Edge2</td>
<td>0.131 ms</td>
<td>0.93</td>
</tr>
<tr>
<td>Cloud</td>
<td>0.098 ms</td>
<td>0.94</td>
</tr>
<tr>
<td>Pooling Layer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td>1.210 ms</td>
<td>0.78</td>
</tr>
<tr>
<td>Edge1</td>
<td>0.845 ms</td>
<td>0.82</td>
</tr>
<tr>
<td>Edge2</td>
<td>0.789 ms</td>
<td>0.83</td>
</tr>
<tr>
<td>Cloud</td>
<td>0.524 ms</td>
<td>0.83</td>
</tr>
<tr>
<td>Activation Layer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td>0.058 ms</td>
<td>0.69</td>
</tr>
<tr>
<td>Edge1</td>
<td>0.029 ms</td>
<td>0.70</td>
</tr>
<tr>
<td>Edge2</td>
<td>0.022 ms</td>
<td>0.72</td>
</tr>
<tr>
<td>Cloud</td>
<td>0.012 ms</td>
<td>0.75</td>
</tr>
<tr>
<td>Fully-connected Layer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td>4.951 ms</td>
<td>0.57</td>
</tr>
<tr>
<td>Edge1</td>
<td>2.098 ms</td>
<td>0.62</td>
</tr>
<tr>
<td>Edge2</td>
<td>1.589 ms</td>
<td>0.66</td>
</tr>
<tr>
<td>Cloud</td>
<td>1.248 ms</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Fig. 6. Optimization of random forest parameters using RMSE.

$R^2$ results for predicting in different layer types and computation nodes. For RMSE, the smaller RMSE indicates the better model’s fitting degree [28], [29]. For $R^2$, an acceptable value of $R^2$ is greater than 0.5 [30], and the closer to 1, the better the model is. Table IV shows that the RMSE of the model is small and $R^2$ is greater than 0.5, illustrating that the prediction model is acceptable. And the high accuracy of prediction model lays the foundation for scheme estimation.

In addition, there are two parameters in random forest: $N_{tree}$, the number of regression trees grown based on a bootstrap sample of the collected layers, and $M_{try}$, the number of different predictors tested at each node. The two parameters ($N_{tree}$ and $M_{try}$) are optimized based on the RMSE of calibration. Take the training of convolution layers on the end device as an example. $N_{tree}$ values from 500 to 4000 with intervals of length 50 were tested, and $M_{try}$ was tested from 1 to 5. The results of random forest parameters ($N_{tree}$ and $M_{try}$) are shown in Fig. 6, which clearly indicates that random forest parameters affect the error of prediction. The optimization was done using the calibration dataset ($n = 297$) and RMSE. The result $N_{tree} = 2000$ and $M_{try} = 3$ yielded the lowest RMSE (0.289 ms). In this case, we chose $N_{tree} = 2000$ and $M_{try} = 3$ as the best parameters.

C. RQ3 Extra Overhead

1) Experimental Settings:
   a) Setting: We use a simple AlexNet [25] application with 24 layers, which is a state-of-the-art DNN for image classification, and simulate five typical offloading schemes, which represent device-cloud, device-edge, and device-edge-cloud offloading, as shown in Table V.

   b) Compared approaches: We evaluate the overhead of DNNOff by comparing the performance of the adaptive offloaded application with the manual-modified offloaded application for five typical offloading schemes. The adaptive offloaded application is dynamically offloaded according to the offloading scheme, which is supported by our framework. The manual modified one is implemented by separating the code according to the offload scheme case by case.

2) Results: We run the application in the five typical offloading schemes and, respectively, record their average response time, as shown in Fig. 7. We can see that the response time of DNNOff is similar to the manual modified one, but with an overhead of 120–150 ms. The slight increase of response time (under 10%) is due to the condition statements of pipes that are needed to go through for each layers execution in our framework. For instance, the overheads in cases 1–3 are all about 120 ms because the cutoff points of three offloading scheme are the same. The overhead in cases 4 and 5 are both 150 ms because there are two cutoff points in each offloading scheme, for which more condition statements need to be executed. Overall, the overhead is acceptable.

V. DISCUSSION

Some issues about applicability need to be further discussed.

A. Online Decision

For online decision, DNNOff uses the estimation model to calculate the response time given an offloading scheme, based on which the problem of online decision can be reduced to a traditional optimization problem. Some algorithms can be used to reduce overhead. For instance, it takes about minutes to determine the offloading decision for the genetic algorithm,
while it just takes milliseconds to determine for the greedy algorithm; considering the performance and overhead of two algorithms, they are suitable to work in different situations [31]. However, this study mainly focuses on supporting DNN-based applications with the offloading capability in an MEC environment, and the issue above is orthogonal to the problem in this study. For future work, some state-of-the-art algorithms can be introduced to enhance our framework, such as deep reinforcement learning [32].

### B. Energy Saving

Complex applications usually have many computation-intensive tasks and consume a great deal of energy. Although the battery capacity of end devices keeps growing continuously, it still cannot keep pace with the growing requirements of intelligent applications. Computation offloading is a popular technique to help reduce the energy consumption of intelligent application as well as improve its performance [10]. Because of space limitation, this article mainly focuses on performance improvement by offloading. For future work, energy consumption can be introduced to the objective function (see Section III-C) of our framework, wherein energy consumption reduced by offloaded computing and extra energy consumption caused by communication should be both considered [13], [14], [33].

