

Prune-Deprune: Adaptive Compression-Aware Split Learning and Inference for Enhanced Network Efficiency

AKRIT MUDVARI, Yale University, USA

ANTERO VAINIO, University of Helsinki, Finland

IASON OFEIDIS, Yale University, USA

SASU TARKOMA, University of Helsinki, Finland

LEANDROS TASSIULAS, Yale University, USA

The growing number of AI-driven applications in mobile devices has led to solutions that integrate deep learning models with the available edge-cloud resources. Due to multiple benefits such as reduction in on-device energy consumption, improved latency, improved network usage, and certain privacy improvements, split learning, where deep learning models are split away from the mobile device and computed in a distributed manner, has become an extensively explored topic. Incorporating compression-aware methods (where learning adapts to compression level of the communicated data) has made split learning even more advantageous. This method could even offer a viable alternative to traditional methods, such as federated learning techniques. In this work, we develop an adaptive compression-aware split learning method ('deprune') to improve and train deep learning models so that they are much more network-efficient, which would make them ideal to deploy in weaker devices with the help of edge-cloud resources. This method is also extended ('prune') to very quickly train deep learning models through a transfer learning approach, which trades off little accuracy for much more network-efficient inference abilities. We show that the 'deprune' method can reduce network usage by 4x when compared with a split-learning approach (that does not use our method) without loss of accuracy, while also improving accuracy over compression-aware split-learning by 4 percent. Lastly, we show that the 'prune' method can reduce the training time for certain models by up to 6x without affecting the accuracy when compared against a compression-aware split-learning approach.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Networks** → *Network management*.

Additional Key Words and Phrases: Deep Learning, Neural Network Splitting, Compression, Pruning, Transfer Learning, Augmented Reality, Internet of Things, Edge Computing, Fog Computing

1 INTRODUCTION

The past decade or so has seen a meteoric rise in the scope and abundance of AI applications with deep learning as a core tool, leading to solutions or better performance in many areas such as speech recognition, computer vision and much more [9, 12, 14, 23, 31]. It is estimated that the global AI market size was 120 billion USD in 2022 and is expected to reach 1.5 trillion dollars by 2030 [42], and even in each of our daily lives the impact of AI is becoming more and more pronounced, i.e., with the recent release of ChatGPT [39]. While beneficial, these methods, especially deep learning, tend to be quite expensive to execute. Supervised learning methods rely on huge datasets to learn, leading to a colossal demand for data acquisition and storage. The increasing demand of computational resources also pushed the development of dedicated hardware platforms, i.e., Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs), whose goal is to speed-up the training tasks. Another significant source of cost tends to be energy consumption in the devices that run the deep learning models, which not only poses an economic challenge but also raises other concerns such as energy scarcity and climate impacts [34]. It has become quite necessary to

Authors' addresses: Akrit Mudvari, akrit.mudvari@yale.edu, Yale University, New Haven, CT, USA, 06511; Antero Vainio, University of Helsinki, Helsinki, Finland, antero.vainio@helsinki.fi; Iason Ofeidis, Yale University, New Haven, CT, USA, 06511, iason.ofeidis@yale.edu; Sasu Tarkoma, University of Helsinki, Helsinki, Finland, sasu.tarkoma@helsinki.fi; Leandros Tassiulas, Yale University, New Haven, CT, USA, 06511, leandros.tassiulas@yale.edu.

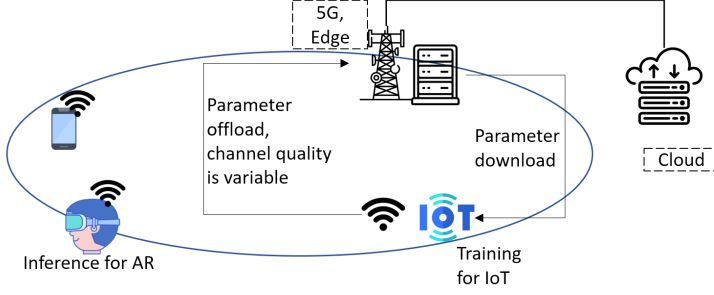


Fig. 1. Illustration of different weaker devices opting for nearby computation sources, for AI tasks

envision and implement these deep learning models in settings that are optimized towards saving different forms of resources.

There have been various efforts resulting in reduction of computational costs, i.e. through the implementation of smaller deep learning models [7, 18, 51] and model pruning including numerous works on parameter pruning [4, 32, 54] as well as some works on feature pruning [13, 15, 36], which allows for utilization of less GPUs and for shorter periods of time (for both inference as well as training). While this helps some devices reduce both energy costs as well as computational latency, it is not adequate for the vast majority of end devices like cell phones, laptops, IoT devices and AR/VR headsets to locally carry out the implementations [16, 24]; as a result the better alternative for these devices tends to be utilizing another more powerful source of computation, i.e. via different forms of distributed learning. This will inevitably involve transferring certain amount of data over the communication networks to locations such as data centers, resulting in a higher network usage as well as introduction of a new source of latency. The ideal solution would therefore involve selecting a method that minimizes network usage and latency.

Different strategies have been proposed for both training and inference of deep learning models in the networked environments. For instance, with the proliferation of data-driven services, such as targeted advertisement, personalized suggestion systems, and more, it has become crucial to find an efficient method of conducting deep learning tasks that is both computation and communication efficient, but also has other features including privacy protection. While one could simply offload all the computation task to the nearby server for further processing, such an approach has a few problems and shortcomings, which split learning and inference are able to overcome. Split learning (at least as we define and discuss in this paper) refers to splitting a neural network into multiple partition so that different parts of the deep learning model can be computed in different ways. For instance, consider a neural network with layers $l_1, \dots, l_m, \dots, l_n$ where $m < n$; here split learning can involve processing l_1, \dots, l_m locally and then offloading l_{m+1}, \dots, l_n to a server for further computations. Such split learning methods can provide us with various benefits over simply offloading the model as a whole. First, unlike offloading raw data, we are able to send a processed output, which is able to remain privacy preserving [17, 50, 53]. Second, the granularity provided by such division of task allows for more optimal scheduling and placement. Third, and perhaps more promising, as evidenced with recent research works [2, 50], it could also help with personalization of learned models by keeping the locally trained parts more personalized, and having a globally shared set of layers for aggregated learning.

Certain other approaches allow for similar benefits in certain cases. Federated learning for instance has emerged as a popular and widely researched method of learning and aggregating the learned information in a privacy preserving manner. However, split learning has emerged as a

promising alternative to federated learning [cite comp]; split learning has been shown to perform better in numerous scenarios including when the number of model parameters is larger, or when more clients are involved [47]. They are also more communication-efficient in many situations [49] as compared with federated learning, and have the capability to be privacy preserving as raw data need not be offloaded. These reasons make exploration of split learning as a method of distributed deep learning quite important and promising.

Previous works have explored split learning, i.e, in [6, 19, 44], which we will explore in the related works section in further details. Some of these works, for example in [6, 44], have explored the feasibility of conducting split learning and inference in different networked scenarios, and have introduced state of the art compression methods to reduce network traffic to a great extent, while minimally affecting the model accuracy. However, while these advances are quite important for making split method useful for day-to-day applications, there are certain shortcomings, which need to be addressed before this method can become widely feasible. While network-efficient split methods can help conduct spit learning and inference, the most glaring weakness of current state-of-the-art approaches is that training has to happen again and again for different split configurations, specially for different compression levels. Various factors like network conditions, operation budget, latency constraints, and tolerance of service level agreement violation, result in the network operator having different preference when it comes to compression and accuracy trade off. Hence providing models suited to differ rent compression-accuracy trade offs/ compression levels becomes very time and resource consuming, or even outright impractical.

In this work, we introduce a method that uses pruning and transfer learning to very quickly train deep learning models with different compression levels for the data being transferred during split learning, allowing for efficient creation of models suitable for different network conditions, computation resources, and service level agreements. This is done by introducing an encoder-decoder module with a method of training across different compression levels and at different positions of splitting. One use case of our approach would be providing AI-driven services like Augmented Reality (AR) with latency constraints to handheld devices in a wireless environment. AI/AR applications requiring object detection may employ models like YOLO [41] or SSD [25] but will not be able to meet the users' latency requirements [1]. With the models developed using our aforementioned approach, we could provide each user with different accuracy-compression trade offs, while split learning allows for optimal distribution of workload and an option to avoid offloading raw user data such as video feed or captured image. Hence, service level agreements can be met more efficiently.

This encoder-decoder module, further explained in section 3, is a tool that uses the learned information from a model with certain compression level, and adaptively learns to optimize for a different compression level. As mentioned above, one of the two goals is to develop these approximately similar models suitable for different user requirements and environments; however, a modified version of the aforementioned approach is also very useful in a second way, where this encoder-decoder approach can be used to adaptively change compression configurations to train in a way where the training can result in effectively same level of accuracy at the end while significantly cutting down on network usage during the distributed training. By training on a high compression level, and then towards the end of the training process, adaptively switching to a lower compression level using the deprune method (discussed further in section ??), we observed that network usage could be reduced without significantly effecting the accuracy of the final learned model. Such training can be specially beneficial when weaker devices, i.e., at the edge of the network, have limited computation capability as well as network cost; and since split learning is involved, raw data does not have to be offloaded.

