Resolution-Aware Deep Multi-View Camera Systems

1nd Zeinab Hakimi
School of Electrical and Computer Engineering
Pennsylvania State University
University Park, PA, USA
zuh17@psu.edu

2nd Vijaykrishnan Narayanan
School of Electrical and Computer Engineering
Pennsylvania State University
University Park, PA, USA
vxn9@psu.edu

***Abstract*—Recognizing 3D objects with multiple views is an important problem in computer vision. However, multi view object recognition can be challenging for networked embedded intelligent systems (IoT devices) as they have data transmission limitation as well as computational resource constraint. In this work, we design an enhanced multi-view distributed recognition system which deploys a view importance estimator to transmit data with different resolutions. Moreover, a multi-view learning-based super-resolution enhancer is used at the back-end to compensate for the performance degradation caused by information loss from resolution reduction. The extensive experiments on the benchmark dataset demonstrate that the designed resolution-aware multi-view system can decrease the endpoint's communication energy by a factor of 5× while sustaining accuracy. Further experiments on the enhanced multi-view recognition system show that accuracy increment can be achieved with minimum effect on the computational cost of back-end system.**

I. Introduction

Classifying images into different categories based on provided visual information is a fundamental problem in intelligent vision systems and the core of computer vision applications such as segmentation, localization and detection. Smart cameras that embed intelligent visual processing either locally or through the ability to communicate to a remote server are becoming pervasive. These embedded systems are becoming increasingly powerful and multiple coordinated smart cameras that enable to recognize 3-dimensional real world objects have become popular in recent times. Such multi-view deep learning object recognition systems have proven to improve the accuracy of object classification by leveraging features captured from various views of an object [1]. However, multi-view classification is challenging in distributed embedded devices due to the resource-constrained nature of these systems [2].

This paper proposes a collaborative approach to mitigate the communication cost in a distributed system. In this method, a communication cost is reduced by proposing the importance estimator method. Using this method, we dynamically select less informative views to send sub-sampled visual data for classification at back-end. However, sub-sampled input images detrimentally impact the classification accuracy of the implemented Deep Neural Network (DNN) in the server. Although deep neural network approaches have shown to be powerful in providing deep discriminative features automatically, their performance is sensitive to the resolution of the image data. Most existing DNN approaches performance rely on the high-resolution images for training and testing tasks. Our proposed method enhances the resolution of images using multi-view super resolution technique in the back-end and reconstructs the necessary discriminative features. We leverage DNN based super resolution methods which significantly improves the quality of images relative to traditional interpolation methods such as Bicubic [5]. Even though the classification accuracy can be increased by DNN-based super resolution method, it increases the latency of the system because of the intensive computations it entails. Therefore, it is crucial to only enhance the resolution of the most informative views at the back-end. To this end, we leverage a saliency and entropy-based metrics to detect informative views. A recent work addressed the communication problem in distributed inference by shrinking each node's feature size before transmission. It either selects the important features among all the views or find one dominant view and only reduce the feature's resolution of non-dominant views [4]. Although they decrease the communication cost by reducing the feature size, they don't use any super-resolution method to compensate the accuracy loss due to the feature reduction. The key contributions of this paper include: (i) Design of view importance estimator in a distributed embedded system to manage data resolutions based on the resource constraints of the front-end and back-end systems while achieving the desirable accuracy; (ii) Accuracy improvement in a multi-view system by using the super-resolution method; (iii) Significant reduction in communication cost and energy consumption by up to 98% with sacrificing only 2% in accuracy by leveraging importance decision making and using multi super resolution in a distributed environment.

II. Background

A. Multi-View Convolutional Neural Network

Multi-view CNN models (MVCNN, hereafter) are the state of the art approaches for 3D object recognition. They leverage the viewpoint-wise deep features of the images to improve object classification accuracy. More viewpoints provide more details of an object and results in better accuracy. MVCNN algorithm's focus is on the feature extraction of an object

This work was supported in part by CBRIC, a center funded by Semiconductor Research Corporation(SRC).

