Joint Architecture and Knowledge Distillation in Convolutional Neural Network for Offline Handwritten Chinese Text Recognition

Zi-Rui Wang, Jun Du

Abstract

The technique of distillation helps transform cumbersome neural network into compact network so that the model can be deployed on alternative hardware devices. The main advantages of distillation based approaches include simple training process, supported by most off-the-shelf deep learning softwares and no special requirement of hardwares. In this paper, we propose a guideline to distill the architecture and knowledge of pre-trained standard CNNs simultaneously. We first make a quantitative analysis of the baseline network, including computational cost and storage overhead in different components. And then, according to the analysis results, optional strategies can be adopted to the compression of fully-connected layers. For vanilla convolution layers, the proposed parsimonious convolution (ParConv) block only consisting of depthwise separable convolution and pointwise convolution is used as a direct replacement without other adjustments such as the widths and depths in the network. Finally, the knowledge distillation with multiple losses is adopted to improve performance of the compact CNN. The proposed algorithm is first verified on offline handwritten Chinese text recognition (HCTR) where the CNNs are characterized by tens of thousands of output nodes and trained by hundreds of millions of training samples. Compared with the CNN in the state-of-the-art system, our proposed joint architecture and knowledge distillation can reduce the computational cost by $>10\times$ and model size by $>8\times$ with negligible accuracy loss. And then, by conducting experiments on one of the most popular data sets: MNIST, we demonstrate the proposed approach can also be successfully applied on mainstream backbone networks.

Index Terms

Zi-Rui Wang, and Jun Du are with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, P. R. China (e-mail: cs211@mail.ustc.edu.cn; jundu@ustc.edu.cn).
Convolutional neural network, acceleration and compression, architecture and knowledge distillation, offline handwritten Chinese text recognition.

I. INTRODUCTION

Convolutional neural networks (CNNs) play an important role in the new wave of artificial intelligence. Since the first-generation CNNs were proposed by LeCun [1], [2] for handwritten character recognition, numerous CNNs have been springing up in different applications, such as (Alex, VGG, GoogLe, Res, Dense)-Nets in natural image recognition [3], [4], [5], [6], [7], DCNN in offline handwritten Chinese text recognition (HCTR) [8], [9], HCCR-CNN in handwritten Chinese character recognition (HCCR) [10], [11], Facenet in face recognition [12] and FCN in speech emotion recognition [13]. Such CNNs share the same basic components, i.e., convolutional layer, pooling layer and fully-connected layer. Although these networks have improved performance dramatically and CNN-based approaches have become mainstream in a wide range of pattern recognition tasks, the trend going deeper and wider for CNN makes them difficult to be deployed on resource-limited devices e.g., mobile phone and embedded chips. Moreover, there is an evident fact that the current state-of-the-art CNNs still mainly depend on massive trial-and-error experiments by handcraft. Both the architecture and the internal knowledge of a CNN should be treasured information for acceleration and compression algorithms. Based on this observation, to simultaneously distill the architecture and the knowledge of CNN into a compact network is the topic of this paper.

The concept of knowledge distillation, which could be traced back to Caruana’s research in 2006 [14], is to transfer the knowledge from cumbersome models into smaller model. The first related work in neural network is completed by Hinton et al. in [15], i.e., knowledge distillation. The authors adopted the soft labels from multiple neural networks to guide the training of a single network. Soon after, Romero et al. [16] improved the algorithm of knowledge distillation by using the outputs of hidden layers and the soft labels from a shallow network with more parameters as hints to instruct a thin deep network. Recently, the authors in [17] inspired by human visual experience, tried to transfer the defined attention map in convolutional layers from one network to another. In [18], the authors defined the flow of solution procedure (FSP) matrix to measure the change of information between two different layers for compressed network to imitate the middle produces of baseline network. More recently, Liu et al. [19] integrated pixel-wise loss, pair-wise loss and generative adversarial loss in knowledge distillation for semantic segmentation. Almost
at the same time, under the framework of knowledge distillation, He et al. [20] extracted more compact middle features by using a pre-trained autoencoder and proposed affinity distillation module to capture the long-range dependency. Both of them utilize the relationship between pixels, which is important to semantic segmentation.

Different from knowledge distillation, the researches of architecture distillation in CNN focus on inventing new efficient convolutions or units to directly replace standard convolutions of baseline CNN. A representative work has been conducted in [21] where the authors reconstructed a lightweight CNN by using multiple efficient compact blocks according to the different locations of baseline CNN. Recent influential works about architecture distillation can be found in [22], [23]. They share a similar concept that the feature maps in standard convolution layer are redundant and can be split into different parts. The authors in [22] proposed heterogeneous convolution (HetConv) with different kernel sizes in each layer to handle the corresponding parts of input feature maps while the feature maps in [23] are factorized into high frequency with fine spatial resolution and low frequency with smaller spatial size.

The realization of distillation can be figuratively described as teacher-student learning in which the network with massive parameters and high performance plays as a teacher and the compressed network is a student. Both the architecture and the internal knowledge of teacher network should be learned by student network. The single consideration usually leads a contradiction between optimal performance and satisfactory compression. Unlike previous distillation algorithms, in this paper, we propose a guideline to simultaneously distill the architecture and knowledge of a pre-trained CNN. Specifically, instead of using multiple acceleration blocks [21], we develop a uniform block named as parsimonious convolution (ParConv) block that only consists of depthwise separable convolution (DSConv) [24] and pointwise convolution in a heterogeneous combination way. In distillation knowledge, a new solving procedure loss (SPL) is added to further improve the performance of student network. The solving procedure is represented by the differences of attention maps between two layers.

The effectiveness of the proposed algorithm is first demonstrated on offline handwritten Chinese text recognition (HCTR). The application has a wide use, such as mail address recognition [25], bank check [26] and document recognition [27]. Although many progresses have been made owing to deep learning [28], it remains a challenging problem due to the following reasons: 1) The text line must be considered as a whole rather than isolated characters; 2) More than 7000 classes in common Chinese vocabulary and large-scale training samples; 3) The unconstrained
writing condition. The involving experiments are conducted on the ICDAR 2013 competition task of CASIA-HWDB database [29], [30] that is one of the most popular benchmark databases. To our best of knowledge, no acceleration and compression approaches in CNN have been validated on offline HCTR. Furthermore, in order to display the generalization ability of the proposed algorithm, we use it to reduce computational costs of the mainstream backbone network on MNIST [2].

The main contributions of this study can be summarized as follows:

- We propose a guideline to distill the architecture and knowledge of pre-trained standard CNNs simultaneously for fast deployability on alternative hardware devices.

- In architecture distillation, we invent a parsimonious convolution block (ParConv) to directly replace vanilla convolution without other adjustments. Compared with LightweightNet [21] and other popular plug-and-play compact convolutional units (DSConv [24], HetConv [22]), the proposed ParConv demonstrates its superiority in recognition performance, computational cost and storage overhead.