2 RELATED WORKS

Multitude of work have focused on extending placement and scheduling problems in an edge-cloud paradigm, allowing for a generic workload to be delegated to servers in a networked environment. For instance, in [38], any service that requires computational offloading is placed at the network edge and defined constraints include storage, computation and communication costs, and in [37] the process is extended to data-intensive tasks. In [35] and [8], algorithms are developed for service placement with guarantees for near optimal solutions, with [35] creating a deterministic algorithm for general service placement and [8] working on polynomial time solution for data-intensive tasks. As we can see in survey papers such as [26, 27, 48], different types of placement and scheduling methods may be utilized for implementing state of the art techniques for efficient allocation of communication and computation resources in the edge-cloud paradigm. While these methods are an acceptable improvement over a naive deployment of deep learning models, for either training or inference, their generalized scope means that they fail to take into consideration unique features of deep learning models, or specific requirement of the clients, that may be leveraged to provide a more efficient allocation scheme for such implementations under edge-cloud environment. And our developed methods can complementarily be implemented together with such placement and scheduling approaches.

In [33], the authors develop a measurement-driven framework that decides which deep learning model to run and where based on the different model-related metrics such as accuracy, frame rate of processed videos, energy consumption, and network utilization. In [11], an approximate model scheduling is developed, where the trade-off between accuracy and resources is harnessed to optimize model accuracy. Various factors such as device energy, cloud costs and capacity, and execution deadline are taken as constraints, and an online variant of the method is also developed. Complementing the scheduling and placement research, there are methods of creating variants of the DNNs that aim to reduce space or computational demands, effectively creating a different NN such as through quantization of the entire network [10], binarization of the models [40], representation of the model as a lower rank representation after singular value decomposition [52], whole model compression through algorithms consisting of low-rank tensor decomposition [20], and more.

Complementary to a lot of such efforts that treat the model as a whole and attempt to efficiently orchestrate NN model deployments, it is possible to add granularity to the methods and also potentially improve privacy [17, 53] by splitting the neural network into different parts for computation at different devices. While methods like deep leakage [55] have haunted federated learning methods for a while now by eliminating privacy, no such method exists for reverse engineering split learning data to the best of our knowledge. Intuitively, there are a lot more parameters being updated in federated learning as opposed to the number of feature maps communicated during split learning, so the process developed to counter split learning is possibly a harder endeavor if not impossible.

In [19], the authors proposed a method for collaborating intelligence among the end devices and the mobile edge by dividing the neural network for partial computation at each end. Here, the authors investigated methods of discerning the best points of splitting and also showed that this method could be used to improve inference latency and mobile device's energy efficiency, for a wide range of deep learning models. This method is now well-understood as split learning (or vertical split learning, but we will use split learning to mean the vertical case). Capitalizing on this observation regarding the viability of the deep learning models' splitting as shown in [19], authors in [6] developed a compression aware training method for image classification neural networks that inserts a 'bottleneck structure' at the point of model splitting for efficient lossy compression. These 'bottleneck' structures use convolution methods to decrease the dimensions of the feature map for compression, and use de-convolution to obtain the larger image once the

communication of compressed data has taken place. The authors show that lossy compression approach is complementary to lossless methods and also show great improvements in end-to-end latency and mobile energy consumption. [44] improves on the compression structure proposed in [6] specifically considering different network channels. Here, they also demonstrate the superiority of such compression and decompression-based training method over the other methods including jpeg [21] method and quantization+Huffman [45] method, towards improving both communication overhead as well as transmission latency. [28] proposes a method for specific classification tasks by running a generalized head network distillation in the first stage and then targeted knowledge distillation in the second state. Other works such as [3, 29, 30] have focused on object detection tasks, where the compression mechanism is analyzed for the object detection. While a lot of these works have focused on implementing a compression-aware split learning that reduces network usage by splitting at certain layer and compressing the data to be transferred, the glaring inadequacy in these methods is that every time a new compression level is needed, or if compression behavior needs to change, the training has to start from the scratch. Once a compression factor has been decided on, i.e., by how much to compress, this value cannot dynamically change. This limits the scope of compression-aware training approaches since it would be unfeasible if not just inefficient to train from scratch for every compression configuration.

In our work, we come up with an adaptive compression-decompression module that creates approximate models for deep learning architectures suitable for split learning and inference, and we introduce two different novel algorithms, 'deprune' and 'prune', that greatly improve the performance during training as well as inference, i.e., by allowing us to train the deep learning model much more quickly and by improving models for faster inference. In the first step, we develop an adaptive compression-decompression method that allows for dynamic adjustment of compression levels through a 'filter' variable that we can change during training. In the second step, we introduce 'deprune' method which uses this module to significantly reduce network usage during the training under a split-learning environment, while also reducing the latency. We also show the authenticity of this method by deploying it in networked testbeds. This approach is highly beneficial for IoT and other weaker end devices with AI tasks, as network usage and training latency can be significantly improved for split learning. Finally, we introduce the 'prune' method that can be used to transfer learned parameters to dynamically create deep learning models for inference that greatly outperform having to train from scratch, in that the training time is significantly reduced. Such a method can create approximate models for different network requirements that trade little accuracy for network performance, and help create deep learning models that the AI tasks can implement in a network constrained environment; i.e., a network manager may reduce performance for AI-driven XR tasks by selecting a model with higher compression level if network traffic becomes an issue.

3 PROPOSED METHODS

In this section, we will describe the split learning architectures, including methods of training and implementing the algorithms. In subsection 3.2, we propose a method where the network communication cost of training is reduced using our method. And in subsection 3.3, we will describe a method of efficiently creating deep learning architectures that are able to quickly train and provide a set of models that offer a trade-off between accuracy and inference speed. But first, we will describe the formulation for deep learning architecture, as well as compression and decompression methods that we implement in our split-learning architecture, in subsection 3.1.

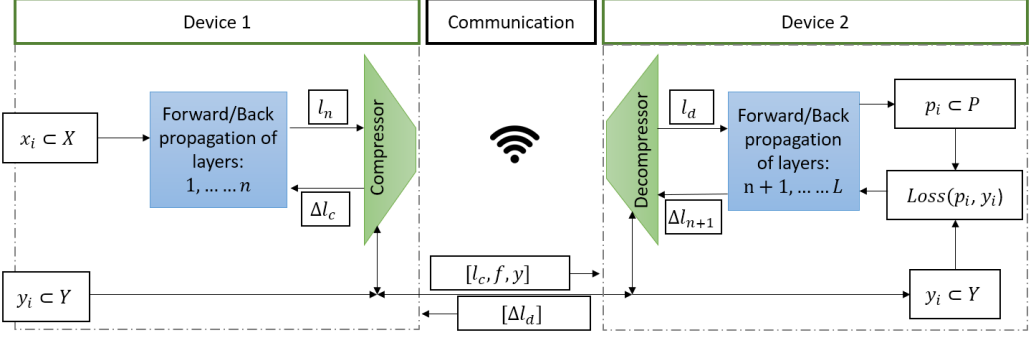


Fig. 2. System Overview: represents split learning across multiple devices with the compression-decompression module involved

3.1 Formulation

Our implementation is designed for general deep learning models, and so we define a model M with an arbitrary L number of layers where output of the layers are represented by $l = l_0, l_1, l_2, \dots, l_L$; in this work, we assume the input to the neural network to be l_0 and the model prediction is l_L . The parameters of the model is then given by $\theta = [\theta_1, \theta_2, \dots, \theta_L]$, where θ_l represents the parameters of model M corresponding to the layer l . The first step in implementing a split learning design is to split the model into different parts so that the different parts can be implemented at different locations, i.e., in a model with 15 layers, the first 3 layers may primarily be computed in the end device which is generating or storing the input data, while the remaining layers may be computed in the edge device or the cloud. Similarly, for training cases, the backpropagation will be calculated at different locations as well, i.e., the last 12 layers may be computed in the edge/cloud for our example case, and then the needed backpropagation information may be sent to the end device for backpropagation calculation for the remaining 3 out of 15 layers.

Figure 2 shows an overview of our architecture for split implementation of a deep learning model. Let us take the model to have L layers, where the splitting takes place after n layers such that layers $1, \dots, n$ are computed at device 1, and the remaining $n+1, \dots, L$ layers are computed at device 2. Any input $x_i \in X$ is generated or provided to device 1, along with the corresponding real output $y_i \in Y$. In the forward propagation phase, the input x_i is computed through the first n layers in device 1 to give us the output of layer n , which is l_n . But since l_n may be very large compared to the desired communication constraints, we introduce a compression mechanism (we will call this compression module) to reduce the size of l_n through a lossy compression method that compresses l_n to l_c . This compression module is also computed on device 1. Once the data is communicated to device 2, a decompression module acts upon l_c to give l_d , which becomes input for layer $n+1$. Then the remaining $L-n$ layers are computed on device 2 to give the prediction output p_i .