978-3-9819263-5-4/DATE21/©2021 EDAA

recognition network and is typically composed of two basic parts of feature extraction (CNN1) and classification (CNN2). To extract viewpoint-wise features, each camera deploys CNN1 independently. The view pooling layer with different algorithms can be applied to aggregate the deep features generated from each view. Once it has fused all the features, the second CNN2 focuses on classification.

B. Learning-Based Super Resolution

Deep learning based techniques have been used for various image enhancement techniques, among them one of the image enhancement techniques uses a GAN (Generative adversarial network) based super resolution. These type of super resolution approaches has produced better quality of images as measured by PSNR over traditional super resolution techniques Bicubic operations. However, it should also be noted that GAN based works are highly compute intensive and slower than the simple interpolation (S-SR) method by the factor of 700. Even in high performance IBM Power server machines, the latency for enhancing the image is high. Therefore, deploying learning-based super resolution (L-SR) methods in a multi-view system is expensive. One of the goals in this paper is to deploy the techniques to selectively choose important views for L-SR reconstruction and reduce the inference latency.

C. View Importance Estimation

In order to select the connection bandwidth or more specifically the resolution for communicating the images, we need to identify the views that introduce more informative features for our classification task. Besides Multi-view super resolution network employs this scheme to selectively reconstruct the low resolutions images at the back-end. Two different methods are proposed for resolution selection.

1) Entropy: In machine learning, Shannon entropy presents a DNN model's optimization by minimizing negative log-likelihood of multivariate probability distribution, which is defined as:

$$\mathit{entropy}(\mathrm{y}) = -\mathbb{E}_{x\sim P}\left[\log(\mathrm{y})\right] = -\sum_{c\in C} y_c \log P(\mathbf{x};\Theta), \tag{1}$$

We attach a residual layer deriving from Shannon entropy to measure the uncertainty associated with sensor's modality ϕ:

$$\xi(\mathrm{y};\phi) = \sum_{c} \sigma(-y_{c;\phi}\log(y_{c;\phi})) \tag{2}$$

where ξ is the likelihood function and y_c is the inferred probability for label c ($\sum_c y_c = 1$). σ is a rectified linear unit (ReLU) activation to suppress negative impact undefined in likelihoods. Entropy can provide a measure for proximity to the training data, because if the probability distribution's uncertainty decreases, DNN generates classification results which are based on the features embedded in training data.

2) Saliency Score: One way to identify the most important parts of an image is saliency map. Image segmentation is an important application of saliency estimation in computer vision. A saliency map is a two-dimensional map in which the most salient pixels are highlighted. The brightness of a pixel in the saliency map is directly proportional to its saliency. To create the saliency map, basic features such as color, orientation, and intensity are extracted from the image. Consequently, the feature maps are combined into three conspicuity maps. Finally, the mean of all three conspicuity maps results in the saliency map. Since the value of each pixel in the feature map indicates the importance of that pixel, the total number of pixels whose value exceeds some threshold represents the value of that viewpoint image for classification. Therefore, the saliency score is defined as:

$$S(x) = \sum_{i=1}^{n} I(x_i > \tau) \tag{3}$$

where S is the saliency score estiamtion function, x is the calculated saliency map for an input image and τ is a threshold to indicate sailent regions.

TABLE I
COMMUNICATION COMPARISON BETWEEN FEATURE SPACE AND RAW IMAGES IN MULTI-VIW SYSTEM DEPLOYING GOOGLNET

Data	**Communicated Data(KB)**
`Standard Image`	$\#sensors \times 196.6$
`Sub-sampled Image`	$\#sensors \times 12.2$
`Feature Size`	$\#sensors \times 404$

III. RESOLUTION-AWARE DEEP INTELLIGENT SYSTEM

In this section, we discuss three proposed architectures: High Resolution MVCNN (HR-MVCNN), Low Resolution MVCNN (LR-MVCNN) and Enhanced MVCNN in a multi-view embedded system environment. In each design, the front-end system comprises n cameras installed to capture objects from n different viewpoints. These images are transmitted wirelessly to the back-end. Our multi view system uses the networks proposed in [2] with reduced input resolution as the baseline (128 × 128). This network partitions the computation between end-devices and the server. The key difference in our approach is that since the image size can be smaller than the feature dimensions, we propose to transmit the raw image data rather than the features, to the back-end. Besides, our design allows to send some view data with low resolution. All of these considerations result in a significant reduction in the communication cost. The data communicated for raw image and features is depicted in Table 1.

