From Data to Knowledge: Deep Learning Model Compression, Transmission and Communication

Ziqian Chen[1], Shiqi Wang[2], Dapeng Oliver Wu[3], Tiejun Huang[1], Ling-Yu Duan[1*]
National Engineering Lab for Video Technology, Peking University, Beijing, China[1]
Department of Computer Science, City University of Hong Kong, Hong Kong, China[2]
Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA[3]
{wzziqian,tjhuang,lingyu}@pku.edu.cn,shiqwang@cityu.edu.hk,dpwu@ieee.org

ABSTRACT

With the advances of artificial intelligence, recent years have witnessed a gradual transition from the big data to the big knowledge. Based on the knowledge-powered deep learning models, the big data such as the vast text, images and videos can be efficiently analyzed. As such, in addition to data, the communication of knowledge implied in the deep learning models is also strongly desired. As a specific example regarding the concept of knowledge creation and communication in the context of Knowledge Centric Networking (KCN), we investigate the deep learning model compression and demonstrate its promise use through a set of experiments. In particular, towards future KCN, we introduce efficient transmission of deep learning models in terms of both single model compression and multiple model prediction. The necessity, importance and open problems regarding the standardization of deep learning models, which enables the interoperability with the standardized compact model representation bitstream syntax, are also discussed.

KEYWORDS

Knowledge communication, deep learning model compression, standardization.

ACM Reference Format:
Ziqian Chen, Shiqi Wang, Dapeng Oliver Wu, Tiejun Huang, Ling-Yu Duan. 2018. From data to knowledge: deep learning model compression, transmission and communication. In 2018 ACM Multimedia Conference (MM '18), October 22-26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240654

1 INTRODUCTION

Over the past decade, the big data is spurring on tremendous growth in the information technologies, especially with the wide deployment of the Internet of Things (*IoT*) [5] which is constantly generating data from edge devices for better sensing the real world. However, the data-centric processing is creating many issues as obtaining the data only is not the ultimate objective. The big data should be converted to descriptive knowledge for effective utilization. As such, an innovative concept of the Knowledge Centric Networking (KCN), which creates a dramatic paradigm shift from big data to big knowledge in communication, was proposed in [50]. In KCN, the knowledge creation, composition and distribution are three vital components.

Recently, the deep learning based algorithms, which can be regarded as the specific techniques in creating knowledge, have shown great potentials in dealing with various tasks in a data-driven manner. With the increase of computational capability and enormous volume of data, numerous powerful deep neural network models have been learned. Moreover, the learned models can be furtherfi netuned to tackle domain-specific problems. Therefore, from the perspective of knowledge extraction, the deep learning models learned from data are creating and conveying valuable knowledge. As such, in additional to the compact feature representation which treats feature descriptors as the special modality of knowledge, the deep learning model communication serves as an indispensable component in KCN as well. However, when the knowledge information is frequently exchanged in terms of deep learning models, there exist unprecedented challenges for effectively representing the large scale deep neural networks, especially in those scenarios where minimal knowledge degradation is expected.

Different from traditional deep learning model compression which aims to produce a lightweight deep neural network, the motivation, application scope as well as the methodology of deep learning model compression should be further extended in the context of knowledge communication, and more important and practical issues regarding deep learning model communication are raised accordingly. Moreover, regarding transmission and communication, there is also an increasing demand towards standardization and interoperability of deep learning model compression. Based on the existing standards of raw data and features, how the interoperability towards compact deep model representation under frequent knowledge exchange environment is enabled should be further investigated.

In this work, motivated from substantially different principles, we investigate the compression, transmission and communication of deep learning models from the perspective of knowledge transmission and communication. The main contributions of this paper are summarized as follows:

*Ling-Yu Duan is the corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240654

- We propose to reformulate the deep model compression, transmission and communication framework in the context of KCN, such that the transmission of knowledge in terms of deep learning model is enabled.
- The inter model prediction from the perspective of communication and transmission is investigated in the scenario of multiple deep model compression, such that better coding efficiency can be ensured.
- The interoperability of the deep model compression is discussed and future standardization towards deep model compression is envisioned.

2 RELATED WORK

2.1 Deep Learning Models

Recent years have witnessed a strong growth of interest and development of the deep neural networks. Since *Alexnet* [30] won the first place on *ImageNet* classification challenges, fantastic progress has been made to design advanced deep learning models, such as *VGG* [44], *Resnet* [25], *GoogleNet* [47], *DenseNet* [27], which have been successfully applied in variousfi elds. Moreover, other deep neural networks such as Recurrent Neural Networks (*RNN*) [8] and Generative adversarial networks (*GAN*) [20, 56] have also shown great potentials to tackle specific problems using their unique network structure. It is widely acknowledged that a deeper and wider network can achieve better representation and gain more knowledge. However, heavy burden is also imposed on the storage of both training data and models, as well as the computational resources.

To address these issues, different deep learning model compression approaches have been proposed towards the specifically designed lightweight deep neural networks for the correspondingly given tasks. More specifically, model parameters pruning [11, 23, 24], matrix factorization [52, 55], quantization [19, 32, 54], filter selection [3, 26, 36, 49] as well as knowledge transfer [33, 37] with network redesign are the commonly adopted approaches. Recently, the work in [34] paid attention to the gradient compression during distribution training period. However, most of these works focus on the knowledge creation, such that the model is compressed or distilled to convey a smaller amount of knowledge. In the context of knowledge transmission and communication, the scope of deep neural compression can be further extended in several ways, including the multiple model transmission and standardization.