As we note in figure 2, the communication from device 1 to 2 not only carries with it l_c but also two other variables, f and y . f is a filter parameter which we will discuss later while talking about compression and decompression modules (we will refer to them together as compression-decompression module.); f is used to define which feature maps are communicated across the devices. While y_i is the real output corresponding to the input x_i , y_i is not needed in case of inference and we conclude after getting the prediction p_i , which may be sent back to device 1 (in any case it is not a communication costly task), but for training, next steps remain. Taking p_i and y_i , we calculate the loss $\delta L(p_i, y_i)$, which is then back-propagated through the neural network and decompression

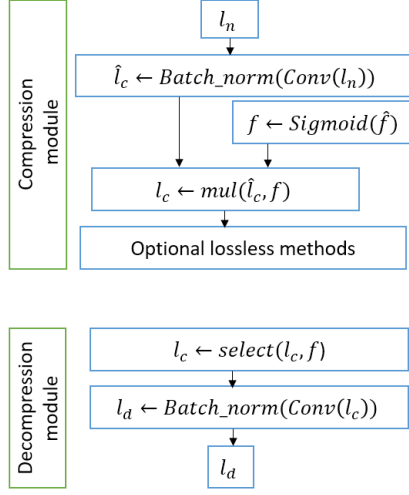


Fig. 3. Illustration of Compression and Decompression modules

module to obtain δl_d . δl_d is transmitted to device 1 where the remaining backpropagation can take place. Thus both parts of the vertically split neural networks are trained in two different devices during the training phase.

While this unit may be a node in a multilayer perceptron or feature maps in CNNs, the method is designed to achieve compression for a range of general NNs; since the names are different in different context, we will refer to this unit as feature map for the sake of brevity in explanation. An equivalent unit would be the nodes at layer n on multilayer perceptron.

Compression and decompression Modules: In this section, we will discuss the compression and decompression modules mention earlier. The first and optional step for the input to the compression model l_n may be to go through an encoder mechanism to achieve a resolution compression for each node/feature map. This is an optional method and dependent on the kind of neural network being implemented. For multi-layer perceptron, the nodes are simply a real number and do not require this kind of compression. In convolutional neural network for a potential resolution compression, this is achieved by controlling the kernel dimension and stride dimension of the convolution layer. We define resolution compression as the compression in size of each feature map of the input layer l_n , and resolution compression factor r as the reduction is size of the feature map. For a given r , to achieve the desired compression, we define kernel height $kc_h = 2 + r$, kernel width $kc_w = 2 + r$ stride value $sc = r$ and padding value $pc = r$. Then the values are passed through a batch normalization layer. We will refer to this output as \hat{l}_c . Notice that compression side of this module gives as a method of reducing the number of feature maps from

Next we will explain the core compression and decompression modules as illustrated in figure 3; we will also refer to them together as compression-decompression module. We begin as input l_n , the output of n^{th} layer, with $\tilde{\phi}$ entries/nodes/feature maps (we will refer to them as nodes). For the output of the compression module above, we obtain ϕ number of nodes where $\phi \leq \tilde{\phi}$, and the compression filter \hat{f} is a vector of size ϕ . At the beginning of training we initialize \hat{f} as a learning parameter of the neural network, with each value $\hat{f}_k \in \mathbb{R}$ where $k = 1 \dots \phi$. After passing through a sigmoid to obtain f from \hat{f} , we multiply each element of f to the corresponding feature map, i.e, for layer output \hat{l}_c with ϕ outputs, $\hat{l}_{c,k}$, the k th feature map, is now $l_{c,k} = f_k \cdot \hat{l}_{c,k}$. We will refer to

this operation as "mul" for convenience as referred in figure 3 as well, and the output is l_c . This method allows for further lossless compression methods, optionally. This compression module can be consider a method of compression similar to encoding but the level of compression is controlled by training the entries in the vector f to either be close to zero or not. The nodes corresponding to zero value need not be used further / communicated further, and we will explain later now f is trained to represent a desired compression level.

Following the compression, the data is communicated over to device 2, where the decompression has to take place, with the decompression module restoring the resolution of the feature maps to what it was during l_n , and gives an output l_d . A communication budget B is used to limit how many of the feature maps are communicated from device 1 to device 2, i.e., the feature map $l_{c,k}$ is communicated if and only if the corresponding f_k is larger than B entries in f .

l_d is then processed further with the remaining $n + 1, \dots, L$ layers to produce prediction p , to be compared against real output y during the training phase. So, during the training phase, we have two learning objectives: first training the traditional learning parameters (for example, weights and biases) of the model, and second learning the values of filter f such that a communication budget B is satisfied. In this work, we define a single loss function, that is used to train both sets of parameters, with the loss function $Loss(p, y)$ being an arbitrary loss function that is traditionally used to train a model, and a pruning loss function $pruneLoss(f, B)$, which is defined as:

$$pruneLoss = \exp\left(\delta \cdot \left[\left(\sum_{i \in 1, \dots, \phi} f_i\right) - B\right]\right) + \lambda \exp\left(-\delta \cdot \left[\left(\sum_{i \in 1, \dots, \phi} f_i\right) - B\right]\right) \quad (1)$$

In equation 1, the exponential functions act as a soft-constraint, or barrier functions, that let the budget be violated but at a great cost, and as the $pruneLoss$ is minimized, the number of nodes selected can come in agreement with the given budget constraint B . There are two exponential functions, with second on to the right (negative exponential) ensuring that the budget constraint is not over-corrected; this is not necessary but improves the results. Note that $0 \leq \lambda \geq 1$.

The total loss $totalLoss$ is defined as:

$$totalLoss = Loss + \epsilon \cdot pruneLoss \quad (2)$$

In equation 1 δ is a value defining how the violation of budget is handled as a soft constraint and in equation 2 ϵ is a value defining how the $pruneLoss$ is weighed against $Loss$ during the training.

The method described above, including the selected approach of adding pruneloss (shown in equation 1) to the loss, affords certain benefits to the process of training and implementing deep learning models in a networked system.

- First, notice that the total loss function allows for both $loss$ (associated with accuracy of the model) as well as the $pruneLoss$ (associated with the satisfaction of compression level) to be learned at the same time. We design and implement the method is such a way that the deep learning approach, i.e. implementation in the GPU, can be done for both kinds of learning, hence they can be trained in parallel.
- Second, notice that the parameter that controls the level of compression, f , is the only parameter in this deep learning architecture that needs to be controlled for different compression levels. So by training for one compression level, i.e., a particular value for $\sum_{i \in 1, \dots, \phi} f_i$, at the beginning, and then changing this value according to budget without resetting other parameters, we are able to quickly retrain for different compression levels through this form of transfer learning.

- Third, by dynamically changing compression levels during learning, we are able to acquire different deep learning architectures suitable for different communication environments.

As we discussed earlier, this is more of a skeleton method that the aforementioned the aforementioned techniques, "Deprune Method" and "prune Method", employ to achieve their goals.

3.2 Deprune Method

When the source of the data, i.e., device 1 (client device) in the formulation above and devices like IoT devices in the real world, are incapable of completing the learning process by themselves, or are avoiding learning in one location for the sake of saving on-device energy, split learning can be employed to share the task with the more powerful connected devices, i.e., device 2 (server) in the formulation above or devices like edge computers. In this instance, if the split learning were to proceed as is without any compression, depending on the model and the location where the splitting is taking place, a large amount of data would have to be transmitted across the network. For instance, with the convolution neural networks, a huge number of feature maps would have to be transported across the network for split learning. A solution to this problem is to implement compression, and with lossy compression, we are able to further increase the compression as discussed earlier. However, there is a loss of accuracy for the model when too much compression is implemented; the work around which lets us achieve both a lower training time and a higher accuracy level is the 'deprune' method based on our formulation.

Algorithm 1 dePrune training at client device

Input: $M, H, n, Budgets, Epochs, l_k, \gamma$
Output: Trained parameters θ_{client}

- 1: Initialize: Parameters θ_1 for M 's layers $\{1, \dots, n\}$
- 2: **for** $B \in Budgets$ **do**
- 3: **for** $i \in Epochs[B]$ **do**
- 4: $selectLr(l_k, \gamma)$
- 5: **for** $X_t, y_t \in batchLoader$ **do**
- 6: $l_n \leftarrow ForwardPass(X_t)$
- 7: $l_c, f \leftarrow CompressionModule(l_n)$
- 8: $Send(l_c, y_t, f)$
- 9:
- 10: $Receive(\Delta l_d)$
- 11: $Backpropagation(\Delta l_d)$
- 12: **end for**
- 13: **end for**
- 14: **end for**

Algorithm 2 dePrune training at server device

Input: $M, n, Budgets, Epochs, \delta, \epsilon, l_k, \gamma$
Output: Trained parameters θ_{server}

- 1: Initialize: Parameters θ_2 for M for M 's layers $\{n+1, \dots\}$
- 2: **for** $B \in Budgets$ **do**
- 3: **for** $i \in Epochs[B]$ **do**
- 4: $selectLr(l_k, \gamma)$
- 5: **while** running() **do**
- 6: $Await(l_c, y, f)$
- 7: $l_d \leftarrow decompressionModule(l_c, f)$
- 8: $p \leftarrow ForwardPass(l_d)$
- 9: $\Delta l_L \leftarrow totalLoss(p, y_t)$ given by equation 2
- 10: $\Delta l_d \leftarrow Backpropagation(l_L)$
- 11: $Send(\Delta l_d)$
- 12: **end while**
- 13: **end for**
- 14: **end for**

In the 'deprune' method, we begin by taking a filter with a much smaller communication budget, i.e., number of 1's in f , guided by B , is much smaller than size of f given by ϕ . For instance, after a certain layers of computation at layer n a VGG neural network has 128 feature maps, but we may want to begin the training by only allowing 4 feature maps to be communicated from device 1 to device 2. After training the neural network for a few epochs, we would then increase the number of feature maps from $b \ll \phi$ to maximum value of ϕ , i.e., in case of the example VGG,

by increasing the allowed number of features maps to be communicated, $b = 4$, to $b = 128$. This method is elaborated in algorithms 1 and 2.