HR-MVCNN Fig. 1 shows the overview of the HR-MVCNN system. In this architecture, the front-end devices capture the image of the object and compute its view's importance using the importance estimation component. Two different algorithms are deployed in this component: entropy and saliency score. Entropy is estimated for each camera by using early features extracted for each view and applying global average pooling (GAP) at the end. In order to decrease the computation cost of importance estimation, an alternate approach is proposed which uses a lighter-weight saliency score method. In this method, the importance of each camera is proportional to the magnitude of salient regions. Using saliency score, the

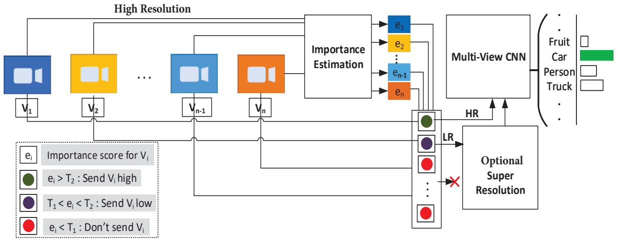
High Resolution
V$_1$
V$_2$
...
V$_{n-1}$
V$_n$
Importance Estimation
e$_1$
e$_2$
e$_{n-1}$
e$_n$
Multi-View CNN
Fruit
Car
Person
Truck
HR
LR
Optional Super Resolution
e$_i$ Importance score for V$_i$
e$_i$ > T$_2$: Send V$_i$ high
T$_1$ < e$_i$ < T$_2$: Send V$_i$ low
e$_i$ < T$_1$: Don't send V$_i$

Fig. 1. Overview of HR-MVCNN

amount of computation is decreased. Two thresholds T_1 and T_2 ($T_2 > T_1$) are considered for data transmission. Threshold T_1 is used for pruning the least informative data. Views with importance score less than T_1 are not transmitted to the back-end. Views with importance greater than T_2 (i.e. views with the most discriminative features) are transmitted with their standard resolution, while the remaining views are down sampled to 64×64 before transmission. Therefore, HR-MVCNN exploits both low and high bandwidth connections to transmit data from sensors with different resolutions. Increasing the number of standard resolution views leads to an increase in the amount of required bandwidth. In other words, front-end consumes more energy for providing more information to the back-end which might be critical when energy resources are limited at the front-end. The required energy for the front-end devices is:

$$E_{\text{front-end}} = E_{\text{importance}} + (n^* - x)E_{\text{cm-low}} + xE_{\text{cm-high}} \quad (4)$$

where x is the number of views with importance score greater than T_2 and $E_{\text{importance}}$ is the energy for view importance calculation. $E_{\text{cm-low}}$ and $E_{\text{cm-high}}$ are the energy needed for communicating low and high resolution images, respectively. Back-end latency in HR-MVCNN is proportional to the latency of performing S-SR up-sampling and running MVCNN for classification. Although the power for entropy estimation is high, it is lower than the power required for transmitting standard images. For instance, the estimated energy for entropy estimation of batch of data with ISAAC accelerator [6] is 0.5 J. In contrast, the communication energy for sending the same batch of images with standard resolution is 2.5 MJ with LTE and 39.4 MJ with 3G connection. As the bandwidth becomes more constrained, there is no room for standard resolution images. So we propose another design LR-MVCNN.

LR-MVCNN The Low Resolution MVCNN architecture aims to further alleviate the communication cost of inference by decreasing all view's resolution (down-sampling) before transmission. In LR-MVCNN, each end device is equipped with importance estimation component to evaluate the feature quality of each image. The importance estimation component receives sensor's data as an input, extracts its relative feature using $CNN1$, and calculates the importance measure, using two methods of entropy and saliency, as a proxy for context representations. This additional data along with the image helps in identifying the most informative viewpoints to optimize on the latency in the back-end system. To this end, each end device down samples its input image to 64×64 resolution, encapsulates it with the importance information and sends it to the back-end via a low bandwidth wireless connection.