- In knowledge distillation, a new solving process loss (SPL) is added to further improve the performance of compressed CNN.

- The effectiveness of the proposed approach is first verified on offline HCTR. No study has investigated whether previous approaches are still feasible in this field.

- Compared with the baseline CNN in HCTR, our proposed joint architecture and knowledge distillation can reduce the computational cost by >10× and model size by >8× with negligible accuracy loss. Applying the algorithm to the mainstream backbone network ResNet18 [6] on MNIST, we can obtain >9× compression rate for both model size and computational cost with almost no accuracy loss.

The remainder of this paper is organized as follows. Section II reviews related work. Section III elaborates on the details of the proposed approach. Section IV reports the experimental results and analyses. Finally, Section V concludes.

II. RELATED WORK

A. Offline HCTR

Offline HCTR can be formulated as the Bayesian decision problem:

$$\hat{C} = \arg \max_C p(C|X)$$  \hspace{1cm} (1)
where $X$ is the feature sequence of a given text line image and $C = \{C_1, C_2, ..., C_n\}$ is the underlying $n$-character sequence. The research efforts for addressing such sequence modelling task can be divided into the three categories: oversegmentation \cite{31}, \cite{32}, \cite{33}, connectionist temporal classification (CTC) \cite{34}, \cite{35}, hidden Markov model (HMM) \cite{36}, \cite{8}, \cite{9}. Almost all of these approaches benefit from the recent progresses of deep learning \cite{28}. The outputs of neural networks in different modelling ways correspond to different concepts. For example, in oversegmentation and CTC-based approaches, the outputs of neural network are related to segmentation identification or character classes while the outputs of network are posterior probabilities of states in HMM-based approaches. In our recent work \cite{9}, each character is modelled by three tied states on average and the deep CNN (DCNN) with 22080 ($7360 \times 3$) output nodes is adopted as character model and trained by hundreds of millions of frame-level images. As shown in Fig. 1, frame-level images are extracted from original images by a left-to-right sliding window and fed into the DCNN. Then, the posterior probabilities of states are utilized in a WFST-based decoder \cite{37} with/without language model (LM) for the final recognition results. In order to fit such a massive training samples, the parameters of DCNN have been up to 124.5 MB and $16.02 \times 10^8$ FLOPs are needed in each inference. More details and analyses of the DCNN are shown in Section IV.
B. Acceleration and Compression

Except for distillation based approaches, many other acceleration and compression algorithms have also been widely investigated such as low-rank decomposition [38], [39], [40], [41], [42], parameters pruning [43], [44], [10], [45], [46], [47], parameters quantization [48], [49], [50], [51], [52], [53] and compact network design [54], [55], [56], [57], [58].

As one of the first attempts for low-rank decompositions of filters, Denton et al. [38] proposed several decomposition designs along different dimensions. In [39], the \( k \times k \) filters are decomposed into \( k \times 1 \) and \( 1 \times k \) filters. A representative work comes from [40], where the authors take the nonlinear units into account in their decomposition algorithm based on the assumption that the filter response lies on a low-rank subspace. The authors in [41], [42] use Tucker decomposition to achieve compression. Such algorithms need to be conducted layer by layer. Once a layer has been decomposed, the whole network is retrained again by back-propagation (BP) algorithm. For large-scale tasks, the repeated decomposition and training is usually time consuming.

The parameters pruning is based on a reasonable intuition that the low weights in a neural network are not important so that they can be safely removed. In [43], [44], the weights were kept or removed by comparing with a fixed threshold. The authors in [10] proposed the adaptive drop-weight (ADW) to dynamically increase the threshold. In [45], Liu et al. exploited channel sparsity regularization to prune channels with small scaling factors. Almost at the same time, the authors in [46] pruned filters based on the reconstruction error of the corresponding next layer by using a greedy algorithm. The fine-grained pruning [43], [10] requires special software/hardware accelerator. Although the channel-level pruning [45], [46] can be directly applied to the existing software platforms, like the low-rank decomposition based algorithms, the requirement of repeated pruning and fine-tuning is time consuming for large-scale tasks. Besides, from a recent research [59], the structure of neural network seems more important than the values of weights. Compared with layer-wise pruning, a global pruning strategy [47] might be more valuable.

For parameters quantization, Han et al. [48] divided the weights of network into different groups by hash algorithm and made the weights in the same group share a value. Vanhoucke et al. [49] used 8-bit type instead of the common 32-bit floating type in network. Courbariaux et al. [50] proposed a binarized neural network in which all weights and outputs are constrained to \( \{1, -1\} \) while the authors in [51] quantized weights into \( \{-1, 0, 1\} \). Such methods can save a large
amount of resources. However, the approach in [48] takes up extra space to store the original positions for shared weights and low-bit approximation usually degrades network performance. Different from constraining the weights to +1 or -1, the authors in [52], [53] replaced the multiply-accumulate operation to one shift or a constrained number of shifts and adds, which can make trade-offs between accuracy and computational consuming.

An efficient and effective network structure can save a lot of memory, computational cost and yield a competitive performance. Many compact blocks have been invented to control the fast increasing of network parameters, such as Fire module in SqueezeNet [54], Inception module in GoogleNet [60]. The basic unit in these networks still consists of canonical convolution. As one low consumption (storage and computational cost) substitute, the depthwise separable convolution (DSConv) is first introduced in [61], [24] and has become a key building block in recent compact networks [55], [56], [57], [58]. Besides, the authors in [62] prove that DSConv is principal components of standard convolution and can approximate the standard convolution in closed form. Although the DSConv is far more efficient than standard convolution, consistent observations can be found in [24], [21]: Simple replacement by using DSconv is not effective. The authors in [24] scale up depthwise separable filters so that the DSConv based network Xception with the same number of parameters as Inception V3 [63] can outperform the Inception V3.

### III. ARCHITECTURE AND KNOWLEDGE DISTILLATION

Given a baseline CNN($W_{fc}, W_{con}$), let $W_{fc}$ represents the weight set of fully-connected layers and $W_{con}$ corresponds to the weight set of convolutional layers. $\ell$ denotes the number of FLOPs or the storage. For fully-connected layers in CNN, only the storage needs to be considered due to the relatively small amount of computational cost. We use the ratio $\gamma(W_{fc}, W_{con}) = \frac{\ell(W_{fc})}{\ell(W_{fc}) + \ell(W_{con})}$ to measure whether a strategy $\pi$ is conducted on the weights of fully-connected layers or not.

