

Efficient Privacy Preserving Edge Computing Framework for Image Classification

Omobayode Fagbohungbe, Sheikh Rufsan Reza, Xishuang Dong, Lijun Qian
 Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT)
 Prairie View A&M University, Texas A&M University System
 Prairie View, TX 77446, USA
 Email: ofagbohungbe, sreza@student.pvamu.edu, xidong@pvamu.edu, liqian@pvamu.edu

Abstract—In order to extract knowledge from the large data collected by edge devices, traditional cloud based approach that requires data upload may not be feasible due to communication bandwidth limitation as well as privacy and security concerns of end users. To address these challenges, a novel privacy preserving edge computing framework is proposed in this paper for image classification. Specifically, autoencoder will be trained unsupervised at each edge device individually, then the obtained latent vectors will be transmitted to the edge server for the training of a classifier. This framework would reduce the communications overhead and protect the data of the end users. Comparing to federated learning, the training of the classifier in the proposed framework does not subject to the constraints of the edge devices, and the autoencoder can be trained independently at each edge device without any server involvement. Furthermore, the privacy of the end users' data is protected by transmitting latent vectors without additional cost of encryption. Experimental results provide insights on the image classification performance vs. various design parameters such as the data compression ratio of the autoencoder and the model complexity.

Index Terms—Deep Learning, Edge Computing, Autoencoder, Convolutional Neural Network, Internet of Things

I. INTRODUCTION AND MOTIVATION

Emerging Technologies such as the Internet of Things (IoT) and 5G networks will add a huge number of devices and new services, as a result, a huge amount of data will be generated in real time. One of the important data types is image data, since many applications involve images and videos such as in video surveillance. In order to take advantage of the “big image data”, data analytics must be performed to extract knowledge from the data. One way to handle the data would be uploading all data from edge devices to the cloud or remote data centers for processing and knowledge extraction [1]. However, as highlighted in Figure 1, there are several factors that may render this practice infeasible: 1) The sheer volume of the images may overwhelm the uplink with limited bandwidth; 2) The uplink may not be always available especially when wireless communications is used due to weather (e.g., for mmWave), distance, or jamming; 3) Proprietary images may need encryption that introduce additional delay; 4) The end users may have concerns about the security and privacy of their images, thus they may not agree to upload raw images that may contain private information. Furthermore, uploading is subject to eavesdropping, interceptions, or other unauthorized access.

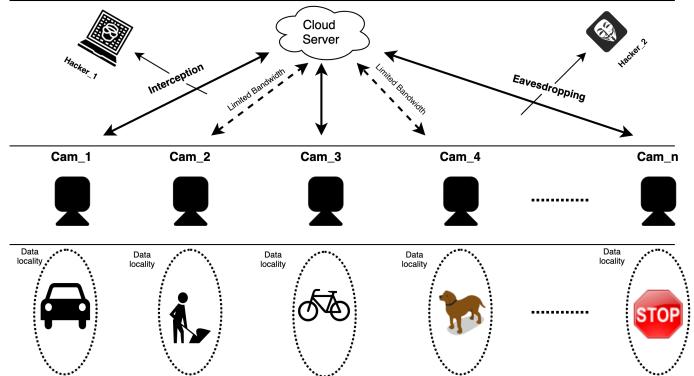


Figure 1: Challenges incurred when uploading all data from edge devices to the cloud.

In order to address these challenges, a novel efficient privacy preserving framework for image classification in edge computing systems is proposed in this paper. Specifically, the large raw data will be processed locally (at the edge) by pre-trained autoencoder; then instead of uploading the raw image, only compressed latent vector that contain critical features learned from the raw image will be uploaded through the access point or hub to the cloud for further processing. The proposed framework is highlighted in Figure 2. It is demonstrated in our experiments that the learning performance of extracting knowledge at the cloud has very little degradation when the compression ratio is not large (e.g., below 16 in our test cases). Furthermore, the raw images can be reconstructed with very small error at the cloud using the pre-trained decoder if needed.

Comparing to traditional source coding (e.g., zip), using autoencoder has the following advantages: 1) Instead of only reducing the redundancy in the raw data as in source coding or traditional data compression, autoencoder will extract critical features in the raw data and encode the features in a compact form (the latent vector), in other words, the encoder performs initial learning at the edge devices; 2) In addition to compressing the data, autoencoder also “encrypt” the data by transforming the raw data into latent vectors, which enhance the security of data. For example, a zipped file can easily be unzipped by an adversary if not encrypted; on the contrary, an

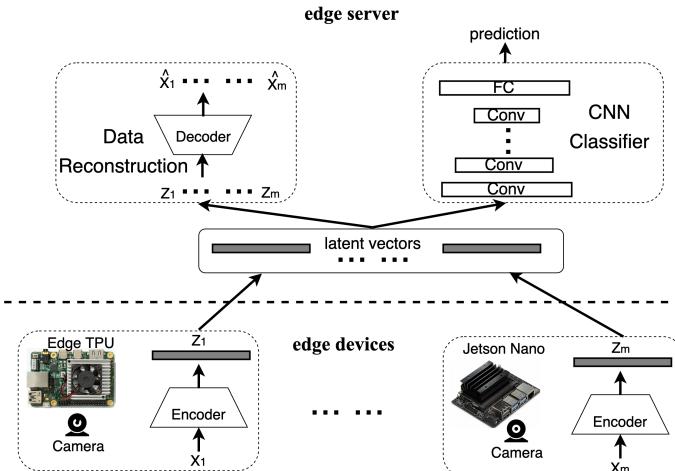


Figure 2: The proposed efficient privacy preserving framework for image classification in edge computing systems. Here x_i is the raw image, z_i is the compressed latent vector, and \hat{x}_i is the reconstructed image.