2.2 Knowledge Centric Networking

The interest of knowledge creation, composition and distribution [50] has been growing at an accelerated pace in the context of Internet of Things (*IoT*) [5], which aims at sensing every physical object with large volume of the data created. Such enormous amount of data inspires better communication modality arrangement , and the innovative concept of KCN was proposed in [50]. In KCN, instead of directly transmitting raw data or the content, the generated knowledge from raw *IoT* sensing data is extracted and transmitted. The digital object index (DOI) technology is applied on the content-based KCN framework to enable search of knowledge [46]. As a consequence, redundancy on the raw data can be largely removed with knowledge extraction, making better utilization of the big data from the perspective offi nal utility in the future smart society. In the context of KCN, the deep learning models, which can be regarded as a specific form of knowledge, are revisited from a novel perspective in terms of compression, transmission and utilization.

2.3 Video Compression Standard

In contrast to knowledge, images and videos serve as the important modalities of raw data, which also offer a digital bridge from real visual world to valuable knowledge. As such, the video compression, transmission and communication have received sufficient interest and a series of video coding standards have been developed, including H.262/MPEG-2 [1], H.264/MPEG-4 AVC [2], H.265/HEVC [45] as well as AVS2/IEEE 1857.4 [6] from ITU-T VCEG, ISO/IEC MPEG, IEEE and AVS. Though the state-of-the-art video coding standards such as H.265/HEVC have dramatically improved the coding performances, such compression efficiency still cannot meet the requirement of the exponential increase of the data volume. This also motivates the knowledge centric solution towards the better utilization of the raw data.

2.4 Feature Based Knowledge Communication Standard

The automatic image and video analysis relies on the extracted features, such that conveying the powerful and discriminative features becomes an alternative way towards future knowledge communication. The analyze-then-compression framework was analyzed in [18, 43], where the advantages and technical challenges were discussed. In [10], the authors focused on enhancing keypoint encoding for improved feature extraction. In [16], the predictive distributed visual analysis was presented. In [17], coordinating distributed algorithms for feature extraction is taken into consideration with limited signaling.

In the view of the importance of compact feature representation, the MPEG hasfi nalized the Compact Descriptors for Visual Search (CDVS) [14]. In CDVS, the handcrafted features (i.e. SIFT) are compactly represented and standardized. The features of CDVS are highly efficient and adaptive to the given bit budget, and superior performance of visual search has been achieved based on the combination of the local and global compact descriptors. With the development towards deep learning based algorithms, deep models show their great abilities of feature extraction. The Compact Descriptors for Video Analysis (CDVA) [15] combines both handcrafted features and deep learning based features. In [13], the future standard towards AI oriented large-scale video management for smart city is also discussed, targeting at utilizing the feature knowledge with ensured interoperability.

3 KNOWLEDGE COMMUNICATION IN TERMS OF DEEP LEARNING MODEL

In this section, we propose to reformulate deep learning model compression, transmission and communication framework as the specific forms of knowledge from the perspective of knowledge creation, transmission and standardization. Instead of considering single model transmission only, the feasibility of multiple model transmission in a consecutive or simultaneous way is also involved.

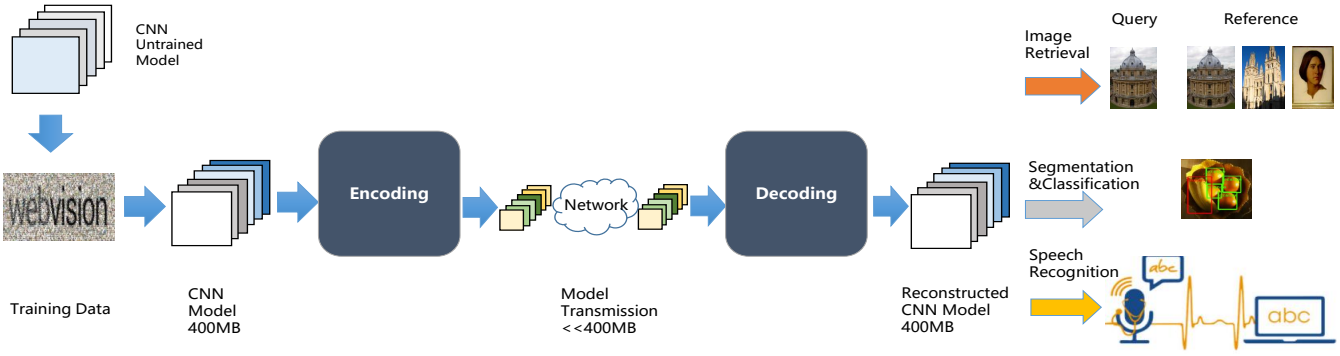


Figure 1: Illustration of knowledge transmission with deep learning model compression, where the decoding is also performed to reconstruct the original model after compression and transmission.

3.1 Knowledge Creation

The improvement of hardware acquisition capabilities has enabled big data to be collected from the real world. As a specific example, the IP video traffic will take up 82% of all consumer Internet traffic by 2021 [35], and the surveillance videos which are constantly generated in real-time have already become the biggest big data [28]. Without certain automatic tools, it is difficult to rely on human beings to analyze such amount of big data. Moreover, the raw big data also imposes heavy burden on network transmission and storage. The rapid development of machine learning, especially deep learning, greatly facilitates the utilization of the big data for further analysis and knowledge extraction. As such, KCN takes a further step towards future multimedia networking, and the knowledge rather than the directly acquired raw data becomes the centric point, which can significantly mitigate the hardware burden and energy issues.

For image and videos, features are regarded as an important modality of knowledge, as it conveys useful information regarding the low-level attributes such as Scale Invariant Feature Transform (SIFT), shape, color, etc., as well as the high level semantic meanings. Hence, in the context of KCN, the feature extraction can be generally treated as a procedure of knowledge creation. With the advances of deep learning and computer vision, features evolve from hand-craft based to deep-learning based in a data-driven manner, and more powerful models as well as representations have been developed to obtain the knowledge from raw data.