As summarized in Algorithm 1, the guideline involves architecture distillation and knowledge distillation. We first analyze the computational cost and storage overhead of baseline CNN($W_{fc}, W_{con}$) and compute the corresponding $\gamma(W_{fc}, W_{con})$. And then, in architecture distillation, if the weights of fully-connected layers occupy non-ignorable consumption of a certain computing resource (i.e., storage), we find a strategy $\pi(W_{fc})$ to construct a new CNN($\pi(W_{fc}), W_{con}$) and make sure $\gamma(\pi(W_{fc}), W_{con}) \leq T$ with the neglected performance loss (even better). In most cases, it’s easy to find such a strategy for fully-connected layers, e.g. global pooling [54],
low-rank decomposition [64], low-dimensional feature [21]. Because we mainly focus on the compression of convolutional layers in this study, a naive solution \( \pi \) that depends on the number of active output targets [64] to find an appropriate bottleneck feature before output layer is adopted. For convolutional layers, the proposed ParConv blocks are used as a direct replacement to build a compact CNN (CCNN). Finally, in order to maintain the performance of CCNN, the knowledge distillation with three kinds of losses, namely, the Kullback-Leibler (KL) divergence loss, the cross entropy (CE) loss and solving process (SP) loss, is adopted to transfer knowledge from standard CNN into the ParConv based CCNN. Fig. 2 illustrates the proposed algorithm. More details of respective parts will be described in the following subsections.

**Algorithm 1** The guideline of joint architecture and knowledge distillation.

**Require:**
- Baseline CNN\((W_{fc}, W_{con})\).
- Threshold \(T\).

1. Analyze the computational cost and storage overhead in baseline CNN\((W_{fc}, W_{con})\) and compute \(\gamma(W_{fc}, W_{con})\).
2. if \(\gamma(W_{fc}, W_{con}) > T\) then
   3. Find a strategy \(\pi(W_{fc})\) to construct a new CNN\((\pi(W_{fc}), W_{con})\) with \(\gamma(\pi(W_{fc}), W_{con}) \leq T\) and neglected performance loss (even better).
   4. Build a compact CNN (CCNN) by using ParConv blocks to replace the convolutional layers in CNN\((\pi(W_{fc}), W_{con})\).
   5. Distill the knowledge of CNN\((\pi(W_{fc}), W_{con})\) into the CCNN.
3. else
   7. Build a CCNN by using ParConv blocks to replace the convolutional layers in CNN\((W_{fc}, W_{con})\).
   8. Distill the knowledge of CNN\((W_{fc}, W_{con})\) into the CCNN.
4. end if
5. return The CCNN

A. Bottleneck Feature

As shown in Fig. 3, if there are \(M\)-dimensional feature from the last conv layer, \(B\)-dimensional bottleneck feature and \(O\) output nodes, the total computational costs (FLOPs) of fully-connected
Fig. 2. The simplified framework of joint architecture and knowledge distillation.

Fig. 3. Using bottleneck layer to reduce the parameters of fully-connected layers.
layers (FCs) is computed as:

\[ FL_{\text{FCs}} = M \times B + B \times O \]  

(2)

For \( O \gg M \), \( FL_{\text{FCs}} \) approximately equals to \( B \times O \), which means the compressional ratio can be controlled by adjusting the dimension of bottleneck feature.

### B. Parsimonious Convolution

In a standard convolutional layer, assuming the input is a square feature map, it can be represented by a three dimensional tensor of size \( D \times D \times C_{\text{in}} \). Here \( D \) is the spatial width and height while \( C_{\text{in}} \) is the number of the input channels. Usually, the corresponding output tensor with the channels \( C_{\text{out}} \) obtained by applying the \( C_{\text{out}} \) filters of size \( K \times K \times C_{\text{in}} \) has the same spatial size \( D \times D \), namely, the size of output is \( D \times D \times C_{\text{out}} \). Therefore the FLOPs at this layer is:

\[ FL_{\text{Conv}} = D^2 \times C_{\text{in}} \times C_{\text{out}} \times K^2 \]  

(3)

The depthwise separable convolution (DSConv) is made up of two components: channel-wise convolution and point-wise convolution. The fundamental hypothesis behind DSConv is that cross-channel correlations and spatial correlations can be decoupled. The channel-wise convolution is used to capture spatial correlations and the point-wise convolution is a \( 1 \times 1 \) standard convolution to combine information from different channels. In channel-wise convolution, each output channel is only associated to one input channel so that the convolutional filters are represented by a 3-D tensor \( K \times K \times C_{\text{in}} \). The FLOPs of DSConv is computed as follows:

\[ FL_{\text{DSConv}} = D^2 \times (C_{\text{in}} \times K^2 + C_{\text{in}} \times C_{\text{out}}) \]  

(4)

Although compared with standard convolution, the DSConv is extremely efficient and is a key building unit for many compact networks [55], [56], [57], [58], it is more likely to rely on residual connection due to the potential increasing of the depth in network. Besides, simple replacement by using DSConv in standard CNN makes the network suffer performance degradation, which may be the reason that the authors in [24] have to scale up the filters in DSConv.

Based on the opinion that the filter can cover the spatial correlation from a part of the input channels, heterogenous convolution (HetConv) is proposed in [22]. Essentially, the input
channels in HetConv are split into two branches, one with $\alpha C_{in}$ ($0 \leq \alpha \leq 1$) channels for standard convolution e.g., $3 \times 3$ kernel size, the another with $(1 - \alpha)C_{in}$ channels for pointwise convolution. A similar idea can be found in Octave convolution [23]. Such a kind of convolution can sufficiently utilize the advantage of parallelization without increasing a latency in system like DSConv. However, a small $\alpha$ corresponds to a considerable compression ratio, which easily leads to a large performance decline. The corresponding FLOPs is computed as follows:

$$FL_{HetConv} = D^2 \times \alpha C_{in} \times C_{out} \times K^2 + D^2 \times (1 - \alpha)C_{in} \times C_{out}$$

(5)

Inspired by the HetConv, the proposed parsimonious convolution (ParConv) adopts DSConv to approximate standard convolution in the branch with $\alpha C_{in}$ channels. Specially, before DSConv, a pointwise convolution with a channel multiplier $\omega$ is added to deeply integrate the information among channels, which is important for DSConv to extract features. In order to promote the flow of information between branches, a channel shuffle operator is conducted before the input feature maps are split into two branches. For simplicity, the $\alpha$ is set to 0.5 in all ParConvs. The FLOPs of ParConv is:
TABLE I

THE FLOPs RATIOS OF COMPACT CONVOLUTIONS TO STANDARD CONVOLUTION

| Type      | FLOPs Ratio                     |
|-----------|---------------------------------|
| DSConv    | $\frac{1}{C_{out}} + \frac{1}{K^2}$ |
| HetConv   | $\alpha + \frac{1-\alpha}{K^2}$ |
| ParConv   | $\frac{1}{2K^2} + \frac{\omega}{2}(\frac{1}{K^2} + \frac{1}{C_{out}} + \frac{C_{in}}{2K^2})$ |

$$FL_{ParConv} = D^2 \times \frac{1}{2}C_{in} \times \frac{\omega}{2}C_{in} + D^2 \times \frac{\omega}{2}C_{in} \times K^2$$

$$+ D^2 \times \frac{\omega}{2}C_{in} \times C_{out} + D^2 \times \frac{1}{2}C_{in} \times C_{out}$$