adversary could not reconstruct the raw data from the latent vector without knowing the structure (e.g., number of layers, number of nodes in each layer) and all the weights of the pre-trained autoencoder (the decoder part to be exact), in fact, it is shown in [2] that autoencoder provides a similar level of security to normal encryption - assuming that the decoder is not shared; (3) Even if the edge device is captured by an adversary, it is very difficult for the adversary to deduce the decoder part from the encoder part on the edge device.

The proposed framework has some similar characteristics such as taking advantage of large diverse data from many edge devices and data locality at each device as in federated learning [3]–[6]. However, comparing to federated learning, the proposed framework has the following advantages:

- 1) In federated learning, the server and the end users (edge devices) train the same model. As a result, the complexity of the model is constrained by the computing capability and storage of the edge device. On the contrary, *in the proposed framework, the training of the classifier is done at the cloud server only, thus it can be very deep and complex if needed, and it is not subject to the constraints of the edge devices.*
- 2) In federated learning, the edge device must rely on the server to update the gradients and train the model. *In the proposed framework, the training of the autoencoder can be done independently at each edge device without any server involvement.*
- 3) In federated learning, the privacy of the end users' data is protected by applying differential privacy schemes [7] or through secure aggregation [8], thus introduce additional cost due to encryption or secret sharing. *In the proposed framework, the privacy of the end users' data is protected by transmitting latent vectors without additional cost of encryption.*

The proposed framework is explained in Section II. Section III gives the design and implementation details. Experimental results and analysis are given in Section IV. Further discussions and related works are reviewed in Section V. Section VI concludes the paper.

II. PROPOSED FRAMEWORK

The proposed efficient privacy preserving framework for image classification in edge computing systems is shown in Figure 2. It has two levels: the edge devices and the edge server. It is assumed that the nodes of the edge devices contain sensors such as cameras and embedded computing devices such as Google edge TPU [9] or NVIDIA Jetson Nano [10]. The edge server is assumed to have strong computational capacity and large storage.

The edge segment of the framework mainly contains the various edge devices of interest and the pre-trained encoder. The server mainly contains the hub, the pre-trained classifier and the pre-trained decoder. We only consider supervised learning in this paper and it is assumed that the training dataset is labeled.

The data from each of the edge devices are passed to the corresponding encoder attached to it. Unique pre-trained encoder is used for each of the edge devices in order to take advantage of the data locality at each device. The function of the pre-trained encoder, which is in the inference mode, is to extract the most important and critical features in the data. The encoder also ensures dimension reduction of the input data by a pre-determined amount. The extracted critical features (latent vectors, or feature maps when the data are images) are then transmitted to the hub at the server. The two major functions at the server are the classification task and the data reconstruction task (recover a copy the original image from the latent vectors). In other words, the latent vectors are input to the pre-trained classifier for prediction and they are also input to the corresponding decoder at the server for the reconstruction of the images.

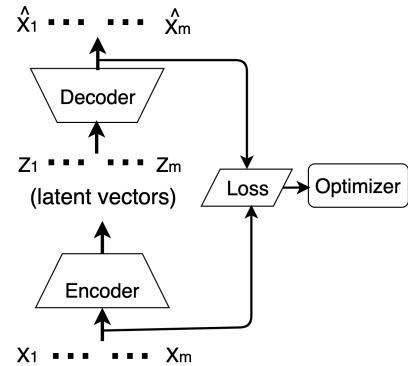


Figure 3: The training for the proposed autoencoder at edge device.

Table I: Information on the CIFAR10 and ImageNet (IMGNETA and IMGNETB) Datasets

Dataset	Image size	# of images	Training Testing ratio	# of classes	Comments
CIFAR10	32*32*3	60 000	5:1	10	
IMGNET-A	256*256*3	13,000	7:3	10	very different images
IMGNET-B	256*256*3	13,000	7:3	10	very similar images

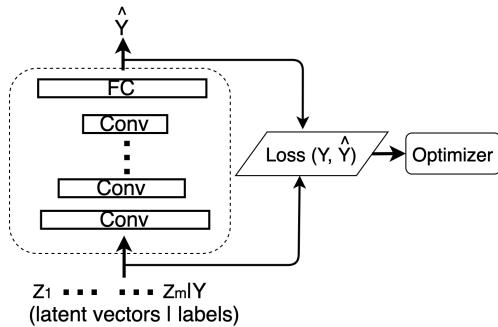


Figure 4: The training for the proposed CNN classifier at the server.