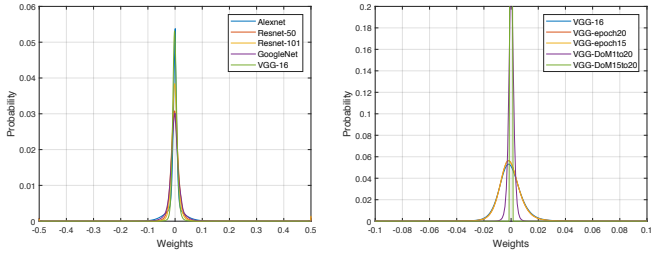
The scope of knowledge should not be limited to features only, and the model that is responsible for the extraction of features is also becoming an essential modality of knowledge. The sophisticatedly trained deep neural networks from the large-scale datasets for feature extraction are also conveying meaningful information regarding the specific tasks. As such, methods have been developed for the neural network visualization to investigate the information that has been learned [51, 53]. In other words, these elaborately trained deep neural network models are highly representative for these training data with abundant knowledge information. In essence, the deep neural network models also align with the concept of KCN, as the model which also serves as the knowledge created from the learning process can be ultimately delivered and distributed. The different levels of knowledge composition contribute to the different specific objectives, and given the model the network can be furtherfi netuned for self-adaption. Taking the models trained from ImageNet as examples, these well-trained models are able to tackle many classification tasks with an ordinaryfi ne-tuning or even without anyfi ne-tuning, as abundant knowledge has been absorbed from the labeled ImageNet data. Nowadays, there is a significant increase on the numbers and varieties of such well-trained or half-trained models in the Internet, which are subsequently subjected to further transmission and communication.

With the advances of computational capabilities of various facilities, the model can be trained at the server or even the front end, leading to more frequent exchange of models in different granularity levels. Generally speaking, the transmission of one model only cannot fully satisfy the knowledge transfer purpose. For example, multiple models need to be transmitted simultaneously, and the update of the models with new training data or training strategy is also regularly encountered. Moreover, to enable interoperability, the standardization of model compression becomes an emerging important topic.

3.2 Standardization of Knowledge Compression

In this subsection, we discuss the potentials and necessities for the standardization of deep model compression. First, we discuss the standardization for raw data and features, where the video texture and visual features serve as two specific examples. Subsequently, we envision the future deep learning model compression standard.

From the standpoint of raw data compression, video coding becomes a specific and noticeable example. Recent years have witnessed a series of video coding standards. It is worth mentioning that most techniques adopted in present video coding standards aim to improve the video compression performance in terms of



(a) (b)

Figure 2: (a) The parameter distributions of different deep learning models. (b) The comparisons between the original model among different versions offi netuned VGG and DoM parameter distributions. Analogously, the x axis is divided with equal space 0.002, and only weights within the range of [-0.5 0.5] in (a) and weights within the range of [-0.1, 0.1] in (b) are displayed.

bit rate and video quality. In spite of the different motivations of video coding and deep learning model compression, the design philosophy in video coding can still benefit to the deep model compression. For example, the common strategy of both video and model compression is to remove the redundant information. The principles of inter-prediction and transform-relevant techniques, which serve as the key modules in video coding, can also promote the performance of deep learning model compression by further exploiting the statistical information of model parameters.

Regarding knowledge transmission in terms of features, the rapid development of deep learning also imposes great challenges to the standardization, as the compact feature representation involves both feature extraction and compression. The recent developments of deep learning have also advanced the compact feature representation [7, 48]. The CDVA [15], during the development of which much more attention has been paid to the combination of deep feature and handcraft feature, has received increasing interest recently. However, the feature extraction models are expected to evolve over the coming years, such that there arise challenging issues along with the explosion of the deep learning models for the standardization.

It is apparent that the compression and standardization of deep learning models possess favorable properties in KCN. Nevertheless, the explosion of the deep learning structures is also bringing challenges. Currently, most of the well-designed deep learning models are organized by different numbers of layers in different types. Taking Convolutional Neural Network (CNN) as an example, it is mostly constituted of convolutional layers, pooling layers, fully connected layers or other self-designed layers. The activation functions such as Relu and Sigmoid are used to introduce the nonlinearities. As such, how to standardize these modules in the bitstream syntax should be carefully considered. Moreover, for video coding the structure of source visual signals is well established in terms of the pixel values. By contrast, the emerging deep learning models may pose new challenges. For example, when a new activation function or type of layer is developed, it may be difficult to accommodate this modified network structure to the existing standard. As such, the extension ability of the model compression standard should also be investigated.

Moreover, for the deep learning model compression standard, the scalability and generalization ability should also be well considered due to the varying requirements and bandwidth conditions. For example, the deep learning models could be compressed ranging from lossless to lossy compression. As illustrated in Figure 1, the pipeline of deep model compression and transmission is demonstrated.

3.3 Deep Learning Model Compression

In this subsection, we will investigate the deep learning model compression to effectively convey knowledge in terms of both single and multiple models. In particular, in the scenario of multiple model compression, the inter model prediction is further studied based on knowledge center.

3.3.1 Single model compression. Let usfi rst analyze deep neural network model compression for the single model case, which serves as the foundation of the multiple model compression. In particular, in the scenario where a brand-new model is required to be compressed and transmitted, the intrinsic structure and statistics of the parameters play important roles. As pointed out by numerous works, the parameters occupy the majority of the storage space. As such, deep learning model compression can be exploited by the model parameter statistics.

From the perspective of parameter distributions, we investigate several state-of-the-art deep learning models trained on ImageNet, including *Alexnet*, *Googlenet*, *VGG-16*, *Resnet-50*, *Resnet-101*. The parameter distributions are shown in Figure 2 (a). We can observe that the peak locates at zero with limited ranges of values. This provides an useful evidence that the parameters can be further compressed, which has been well validated [11, 23, 24]. Moreover, the parameter precision should also be taken into consideration as most of the models are designed with *single-precision(float32)*. As indicated in the recent research [12], thefi nal loss of whole model output is bounded by the weighted accumulated of layer-wise loss.