(6)

Fig. 4 shows the different types of convolutions and Table I lists the FLOPs ratios of these compact convolutions to standard convolution. From Table I it can be observed clearly that the DSConv can approach $K^2$ times less computation than standard convolution while the reduction of computation in HetConv is controlled by the coefficient $\alpha$. If $\alpha = 1$, the HetConv degenerates into standard convolution. For ParConv, under the reasonable assumption that $C_{in} = C_{out}$ and $C_{out} \gg K^2$, the FLOPs ratio to standard convolution can be rewritten as follows:

$$\frac{1}{2K^2} + \frac{3\omega}{4K^2}$$

(7)

Obviously, it can be free to adjust the computational cost by changing the value of channel multiplier $\omega$.

C. Knowledge Distillation with Multiple Losses

In order to repair the performance gap between the standard CNN and the corresponding ParConv based compact CNN (CCNN), the knowledge distillation is necessary. Three kinds of training losses are included in the process of knowledge distillation, i.e., Kullback-Leibler (KL) divergence loss, cross entropy (CE) loss and solving procedure (SP) loss. The final loss is formulated as the weighted sum of these losses:

$$l = \mu l_{KL} + \beta l_{CE} + \lambda l_{SP}$$

(8)
The CE loss with so-called hard labels is the most common training criterion in classification tasks and is defined as follows:

\[ l_{CE} = - \sum_t \log p(s_t|x_t) \]  \hspace{1cm} (9)

where \( \log p(s_t|x_t) \) is the estimated posterior probability of the target class \( s_t \) from CNN output given the input \( x_t \).

The KL divergence is a measure of how one probability distribution is different from another one. In our approach, it is used to compute the difference between the output distribution of standard CNN \( p_S(s|x_t) \) and the corresponding distribution from CCNN \( p_C(s|x_t) \):

\[ l_{KL} = \sum_t \sum_s p_S(s|x_t) \log \frac{p_S(s|x_t)}{p_C(s|x_t)} = \sum_t \sum_s [p_S(s|x_t) \log p_S(s|x_t) \]  \hspace{1cm} (10)

\[ -p_S(s|x_t) \log p_C(s|x_t)] \]

Because we only optimize the CCNN, the KL loss in Eq. (10) can be rewritten to retain:

\[ l_{KL} = - \sum_t \sum_s p_S(s|x_t) \log p_C(s|x_t) \]  \hspace{1cm} (11)

Essentially, the KL loss in Eq. (11) is simplified to CE loss with soft labels. The weighted sum of KL loss and CE loss is:

\[ \mu l_{KL} + \beta l_{CE} = - \sum_t \sum_s \mu p_S(s|x_t) \log p_C(s|x_t) \]  \hspace{1cm} (12)

\[ -\sum_t \beta \log p_C(s_t|x_t) \]

From the above formula, it’s clear that the CE plays an assistant role to help the model focus on the important parts by providing prior knowledge (ground truth of inputs).

Furthermore, it’s not enough to tell the CCNN the answers of problems from standard CNN. A better teacher always explains the solving procedures of problems so that the students can
handle such kind of problems from learning one certain example. In CNN, we define a series of solving procedures matrices (SPMs). A SPM is the result of element-wise subtraction between the extracted attention feature maps from two layers, which is intuitively reasonable by using the change of outputs between two layers to represent the solving procedure. The attention map in a layer needs to emphasize valuable information for the following flow. Naturally, assuming a convolutional layer has the output tensor \( O \in \mathbb{R}^{D \times D \times C} \) with each feature map \( O_c \in \mathbb{R}^{D \times D} \), the attention map can be simply computed as \[ (13) \] 
\[
A = \sum_{c=1}^{C} O_c \]

The SPM \( S \) for the \( i \)-th layer and the \( j \)-th layer \( (j > i) \) is:

\[
S = A_j - A_i \tag{14}
\]

Finally, the SP loss can be obtained as follows:

\[
l_{SP} = \sum_t \sum_{n=1}^{N} \frac{1}{N} \times \left\| \frac{S_{Sn}(x_t)}{\|S_{Sn}(x_t)\|_F} - \frac{S_{Scn}(x_t)}{\|S_{Scn}(x_t)\|_F} \right\|_F^2 \tag{15}\]

where \( N \) is the total number of SPMs in CNN, \( S_{Sn} \) is the \( n \)-th SPM in standard CNN while the \( S_{Scn} \) is the corresponding SPM in ParConv based CCNN, and \( \| \cdot \|_F \) is standard Frobenius norm.

IV. EXPERIMENTS

The proposed distillation algorithm is mainly validated on offline handwritten Chinese text recognition (HCTR) using the CASIA database \[29], \[30]. Besides, in order to accurately observe the performance changes of CNNs, a 5-gram LM \[65\] is only added in our final results. Pytorch \[66\] is used as deep learning platform in all experiments.

A. DCNN on CASIA

The baseline DCNN architecture in \[9\] is adopted. Please note except for the categories of output layers, the CNNs in \[8], \[9\] have the same architecture. According to the configuration in \[9\], both offline isolated handwritten Chinese character datasets (HWDB1.0, HWDB1.1 and HWDB 1.2) and the training sets of offline handwritten Chinese text datasets (HWDB2.0, HWDB2.1 and HWDB2.2) are used. In total, there are 7,360 classes (Chinese characters, symbols, garbage)
and 3,932,197 images. After extracting frame-level images from the original datasets, there are 148,648,249 training samples for the training of DCNN. The ICDAR 2013 competition set is adopted as the evaluation set \[30\]. The CER is computed as:

\[
\text{CER} = \frac{N_s + N_i + N_d}{N}
\]

where \(N\) is total number of character samples in the evaluation set. \(N_s, N_i\) and \(N_d\) denote the number of substitution errors, insertion errors and deletion errors, respectively. In this study, we don’t use additional language models because we focus on the performance of CNN.

Each class is modelled by 3 tied HMM states on average. The input of DCNN is a normalized frame-level image of size 40×40 extracted from original images while the output layer has 22080 (7360×3) output nodes. In DCNN architecture, there are 14 convolutional layers which use standard 2D convolution and are followed by the batch normalization (BN) and nonlinearity activation ReLU. The number of channels continuously increases from 100 to 700. Except for the first and last convolutional layers without padding, other convolutional layers have the same padding value 1. The stride is set to 1 for all convolutional layers, while the stride of all max pooling layers is 2 with a 3×3 window.