A. Training Stage

The dataset collected at each edge device is used to train an autoencoder for the corresponding device. This is done to take advantage of the data locality at each devices. Autoencoders are generative models where an artificial neural network is trained to reconstruct its own input in an unsupervised way. Figure 3 illustrates all the components of an autoencoder and the training process. It is made up by two main blocks which are the encoder and the decoder [11], [12]. The encoder compresses the input X into a low dimensional representation of pre-determined size, called the latent vector denoted by Z that contains the most important features in the data. When the input data are images, Z will be the corresponding feature maps. The decoder then tries to reconstruct the original data from the latent vector Z . The reconstructed data obtained at the decoder output is denoted by \hat{X} . It should be noted that an autoencoder is a lossy network as the original image will not be fully recovered. However, it is expected that the critical features will remain in the recovered image.

The Autoencoder achieves the proper training of the encoder and decoder by minimizing the differences between original input and the reconstructed input. This is achieved by the use of the mean square error loss function or any other loss function. After the training of the autoencoder, the encoder part of the autoencoder is then extracted, deployed in the inference mode on the edge device, and then used to generate the latent vector Z . Hence, the dataset is transformed from $[X, Y]$ to $[Z, Y]$ where Y are the labels.

The latent vectors and the corresponding labels are then aggregated at the hub and they are used for training a classifier on the cloud in a supervised manner as shown in Figure 4. The type of classifier on the cloud is determined by the nature or type of supervised task to be done. The most common type of classifier used for image dataset is the convolutional neural

network (CNN) and it uses the cross entropy loss function.

B. Inference Stage

In this stage, the pre-trained encoder, decoder and classifier are deployed in the inference mode. The data X from the edge device is fed to the corresponding pre-trained encoder attached to it. The encoder then transforms the data X to a latent vector Z which represents the most critical feature in X . The latent vector Z , which is smaller than X by a pre-determined ratio, is then transmitted to the cloud. At the cloud server, the latent vector Z is then fed into the pre-trained classifier and predict a label \hat{Y} . In situations where the original data is needed on the cloud, say we want to see a copy of the original image, the latent vector Z is also fed into the input of the corresponding decoder and the estimated data is obtained.

III. EXPERIMENTS

A. Dataset Description

The result in this work was generated using three different datasets summarized in Table I. These datasets are from the Canadian Institute For Advanced Research dataset (CIFAR10) [13] and the ILSVRC (ImageNet) 2012 datasets [14].

1) **Canadian Institute For Advanced Research (CIFAR10):** This dataset containing 60,000 color images, is a subset of about 80 million labeled but tiny images. The dataset is further divided into 50,000 training samples and 10,000 testing samples. It has about 10 classes which are mutually exclusive and there's no semantic overlaps between images from different classes.

2) **ILSVRC (ImageNet) 2012:** The original ILSVRC 2012 dataset contain about 1.2 million color images of different sizes across about 1,000 classes. The 1,000 classes are either internal or leaf nodes but do not overlap. Two subsets of the ILSVRC 2012 dataset termed IMGNET-A and IMGNET-B were used in this work. Each subset contains about 13,000 images each resized to a dimension 256×256 , spanning 10 classes. The images in each subset is further divided into training samples and testing samples with ratio 7:3. The difference between the two subsets lies in the type of nodes they contain. The IMGNET-A subset contains images from 10 different leaf nodes (diverse images) while IMGNET-B contains 10 child nodes from a single leaf node (similar images).

B. Deep Learning Model Design and Training Strategy

The autoencoder for the edge devices and the classifier at the edge server were chosen such that the autoencoder is optimized for feature extraction and the classifier is optimized for image classification.

Table II: The deep learning models and the dataset used in training the models

		CIFAR10	IMGNET-A	IMGNET-B
Vanilla Model	Model-A	x	-	-
	Model-B	-	x	x
Transfer Model	Model-C	-	x	x

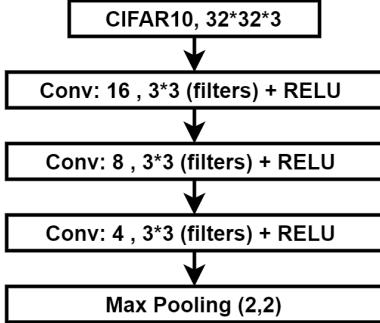


Figure 5: Details of an encoder model for compression size of 4 using CIFAR10 dataset

Table III: The architecture of the vanilla model for CIFAR10 dataset (Model-A)

Vanilla Model For Imagenet Dataset
Conv2D, Filter Size = 3*3, No of Filters = 32, Stride = 1*1, No Padding Activation Layer (Relu)
Max Pooling, Pool Size = 2*2, Stride = 1, No Padding
Conv2D, Filter Size = 3*3, No of Filters = 32, No Padding Activation Layer (Relu)
Max Pooling, Pool Size = 2*2, Stride = 1*1, No Padding
Conv2D, Filter Size = 3*3, No of Filters = 32, Stride = 1*1, No Padding Activation Layer (Relu)
Max Pooling, Pool Size = 2*2, Stride = 1*1, No Padding
Flatten
Dense(64)
Activation(Relu)
Dropout(0.5)
Dense(10)
Activation(Softmax)

1) **Autoencoder Design and Training Strategy:** The autoencoder architecture is affected by the type of images and the compression ratio. For instance, the model architecture for CIFAR10 dataset for compression ratios 4 and 8 are different. This also applies for compression ratio 4 for IMGNET-A and CIFAR10 dataset. Hence, different models were developed across several edge devices, compression ratio and datasets.