Here we perform a straightforward compression in generating the baseline for single model. It is worth mentioning that we are not attempting to adopt advanced solution such as trained quantization with k-means clustering [19, 24], as the performance of the naive compression strategy is also expected to provide the baseline for further investigation. Generally speaking, the limited decimal precision loss in a certain range will barely introduce the performance degradation [34, 42]. Therefore, we adopt the received parameter decimal precision to control the compression ratio, and the represented decimal precision is termed as Representation Precision (RP). Suppose the decimal precision of parameters w is limited to p, which indicates that the parameters w are converted to the corresponding integers by multiplying 10^p in the encoding process. Subsequently,fi x-point followed by binarization and lossless encoding are performed.

More specifically, given the learned model and specified RP, we convert all the parameters to the integer with $w' = int64(w * 10^p)$, which are then binarized as $w'_T = Binary(w')$ and encoded into the bitstream $w'_{compress} = Encoding_{entropy}(w'_T)$. The lzma algorithm [41] is chosen as the coding method due to its promising

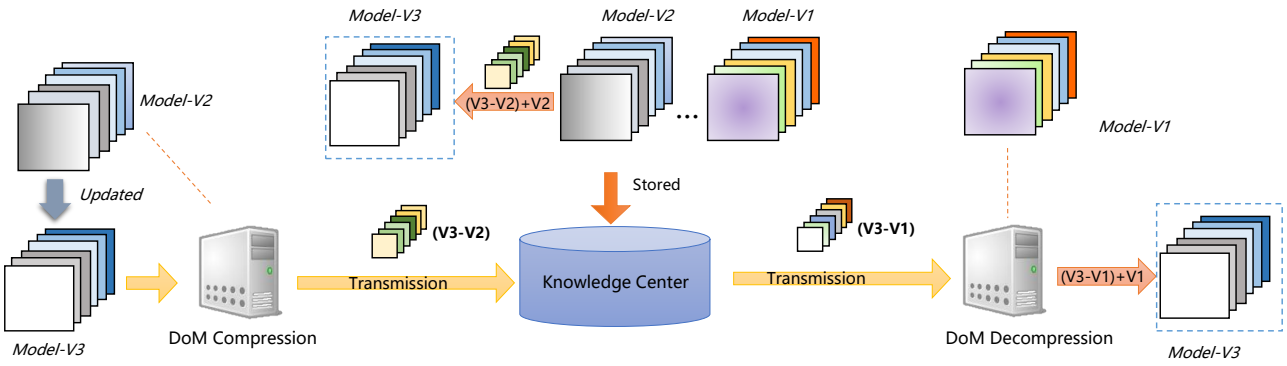


Figure 3: Illustration of multi-model compression with knowledge center. When model-V3 is transmitted towards the receiver side which already has model-V1, the concept of DoM is applied for better compression efficiency. In particular, wefi rst transmit the DoM between model-V2 and model-V3 to the knowledge center, and distribute the DoM between model-V1 and model-V3 to the receiver to recover model-V3.

performance [31]. In the receiver side, thefi nal parameters can be obtained by $w'_{receiver} = float32(w'/10^p)$.

As such, when the RP is large enough to ensure the decimal precision, the differences between the decoded parameters and original parameters $\Delta w = |w - w'_{receiver}| << w$ can be ignorable. Suppose the parameters lie on the convolutional layer, and Y and X are the input and output of the layer, respectively, the output can be calculated by $Y = w * X + b$. As a consequence, the loss in one layer can be calculated by $\Delta Y = \Delta w * X + b << Y$, which can also be ignored. Therefore, it provides a basic and generalized near-lossless solution to the single model compression, serving as the foundation of the multiple model prediction.

3.3.2 Multiple model compression. The knowledge communication further extends the scope of deep learning model compression from a single model to multiple models, especially when considering the frequent model communication scenarios. In particular, with the rapid advances on computational capability for knowledge creation in the intelligent front ends, thefl ow of the delicate deep models is more frequent than ever before. With the gradually generated data, the updated models will become more generalized and representative as more knowledge will be absorbed. As such, there is a strong desire to transmit these updated models. However, frequently transmitting these models without any inter prediction between different models may waste tremendous resources as strong redundancy exists between existing and the to-be-transmitted models. In other words, given the existing knowledge at the receiver side, the updated knowledge will become light weight such that only certain modifications need to be conveyed. Another commonly encountered scenario is that multiple models which may be trained for multiple tasks need to be simultaneously transmitted. However, due to the similarities between different tasks (e.g., object recognition and visual search), there may exist high correlations between the learned models, such that effective model prediction is also necessary to further improve the coding efficiency.

The straightforward solution of the multiple model compression is extracting the differences between the former and current deep learning models, and only the differences are signaled in the bitstream. However, if there is a mismatch between the existing model at the receiver end and the one used for prediction in the encoding process, error drifting will be introduced. Therefore, we introduce the concept of knowledge center, which stores multiple models in the cloud for deep learning model distribution. To exchange knowledge between the edge node and the center, the standardization is also required to ensure the interoperability.

Under such circumstance, the best prediction model in the center as well as the encoder side should befi rstly identified. An alternative strategy is that the prediction model can be generated by several commonly shared models with clustering techniques. Differences of Models (DoM) between the to-be-compressed model and the prediction model is extracted, compressed and transmitted to the center side. Given the DoM, the new model can be generated based on the existing prediction model in the center. Moreover, the newly generated model can be delivered to the receiver side using the identified prediction model in both the center and the receiver.