Algorithm 1 takes place in the client device, or as we referred earlier, device 1. For a model M with training hyper-parameters H , our goal will be to train the parameters corresponding to the layers $\{1, \dots, n\}$, with the help of two lists: *Budgets*, where each entry gives the budget B (as defined earlier) under which the compression is supposed to operate at a given instance of training, and *Epochs* where each entry corresponds to the number of epochs to be trained for a given budget B . The end goal here is to obtain the trained parameters in client device given by θ_{client} . Please note that this value consists of training parameters for first n layers as well as the parameters for compression module, i.e., f . The data are obtained for certain batch sizes as defined by H through a batch-loading mechanism (any sort of) in the form of (X_t, y_t) , and then we perform a forward pass on input X_t through the first n layers to obtain the output l_n .

Next step is to perform compression on l_n using the compression module as discussed during the formulation section and illustrated in figure 3, which as we discussed gives as output the parameters l_c (output of compression) and f (the filter). Then (l_c, y, f) is sent to the server device by the client device.

Once (l_c, y, f) is received by the server device, which is awaiting a request from client as shown in 2, l_c is decompressed using the decompression module as explained in the formulation section 3.1; output l_d then goes through forward pass for the remaining layers to obtain the prediction p . Using equation 2, we compute the error for the given forward pass, and then the obtained Δl_L is back-propagated through $L, \dots, n + 1$ layers as well as the decompression module. Then the output of backpropagation on server device may be sent to the client as Δl_d . Client device, which is until now waiting to receive Δl_d will now use this value from the server device to finish the training step.

An important function that is yet to be explained for the deprune methods as shown in algorithms 1 and 2 is the dynamic selection of learning rate (*selectLr*(l_k)). During *selectLr*(l_k) for each non-initial budget level, i.e., when we transfer to higher budget level, the learning rate is set to a higher value by a factor of γ for the first l_k epochs. This approach was observed to significantly help avoid unwanted local minima during training and help achieve higher accuracy as we moved to higher budget levels.

3.3 Prune Method

When the source of the data, i.e., device 1 (client device) from the formulation section and devices like XR devices in the real world, are not capable of completing the inference process in a timely manner or doing so in an energy-permissible way, split learning can be employed to share the task with the more powerful devices in the network, i.e., device 2 (server) in the formulation section or devices like edge computers. In such an instance, if the split learning were to proceed without any compression, depending on the model and other requirements like latency constraints for inference, it might not be feasible to complete the task with just the end device running the inference model. Hence, we could again rely on compression, including lossy compression to achieve network performance (such as lower latency and network usage) that would otherwise be impossible when relying only on a weak end device and/or split learning without compression.

As we implement more compression, we expect the model accuracy to fall, generally speaking, as less information can be transmitted from the source device/ client device to the server for further inference. In any case, for different levels of compression, we will end up achieving different levels of model efficiency [6]. However, depending on latency/ network usage requirements, the client may want to opt for a different level of this accuracy-compression trade-off. While the lossy compression model is shown to be efficient, it does take quite a bit of time to train for a given

compression level as training needs to start again from the beginning for that compression level. Therefore, if we want to train for different compression levels, it will take a lot of time to finish training each process. With the help of our method, prune, based on the deep learning architecture as defined in the formulation section, we are able to implement transfer learning approach and generate multiple approximate models with similar accuracy (or, slightly reduce accuracy), but significantly reduced network costs and computation costs at the client devices.

In the pruning method, we begin by taking a filter with the largest possible filter size, i.e., where the communication budget is maximum, i.e., all the feature maps may be communicated so $B = \text{len}(f)/\phi$. For this particular model, we allow the training to fully proceed, and then let this be the "base model" upon which transfer learning may take place. Let *Budgets* be a vector containing each desired feature map compression levels with k th compression level given by B_k . Then the model corresponding to k th compression level is given by M_k . Hence, after the full model with no compression ($B = \phi$) is trained to obtain parameters for M_0 , we do the same for different k compression levels as defined in *Budgets* using the transfer learning approach as defined in algorithm 3 shown in Appendix ??, which is designed to give us a set of trained parameters Θ where each element θ_k is parameters for model M_b corresponding to compression level B_k .

Only for budget $B = \max(\text{Budgets})$ where the feature map compression is not implemented, the set of parameters θ_b is randomly initialized as if training a new deep learning architecture; for every other compression level, we transfer the parameters learned from previously trained model. Then for a certain epochs given by $\text{Epochs}[B]$, we train the deep learning architecture. Later while discussing results we will show that $\text{Epochs}[B]$ only needs to be larger when $B = \max(\text{Budgets})$ but can be much smaller for the remaining compression levels. This way, the transfer learning approach can help us train many subsequent approximate models quickly. Notice that the models are still split into two parts, with one side computing the first n layers, i.e. 1, ..., n along with a compression module, and the other side computing the remaining $L - n$ layers after going through decompression module.

This still is in keeping with our split learning formulation, but the major difference here when compared against the deprune method is that the training phase need not be run in different devices, but may be run in one powerful device that trains the different deep learning architectures with different network cost and performance trade-offs. Once when the training is completed, the developed models represented by output of the prune method Θ may be used by the client and server devices. In that case, the inference method is equivalent to the forward pass in the client device (as shown in algorithm 1) followed by the forward pass in the server device (as shown in algorithm 2). The client would in that case only expect the prediction P from the server device; needless to say the learning parts including loss predictions are not necessary for these inferences.

One important step after completing the training for each budget level is *resetprune* as shown in algorithm 3, where the filter parameters are reset. Notice that f is always obtained by passing through sigmoid, and after training at certain budget level b_{k1} , we then move to budget level $b_{k2} < b_{k1}$. So here by making sure that the entries in f that were already greater than 1 during training for b_{k1} is randomized during the training for b_{k2} , we make sure that those filters can potentially be trained to approach 0, or else further compression would not be likely. Another important remark would be this prune algorithm may be run for different resolution compression in case of CNN training. As was the case for deprune method, we also employ the dynamic selection of learning rate ($\text{selectLr}(l_k)$). During $\text{selectLr}(l_k)$ for each non-initial budget level, i.e., when we transfer to lower budget level, the learning rate is set to a higher value by a factor of γ for the first l_k epochs.

Algorithm 3 summarizes our prune method as an algorithm as discussed in section 3.3.

Algorithm 3 prune training**Input:** $M, H, n, B, Epochs, l_k, \gamma$ **Output:** Trained parameters set Θ

```

1: for  $B \in Budgets$  do
2:   if  $b == \max(Budgets)$  then
3:     Initialize: Parameters  $\theta_b$  for  $M_0$ 
4:     Initialize  $\Theta \leftarrow \{\}$ 
5:   else
6:     Transfer:  $\theta_b \leftarrow \theta$ 
7:   end if
8:   for  $i \in Epochs[B]$  do
9:     for  $X_t, y_t \in batchLoader$  do
10:       $l_n \leftarrow ForwardPass[l_1, \dots, l_n](X_t)$ 
11:       $l_c, f \leftarrow CompressionModule(l_n)$ 
12:       $l_d \leftarrow decompressionModule(l_c, f)$ 
13:       $P \leftarrow ForwardPass[l_{n+1}, \dots, l_L](l_d)$ 
14:       $\Delta l_L \leftarrow totalLoss(P, y_t)$  given by equation 2
15:       $\Delta l_d \leftarrow Backpropagation(l_L)$ 
16:       $Backpropagation(\Delta l_d)$ 
17:    end for
18:  end for
19:   $\theta \leftarrow \theta_b$ 
20:   $\Theta[M_b] \leftarrow \theta_b$ 
21:  resetprune
22: end for

```

4 EVALUATION

In this section, we will show the results demonstrating the efficiency of our method at creating the deep learning architectures in network system paradigms, with a focus on overcoming network costs, i.e., reduced network usage and decreased communication latency. But we will also focus on overall processing time. In subsection 4.1, we will evaluate the depruning method, alongside its implementation on testbeds. And in subsection 4.2, we will present and describe the benefits of the pruning method.

4.1 Results: dePrune Method

A primary goal of our deep learning architecture, and of the deprune method, is to reduce bandwidth usage and improve latency during the training process; it is then by extension that the benefits are afforded to the end device in the form of manageable reduction in computation cost/energy consumption through split learning. We begin by showing that our method is able to significantly reduce bandwidth consumption on a realistic testbed, while also being capable of improving learning time/latency. Then we run simulations over python implementations to demonstrate the reduction in the network usage as we employ the deprune method on different models, datasets, split locations and resolution compressions.¹

In order to show that our method allows training to continue with a reduction in bandwidth usage over time, we show that less bytes have to be communicated over the training period (to reach the

¹Implementation codes will be provided after the decision, as per the conference recommendation.

same level of accuracy in the same number of epochs) with our method added to a distributed/split learning paradigm. In order to show that the end-to-end latency of learning episodes can be reduced on average, we show that more samples are processed using our method over time, when a device has to offload certain fraction of the deep learning task to a server. Notice that offloading raw data (i.e., image) can be considered a special case where offloading takes place before layer 1 as opposed to certain arbitrary layer $n + 1$.