Figure 5 shows the model architecture for an autoencoder designed for CIFAR10 dataset for compression ratio of 4. The model contains a mix of convolutional (same padding), max pooling, and upsampling layers. The relu function was used as the activation function for all layers except the last layer where the sigmoid function was used.

The models were trained from scratch using glorot-uniform method as initializer, mean square error as the cost function and rmsprop optimization algorithm as the optimizer. After the convergence of the autoencoder model during training process, encoder part of the autoencoder was then used in the inference mode to compress all the images to obtain the latent variables needed to train the classifier.

C. Training Stage

1) **Classifier Design:** The convolutional neural network (CNN) models were used in this work. CNNs are well suited for image processing applications and other grid-like data [12]. They are more computationally efficient than the dense deep neural network thus reducing the memory usage. Using the filters, CNNs find and extract meaningful features from the images and preserve spatial relations. Three different CNN classifiers, denoted Model-A, Model-B and Model-C as listed in Table II, were used in this work.

a) **Model-A and Model-B:** Model-A and Model-B are considered to be vanilla models because they were trained from scratch. Model-A and Model-B were specifically designed for the original input image and feature maps of CIFAR10 dataset and ImageNet dataset, respectively. The detailed CNN architecture of Model-A and Model-B are shown in Tables III and IV, respectively. The models contain a mix of convolutional, max pooling, and fully connected layers and relu and softmax activation functions. Furthermore, the models

Table IV: The architecture of the vanilla model for Imagenet dataset (Model-B)

Vanilla Model For CIFAR10 Dataset
Conv2D, Filter Size=3*3, No of Filters=32, Stride=1*1,Padding Activation Layer (Relu)
Conv2D, Filter Size=3*3, No of Filters=32, Stride=1*1, Padding Activation Layer (Relu)
Max Pooling, Pool Size = 2*2,Stride = 1*1, Padding Dropout(0.25)
Conv2D, Filter Size = 3*3, No of Filters = 64,Stride = 1*1, Padding Activation Layer (Relu)
Conv2D, Filter Size = 3*3, No of Filters = 64,Stride = 1*1, Padding Activation Layer (Relu)
Max Pooling,Pool Size = 2*2,Stride = 1*1, Padding Dropout(0.25)
Flatten
Dense(512)
Activation(Relu)
Dropout(0.5)
Dense(10)
Activation(Softmax)

also contain some dropout layer in order to prevent over-fitting. The differences between Model-A and Model-B lies in the number of the different layers used and the use of padding in the convolutional layer of Model-A.

$$\text{Base-Layer} + \text{Top-Layer} = \text{Base-Layer} \quad \text{Top-Layer}$$

Figure 6: The Transfer Learning Model Block (Model-C)

The models were trained from scratch with the aim of minimizing the difference between the labels (ground truth)

Table V: The architecture of the transfer learning model for Imagenet datasets (Model-C)

	Model_1	Model_2	Model_3	Model_4	Model_5
Base layer	VGG16	VGG19	InceptionV3	InceptionResnetV2	Resnet50
Top layer	Dense(256)	Dense(256)	Dense(256)	Dense(256)	Dense(50)
	Activation(Relu)	Activation(Relu)	Activation(Relu)	Activation(Relu)	Activation(Relu)
	Dense(10)	Dense(10)	Dense(10)	Dense(10)	Dense(10)
	Activation(Softmax)	Activation(Softmax)	Activation(Softmax)	Activation(Softmax)	Activation(Softmax)

and the predicted labels. This was achieved by the use of glorot-uniform method as initializer, categorical cross entropy as the loss function and adams optimization algorithm as the optimizer. Data augmentation was also used during the training process to mitigate overfitting due to the small quantity of the datasets. It should be stated that each classifier were trained with their respective original image and the feature maps (compressed images).

b) *Model-C*: Model-C is a transfer learning based model designed specifically for the ImageNet dataset in this work. The CIFAR10 version was not presented in this work because of poor performance when used with the feature maps (compressed image) of CIFAR10 which can be attributed to its small dimension.

The block diagram of the model is shown in Figure 6. Model-C can be divided into two parts: The base layer and the top layer. The base layer is a pre-trained layer of other standard deep learning model (without the fully connected layer) that has been trained with data similar to the Imagenet data and achieved better performance. By using this pre-trained models, the excellent feature extracting property of the standard model is being leveraged to achieve better performance. Furthermore, it also complements data augmentation in training decent model in situations where datasets are limited. VGG16, VGG19, InceptionV3, InceptionResnetV2 and Resnet50 pre-trained models [15] were used as base models for Model-C.

The details of the top layer used for this work is shown in Table V. It should be noted that the first dense layer of the top layer in model 5 is smaller than that of the other models. This is because model 5 was overfitting if the number of neurons in the first layer was 256, the same number used in the other models. Hence, the size of the dense layer was reduced to reduce overfitting and achieve good performance.