To encode the DoM parameters, we also investigate the parameter distribution, as shown in Figure 2 (b), from which we can observe that the DoM has a narrower range such that the coefficient energy is consequently greatly reduced. In this manner, significant bit rate savings can be achieved. Moreover, better single model compression method can be applied upon the DoM strategy as well and more adaptive representation and compression algorithms need to be developed in the future to better utilize the narrower parameter distribution.

In Figure 3, a motivating example is provided when transmitting model-V3 to the receiver side using DoM. In particular, as model-V2 remains in the sender side while model-V1 is in the receiver side, directly transmitting the DoM between model-V2 and model-V3 may cause significant errors. Our strategy is to transmit the DoM between model-V2 and model-V3 to the knowledge center,

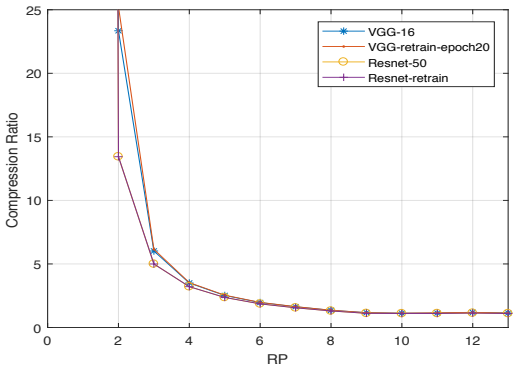


Figure 4: Relationship between RP and the compression ratio.

and distribute the DoM between model-V1 and model-V3 to the receiver, which can successfully avoid such problem.

3.4 Comparisons with Traditional Deep Learning Model Compression

Recently, the deep model compression has been widely studied in an effort to obtain the lightweight deep neural networks. Most of these works focus on removing potential redundancy in models for a specific task. Here, we aim to extend the scope of deep model compression and take the future transmission and communication into consideration, such that more attention has been paid towards the redundancy removal between different models with identical network structures. Moreover, from the perspective of the ultimate utility, not only thefi nal analysis performance is desired to be maintained, but also the knowledge conveyed in the deep learning model should be preserved in the compression process. In the future, the knowledge itself or the missing knowledge from one end to another can be carried with the presented concept of deep learning model compression. As such, the existing deep learning model compression methods, as well as newly motivated methods, should be thoroughly studied for the potential standardization of deep learning model compression.

Recently, MPEG is taking further steps towards the compact representation of deep neural networks [9], aiming at facilitating deep learning model transmission and communication with ensured interoperability. In particular, different use cases have been perceived, ranging from different transmission frequencies, bandwidth requirements as well as network models. Typical use cases include camera applications with object recognition, translation, public surveillance, etc. It is anticipated that the deep learning model compression and standardization can greatly benefit the future deep learning model communication, and play important roles in the establishment of the basic infrastructure of KCN.

4 VALIDATIONS

4.1 Experimental setup

We use image retrieval as the specific example to demonstrate the proposed strategies. Compared with classification and other general tasks with deep learning models, for image retrieval, slight changes on the deep neural networks will impact on the accuracy. Moreover, as image retrieval serves as the foundation for many visual analysis applications, here we investigate this specific task for performance validations. Landmark datasets such as *Holidays* [29], *Oxford5K* [38] and *Paris6K* [39] which have been widely accepted to evaluate the performances of image retrieval are used, and multiple deep learning models which have become the state-of-the-art feature extractors are adopted. In particular, we compare the performance with different RPs , to investigate the trade-off between compression ratio and the knowledge utility in terms of retrieval accuracy.

Table 1: The descriptions of the models and the corresponding retrieval accuracy in terms of mAP.

Model name	Holidays	Oxford5K	Paris6K
VGG-16	0.78066	0.51699	0.65799
VGG-retrain-epoch15	0.78192	0.68107	0.69101
VGG-retrain-epoch20	0.77720	0.69614	0.69929
Resnet-50	0.84925	0.44315	0.64789
Resnet-retrain	0.88957	0.687493	0.79061

For deep learning models, we choose pretrained *VGG-16* and *Resnet-50* deep neural network models as the base models along with differentfi ne-tuned versions. The descriptions are shown in Table 1. The performances without Principal Component Analysis (PCA) in terms of mean average precision (mAP) are illustrated. It is also worth noting that we retrain the *VGG-16* and *Resnet-50* models to obtain the three updated models, including *VGG-retrain-epoch15*, *VGG-retrain-epoch20* and *Resnet-retrain*. In particular, for *Resnet-retrain* we only retrain *Resnet-50* for the last four convolution blocks such that certain parts of the model remain the same. During the model training, we use 3D-Landmark dataset as the training dataset [40] and choose triplet loss [4, 21, 22] as the loss function which has been widely adopted in image retrieval tasks. Following the work in [40], similar sampling approach is applied and each training tuple contains 1 query, 1 positive and 5 negative images. We set triplet margin to 0.1, learning rate to 0.001, moment with 0.9 and epoch with 20.

4.2 Performance of Single Model Compression

First, we investigate the relationship between the RPs and the compression ratio, which is obtained by the ratio between the original model size and the compressed model size. The relationship is shown in Figure 4, from which we can observe that there is a monotonically decreasing relationship between the compression ratios and the RP. Moreover, the compression ratios increase dramatically when the RP is reduced to be lower than 3, as much less information needs to be encoded. It is also obvious that the relationship is not influenced by thefi ne-tuning process.