According to our proposed guideline in Algorithm 1, we first analyze the computational costs and storages of convolutional layers (Convs) and fully-connected layers (FCs) in DCNN. The details of DCNN and statistical results are shown in Table II. Here, we don’t consider the consuming of batch normalization (BN) operation, max pooling (MaxPooling) operation because they occupy a negligible part. Based on analyzed results, it’s necessary to reduce the parameters in fully-connected layers due to a large proportion of storage (35%). Besides, we can observe almost all computational costs are generated by convolutional layers, especially, the front layers.

**B. Experiments on Architecture Distillation**

In order to make a fair comparison, the setting of hyperparameters in the training stage are the same for all experiments in this part. The minibatch size is 1,000, the momentum is 0.9 and the weight decay is 0.0001. The learning rate is initially set to 0.1 and decreased by 0.92 after every 4,000 batches. Two epoches are conducted. For fully-connected layers, although there are many algorithms can be used, we choose the simplest one that controls the weights of fully-connected layers by adjusting the feature dimension before output layer. The effectiveness of this strategy has also been reported in \[21\]. In Table III different bottleneck features are compared.
TABLE II
ARCHITECTURE AND QUANTITATIVE ANALYSIS OF DCNN CHARACTER MODEL. THE ABBREVIATIONS F, K, S, P REPRESENT THE NUMBER OF FEATURE MAPS, KERNEL SIZE, STRIDE LENGTH AND PADDING SIZE, RESPECTIVELY.

| Layer  | Configurations     | FLOPs ($\times 10^8$) | Fraction | Storage (MB) | Fraction |
|--------|--------------------|------------------------|----------|--------------|----------|
| FC2    | 500×20280          | 0.1104                 | 0.69%    | 42.1143      | 33.88%   |
| FC1    | 700×500            | 0.0035                 | 0.02%    | 1.3351       | 1.07%    |
| Conv5  | F:700, K:1×1, S:1, P:0 | 0.0049              | 0.03%    | 1.8692       | 1.50%    |
| MaxPooling | K:3×3, S:2     |                        |          |              |          |
| Conv4  | F:700, K:3×3, S:1, P:1 | 0.7056              | 4.40%    | 16.8228      | 13.54%   |
| Conv4  | F:700, K:3×3, S:1, P:1 | 0.6048              | 3.78%    | 14.4196      | 11.60%   |
| Conv4  | F:600, K:3×3, S:1, P:1 | 0.4320              | 2.70%    | 10.2997      | 8.29%    |
| Conv4  | F:500, K:3×3, S:1, P:1 | 0.3600              | 2.25%    | 8.5831       | 6.91%    |
| MaxPooling | K:3×2, S:2     |                        |          |              |          |
| Conv4  | F:500, K:3×3, S:1, P:1 | 2.2500              | 14.04%   | 8.5831       | 6.91%    |
| Conv3  | F:500, K:3×3, S:1, P:1 | 1.8000              | 11.23%   | 6.8665       | 5.52%    |
| Conv3  | F:400, K:3×3, S:1, P:1 | 1.0800              | 6.74%    | 4.1199       | 3.31%    |
| Conv3  | F:300, K:3×3, S:1, P:1 | 0.8100              | 5.06%    | 3.0899       | 2.49%    |
| MaxPooling | K:3×3, S:2     |                        |          |              |          |
| Conv2  | F:300, K:3×3, S:1, P:1 | 3.9204              | 24.47%   | 3.0899       | 2.49%    |
| Conv2  | F:300, K:3×3, S:1, P:1 | 2.6136              | 16.31%   | 2.0599       | 1.66%    |
| Conv2  | F:200, K:3×3, S:1, P:1 | 0.8712              | 5.44%    | 0.6866       | 0.55%    |
| Conv2  | F:100, K:3×3, S:1, P:1 | 0.4356              | 2.72%    | 0.3433       | 0.28%    |
| MaxPooling | K:3×3, S:2     |                        |          |              |          |
| Conv1  | F:100, K:3×3, S:1, P:0 | 0.0190              | 0.12%    | 0.0034       | 0.00%    |
| Input  | Frame-level image 40×40 |                   |          |              |          |

The dimension 500 corresponds to the baseline DCNN. Three observations can be found: First, because the FLOPs of fully-connected layers occupy a small proportion (less than 1%), the total FLOPs almost keep unchanged when the dimension is below 100. Second, with the dimension changing from 500 to 50, the storage reduces a lot and the CER has a small fluctuation, which means most parameters in fully-connected layers are redundant and can be safely ignored [67]. Finally, when the dimension is smaller than 50, the storage tends to be stable as most storage consuming is caused by convolutional layers, however, it’s reasonable to observe that the CER begins to increase due to very small number of parameters in bottleneck feature leading to the information missing. Considering the trade-off between storage ratio $\gamma$ about fully-connect layers (see Algorithm [1]) and CER, we choose dimension 50 in the following experiments.

Based on the DCNN with bottleneck feature 50 (DCNN_LF50), except for the initial and last convolutional layers, we replace all remaining 12 standard convolutional layers by our ParConv blocks (same channel multiplier $\omega$ for all 12 layers). The reconstructed compact CNN is notated as ParCNN$_{\omega}$. For example, the ParCNN$_{\omega}0.5$ means the value of $\omega$ in all ParConv blocks is set...
to 0.5. In order to demonstrate the proposed ParConv is a more efficient and effective replacement of standard convolution, we compare it with depthwise separable convolution (DSConv), heterogeneous convolution (HetConv) with different splits $\alpha$ and the architecture distillation algorithm LightweightNet proposed in [21]. For DSConv and HetConv, we also directly adopt them to replace the same 12 standard convolutional layers and build the corresponding compact CNN: DSCNN and HetCNN$_\alpha$, respectively. Table IV lists all related results. The notation *_Res means we add residual connections for all corresponding compact blocks, namely, there is another path directly connecting the input and output of the compact block.

• Comparison with different values of $\omega$

We change the value of $\omega$ from 0.5 to 4. As shown in Table IV, the CERs of ParCNN, ParCNN$_\text{Res}$ consistently decrease from 10.44% to 9.59%, 10.03% to 9.53%, respectively. Naturally, the computational resources also increase with the increment of $\omega$. Compared with the network DCNN$_\text{LF50}$, the ParCNN without residual connection can achieve 10.21× to 2.62× FLOPs based improvement and 6.35× to 2.42× in storage reduction, while the corresponding ParCNN$_\text{Res}$ can reduce FLOPs from 6.69× to 2.31× and storage overhead from 4.86× to 2.17×. By comparing the results of ParCNN and ParCNN$_\text{Res}$, we can observe the residual connection always yields a performance improvement. Meanwhile, the residual connection introduces extra computational resources (about 0.82×10$^8$ FLOPs and 4.33MB) due to pointwise convolution necessary for the special situation where the number of input and output channels of a compact block is different. Actually, in most other typical CNNs where the number of input and output channels are the same for most convolutional layers, the residual connection doesn’t lead too much extra consuming.