A 2-stage training method was used for the transfer learning model to minimize the error between the ground truth labels and the predicted labels. This approach is different from the training approach used for Model-A and Model-B which were trained from scratch. In the first stage, the base layer was fixed while the fully connected top layer was trained using the Adam optimizer after being initialized using the glorot-uniform method. This was done in order to initialize the weight of the top layer close to the weight of the base layer. Thereafter, the whole model was retrained using the stochastic gradient descent (SGD) with momentum optimizer in order to tune the whole weight of the model appropriately. SGD with momentum was used because it is less aggressive than the adam optimizer as the use of an aggressive optimizer in the second step might cause the information in the base layer to

be significantly eroded or lost. The binary categorical entropy was used as the cost function in the entire training process.

D. Experiments

This work seeks to propose a new approach to design and implement deep learning models for distributed systems without compromising on data privacy and security. It achieves this by extracting the most important/critical machine features intelligible yet human unintelligible features from the dataset. These features are then transmitted across the communication network from the edge devices to the edge server where they are aggregated and used to train a classifier. The experimental methods, performance metrics and tools used in validating our proposed framework is explained in this section.

1) **Experimental method**: A 2-stage methodology was used to validate our proposed framework and this method is the same irrespective of the type of dataset or model used. In the first stage, the original training set of the original input dataset (uncompressed images) was used to train the classifier. Thereafter, the test set was then used to obtain the needed performance metric in order to set the baseline.

In second the stage, the training set of the feature maps (compressed images of the dataset used in stage 1), is used to train the same classifier model. The feature map, which is smaller than the original image by a pre-determined factor, is obtained by passing the original dataset through the encoder of the autoencoder. Thereafter, the performance metric of the classifier is obtained using the test set of the feature maps and the performance compared to the baseline.

2) **Performance Metric**: The effectiveness of the framework is assessed using a simple classification task. The test accuracy of the model obtained after the training process is used as the primary performance metric. Furthermore, the effect of our proposed method on the training and testing time, and number of model parameters were also investigated.

3) **Software and Hardware**: The design, training and testing of the deep learning models (Autoencoders and CNN Classifiers) were implemented using Keras deep learning framework on TensorFlow backend, running on a NVIDIA Tesla P100-PCIE-16GB GPU.

IV. RESULTS AND ANALYSIS

The results of the experimental work are presented in this section. The performance of the proposed framework is compared with our baseline using the performance metrics stated in Section III-D2 above. The baseline performance is represented by compression ratio 1 and it is synonymous to using the uncompressed image to test our various models.

Furthermore, it should be noted that the vanilla model for the CIFAR10 and ImageNet datasets are different as stated in section III-D

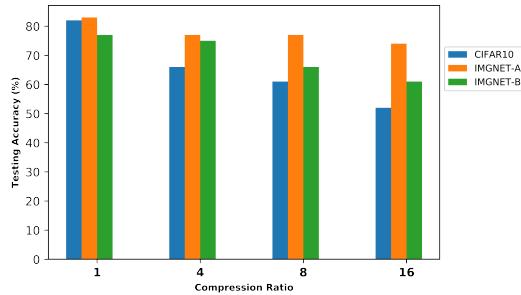


Figure 7: Comparison of the testing accuracy of the vanilla models for the original dataset (compression ratio =1) and compressed dataset (latent variables) with compression ratio = 4, 8, 16.

A. Effect on Test Accuracy

Figure 7 shows the testing accuracy of vanilla CNN Classifiers (Model-A and Model-B) when trained and tested with compressed and uncompressed CIFAR10 and Imagenet datasets. The testing accuracy for the compression ratio 1 (uncompressed images), representing our baseline, is highest across all the cases, as expected. This is because all features in the raw images can be used for classification. Furthermore, the testing accuracy for IMGNET-A is higher than that of IMGNET-B. The differences in performance can be attributed to the very close similarity in the images in IMGNET-B, as classifying such images is a much more difficult classification task as compared to classifying images in IMGNET-A.

A general degradation in the testing accuracy is observed in Figure 7 as the compression ratio is increased, although the rate of reduction varies across the models used for the 3 datasets. The rate of degradation of the testing accuracy of the model tested for CIFAR10 dataset is the highest for all the compression ratios. This is because of the small dimension of the CIFAR10 images (32*32), implies that the amount of features needed to perform a classification task is even smaller when compressed.

Furthermore, the rate of degradation of the testing accuracy for IMGNET-A dataset was very modest across all the compression ratios. However, similar performance was not observed in IMGNET-B, particularly for compression ratios 8 and 16 despite having the same image dimension (256*256). The bigger degree of degradation observed in the case of IMGNET-B for compression ratios 8 and 16 is due to the complexity of the classification task. This is because of the similarities in the images that make up the various classes in IMGNET-B, unlike IMGNET-A where the the images that make up the classes are very different.Hence, the complexity of its classification task means it requires a lot more features than that of IMGNET-A. Furthermore, bigger compression

ratio also means the model has less amount of features to make a classification decision.