Subsequently, the relationships between mAP and RP in the adopted three datasets are shown in Figure 5, from which we can observe that the retrieval accuracy remains strictly constant when the RP is lower than 6. Besides, thefi rst changes in performances mostly appear in the RP of 5 when the mAP variation threshold is set to be 0.00001. The RPs and corresponding compression ratio of

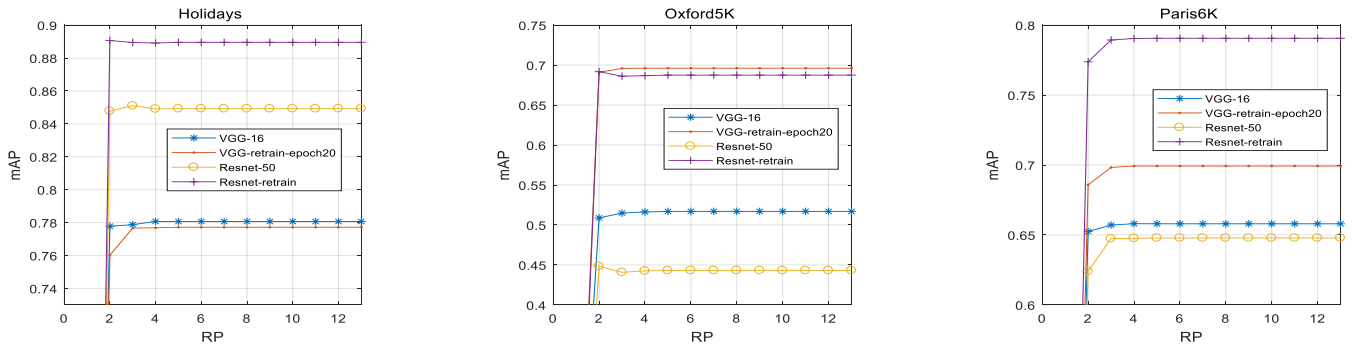


Figure 5: The relationships between the RP and the retrieval performance of three datasets (Holidays, Oxford5K, Paris6K) in terms of mAP.

Table 2: The corresponding compression ratio when the RP is 6, which is derived based on the threshold of performance variation.

Model	RP	Compression Ratio
VGG-16	6	1.9759
VGG-retrain-epoch20	6	1.9759
Resnet-50	6	1.8582
Resnet-retrain	6	1.8582

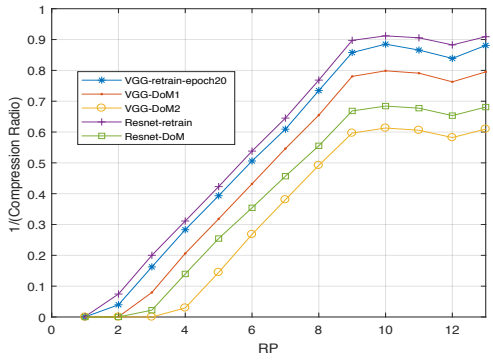


Figure 6: Relationship between the RP and compression ratio.

single model compression without retrieval accuracy degradation is illustrated in Table 2, from which it is obvious that significant compression performance can be even obtained with the naive compression strategy. Moreover, though significant compression ratio is achieved when the RP is lower than 3, from Figure 5 we can observe that the retrieval performance is also greatly degraded in this scenario.

4.3 Multiple Deep Model Compression

In this subsection, we investigate the performance of multiple model compression based on the concept of DoM. In particular, different versions of the same model structure are involved in the comparison, and three different scenarios are considered.

First, we compress the *DoM* between the original *VGG-16* and *VGG-retrain-epoch20*, which is denoted by *VGG-DoM1*. Secondly, the DoM between *VGG-retrain-epoch15* and *VGG-retrain-epoch20*, which is denoted by *VGG-DoM2*, is further compressed based on the assumptions that the models are frequently updated within a short period of time. Thirdly, we also investigate the model compression when partial parameters of the models are updated. The DoM between *Resnet-50* and *Resnet-retrain* is denoted by *Resnet-DoM*. In Figure 6, the compression ratios in terms of different RPs are plotted. Compared to the original single model compression strategy, the DoM based scheme can achieve significantly better performance with higher compression ratios under the same RP requirement.

In Table 3, the compression performance and retrieval accuracy for different RPs are provided to compare the performance of the updated model and DoM compression strategy. As there exist significant differences between *VGG-retrain-epoch20* and the *VGG-16*, the *VGG-DOM1* cannot achieve significant compression performance improvement compared to *VGG-DoM2*. However, the *VGG-DoM1* still achieves much better compression performance compared to the direct model compression strategy when the retrieval accuracy is very close. Moreover, *VGG-DoM-2* achieves more than twice of the compression ratios when the RP is 4. It is also very interesting to see that less performance drop is observed for *VGG-DoM-2* with more than 10 times compression ratio achieved compared to directly transmitting *VGG-retrain-epoch20*. When the additional information (or DoM) between VGG-retrain-epoch15 and *VGG-retrain-epoch20* is barely preserved, extremely high compression ratio (28740) with significant performance loss is observed.

It is also intuitive that when the model is partially updated, less updated information is required to convey. One specific example is the *Resnet-DoM* in Table 3, from which it is obvious that though significant performance difference lies between the original Resnet and Resnet-retrained, as illustrated in Table 1, *Resnet-DoM* achieves

Table 3: Performance comparisons in terms of compression ratios and retrieval accuracy for different RPs. The ΔmAP is obtained based on the maximal differences between the mAP with the corresponding RP and the mAP when the RP is 6 of the target model (*VGG-retrain-epoch20* and *Resnet-retrain*) for the three datasets (Holidays, Oxford5K and Paris6k).

Model	ΔmAP RP=6	Compression Ratio	ΔmAP RP=5	Compression Ratio	ΔmAP RP=4	Compression Ratio	ΔmAP RP=3	Compression Ratio
VGG-retrain-epoch20	0.00000	1.9759	-0.00004	2.5419	-0.00033	3.5311	-0.00100	6.1535
VGG-DoM1	0.00000	2.3156	-0.00004	3.1417	-0.00004	4.8506	-0.00174	12.5584
VGG-DoM2	0.00000	3.7322	0.00000	6.9128	-0.00018	33.9115	-0.01214	28740
Resnet-retrain	0.00000	1.8582	-0.0002	2.3645	-0.00069	3.2119	-0.00135	4.9974
Resnet-DoM	0.00000	2.8239	0.00000	3.9300	0.00000	7.1607	-0.00157	45.2715

Table 4: Comparisons between the updated model and the DoM. The RP is obtained when the retrieval performance is not varied.