The testbed consists of an embedded GPU serving offload requests, and a laptop hosting data sources which offload processing. The embedded GPU uses CUDA for most of the tensor processing, and CPU for the rest of the computations, including network message serialization and de-serialization. The laptop uses CPU for all of the computations, including tensor processing. The laptop uses a commodity CPU with 8 cores and a base clock frequency of 1.60 GHz. The processing speed of the GPU according to the manufacturer’s specification is 22 TOPS (INT8 precision). The laptop and the embedded GPU communicate over a physical Wi-Fi 6 channel.

The software framework used in benchmarks is open-sourced². In the benchmarks, the laptop runs multiple parallel data source processes, each of which share the same wireless channel. The processes were containerized, and their deployment was automated by using Kubernetes, which allows for much better orchestration and scaling among other benefits. To emulate the constrained processing power of mobile devices, and to avoid physical hardware saturation from other applications running in the laptop affecting the measurements, each of the data sources’ CPU allocation was limited virtually. Furthermore, the number of data source processes running in the laptop was set to eight, which the laptop had enough RAM to run without having to use swap memory.

In the benchmarks, we measure the task throughput in terms of samples processed, as well as the data traffic in terms of bytes sent over the wireless channel. The tasks throughput is measured per data source, while the data traffic in the wireless channel is the sum of all of the data sources’ traffic. The benchmark measurement starts after processing the first batch in order to avoid the cold start resulting from e.g. CUDA JIT compilation.

We run benchmarks with eight data sources, limiting the CPU usage of each data source to 40% cores. The resource usage of the embedded GPU was not limited. Each data source processed samples in batches of 16 samples, and processed 32 batches in 4 epochs. Split training without pruning does not include the additional neural network layers, and the compression/de-compression modules, resulting in less overall processing.

We run the experiment for CIFAR10 dataset [22] running VGG11 model [46] with our compression-decompression module (referred as ‘with module’) and compare against implementation that does the split learning without any compression-decompression module [19] (referred as ‘without module’); in both cases splitting takes place at 5th layer with the compression module having $f = 4$ in case of our method. These layer selections are arbitrary and later we show that our method works across different layers. As seen in figure 4, the network usage is reduced significantly with our method when compared against split learning (labelled ‘without module’), i.e., the improvement is about 25 times less network usage; and the training throughput as shown in figure 5 is reduced by about 2.5 times. The goal here was to demonstrate that the deployment of our method in a wireless environment can efficiently reduce the network usage, and increase task throughput, i.e., provide improved total latency for the completion of the learning task.

In the next case, we run the experiments with a single data source running CIFAR10 dataset and learning on VGG11 model, with a batch size of 64, and other similar settings as compared with testbed results in section 4.1. The main difference here is the hardware being used, with both

²to be added after publication

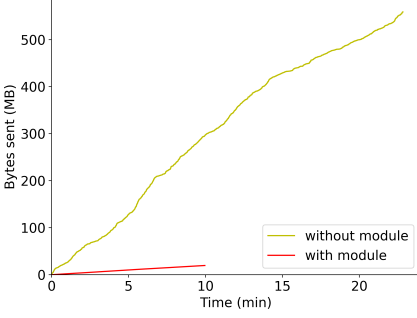


Fig. 4. Data traffic in the Wi-Fi channel during the benchmark execution.

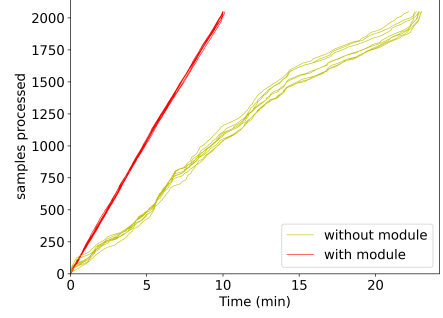


Fig. 5. Throughput for each data source during the benchmark execution. The plots are interpolated linearly to make them more readable.

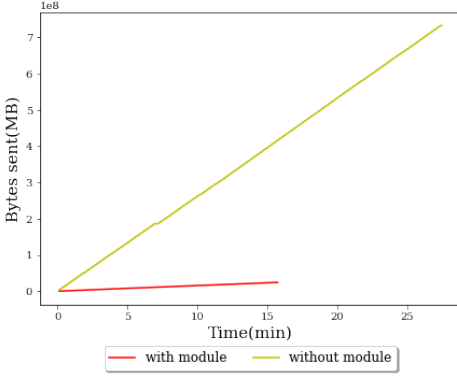


Fig. 6. Data traffic in the Ethernet channel during the benchmark execution.

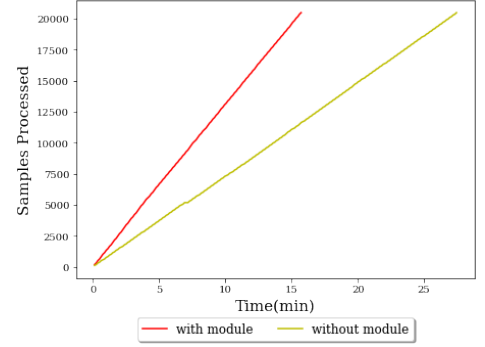


Fig. 7. Throughput for each data source during the benchmark execution.

the client and the server devices being CUDA-enabled nodes and connected to the network over Ethernet. Specifically, the client and the server nodes are set up with an NVIDIA GeForce RTX 3090 and an NVIDIA GeForce RTX 4090 respectively, and the Round Trip Time between these two nodes over Ethernet connection is $0.670 \pm 0.219ms$. As shown in figure 6, our method ('with module') is able to reduce the total network usage over split learning without our module ('without module') by 30x, while as shown in figure 7, the total throughput is reduced by 1.7 times. Of course, this throughput performance was a case where computation was the bottleneck and not the network, since Ethernet was used; when network is a bottleneck, it is shown that the method is more useful (i.e., when WiFi or cellular network is involved, or when network is congested, the method becomes more useful). Of course, this throughput performance was a case where computation was the bottleneck and not the network, since Ethernet was used; when network is a bottleneck, it is shown that the method is more useful (i.e., when WiFi or cellular network is involved, or when network is congested, the method becomes more useful).

In figures 4 or 6, the cumulative bytes of network traffic used up can be interpret in two ways: the instantaneous slope can be used to see that the bandwidth consumption was much lower for our method most of the times, signifying that the bandwidth consumption (bytes sent per unit time) was lower. The red line (our method, 'with module') ends much faster, signifying that for same

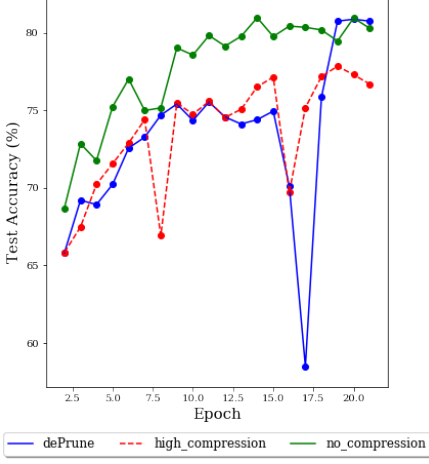


Fig. 8. Progression of training under deprune method, cifar10 dataset

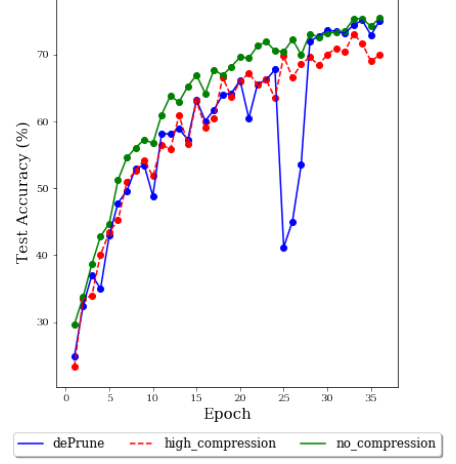


Fig. 9. Progression of training under deprune method, imagenet dataset

level of accuracy and same number of epochs run, the training session was finished faster. Similarly, the slope of the lines in figures 5 and 7, show that our method (red lines, 'with module'), have a higher slope, denoting faster completion of each training sample, denoting a better end-to-end processing latency for the training steps. These experiments have shown that the method could work in different network and hardware settings. Furthermore, the use of multiple workers in case of WiFi network also demonstrates the usefulness of this method for a scaled deployment (with multiple users participating in the learning environment).

Next, we show how the method runs with less bandwidth consumption, while reaching the same level of accuracy as non-compressed offloading of the tasks to servers.