• Comparison with depthwise separable convolution
First, from the results of DSCNN and DSCNN_Res, an interesting and reasonable observation can be made. Without the residual connection, the recognition performance of DSCNN is significantly declined with a CER of 19.44%. In essence, directly using the DSConv to replace standard convolution doubles the depth of network, which easily leads to the degradation of network [6]. After adding the residual connection, the recognition performance of DSCNN_Res returns to a normal value (10.07%). This phenomenon reflects another advantage of our ParConv, i.e., relaxing the requirement of residual connection, which can reduce the possibility of additional computations. In the case of similar CERs (10.0%), the proposed ParConv based compact CNNs (ParCNN_\omega 0.5_Res, ParCNN_\omega 1) consume significantly less computing resources (2.38×10^8, 2.21×10^8 vs. 2.67×10^8 in FLOPs and 18.46MB, 17.41MB vs. 19.79MB in storage).

- Comparison with heterogeneous convolution

For HetCNN, the smaller \alpha is, the more channels in a HetConv layer are fed into pointwise convolutional branch, which reduces both recognition performance and parameters. For example, the CER of HetCNN_\alpha 0.25 is 9.75% while the CER of HetCNN_\alpha 0.125 increases to 10.07%. In Table [IV] the networks HetCNN_\alpha 0.125 and ParCNN_\omega 2 almost have the same

| Model                  | FLOPs (×10^8) | Storage (MB) | CER (%) |
|------------------------|---------------|--------------|---------|
| DCNN_LF50              | 15.92         | 89.74        | 9.01    |
| DSCNN                  | 1.84          | 15.46        | 19.44   |
| DSCNN_Res              | 2.67          | 19.79        | 10.01   |
| LightweightNet          | 2.12          | 23.41        | 10.30   |
| HetCNN_\alpha 0.25     | 5.33          | 32.77        | 9.75    |
| HetCNN_\alpha 0.125    | 3.55          | 23.95        | 10.07   |
| ParCNN_\omega 0.5      | 1.56          | 14.14        | 10.44   |
| ParCNN_\omega 0.5_Res  | 2.38          | 18.46        | 10.03   |
| ParCNN_\omega 1        | 2.21          | 17.41        | 10.00   |
| ParCNN_\omega 1_Res    | 3.03          | 21.74        | 9.80    |
| ParCNN_\omega 2        | 3.50          | 23.95        | 9.72    |
| ParCNN_\omega 2_Res    | 4.32          | 28.29        | 9.54    |
| ParCNN_\omega 4        | 6.07          | 37.04        | 9.59    |
| ParCNN_\omega 4_Res    | 6.90          | 41.37        | 9.53    |
FLOPs and storage, the CER of ParCNN_ω2 is 9.72% which is much better than 10.07% in HetCNN_α0.125. Besides, no matter for recognition performance or computing resources, the network ParCNN_ω2_Res can outperform HetCNN_α0.25, which demonstrates the ParConv is more efficient and effective than HetConv.

- **Comparison with LightweightNet**

We also reproduce the architecture distillation algorithm [21]. The reconstructed LightweightNet needs $2.12 \times 10^8$ FLOPs and occupies 23.41MB while the corresponding CER is 10.30%. The network ParCNN_ω0.5 with a comparable recognition performance can apparently outperform it in FLOPs and storage. Meanwhile, the networks ParCNN_ω0.5_Res and ParCNN_ω1 have similar FLOPs with LightweightNet, but lower storages and CERs.

**C. Experiments on Architecture and Distillation Distillation**

As shown in Table [IV], a smaller value $\omega$ can obtain a larger compression ratio but suffer from worse recognition performance. In order to reduce the performance gap between baseline CNN and the compact CNN, it is necessary to introduce knowledge distillation simultaneously. In knowledge distillation, i.e., based on Eq. (8), the weight $\mu$ is set to 0.8, $\beta$ equals to 0.2 and $\lambda$ is 0.1. Except the batch size is set to 700, all other initial training hyper parameters are the same with the parameters in architecture distillation.

In order to excavate the best capability of the proposed approach, we first combine knowledge distillation to improve the recognition performance of the smallest network ParCNN_ω0.5. From the results of Table [V] we can observe that the knowledge distillation can yield remarkable reductions of CER: from 10.44% to 9.79% (+KL Loss), 9.94% (+SP Loss) and 9.68% (+KL Loss & +SP Loss), which demonstrates the effectiveness of the SP loss and the necessity of knowledge distillation. Compared with the baseline DCNN, our proposed joint architecture and knowledge distillation can achieve $10 \times$ reduction of computational cost and $9 \times$ storage compression with only a 0.51% increment in CER, i.e., relative CER increment of 5.6%.

Table [VI] shows the results for different acceleration and compression ratios based on channel multiplier $\omega$. It can be observed that the value of $\omega$ can effectively control the acceleration and compression ratio and recognition performance. When the $\omega$ is set to 1, compared with the baseline DCNN, the proposed approach can reduce the computational cost and model size by $>7 \times$ with relative CER increment of 2.2%. If we further increase the value of $\omega$ to 2, the
TABLE V
THE RESULTS OF JOINT ARCHITECTURE AND KNOWLEDGE DISTILLATION FOR DCNN.

| Model           | FLOPs ($\times 10^8$) | Storage (MB) | CER (%) |
|-----------------|-----------------------|--------------|---------|
| DCNN            | 16.02                 | 124.5        | 9.17    |
| ParCNN_ω0.5     | 1.56                  | 14.14        | 10.44   |
| +KL             |                       |              | 9.79    |
| +SP             |                       |              | 9.94    |
| +KL+SP          |                       |              | 9.68    |

TABLE VI
THE RESULTS OF PROPOSED APPROACH FOR DIFFERENT ACCELERATION AND COMPRESSION RATIOS BASED ON CHANNEL MULTIPLIER ω.

| Model           | FLOPs ($\times 10^8$) | Storage (MB) | Without KD | With KD |
|-----------------|-----------------------|--------------|------------|---------|
| DCNN            | 16.02                 | 124.5        | 9.17       | -       |
| ParCNN_ω0.5     | 1.56                  | 14.14        | 10.44      | 9.68    |
| ParCNN_ω1       | 2.21                  | 17.41        | 10.00      | 9.37    |
| ParCNN_ω2       | 3.50                  | 23.95        | 9.72       | 9.09    |

compact network ParCNN_ω2 with >4× acceleration ratio and >5× compression ratio can even obtain a better performance (9.09% vs. 9.17%).