Model	RP	Compression Ratio
VGG-retrain-epoch20	6	1.9759
VGG-DoM1	6	2.3156
VGG-DoM2	5	6.9128
Resnet-retrain	6	1.8582
Resnet-DoM	4	7.1607

Table 5: Comparisons of DoM compression efficiency between the corrupted prediction and the original prediction (RP=6). The ΔmAP is obtained as the maximal differences between the mAP with RP equaling to 4 and the mAP when the RP equals to 6 of the target model (*Resnet-retrain*) for the three datasets (Holidays, Oxford5K and Paris6k).

Model	ΔmAP	Compression Ratio
Resnet-retrain	-0.00069	3.2119
Resnet-DoM-U	-0.00051	6.4760
Resnet-DoM	0.00000	7.1607

much higher compression ratio with similar performance. In particular, the RP and corresponding compression ratios comparisons are shown in Figure 4, and less RP is needed when no performance degradation is observed for the DoM strategy.

Moreover, we investigate the scenario when the prediction model for the DoM calculation is corrupted. Here, we take the partially updated model (*Resnet-retrain*) as an example. Supposing that the reconstructed *Resnet-50* with RP equaling to 3 is used as the prediction model, such that to faithfully obtain the updated model with higher RP based on the DoM strategy, more bits are required to represent the information of differences. Here, assume that the DoM between the *Resnet-50* and *Resnet-retrain* with RP equaling to 3 is denoted to be *Resnet-DoM-U*. In *Resnet-DoM-U*, the missing information in the unupdated layer should also be transmitted due to the corrupted prediction. The comparison results whenfi xing the RP to be 4 are shown in Table 5, from which it is obvious that the performance of compression ratio of *Resnet-DoM-U* lies between *Resnet-DoM* and *Resnet-retrain* when the retrieval accuracy remains to be very close. Moreover, it is also observed that the compression ratio of *Resnet-DoM-U* is closer to *Resnet-DoM*, as only a minor proportion of the original model is required to be transmitted. The results further provide us the evidence that even if the prediction model contains significant distortion, the DoM strategy may still be effective as the knowledge is preserved in the prediction model. As such, the corrupted prediction model should be examined for potential utilization before it is discarded.

5 DISCUSSIONS

In this paper, we have discussed the traditional approaches for deep model compression, and have enumerated its limitations in the context of KCN. The application scope and methodology of deep learning model compression have been extended as an alternative motivating example for knowledge transmission. To demonstrate the proposed concept, both single and multiple model compression methods are validated through extensive experiments. In the future, it is anticipated that the deep model compression and the relevant standards can play important roles in the communication of knowledge, and bring further impacts to the new KCN infrastructure.

There are also a number of issues that are worth investigation with regard to the deep model compression and standardization. First, the optimal compression strategy for single model needs to be further investigated based on the principle of knowledge preservation. Here, we adopt the very basic approach to investigate the single deep model compression. Moreover, how to develop the corresponding quantization and entropy coding methods based on the parameter statistics and knowledge preservation task need to be intensively studied. Second, more advanced methods are required to be developed for efficient model prediction in multiple model compression. In particular, how to select or generate the best prediction model should be particularly paid attention to, as it may greatly influence the DoM energy and coding efficiency. Finally, there are still several open issues need to be addressed regarding the standardization of deep model compression, especially in the area when the deep learning algorithms evolve rapidly.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (61661146005, U1611461, 61390515) and the National Key Research and Development Program of China (No. 2016YFB1001501), and in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation.

REFERENCES

[1] ISO/IEC JTC 1. 2000. Information Technology : Generic Coding of Moving Pictures and Associated Audio Information : Video. *ITU-T Rec.H262* (2000).

[2] ISO/IEC JTC 1. 2003. Advanced video coding for generic audiovisual services. *Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC) JTC 1, Rec. H.264 and ISO/IEC 14 496-10 (MPEG-4) AVC, 2003* (2003).

[3] Alireza Aghasi, Afshin Abdi, Nam Nguyen, and Justin Romberg. 2017. Net-Trim: Convex Pruning of Deep Neural Networks with Performance Guarantee. In *Advances in Neural Information Processing Systems*. 3180–3189.

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5297–5307.

[5] Luigi Atzori, Antonio Iera, and Giacomo Morabito. 2010. The internet of things: A survey. *Computer networks* 54, 15 (2010), 2787–2805.

[6] AVS2/IEEE. 2014. urlhttp://www.ieee1857.org/1857.4.asp. (2014).

[7] Artem Babenko and Victor Lempitsky. 2015. Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision*. 1269–1277.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[9] Werner Bailer. 2018. Use cases and requirements for coded representation of neural networks. ISO/IEC JTC1/SC29/WG11/N17338, Gwangju, Koera.

[10] Jianshu Chao and Eckehard Steinbach. 2016. Keypoint encoding for improved feature extraction from compressed video at low bitrates. *IEEE Transactions on Multimedia* 18, 1 (2016), 25–39.

[11] Xin Dong, Shangyu Chen, and Sinno Pan. 2017. Learning to prune deep neural networks via layer-wise optimal brain surgeon. In *Advances in Neural Information Processing Systems*. 4860–4874.

[12] Xin Dong, Shangyu Chen, and Sinno Pan. 2017. Learning to prune deep neural networks via layer-wise optimal brain surgeon. In *Advances in Neural Information Processing Systems*. 4860–4874.