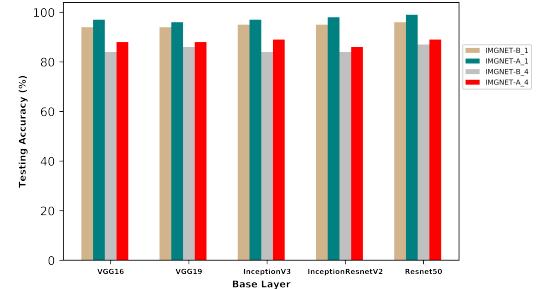


Figure 8: Testing accuracy of the transfer learning based model (Model-C) using different base models for the ImageNet dataset with compression ratio = 4.

The testing accuracy of transfer learning based model (Model-C) for the ImageNet dataset compressed by a factor of 4, using different base models, is shown in Figure 8. The transfer learning model was not designed and trained using the CIFAR10 datasets as its performance was poor with the compressed images. The poor performance can be attributed to the very deep nature of the transfer learning based model leaving inadequate number of features available at the beginning of the fully connected layer (top layer) where classification takes place. The same reason also explains why the transfer learning model was only designed and tested with ImageNet dataset with compression ratio 1 and 4 only.

The testing accuracy of the transfer learning model across different base models for IMGNET-A and IMGNET-B datasets at compression ratio 1 (baseline) and 4 is higher than the corresponding performance of the vanilla model (Model-A and Model-B). This performance can be attributed to the powerful feature extraction property of the different base layer used. However, the rate of degradation in the testing accuracy for compression ratio 4 is higher than what was observed for the vanilla models. This can also be attributed to the very deep nature of the transfer learning models as a small amount of information/features left is not distinct enough to make accurate classification.

B. Effect on Number of Parameters

The number of parameters in a convolutional neural network is determined by many factors such as the filter size, the number of filters, the size of the input data, the number and type of hidden layers, etc. Hence, reduction in the number of parameters can be achieved by reducing the size of the input data. The relationship between the normalized number of parameters in the vanilla model for the CIFAR10 and ImageNet datasets vs. the compression ratio is shown in Figure 9. It can be observed that for the same compression ratio, the rate of reduction in the normalized number of parameters of the vanilla model for the ImageNet dataset is bigger than that for the CIFAR10 dataset. The bigger rate of reduction can be attributed to the size of the image, and in turn, the number

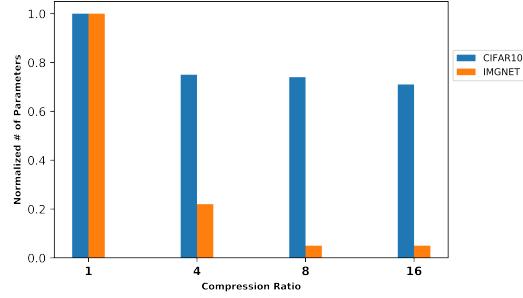
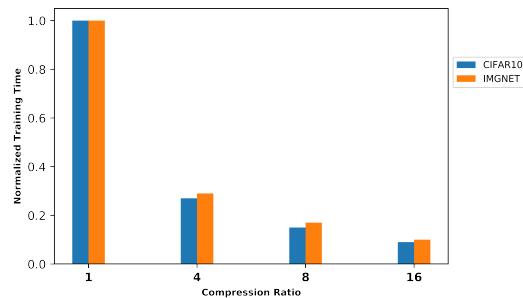


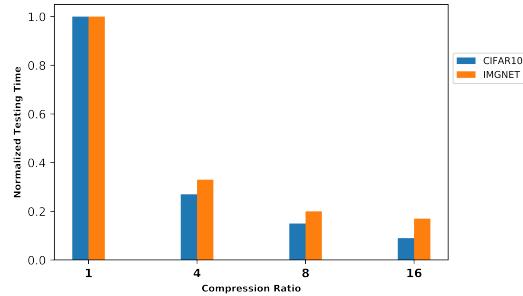
Figure 9: Comparison of the normalized number of vanilla model parameters vs. data compression ratio

of features, which impacts the number of parameters in the fully connected layers of the model. It was also observed that the rate of reduction in the normalized number of parameters for each model appears to be flat after a certain compression ratio which varies from model to model. The flatness is due to the reduction in the significance of the fully connected layer to impact on the number of parameters, as the number of features reduces below a certain point.

C. Effect on Testing and Training Time



(a) Comparison of the normalized testing time



(b) Comparison of the normalized training time

Figure 10: Comparison of the normalized testing time and training time of the vanilla models for various compression ratios for CIFAR10 and ImageNet datasets.

Figure 10 show the normalized amount of time required for testing and training, respectively, of the vanilla models for various compression ratios for CIFAR10 and ImageNet datasets. A reduction in the amount of training and testing time was observed across the compression ratios and the models.

The reduction can be attributed to the decreasing size of the input data and smaller number of parameters to learn, when compression ratio increases.

V. DISCUSSION AND RELATED WORKS

There are several methods proposed in the literature to address the privacy and security concerns associated with data used for training deep learning models. Examples of popular approaches include homomorphic encryption [16], differential privacy [17], [18] and secure multiparty computation [19]. Despite the successes of these methods, some issues remain such as performance degradation, non-trivial overhead or limited application [20]–[22].