In the back-end system, identifying the most informative viewpoints allows to perform the time-consuming deep learning based super-resolution only on selected viewpoints rather than all viewpoints. Hence, the informative viewpoints are enhanced to higher resolution with the help of L-SR model and the rest of the viewpoints are enhanced through S-SR. Threshold T is considered to find the optimal number of informative views. Finally, the context-aware collaborative MVCNN [7] is used to exploit the context of each view and aggregate view-wise features with respect to their importance for final classification layers.

The required energy for front-end devices in this system is calculated as:

$$E_{\text{front-end}} = E_{\text{importance}} + n^* E_{\text{cm-low}} \quad (5)$$

where $E_{\text{importance}}$ is the energy for computing the entropy and $E_{\text{cm-low}}$ is the required energy for communicating low resolution images which is proportional to the number of transmitted bits and the distance between the front-end and the back-end systems.

Enhanced MVCNN In order to further improve our design and increase the accuracy, we consider Enhanced-MVCNN scheme. In this design, we leverage the benefits of both designs: transmitting information using the same method as HR-MVCNN in the front-end, and reconstructing the information with LR-MVCNN schem at the back-end. This enables us to limit accuracy drop with reconstructed information while keeping the communication cost as low as HR-MVCNN method.

IV. Experimental Setup

We use multi-view iLab-80M [8] dataset to evaluate our work. The standard split of 2,377 training, 600 testing are used with respect to each view. In front-end devices, a partial part of the GoogleNet which consists of the first three convolution layers and six inception layers are used. TensorFlow tool is used to train and test DNN models in multi-view systems. Our server platform, IBM 24-core Power9, is equipped with NVIDIA Tesla V100 GPUs which is used in our experiments. Further, the accuracy, energy and latency of the models are evaluated at the time of inference.

V. Evaluation

The key factor which determines importance score thresholds T_2 in HR-MVCNN design is the trade-off between accuracy and the energy each front-end sensor consumes for communicating views. To find an optimal threshold, Fig. 3 and Fig. 4 plot both accuracy and front-end energy cost as a function of entropy and saliency threshold with 3G connection respectively. As we observe in Fig. 3, the least-informative views (which only provide redundant features) are pruned by a threshold $T1$ of 40 as the accuracy drop is negligible for the $T1$ less than this point(around 1% drop). Fig. 3 also indicates that while the

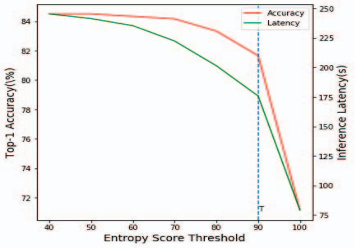

Fig. 2. Accuracy and Latency of LR-MVCNN on iLab-80M dataset [8]

TABLE II
COMPARISON OF FRONT-END ENERGY AND ACCURACY FOR DIFFERENT METHODS. THE PROPOSED METHODS REDUCED THE ENERGY WHILE SUSTAINING THE ACCURACY

Method	Accuracy	Energy(KJ) 3G	Energy(KJ) LTE
MVCNN_Wavg [2]	90	54.9	9.8
Baseline	84.5	28.7	1.8
HR-MVCNN	82.5	9.96	0.97
LR-MVCNN	81.5	**0.46**	**0.37**
Enhanced	**88.8**	9.96	0.97

system sustains its accuracy in the highest band (i.e. between 82% and 84%), we can significantly save on the energy cost by sending only views with entropy greater than 80 with high resolution. Fig. 4 shows that choosing threshold $T2$ for saliency decreases the energy by 20%. Although this result indicates that the saliency score is good to select informative views, it is not as effective as the entropy method to turn off the cameras which leads to more front-end energy cost.

Fig. 2 describes the result for the top-1 accuracy and the latency of deploying LR-MVCNN. In this figure, it can be observed that as more view points are improved by L-SR method(with decreasing T threshold), the classification accuracy decreases (red curve), however, at the cost of the latency (green curve). For instance if we decrease threshold T from 90 to 50, the top-1 classification accuracy increases by 3.16 percentage point, but at the cost of 65.64 s increase in latency. Therefore, there is a trade-off between accuracy and latency of the system. In the LR-MVCNN design, the front-end energy cost is static for different images and it is proportional to down-sampled resolution.