[13] Lingyu Duan, Yihang Lou, Shiqi Wang, Wen Gao, and Yong Rui. 2017. AI Oriented Large-Scale Video Management for Smart City: Technologies, Standards and Beyond. *arXiv preprint arXiv:1712.01432* (2017).

[14] Ling-Yu Duan, Vijay Chandrasekhar, Jie Chen, Jie Lin, Zhe Wang, Tiejun Huang, Bernd Girod, and Wen Gao. 2016. Overview of the MPEG-CDVS Standard. *IEEE Transactions on Image Processing* 25, 1 (2016), 179–194.

[15] Ling-Yu Duan, Vijay Chandrasekhar, Shiqi Wang, Yihang Lou, Jie Lin, Yan Bai, Tiejun Huang, Alex Chichung Kot, and Wen Gao. 2017. Compact Descriptors for Video Analysis: the Emerging MPEG Standard. *arXiv preprint arXiv:1704.08141* (2017).

[16] Emil Eriksson, György Dán, and Viktoria Fodor. 2016. Predictive distributed visual analysis for video in wireless sensor networks. *IEEE Transactions on Mobile Computing* 15, 7 (2016), 1743–1756.

[17] Emil Eriksson, György Dán, and Viktoria Fodor. 2017. Coordinating Distributed Algorithms for Feature Extraction Offloading in Multi-Camera Visual Sensor Networks. *IEEE Transactions on Circuits and Systems for Video Technology* (2017).

[18] Bernd Girod, Vijay Chandrasekhar, David M. Chen, and Ngai Man Cheung. 2011. Mobile Visual Search. *IEEE Signal Processing Magazine* 28, 4 (2011), 61–76.

[19] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. 2014. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115* (2014).

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[21] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*. Springer, 241–257.

[22] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. 2017. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision* 124, 2 (2017), 237–254.

[23] Yiwen Guo, Anbang Yao, and Yurong Chen. 2016. Dynamic network surgery for efficient dnns. In *Advances In Neural Information Processing Systems*. 1379–1387.

[24] 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).

[25] 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.

[26] Yihui He, Xiangyu Zhang, and Jian Sun. 2017. Channel pruning for accelerating very deep neural networks. In *International Conference on Computer Vision*.

[27] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Vol. 1. 3.

[28] Tiejun Huang. 2014. Surveillance video: The biggest big data. *Computing Now* 7, 2 (2014), 82–91.

[29] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2008. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*. Springer, 304–317.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[31] E Jebamalar Leavline and D Asir Antony Gnana Singh. 2013. Hardware implementation of LZMA data compression algorithm. *International Journal of Applied Information Systems (IJAIS)* 5, 4 (2013), 51–56.

[32] Cong Leng, Hao Li, Shenghuo Zhu, and Rong Jin. 2017. Extremely low bit neural network: Squeeze the last bit out with admm. *arXiv preprint arXiv:1707.09870* (2017).

[33] Ji Lin, Yongming Rao, Jiwen Lu, and Jie Zhou. 2017. Runtime Neural Pruning. In *Advances in Neural Information Processing Systems*. 2178–2188.

[34] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. 2017. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. *arXiv preprint arXiv:1712.01887* (2017).

[35] Ingrid Lunden. 2017. (2017). https://techcrunch.com/2017/06/08/cisco-ip-traffic-shoots-up-to-3-zettabytes-by-2021-video-will-be-80-of-it/

[36] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. 2017. Thinet: Afi lter level pruning method for deep neural network compression. *arXiv preprint arXiv:1707.06342* (2017).

[37] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, Xiaoou Tang, et al. 2016. Face Model Compression by Distilling Knowledge from Neurons.. In *AAAI*. 3560–3566.

[38] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007*. IEEE.

[39] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition, 2008*. IEEE, 1–8.

[40] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2016. CNN image retrieval learns from BoW: Unsupervisedfi ne-tuning with hard examples. In *European Conference on Computer Vision*. Springer, 3–20.

[41] N Ranganathan and Selwyn Henriques. 1993. High-speed VLSI designs for Lempel-Ziv-based data compression. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing* 40, 2 (1993), 96–106.

[42] Brandon Reagen, Udit Gupta, Robert Adolf, Michael M Mitzenmacher, Alexander M Rush, Gu-Yeon Wei, and David Brooks. 2017. Weightless: Lossy Weight Encoding For Deep Neural Network Compression. *arXiv preprint arXiv:1711.04686* (2017).

[43] Alessandro Redondi, Luca Baroffio, Lucio Bianchi, Matteo Cesana, and Marco Tagliasacchi. 2016. Compress-then-Analyze vs Analyze-then-Compress: what is best in Visual Sensor Networks? *IEEE Transactions on Mobile Computing* 15, 12 (2016), 3000–3013.

[44] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[45] Gary J Sullivan, Jens Ohm, Woo-Jin Han, and Thomas Wiegand. 2012. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology* 22, 12 (2012), 1649–1668.

[46] Handle System. 2017. http://www.handle.net/. (2017).

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.

[48] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2015. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879* (2015).

[49] Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Beyondfilters: Compact feature map for portable deep model. In *International Conference on Machine Learning*. 3703–3711.

[50] Dapeng Wu, Zhenjiang Li, Jianping Wang, Yuanqing Zheng, Mo Li, and Qiuyuan Huang. 2017. Vision and Challenges for Knowledge Centric Networking (KCN). *arXiv preprint arXiv:1707.00805* (2017).

[51] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[52] Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. 2017. On compressing deep models by low rank and sparse decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7370–7379.

[53] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[54] Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. 2017. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044* (2017).

[55] Hao Zhou, Jose M Alvarez, and Fatih Porikli. 2016. Less is more: Towards compact cnns. In *European Conference on Computer Vision*. Springer, 662–677.

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* (2017).