We ran simulations to show the network usage improvement under different circumstances. In figure 8 we train a VGG11 [46] NN with CIFAR10 dataset [22] using 3 different approaches. 'no-Compression' method refers to a process in which the filter size f as defined during the formulation is maximum, which means all the data/feature maps are sent from client device to server device. When we consider the layer to be 0_{th} /before first layer, it represents simply offloading the raw data to the server (i.e., cloud). On the other hand, 'high-compression' refers to the case where the compression is implemented so that very few number of feature maps are allowed to be communicated. Finally, 'deprune' is our method as described earlier in subsection 3.2. deprune is a method where for a much greater part of the training epochs, we will only train under the lower compression level / lower value of f , where much fewer filters are communicated across the network. In case of figure 8, we will train with the filter size $f = 4$ out of 128 for the first 15 epochs, and then we may train with full feature size $f = 128$ for the remaining time. Here we see that the level of accuracy towards the end is similar to training with full budget, given by test accuracy in percentage in figure 8, reached by epoch 18. High compression option here is with $f = 4$ and no compression option always has $f = 128$, i.e., full budget. Here, we realized that while the accuracy level achieved by 'deprune' is similar to that of the accuracy achieved by 'no-compression' approach by 18 epochs, the improvement in network utilization would be by 420 percent. The accuracy is improved by 4 percent as compared to high compression case by then. In case of this illustration figure 8, the hyperparameters we used were: Stochastic Gradient Descent (SGD) as optimizer, learning rate l of $1e-5$, weight decay γ of $5e-4$, and split layer $n = 5$.

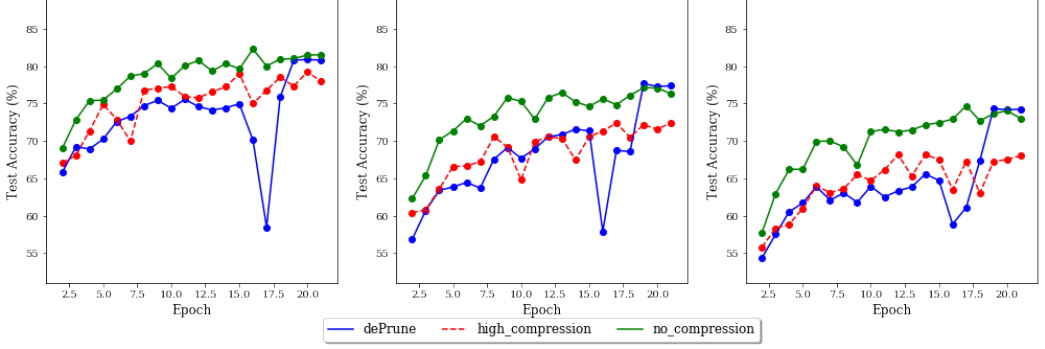


Fig. 10. deprune with different resolution compression factors. Left: no resolution compression, Center: by 2x2, Right: by 3x3

Next, we tested the experiment for a different dataset, Imagenet100, which is a subset of ImageNet-1k dataset[43] that contains 100 out of the 1000 classes in the original dataset, using the same approach as earlier in case of CIFAR10 dataset. The result is illustrated in 9, where comparison is once again made with no-compression and high-compression cases. Here again, training is done for a lower value of $f = 4$ for 25 epochs, before switching to $f = 128$ for the remaining of the training. We see that by epoch 30, the accuracy of 'deprune' method is already similar to the accuracy when no compression happens. Without a loss of obtained accuracy, within these 30 epochs, the network usage is improved by 418 percent. While by then the accuracy has improved by 4 percent as compared against high-compression case. In case of this illustration figure 9, the hyperparameters we used were: Stochastic Gradient Descent (SGD) as optimizer, learning rate l of $1e-5$, weight decay γ of $5e-4$, and split layer $n = 5$.

We also repeated this same experiment with different resolution compression levels for CIFAR10 dataset, as shown in figure 10, which shows that these compression mechanisms work for different levels of resolution compression. As our algorithm uses a convolution method during compression, by increasing the convolution kernel for compression from 1×1 to 2×2 (center plot of 10) and by increasing to 3×3 (left plot of figure 10), we observe that our method of compression can work complementary to resolution compression methods. As expected in each case, as the compression factor increases from 1×1 to 2×2 and 3×3 , the accuracy falls for each case, but in each case our method efficiently matches the maximum test accuracy (no-compression case) for given resolution compression level.

As shown in figure 14, appendix A, the addition of the compression-decompression module did not have a significant impact on the performance of the deep learning model, and the addition of this module at any location within the model did not change the behavior in any significant way either; the test accuracy was similar after any number of epochs for each of the cases. To make sure that the addition of compression-decompression module after any arbitrary layer results in similar efficient behavior, we ran the experiments again for cifar10 dataset with different locations of split layer (in addition to original experiment where splitting occurs at layer 5), as shown in figure 11. In cases where the compression-decompression module were kept at the 0th layer (i.e. compressing the input itself) as shown in figure 11 (left), after 6th layer as shown in 11 (center), and after 9th layer as shown in 11(right), the efficient behavior from our deprune method as explained earlier was always observed. This effectively proves that the module could be added anywhere, as well as that the 'deprune' method can be run in any of these cases.

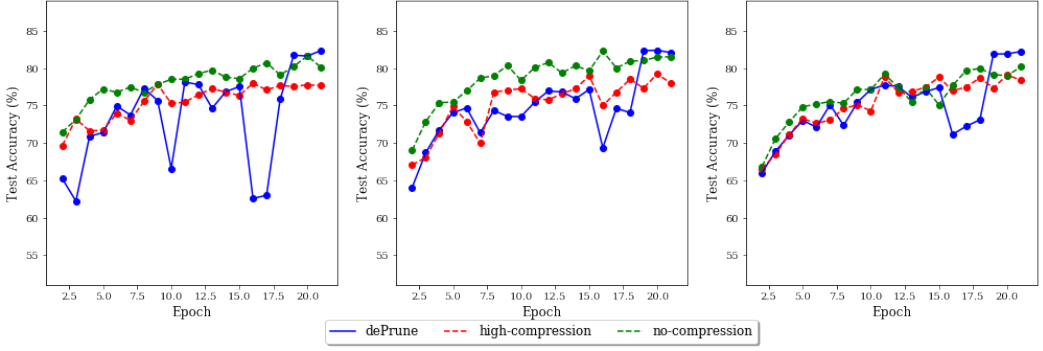


Fig. 11. Deprune with compression-decompression module after different layers. Left: before layer 1/Image, Center: after layer 6, Right: alter layer 9

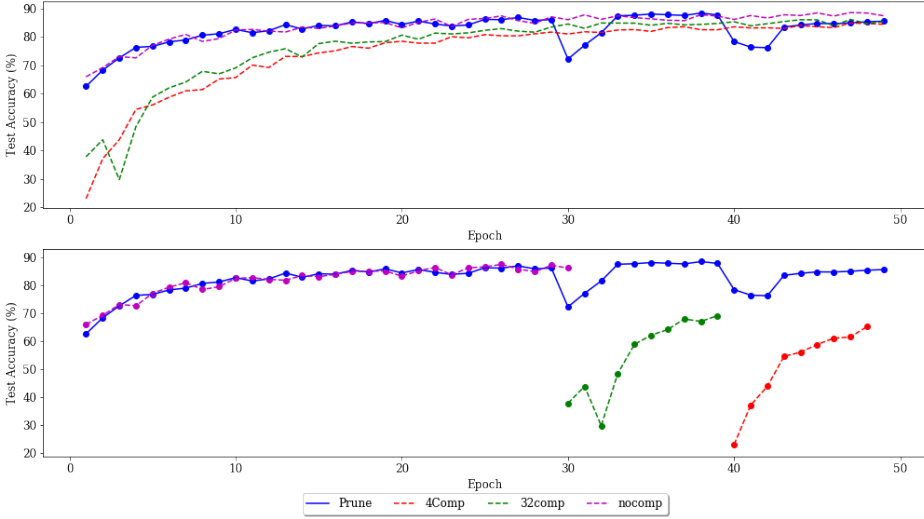


Fig. 12. Progression of training under prune method, CIFAR dataset

This shows that the method can be deployed in different ways and for different data, so that network resources are saved while training deep learning models in a split-learning environment.

4.2 Results: Prune Method

As discussed earlier in the related works and methods sections, compression aware training, where we train deep learning models to split the tasks in a network and latency-efficient way, requires a lot of training time. This is because we need to train the model for every given compression level. Our solution based on the compression-decompression module, 'prune', is able to generate models with different compression-latency trade-off by training the largest model first and then using a transfer learning approach to quickly generate approximate models, as described in subsection 3.3. In this subsection, we will demonstrate the results of our method in reducing the training times for these approximate models in an efficient way.