The use of collaborative deep learning method, such as federated learning, has been introduced in recent years to solve the problem of data privacy. Federated learning is a type of machine learning where the goal is to train a high quality centralized model while the data remains distributed over a large number of clients [3]. It involves the sharing of model parameters and model gradients through a parameter server without sharing their local data. Federated learning is based on an iterative model averaging and it is robust to unbalanced data and non-i.i.d. data distribution. Federated learning has been applied to mobile keyboard prediction, vocabulary word learning and google keyboard query suggestions improvement [23]–[25]. Federated learning may be viewed as an extension of the idea discussed in [26], [27] that stochastic gradient descent can be implemented in parallel and asynchronously.

Federated learning may suffer from non-trivial communication cost. To deal with the high communications cost, an efficient multi-objective evolutionary algorithm, based on a scalable network connectivity encoding method, was proposed in [5]. To help reduce the uplink communication bottleneck, the use of structured and sketched updates were introduced in [4].

Federated learning may also suffer from security/privacy issues due to the need to communicate the model parameters to the central server. One recent study showed potential security/privacy issues due to the possibility of reconstructing original data from the shared gradient [28]. Secure aggregation, a type of secure multi-party computation algorithm for federated learning was introduced in [29]. This helps guarantee the privacy of data used in generating gradients shared by each model and improve communication efficiency. Furthermore, it was observed that federated learning performs poorly when the data distributed across the training center is strictly non-i.i.d. of a single class. This statistical challenge was resolved by creating and using a small subset of data which is globally shared between all the edge devices [30] or adopting a multi-task learning approach [31].

Autoencoder has been applied to address data privacy concerns in several recent works [2], [32], [33]. In [32], a convolutional autoencoder that perturbs an input face image to impart privacy to a subject is proposed. It is shown the method can protect gender privacy of face images. A proof-of-concept study has been performed in [2] to use an autoencoder

for preserving video privacy, especially when non-healthcare professionals are involved. A modified sparse denoising autoencoder has been applied in [33] to reduce the sparsity and denoise the data. Then a 3-class classification is performed on the reconstructed data from the autoencoder and it is shown that the classifier can classify the original black class data as the transformed gray class data.

Although autoencoder has been used to address data privacy concerns, this work is the first in the use of autoencoder for addressing privacy concerns, communication cost, and deep learning efficiency associated with mobile edge computing systems with large number of edge devices. This is achieved by using the autoencoder to extract human unintelligible but machine intelligible features from the data. The features or latent vectors are then used to train the classifier. Furthermore, the proposed approach comes with the added advantage of reducing the dimensionality of data needed to be transmitted, thus reducing the communication cost and the number of model parameters, as well as training and inference time. This approach does not suffer from leaking gradient problem associated with federated learning [28].

VI. CONCLUSIONS

A novel edge computing framework for designing and implementation of privacy preserving image classification models is proposed in this work. The proposed framework provides 1) flexibility of training autoencoder at each edge device individually, thus protect data privacy of end users because raw data is not transmitted at any time; 2) after the training of autoencoder is complete, raw data is “compressed” and “encrypted” by the encoder before transmitting to the edge server, and this will reduce the communications cost and further protect the data privacy and security; 3) the autoencoder will provide features to the classifier at the server, thus allow the classifier to be trained on the features with less and controlled dimensions; 4) the decoupling of the training of the autoencoder at the edge devices and the training of the classifier at the edge server relaxes the frequent communications requirement between edge devices and edge server. Experiments have been carried out using CIFAR10 and ImageNet datasets, and detailed analysis of tradeoff between classifier accuracy, dimensionality of data, compression ratio and different choice of classifiers has been given to provide benchmark and insights on the proposed scheme.

For future work, comparison with federated learning in terms of classifier performance vs the communications cost and model complexity will be carried out for image classification tasks. This will help quantify the pros and cons of the proposed approach. Furthermore, the use of other types of autoencoder to extract latent variables and use of knowledge distillation to help mitigate the reduction in the model accuracy will be explored.

VII. ACKNOWLEDGMENTS

This research work is supported in part by the U.S. Dept. of Navy under agreement number N00014-17-1-3062 and the

U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under agreement number FA8750-15-2-0119. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Dept. of Navy or the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) or the U.S. Government.