### VI. INDUSTRIAL APPLICATIONS

Recently, unmanned aerial/ground vehicles have begun to be applied in industrial IoT scenarios [34], such as patrolling the forest and delivering meals. These intelligent IoT applications have to rely on computer vision, whose cores are large-scale and complex DNNs, and thus, they commonly require sufficient resources and lead to high energy consumption. In the MEC environment, some computationally complex DNN layers are offloaded to the cloud or edges, while other tasks with simpler DNN layers are processed locally. This paradigm can improve performance of DNN-based intelligent IoT applications.

DNNOff first automatically translates the DNN-based application to a target program that is easier to offload. As the unmanned vehicle shifts during the day, its context (e.g., locations, network conditions, and available mobile edges) keeps changing. When the context changes, DNNOff synthesizes an optimal scheme for the intelligent application and then offloads its DNN layers according to the scheme.

### VII. CONCLUSION

The DNN has become increasingly popular in intelligent IoT applications. Due to limited resources about computation and storage on end devices, complex DNN-based applications cannot be directly run on end devices. Although many researchers have considered partitioning DNN models between end devices and the cloud, we believe that it can completely release the potential of offloading DNN if an application can be partitioned at more cut-points and determine which parts shall be offloaded to MEC servers. With this insight, this article presents DNNOff, a novel approach that supports offloading DNN-based applications in MEC. DNNOff can enable a DNN-based application to run its different parts over the end device, the cloud, and edges and automatically determine the offloading scheme based on cost estimation. We evaluate DNNOff on a real-world intelligent IoT application with three DNN models. Results show that DNNOff can significantly reduce the response time.

### REFERENCES

and Technological Progress, separately, in 2018 and 2020.


Dr. Chen was awarded two First Class Prizes for Provincial Scientific and Technological Progress, separately, in 2018 and 2020.

Xing Chen (Member, IEEE) received the B.S. and Ph.D. degrees from Peking University, Beijing, China, in 2008 and 2013, respectively. Since 2020, he has been a Professor with Fuzhou University, Fuzhou, China, where he is also the Deputy Director of the Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing and leads the Systems research group. His current projects cover the topics from self-adaptive software, computation offloading, model-driven approach, and so on. He has authored or coauthored more than 50 journal and conference articles. His research interests include software systems and engineering approaches for cloud and mobility.

Dr. Chen was awarded two First Class Prizes for Provincial Scientific and Technological Progress, separately, in 2018 and 2020.

Xing Chen (Member, IEEE) received the B.S. and Ph.D. degrees from Peking University, Beijing, China, in 2008 and 2013, respectively.

Since 2020, he has been a Professor with Fuzhou University, Fuzhou, China, where he is also the Deputy Director of the Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing and leads the Systems research group. His current projects cover the topics from self-adaptive software, computation offloading, model-driven approach, and so on. He has authored or coauthored more than 50 journal and conference articles. His research interests include software systems and engineering approaches for cloud and mobility.

Dr. Chen was awarded two First Class Prizes for Provincial Scientific and Technological Progress, separately, in 2018 and 2020.

Ming Li received the B.S. degree in computer science and technology in 2019 from Fuzhou University, Fujian, China, where he is currently working toward the M.S. degree in computer technology with the College of Mathematics and Computer Science.

Since September 2019, he has also been a part of the Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University. His current research interests include software engineering and edge computing.

Ming Li received the B.S. degree in computer science and technology in 2019 from Fuzhou University, Fujian, China, where he is currently working toward the M.S. degree in computer technology with the College of Mathematics and Computer Science.

Since September 2019, he has also been a part of the Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University. His current research interests include software engineering and edge computing.

Dr. Zhong is a recipient of the ACM SIGSOFT Distinguished Paper Award 2009, the Best Paper Award of the 2009 IEEE/ACM International Conference on Automated Software Engineering, and the Best Paper Award of the 2008 Asia-Pacific Software Engineering Conference. His Ph.D. dissertation was nominated for the distinguished Ph.D. dissertation award of China Computer Federation. He is a Member of the ACM.

Yun Ma (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2011 and 2017, respectively. He is currently a Postdoctoral Researcher with the School of Software, Tsinghua University, Beijing. Currently, he focuses on synergy between the mobile and the Web, trying to improve the mobile user experience by leveraging the best practices from native apps and Web apps. His research interests include mobile computing, Web technologies, and service computing.

Ching-Hsien Hsu (Senior Member, IEEE) is currently a Chair Professor and the Dean of the College of Information and Electrical Engineering, Asia University, Taichung, Taiwan. He is also a Professor with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan. He has authored or coauthored over 200 papers in top journals such as IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON SERVICES COMPUTING, ACM Transactions on Multimedia Computing, Communications, and Applications, IEEE TRANSACTIONS ON CLOUD COMPUTING, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, IEEE SYSTEM, and IEEE NETWORK, top conference proceedings, and book chapters in these areas. He has been acting as an Author/Co-Author or an Editor/Co-Editor for ten books from Elsevier, Springer, IGI Global, World Scientific, and McGraw-Hill. His research interests include high-performance computing, cloud computing, parallel and distributed systems, big data analytics, and ubiquitous/pervasive computing and intelligence. Prof. Hsu is a Fellow of the Institution of Engineering and Technology.