In order to better understand why more parameters can yield better performance, we draw the learning curves about multiple losses during training for ParCNN_ω0.5, ParCNN_ω1 and ParCNN_ω2 in Fig. 5. It can be observed that all kinds of losses become lower with the increment of ω, which is in line with our expectations. Another interesting observation is the relative gap of SP loss between ParCNN_ω1 and ParCNN_ω2 is larger than other kinds of losses. This indicates that the SP loss should play an important role in the training of ParCNN_ω2.

Finally, considering the LM plays an important role in HCTR, we add the same 5-gram LM [9] to compare the final results of DCNN and ParCNN_ω0.5. As shown in Table VII, it is reasonable to observe the performance gap is almost fixed by LM, which indicates the proposed algorithm can yield a remarkable compression ratio with negligible accuracy loss. We also test actual runtime (millisecond per batch) for DCNN and proposed ParCNN_ω0.5. All models with batch size 120 are run 10 times in the same machine that is equipped with Pytorch (version 1.0.1) with GeForce RTX 2080, CUDA version 10.0.130 and CUDNN [68] version 7402. Although FLOPs reduction (theoretical) is amazing, the practical speedup (2×) is limited. The main reason
Fig. 5. The comparison of multiple losses for ParCNN$_{\omega 0.5}$, ParCNN$_{\omega 1}$ and ParCNN$_{\omega 2}$. For simplicity, we use $\omega 0.5$, $\omega 1$, $\omega 2$ to represent respective networks in all figures.
TABLE VII
THE COMPARISON OF FINAL RESULTS AFTER ADDING THE SAME 5-GRAM LM.

| Model         | FLOPs ($\times 10^8$) | Storage (MB) | GPU time (ms/batch) | GPU Occupancy (MB) | CER (%) |
|---------------|------------------------|-------------|---------------------|-------------------|---------|
| DCNN [9]      | 16.02                  | 124.5       | 39.3                | 3353              | 3.52    |
| ParCNN$_{\omega 0.5}$ | 1.56                  | 14.14       | 19.0                | 759               | 3.55    |

is the 1×1 convolutions and depthwise convolutions in Pytorch are relatively slow and the latest CUDNN library is specially optimized for 3×3 convolutions. However, we observe that when running the same batch, the DCNN consumes 3353 MB of GPU memory while the ParCNN$_{\omega 0.5}$ only needs 759 MB.

D. Experiments on MNIST

In order to demonstrate the proposed approach can also be successfully applied on mainstream backbone networks, according to the structure of Res18, we build the corresponding compressed network ParRes18 based on the proposed parsimonious convolution with channel multiplier $\omega = 0.5$ (see Fig. 4). Fig. 6 shows the differences between Res18 and ParRes18. And then we conduct experiments on one of the most popular datasets: the MNIST dataset that includes 60,000 training images and 10,000 test images. Each image is resized to 28×28 and labelled as a digit (0-9). We first train the three kinds of mainstream neural networks i.e., AlexNet [3], VGG19 [4], Res18 [6]. These network prototypes are provided by Pytorch and batch normalization is used for all convolutional layers. We use the same training criterion to train all networks: the minibatch size is 64, the momentum is 0.9, the weight decay is 0.0001 and learning rate is set to 0.01. From Table VIII we can observe that compared with the Res18, the ParRes18 can obtain >9× acceleration ratio and compression ratio with a similar performance. Besides, it has obvious advantages over AlexNet and VGG19.

V. CONCLUSION

In this paper, we propose a guideline to distill the architecture and knowledge of pre-trained standard CNNs simultaneously. The proposed algorithm is first verified on offline handwritten Chinese text recognition. In architecture distillation, we invent a parsimonious convolution block (ParConv) to directly replace vanilla convolution without any other adjustments. Compared with other popular plug-and-play compact convolutional units, the proposed ParConv demonstrates
its superiority in recognition performance, computational cost and storage overhead. To further reduce the gap between the baseline CNN and the corresponding compact CNN, knowledge distillation with multiple losses is adopted. And then, by conducting experiments on one of the most popular data sets: the MNIST, we demonstrate the proposed approach can also be successfully applied on mainstream backbone networks. For future work, we will combine other compression and acceleration algorithms in our task.

**TABLE VIII**

THE OVERALL COMPARISON FOR DIFFERENT NETWORKS ON MNIST.

| Model     | FLOPs ($\times 10^7$) | Storage (MB) | CER (%) |
|-----------|------------------------|--------------|---------|
| AlexNet [3] | 2.15                   | 77.57        | 1.02    |
| VGG19 [4]  | 27.65                  | 148.67       | 0.34    |
| Res18 [6]  | 45.58                  | 42.68        | 0.33    |
| ParRes18   | 4.86                   | 4.47         | 0.34    |
VI.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grant Nos. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, and the MOE-Microsoft Key Laboratory of USTC.

REFERENCES

[1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Neural computation, vol. 1, no. 4, pp. 541–551, 1989.

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner et al., “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[8] Z.-R. Wang, J. Du, W.-C. Wang, J.-F. Zhai, and J.-S. Hu, “A comprehensive study of hybrid neural network hidden markov model for offline handwritten Chinese text recognition,” International Journal on Document Analysis and Recognition (IJDAR), vol. 21, no. 4, pp. 241–251, 2018.

[9] Z.-R. Wang, J. Du, and J.-M. Wang, “Writer-aware cnn for parsimonious hmm-based offline handwritten chinese text recognition,” Pattern Recognition, p. 107102, 2019.

[10] X. Xiao, L. Jin, Y. Yang, W. Yang, J. Sun, and T. Chang, “Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition,” Pattern Recognition, vol. 72, pp. 72–81, 2017.

[11] X.-Y. Zhang, Y. Bengio, and C.-L. Liu, “Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark,” Pattern Recognition, vol. 61, pp. 348–360, 2017.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[13] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, “Attention based fully convolutional network for speech emotion recognition,” in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018, pp. 1771–1775.

[14] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, “Model compression,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006, pp. 535–541.
[15] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, 2015.

[16] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” arXiv preprint arXiv:1412.6550, 2014.

[17] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” arXiv preprint arXiv:1612.03928, 2016.

[18] J. Yim, D. Joo, J. Bae, and J. Kim, “A gift from knowledge distillation: Fast optimization, network minimization and transfer learning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4133–4141.

[19] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, “Structured knowledge distillation for semantic segmentation,” arXiv preprint arXiv:1903.04197, 2019.

[20] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, “Knowledge adaptation for efficient semantic segmentation,” arXiv preprint arXiv:1903.04688, 2019.

[21] T.-B. Xu, P. Yang, X.-Y. Zhang, and C.-L. Liu, “Lightweightnet: Toward fast and lightweight convolutional neural networks via architecture distillation,” Pattern Recognition, vol. 88, pp. 272–284, 2019.