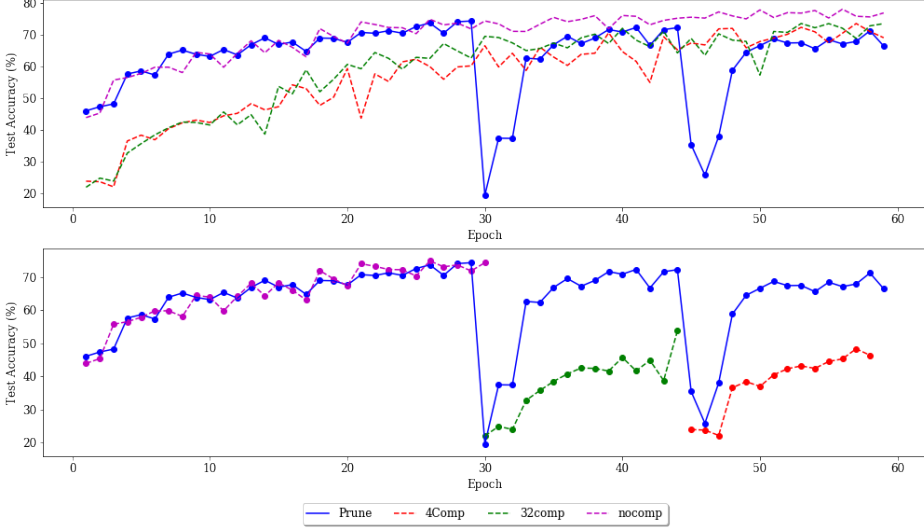


Fig. 13. Progression of training under prune method, STL10 dataset

In figure 12, we train a VGG11 [46] neural network with CIFAR10 dataset [22] using our method, 'prune'. In this case, we begin by training the model first with no compression, i.e., filter size $f = 128$ so all 128 feature maps in the layer are to be sent across network for the first 30 epochs. Then we reduce the number of filters to $f = 32$, and then to $f = 4$, training for each f for a few epochs. In figure 12(top), we compare our method Prune with situations where compression levels are no compression (labelled 'nocomp'), $f = 32$ (labelled '32comp', representing a compression by 4 times), and $f = 4$ (labelled '4comp', representing a compression by 32 times). The top image, figure 12(top), is showing that our method could generate results that are comparable to the results generated by 'training from start', with compression by 4 times called 32 comp, compression by 32 times labelled 4comp, and no compression-case labelled nocomp. In these cases, which we call 'training from scratch' compression level is predefined and the model is trained without transferring any learned information from any other model, for each compression level. For our approach prune, training happens in epochs 0-30 for $f = 128$, epochs 30-40 for $f = 32$, and epochs 40-50 for $f = 4$, and in all of those cases, accuracy to the 'training from scratch' approach is matched.

In figure 12(bottom), the bottom figure, we are able to see that after training for $f = 128$ and transferring to $f = 32$, the training time is significantly reduced, i.e. the accuracy after training for just 5 epochs has reached the maximum, while for 'training from scratch', this value is only achieved (inferred from figure 12) after about 30 epochs; after 5 epochs accuracy is only 70 percent of maximum for the 'training from scratch' approach. Hence, this is an improvement of 6 times better training speed. Similarly, for the compression level of $f = 4$, the training time was an improvement of about 6 times as well. In case of this illustration figure 12, the hyperparameters we used were: Stochastic Gradient Descent (SGD) as optimizer, learning rate l of $1e-5$, weight decay γ of $5e-4$, and split layer $n = 5$.

We repeated this experiment with STL10[5], as illustrated in figure 13. The same method as described in case of CIFAR10 above was implemented again. This time for our approach prune, training happens in epochs 0-30 for $f = 128$, epochs 30-45 for $f = 32$, and epochs 45-60 for $f = 4$ as evident in figure 13 bottom or top. As seen in 13(top), our method was able to match

performance of 'training from scratch' with 4 times compression (labelled '32Comp' in figure 13) by epoch 7 as opposed to taking about 30 epochs to reach the maximum performance, which is a improvement of 4 times. As seen in 13(bottom), STL10 was even much slower than the CIFAR10 case of 'training from scratch'; here by epoch 7, in either cases of compression, 4 times compression with $f = 32$ or 32 times compression with $f = 4$, within 7 epochs, the training in case of 'training from scratch' had only reached about 60 percent of our method. In case of this illustration figure 13, the hyperparameters we used were: Stochastic Gradient Descent (SGD) as optimizer, learning rate l of $1e-5$, weight decay γ of $5e-4$, and split layer $n = 5$.

In Appendix B, we discuss further configurations for trying out the prune methods. So this section shows that we are able to obtain network-efficient deep learning architectures quickly. The next straightforward step in deploying our models, after training, would be to have them in edge-cloud split environment for quick inference, while giving the network management entity the choice of selecting between accuracy and latency; this could help make decisions to optimize service quality, fulfill latency constraint, or help with any other utility maximization.

5 CONCLUSION

In this work, we developed a novel approach to training deep learning models that improve the network usage and the training latency. We designed a compression-decompression module, that when combined with our deprune algorithm, is able to significantly reduce network usage and make learning faster without significantly effecting accuracy; network usage for learning was improved by 4x when compared with split learning approach, and the accuracy was improved by 4 percent when compared with compression-aware split learning. Such algorithms will be very beneficial to IoT and other resource-constrained end devices in the next generation network architectures that must provision AI tasks. We also combined the compression-decompression module with our prune method to significantly reduce the training time for deep learning models that trade little accuracy for significant reduction in network consumption; here we showed that the training time could be improved by up to 6 times when compared with compression-aware split learning. Such methods could help develop orchestration options for AI task like AR/XR in the network constrained environments.

REFERENCES

- [1] 2019. <https://www.gsma.com/futurenetworks/wiki/cloud-ar-vr-whitepaper>
- [2] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818* (2019).
- [3] Juliano S Assine, Eduardo Valle, et al. 2021. Single-training collaborative object detectors adaptive to bandwidth and computation. *arXiv preprint arXiv:2105.00591* (2021).
- [4] Miguel A Carreira-Perpinán and Yerlan Idelbayev. 2018. "learning-compression" algorithms for neural net pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8532–8541.
- [5] Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 215–223.
- [6] Amir Erfan Eshratifar, Amirhossein Esmaili, and Massoud Pedram. 2019. Bottlenet: A deep learning architecture for intelligent mobile cloud computing services. In *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 1–6.
- [7] Wei Fang, Lin Wang, and Peiming Ren. 2019. Tinier-YOLO: A real-time object detection method for constrained environments. *IEEE Access* 8 (2019), 1935–1944.
- [8] Vajiheh Farhadi, Fidan Mehmeti, Ting He, Thomas F La Porta, Hana Khamfroush, Shiqiang Wang, Kevin S Chan, and Konstantinos Poularakis. 2021. Service placement and request scheduling for data-intensive applications in edge clouds. *IEEE/ACM Transactions on Networking* 29, 2 (2021), 779–792.
- [9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2015. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence* 38, 1 (2015),

- 142–158.
- [10] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).
 - [11] Seungyeop Han, Haichen Shen, Matthai Philipose, Sharad Agarwal, Alec Wolman, and Arvind Krishnamurthy. 2016. Mcdnn: An approximation-based execution framework for deep stream processing under resource constraints. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. 123–136.
 - [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
 - [13] Yihui He, Xiangyu Zhang, and Jian Sun. 2017. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*. 1389–1397.
 - [14] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.
 - [15] Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. 2016. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250* (2016).
 - [16] Zhaohui Huang and Vasilis Friderikos. 2021. Proactive edge cloud optimization for mobile augmented reality applications. In *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 1–6.
 - [17] Hyuk-Jin Jeong, InChang Jeong, Hyeon-Jae Lee, and Soo-Mook Moon. 2018. Computation offloading for machine learning web apps in the edge server environment. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1492–1499.
 - [18] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351* (2019).
 - [19] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News* 45, 1 (2017), 615–629.
 - [20] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. 2015. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530* (2015).
 - [21] Jong Hwan Ko, Taesik Na, Mohammad Faisal Amir, and Saibal Mukhopadhyay. 2018. Edge-host partitioning of deep neural networks with feature space encoding for resource-constrained internet-of-things platforms. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.
 - [22] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
 - [23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.
 - [24] Qiang Liu, Siqi Huang, Johnson Opadere, and Tao Han. 2018. An edge network orchestrator for mobile augmented reality. In *IEEE INFOCOM 2018-IEEE conference on computer communications*. IEEE, 756–764.
 - [25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14. Springer, 21–37.
 - [26] Qu Yuan Luo, Shihong Hu, Changle Li, Guanghui Li, and Weisong Shi. 2021. Resource scheduling in edge computing: A survey. *IEEE Communications Surveys & Tutorials* 23, 4 (2021), 2131–2165.
 - [27] Hadi Tabatabaee Malazi, Saqib Rasool Chaudhry, Aqeel Kazmi, Andrei Palade, Christian Cabrera, Gary White, and Siobhán Clarke. 2022. Dynamic service placement in multi-access edge computing: A systematic literature review. *IEEE Access* (2022).
 - [28] Yoshitomo Matsubara, Davide Callegaro, Sameer Singh, Marco Levorato, and Francesco Restuccia. 2022. Bottleneck: Learning compressed representations in deep neural networks for effective and efficient split computing. In *2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 337–346.
 - [29] Yoshitomo Matsubara and Marco Levorato. 2020. Split computing for complex object detectors: Challenges and preliminary results. *arXiv preprint arXiv:2007.13312* (2020).
 - [30] Yoshitomo Matsubara and Marco Levorato. 2021. Neural compression and filtering for edge-assisted real-time object detection in challenged networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2272–2279.
 - [31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).
 - [32] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. 2019. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11264–11272.
 - [33] Venkatesh N Murthy, Vivek Singh, Terrence Chen, R Manmatha, and Dorin Comaniciu. 2016. Deep decision network for multi-class image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2240–2248.