REFERENCES

- [1] R. Ramesh, “Predictive analytics for banking user data using aws machine learning cloud service,” in *2017 2nd International Conference on Computing and Communications Technologies (ICCCT)*, Feb 2017, pp. 210–215.
- [2] M. D’Souza, M. Johnson, J. Dorn, C. Van Munster, M. Diederich, C. Kamm, S. Steinheimer, K. Kravalis, J. Boisvert, I. Ormesher, L. Walsh, A. Sellen, F. Dahlke, B. Uitdehaag, and L. Kappos, “Autoencoder - a new method for keeping data privacy when analyzing videos of patients with motor dysfunction (p4.001),” *Neurology*, vol. 90, no. 15 Supplement, 2018. [Online]. Available: https://n.neurology.org/content/90/15_Supplement/P4.001
- [3] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, “Federated learning of deep networks using model averaging,” *CoRR*, vol. abs/1602.05629, 2016. [Online]. Available: <http://arxiv.org/abs/1602.05629>
- [4] J. Konecny, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. [Online]. Available: <https://arxiv.org/abs/1610.05492>
- [5] H. Zhu and Y. Jin, “Multi-objective evolutionary federated learning,” *CoRR*, vol. abs/1812.07478, 2018. [Online]. Available: <http://arxiv.org/abs/1812.07478>
- [6] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, Jan. 2019. [Online]. Available: <https://doi.org/10.1145/3298981>
- [7] R. C. Geyer, T. Klein, and M. Nabi, “Differentially private federated learning: A client level perspective,” *CoRR*, vol. abs/1712.07557, 2017. [Online]. Available: <http://arxiv.org/abs/1712.07557>
- [8] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for federated learning on user-held data,” in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.
- [9] “Google Edge Tpu,” <https://coral.ai/docs/edgetpu/faq/>.
- [10] “Jetson Nano Developer Kit,” <https://developer.nvidia.com/embedded/jetson-nano-developer-kit>.
- [11] M. Maggipinto, C. Masiero, A. Beghi, and G. A. Susto, “A convolutional autoencoder approach for feature extraction in virtual metrology,” *Procedia Manufacturing*, vol. 17, pp. 126 – 133, 2018, 28th International Conference on Flexible Automation and Intelligent Manufacturing (FAIM2018), June 11-14, 2018, Columbus, OH, USAGlobal Integration of Intelligent Manufacturing and Smart Industry for Good of Humanity. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2351978918311399>
- [12] I. J. Goodfellow, Y. Bengio, and A. C. Courville, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.
- [13] A. Krizhevsky, “Learning multiple layers of features from tiny images,” *University of Toronto*, 05 2012.
- [14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [15] F. Chollet, *Deep Learning with Python*, 1st ed. USA: Manning Publications Co., 2017.
- [16] P. Xie, M. Bilenko, T. Finley, R. Gilad-Bachrach, K. E. Lauter, and M. Naehrig, “Crypto-nets: Neural networks over encrypted data,” *ArXiv*, vol. abs/1412.6181, 2014.

- [17] C. Dwork, “Differential privacy: A survey of results,” in *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19.
- [18] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” *ArXiv*, vol. abs/1607.00133, 2016.
- [19] J. Liu, M. Juuti, Y. Lu, and N. Asokan, “Oblivious neural network predictions via minionn transformations,” *IACR Cryptology ePrint Archive*, vol. 2017, p. 452, 2017.
- [20] H. Bae, J. Jang, D. Jung, H. Jang, H. Ha, and S. Yoon, “Security and privacy issues in deep learning,” *ArXiv*, vol. abs/1807.11655, 2018.
- [21] J. Zhao, Y. Chen, and W. Zhang, “Differential privacy preservation in deep learning: Challenges, opportunities and solutions,” *IEEE Access*, vol. 7, pp. 48 901–48 911, 2019.
- [22] R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1310–1321. [Online]. Available: <https://doi.org/10.1145/2810103.2813687>
- [23] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, “Applied federated learning: Improving google keyboard query suggestions,” *CoRR*, vol. abs/1812.02903, 2018. [Online]. Available: <http://arxiv.org/abs/1812.02903>
- [24] A. Hard, K. Rao, R. Mathews, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, “Federated learning for mobile keyboard prediction,” *CoRR*, vol. abs/1811.03604, 2018. [Online]. Available: <http://arxiv.org/abs/1811.03604>
- [25] M. Chen, R. Mathews, T. Ouyang, and F. Beaufays, “Federated learning of out-of-vocabulary words,” *CoRR*, vol. abs/1903.10635, 2019. [Online]. Available: <http://arxiv.org/abs/1903.10635>
- [26] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. aurelio Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng, “Large scale distributed deep networks,” in *Advances in Neural Information Processing Systems* 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1223–1231. [Online]. Available: <http://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks.pdf>
- [27] J. Chen, R. Monga, S. Bengio, and R. Józefowicz, “Revisiting distributed synchronous SGD,” *CoRR*, vol. abs/1604.00981, 2016. [Online]. Available: <http://arxiv.org/abs/1604.00981>
- [28] L. Zhu, Z. Liu, and S. Han, “Deep leakage from gradients,” *CoRR*, vol. abs/1906.08935, 2019. [Online]. Available: <http://arxiv.org/abs/1906.08935>
- [29] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in *CCS ’17*, 2017.
- [30] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, “Federated learning with non-iid data,” *CoRR*, vol. abs/1806.00582, 2018. [Online]. Available: <http://arxiv.org/abs/1806.00582>
- [31] V. Smith, C. Chiang, M. Sanjabi, and A. Talwalkar, “Federated multi-task learning,” *CoRR*, vol. abs/1705.10467, 2017. [Online]. Available: <http://arxiv.org/abs/1705.10467>
- [32] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, “Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images,” *2018 International Conference on Biometrics (ICB)*, Feb 2018. [Online]. Available: <http://dx.doi.org/10.1109/ICB2018.2018.00023>
- [33] R. M. Alguliyev, R. M. Aliguliyev, and F. J. Abdullayeva, “Privacy-preserving deep learning algorithm for big personal data analysis,” *Journal of Industrial Information Integration*, vol. 15, pp. 1 – 14, 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2452414X18301456>