[22] P. Singh, V. K. Verma, P. Rai, and V. P. Namboodiri, “Hetconv: Heterogeneous kernel-based convolutions for deep cnns,” arXiv preprint arXiv:1903.04120, 2019.

[23] Y. Chen, H. Fang, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, and J. Feng, “Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution,” arXiv preprint arXiv:1904.05049, 2019.

[24] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[25] Q. Fu, X. Ding, T. Liu, Y. Jiang, and Z. Ren, “A novel segmentation and recognition algorithm for Chinese handwritten address character strings,” in 18th International Conference on Pattern Recognition (ICPR’06), vol. 2. IEEE, 2006, pp. 974–977.

[26] M. Yu, P. C. Kwok, C. H. Leung, and K. Tse, “Segmentation and recognition of Chinese bank check amounts,” International Journal on Document Analysis and Recognition, vol. 3, no. 4, pp. 207–217, 2001.

[27] H. Fujisawa, “Forty years of research in character and document recognitionan industrial perspective,” Pattern Recognition, vol. 41, no. 8, pp. 2435–2446, 2008.

[28] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, p. 436, 2015.

[29] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, “CASIA online and offline Chinese handwriting databases,” in 2011 International Conference on Document Analysis and Recognition. IEEE, 2011, pp. 37–41.

[30] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, “ICDAR 2013 Chinese handwriting recognition competition,” in 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013, pp. 1464–1470.

[31] Q.-F. Wang, F. Yin, and C.-L. Liu, “Handwritten Chinese text recognition by integrating multiple contexts,” IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 8, pp. 1469–1481, 2012.

[32] S. Wang, L. Chen, L. Xu, W. Fan, J. Sun, and S. Naoi, “Deep knowledge training and heterogeneous cnn for handwritten Chinese text recognition,” in 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2016, pp. 84–89.

[33] Y.-C. Wu, F. Yin, and C.-L. Liu, “Improving handwritten Chinese text recognition using neural network language models and convolutional neural network shape models,” Pattern Recognition, vol. 65, pp. 251–264, 2017.

[34] R. Messina and J. Louradour, “Segmentation-free handwritten Chinese text recognition with LSTM-RNN,” in 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015, pp. 171–175.
[35] Y.-C. Wu, F. Yin, Z. Chen, and C.-L. Liu, “Handwritten Chinese text recognition using separable multi-dimensional recurrent neural network,” in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1. IEEE, 2017, pp. 79–84.

[36] T.-H. Su, T.-W. Zhang, D.-J. Guan, and H.-J. Huang, “Off-line recognition of realistic Chinese handwriting using segmentation-free strategy,” Pattern Recognition, vol. 42, no. 1, pp. 167–182, 2009.

[37] M. Mohri, F. Pereira, and M. Riley, “Weighted finite-state transducers in speech recognition,” Computer Speech & Language, vol. 16, no. 1, pp. 69–88, 2002.

[38] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, “Exploiting linear structure within convolutional networks for efficient evaluation,” in Advances in neural information processing systems, 2014, pp. 1269–1277.

[39] M. Jaderberg, A. Vedaldi, and A. Zisserman, “Speeding up convolutional neural networks with low rank expansions,” arXiv preprint arXiv:1405.3866, 2014.

[40] X. Zhang, J. Zou, K. He, and J. Sun, “Accelerating very deep convolutional networks for classification and detection,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 1943–1955, 2016.

[41] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin, “Compression of deep convolutional neural networks for fast and low power mobile applications,” arXiv preprint arXiv:1511.06530, 2015.

[42] H. Ding, K. Chen, Y. Yuan, M. Cai, L. Sun, S. Liang, and Q. Huo, “A compact CNN-DBLSTM based character model for offline handwriting recognition with Tucker decomposition,” in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1. IEEE, 2017, pp. 507–512.

[43] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural network,” in Advances in neural information processing systems, 2015, pp. 1135–1143.

[44] Y. Guo, A. Yao, and Y. Chen, “Dynamic network surgery for efficient dnns,” in Advances In Neural Information Processing Systems, 2016, pp. 1379–1387.

[45] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, “Learning efficient convolutional networks through network slimming,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2736–2744.

[46] J.-H. Luo, J. Wu, and W. Lin, “Thinet: A filter level pruning method for deep neural network compression,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5058–5066.

[47] A. Gordon, E. Eban, O. Nachum, B. Chen, H. Wu, T.-J. Yang, and E. Choi, “Morphnet: Fast & simple resource-constrained structure learning of deep networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1586–1595.

[48] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” arXiv preprint arXiv:1510.00149, 2015.

[49] V. Vanhoucke, A. Senior, and M. Z. Mao, “Improving the speed of neural networks on cpus,” 2011.

[50] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1,” arXiv preprint arXiv:1602.02830, 2016.

[51] F. Li, B. Zhang, and B. Liu, “Ternary weight networks,” arXiv preprint arXiv:1605.04711, 2016.

[52] R. Ding, Z. Liu, R. Shi, D. Marculescu, and R. Blanton, “Lightnn: Filling the gap between conventional deep neural networks and binarized networks,” in Proceedings of the on Great Lakes Symposium on VLSI 2017. ACM, 2017, pp. 35–40.

[53] R. Ding, Z. Liu, T.-W. Chin, D. Marculescu et al., “Flightnns: Lightweight quantized deep neural networks for fast and accurate inference,” arXiv preprint arXiv:1904.02835, 2019.

[54] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size,” arXiv preprint arXiv:1602.07360, 2016.
[55] X. Zhang, X. Zhou, M. Lin, and J. Sun, “Shufflenet: An extremely efficient convolutional neural network for mobile devices,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.

[56] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “Shufflenet v2: Practical guidelines for efficient cnn architecture design,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 116–131.

[57] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861, 2017.

[58] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[59] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, “Rethinking the value of network pruning,” arXiv preprint arXiv:1810.05270, 2018.

[60] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[61] L. Sifre and S. Mallat, “Rigid-motion scattering for image classification,” PhD thesis, Ph. D. thesis, vol. 1, p. 3, 2014.

[62] J. Guo, Y. Li, W. Lin, Y. Chen, and J. Li, “Network decoupling: From regular to depthwise separable convolutions,” arXiv preprint arXiv:1808.05517, 2018.

[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[64] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, “Low-rank matrix factorization for deep neural network training with high-dimensional output targets,” in 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013, pp. 6655–6659.

[65] S. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” IEEE transactions on acoustics, speech, and signal processing, vol. 35, no. 3, pp. 400–401, 1987.

[66] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[67] M. Denil, B. Shakibi, L. Dinh, N. De Freitas et al., “Predicting parameters in deep learning,” in Advances in neural information processing systems, 2013, pp. 2148–2156.

[68] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, “cudnn: Efficient primitives for deep learning,” arXiv preprint arXiv:1410.0759, 2014.