Table II reports the inference accuracy and front-end energy consumption across various designs. Front-end energy is calculated for both the computation and communication of end-devices. As we can observe, energy consumption of HR-MVCNN and Enhanced-MVCNN are equal because they use the same method for transferring information. However, the accuracy drops less (only 1.2%) in Enhanced-MVCNN than HR-MVCNN due to the information reconstruction at the back-end. LR-MVCNN saves the most energy and still has a comparable accuracy. Energy saving in this scheme is the result of reducing communication bandwidth as much as possible. Therefor LR-MVCNN is suitable for low speed IoT network connections such as 3G in which case the energy cost is only 0.46 KJ.

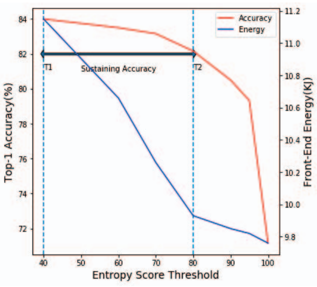

Fig. 3. Accuracy and front-end energy cost of HR-MVCNN using entropy with 3G connection.

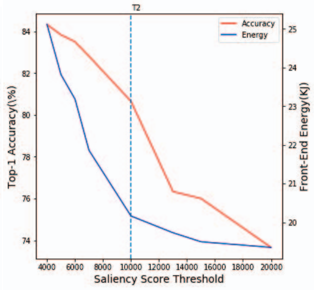

Fig. 4. Accuracy and front-end energy cost of HR-MVCNN using saliency with 3G connection.

VI. CONCLUSION

There has been a proliferation of embedded smart camera systems in a wide variety of applications. In many applications, the cameras embed intelligence by adding compute power or by working in consonance with a server to analyze the images using algorithms such as object detection. However, most of these systems are limited in both their compute and communication abilities, significantly due to energy constraints. While adapting the image resolution is a simple knob to adjust the communication needs, the resolution of the images has significant impact on the data analytic that can be performed at the server. Our work demonstrates how entropy and saliency score can be used as a proxy for the most informative views when dealing with transmission or reconstruction of high resolution images with super resolution method.

REFERENCES

[1] Su, Hang and Maji, Subhransu and Kalogerakis, Evangelos and Learned-Miller, Erik, *Multi-view convolutional neural networks for 3d shape recognition*, Proceedings of the IEEE international conference on computer vision, 2015.

[2] J. Choi and Z. Hakimi and P. W. Shin and J. Sampson and V. Narayanan, *Context-Aware Convolutional Neural Network over Distributed System in Collaborative Computing*, 2019 56th ACM/IEEE Design Automation Conference (DAC),, 2019, pp. 1-6.

[3] Kang, Yiping and Hauswald, Johann and Gao, Cao and Rovinski, Austin and Mudge, Trevor and Mars, Jason and Tang, Lingjia, *Neurosurgeon: Collaborative intelligence between the cloud and mobile edge*,ACM SIGARCH Computer Architecture News, vol. 45, no. , pp. 615–629, 2017.

[4] Singhal, Manik and Raghunathan, Vijay and Raghunathan, Anand, *Communication-efficient view-pooling for distributed multi-view neural networks*, DATE, 2020.

[5] Xintao Wang and Ke Yu and Shixiang Wu and Jinjin Gu and Yihao Liu and Chao Dong and Chen Change Loy and Yu Qiao and Xiaoou Tang, *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks*, CoRR, 2018.

[6] Shafiee, Ali and Nag, Anirban and Muralimanohar, Naveen and Balasubramonian, Rajeev and Strachan, John Paul and Hu, Miao and Williams, R Stanley and Srikumar, Vivek, *ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars*, ACM SIGARCH Computer Architecture News, 2016, pp 14-26.

[7] J. Choi and Z. Hakimi and J. Sampson and V. Narayanan, *Byzantine-Tolerant Inference in Distributed Deep Intelligent System: Challenges and Opportunities*, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, pp. 509-519, 2019.

[8] Borji, Ali and Izadi, Saeed and Itti, Laurent, *ilab-20m: A large-scale controlled object dataset to investigate deep learning*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2221–2230, 2016.