- [34] Anders Nordgren. 2022. Artificial intelligence and climate change: ethical issues. *Journal of Information, Communication and Ethics in Society* (2022).
- [35] Stephen Pasteris, Shiqiang Wang, Mark Herbster, and Ting He. 2019. Service placement with provable guarantees in heterogeneous edge computing systems. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 514–522.
- [36] Hanyu Peng, Jiaxiang Wu, Shifeng Chen, and Junzhou Huang. 2019. Collaborative channel pruning for deep networks. In *International Conference on Machine Learning*. PMLR, 5113–5122.
- [37] Konstantinos Poularakis, Jaime Llorca, Antonia M Tulino, and Leandros Tassioulas. 2020. Approximation algorithms for data-intensive service chain embedding. In *Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*. 131–140.
- [38] Konstantinos Poularakis, Jaime Llorca, Antonia M Tulino, Ian Taylor, and Leandros Tassioulas. 2020. Service placement and request routing in MEC networks with storage, computation, and communication constraints. *IEEE/ACM Transactions on Networking* 28, 3 (2020), 1047–1060.
- [39] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [40] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*. Springer, 525–542.
- [41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
- [42] Precedence Research. 2023. *Artificial Intelligence (AI) Market Size, Growth, Report 2022-2030*. <https://www.precedenceresearch.com/artificial-intelligence-market>
- [43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.
- [44] Jiawei Shao and Jun Zhang. 2020. Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 1–6.
- [45] Wenqi Shi, Yunzhong Hou, Sheng Zhou, Zhisheng Niu, Yang Zhang, and Lu Geng. 2019. Improving device-edge cooperative inference of deep learning via 2-step pruning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 1–6.
- [46] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [47] Abhishek Singh, Praneeth Vepakomma, Otkrist Gupta, and Ramesh Raskar. 2019. Detailed comparison of communication efficiency of split learning and federated learning. *arXiv preprint arXiv:1909.09145* (2019).
- [48] Balázs Sonkoly, János Czentye, Márk Szalay, Balázs Németh, and László Toka. 2021. Survey on placement methods in the edge and beyond. *IEEE Communications Surveys & Tutorials* 23, 4 (2021), 2590–2629.
- [49] Chandra Thapa, Pathum Chamikara Mahawaga Arachchige, Seyit Camtepe, and Lichao Sun. 2022. Splitfed: When federated learning meets split learning. In *Proceedings of the AAI Conference on Artificial Intelligence*, Vol. 36. 8485–8493.
- [50] Manas Wadhwa, Gagan Raj Gupta, Ashutosh Sahu, Rahul Saini, and Vidhi Mittal. 2023. PFSL: Personalized & Fair Split Learning with Data & Label Privacy for thin clients. *arXiv preprint arXiv:2303.10624* (2023).
- [51] Alexander Womg, Mohammad Javad Shafiee, Francis Li, and Brendan Chwyl. 2018. Tiny SSD: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection. In *2018 15th Conference on Computer and Robot Vision (CRV)*. IEEE, 95–101.
- [52] Jian Xue, Jinyu Li, Dong Yu, Mike Seltzer, and Yifan Gong. 2014. Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6359–6363.
- [53] Xin Yang, Jiankai Sun, Yuanshun Yao, Junyuan Xie, and Chong Wang. 2022. Differentially private label protection in split learning. *arXiv preprint arXiv:2203.02073* (2022).
- [54] Chenglong Zhao, Bingbing Ni, Jian Zhang, Qiwei Zhao, Wenjun Zhang, and Qi Tian. 2019. Variational convolutional neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2780–2789.
- [55] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. *Advances in neural information processing systems* 32 (2019).

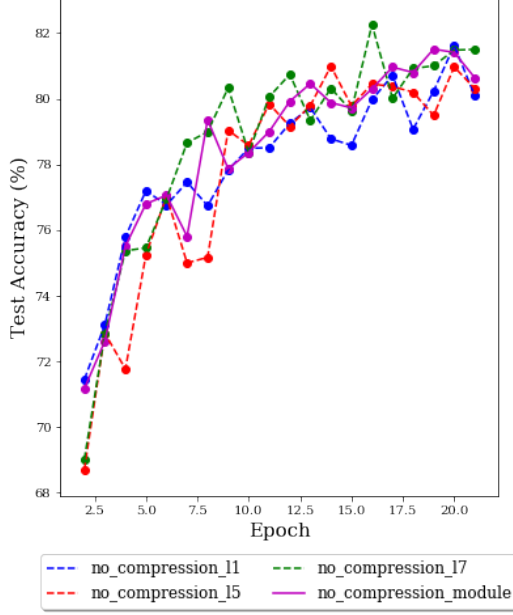


Fig. 14. Progression of training for models with compression-decompression modules after different layers along with a model without the module

APPENDIX

A EFFECT OF ADDING COMPRESSION-DECOMPRESSION MODULE TO A NEURAL NETWORK

In figure 14, we have shown that putting the compression-decompression module after different layers produces very similar learning behavior to not putting them anywhere, i.e., the original model. The results are demonstrated after putting the module after no layers (i.e, image), as 5th layer and as 7th layer.

B EXTENDED EVALUATION FOR PRUNE METHOD

As we are able to observe in figure 15, we see a similar improvement at speeding up training with prune method, as was the case in figure 12 in subsection 4.2, with the difference here being that the compression method is also complemented with a resolution compression factor of 3x3 using convolution method discussed in 3.1. Here again, the maximum value for compression by 4 times is reached within 4 epochs for our method, but takes 34 epochs for 'training from scratch' (32 comp). The results are similar for compression factor of 32. We were able to observe these over-performances consistently. We did note that while for a lot of cases the obtained accuracy was similar albeit with different learning speeds, as is a known case for deep learning training methods, there is uncertainty and variability: sometimes our method gave accuracy levels different than 'training from scratch'. And sometimes the learning could be slower towards reaching the desired accuracy and sometimes faster; but such variances are expected in learning tasks like deep learning for vision. As shown in figure 16, in case of compression by 64 times for CIFAR10 dataset, and resolution compression by 3x3 times, both method, our method as well as training from scratch gave output where accuracy varied noticeably across the epochs. As a future potential

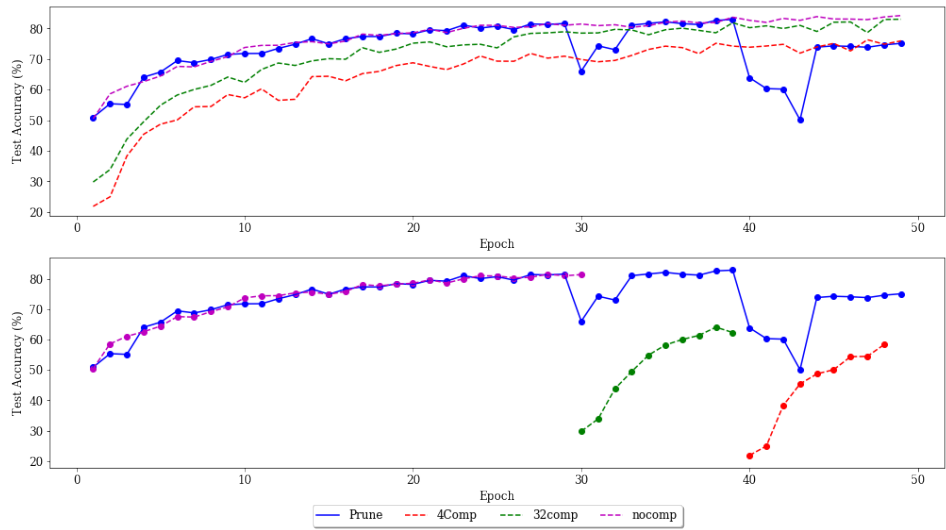


Fig. 15. Progression of training under prune method, for CIFAR10 dataset with resolution compression factor of 3x3

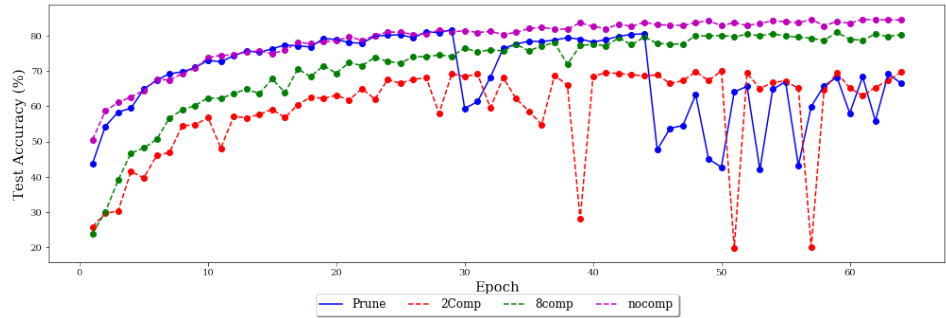


Fig. 16. Progression of training with prune method, for CIFAR10 dataset, when compression can get very high

direction of research, we could find a way to automatically explore right and minimum compression possibilities.