

DWM: A Decomposable Winograd Method for Convolution Acceleration

Di Huang,^{1,2,3} Xishan Zhang,^{1*} Rui Zhang,¹ Tian Zhi,¹ Deyuan He,^{1,2,3} Jiaming Guo,^{1,2,3}
Chang Liu,^{1,2,3} Qi Guo,¹ Zidong Du,¹ Shaoli Liu,³ Tianshi Chen,³ Yunji Chen^{1,2,4,5,6}

¹SKL of Computer Architecture, Institute of Computing Technology, CAS

²University of Chinese Academy of Sciences, ³Cambricon Tech. Ltd

⁴Institute of Brain-Intelligence Technology, Zhangjiang Laboratory

⁵Shanghai Research Center for Brain Science and Brain-Inspired Intelligence

⁶CAS Center for Excellence in Brain Science and Intelligence Technology

{huangdi18s, zhangxishan, zhangrui, zhitian, hedeyuan18s, guojiaming18s, liuchang18s, guoqi, duzidong, cyj}@ict.ac.cn
{liushaoli, tchen}@cambricon.com

Abstract

Winograd’s minimal filtering algorithm has been widely used in Convolutional Neural Networks (CNNs) to reduce the number of multiplications for faster processing. However, it is only effective on convolutions with kernel size as 3×3 and stride as 1, because it suffers from significantly increased FLOPs and numerical accuracy problem for kernel size larger than 3×3 and fails on convolution with stride larger than 1. In this paper, we propose a novel Decomposable Winograd Method (DWM), which breaks through the limitation of original Winograd’s minimal filtering algorithm to a wide and general convolutions. DWM decomposes kernels with large size or large stride to several small kernels with stride as 1 for further applying Winograd method, so that DWM can reduce the number of multiplications while keeping the numerical accuracy. It enables the fast exploring of larger kernel size and larger stride value in CNNs for high performance and accuracy and even the potential for new CNNs. Comparing against the original Winograd, the proposed DWM is able to support all kinds of convolutions with a speedup of ~ 2 , without affecting the numerical accuracy.

Introduction

Deep Convolutional Neural Networks (CNNs) have shown excellent performance on many machine learning tasks but has been plagued by the huge amount of computations for a long time. Recently, CNNs are increasingly larger to achieve better accuracy but also increase the amount of computations. *E.g.*, AlexNet (Krizhevsky, Sutskever, and Hinton 2012), has 7×10^8 multiplications, VGG-16 (Simonyan and Zisserman 2014) contains 1.5×10^{10} multiplications, and SENet-154 (Hu, Shen, and Sun 2018) contains 2.1×10^{10} multiplications. The massive amount of computations blocks the application of CNNs. Therefore, reducing the multiplications in CNNs is essential and meaningful.

Lavin (2016) applied Winograd’s minimal filtering algorithm (Winograd 1980) to half reduce the number of multiplications in convolutions. Unfortunately, Winograd’s minimal filtering algorithm is only effective on 3×3 kernels with

*Xishan Zhang (zhangxishan@ict.ac.cn) is the corresponding author.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

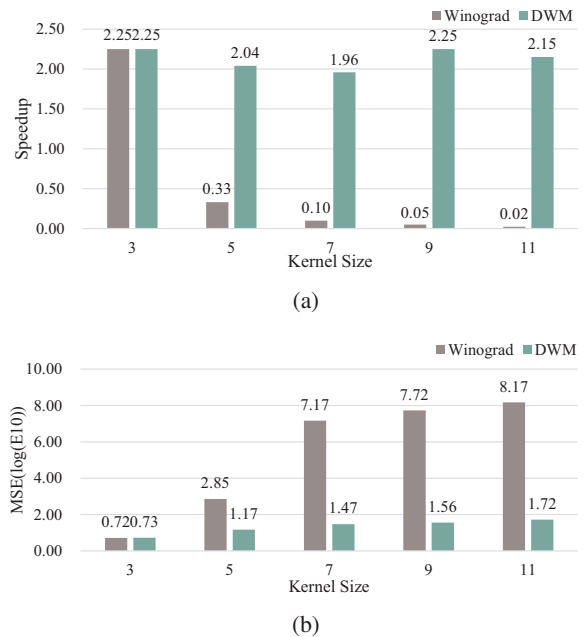


Figure 1: (a): The speedup comparison between two accelerating algorithms, measured by FLOPs. Baseline is the regular convolution result of FP64; (b): The numerical accuracy of two accelerating algorithms, which is the mean squared error (MSE), scaled by $\log_{10}(\text{error}) + 10$ to make it positive.

stride 1. When the kernels are larger than 3×3 , the transform matrices of Winograd’s minimal filtering algorithm will introduce much more decimal multiplications causing precision and efficiency problems. Another problem is that Winograd’s minimal filtering algorithm cannot be used on convolutions with stride > 1 . Since convolutions of kernel size larger than 3×3 and stride > 1 are frequently used in CNNs, these two restrictions severely limited the application of Winograd’s minimal filtering algorithm.

To tackle the above drawbacks, we propose the Decomposable Winograd Method (DWM) to extend the Winograd’s

minimal filtering algorithm into the cases of large kernels and stride > 1 . First, for kernels larger than 3, DWM decomposes the original kernel into several kernels smaller than 3, on which we can apply Winograd’s minimal filtering algorithm separately. Thus, for large kernels situation, DWM can still reduce the number of multiplications by 50% and keep the numerical accuracy the same as the original convolution. Second, for stride larger than 1 situation, DWM splits the kernels into several parts to apply Winograd’s minimal filtering algorithm. *E.g.*, with stride 2, DWM is equal to split the polynomials of Winograd into odd ones and even ones and compute them respectively. Therefore, DWM break through the kernel size and stride restriction of Winograd’s minimal filtering algorithm. DWM has the advantages of both computation and numerical accuracy. DWM can efficiently reduce the multiplications of regular convolutions. As shown in Figure 1(a), with the kernel size increasing from 3 to 11, the speedup of Winograd’s minimal filtering algorithm decreases seriously, while the speedup of DWM keep to ~ 2 . Furthermore, DWM can keep the final result totally equivalent to the result of regular convolutions, which makes it suitable to be applied to actual products. As shown in Figure 1(b), with the kernel size increasing from 3 to 11, the accuracy error of DWM keep small but the error of Winograd’s minimal filtering algorithm increases quickly. These advantages of DWM enables the fast exploring of larger kernel size and larger stride value in CNNs for high performance and accuracy and even the potential for new CNNs. Experiments show that DWM achieves $\sim 2\times$ acceleration while keeping the numerical error under E-07, which is close to the numerical accuracy of FP32 convolution.

The contribution of this paper has three aspects:

- We identify the limitation of the Winograd’s minimal filtering algorithm that it suffers from significantly increased FLOPs and numerical accuracy problem for kernel size larger than 3×3 and fails on convolution with stride larger than 1.
- We propose a novel DWM method which decomposes kernels with large size or large stride to several small kernels with stride as 1 for further apply Winograd method, so that DWM can reduce the number of multiplications while keeping the numerical accuracy.
- We evaluate the proposed DWM method on convolutions with kernel size varying from 3 to 11 and stride from 1 to 2. Experimental results show that DWM is able to support all kinds of convolutions with a speedup of ~ 2 , without affecting the numerical accuracy.

Related Works

So far, with the success of convolutional neuron networks, many researchers focus on accelerate convolution using linear algebra property. Cong and Xiao (2014) save 47% amount of multiplications by utilizing the linear algebra property at the sub-matrix block level. Mathieu et al. (1997) first proposed the method of using Fast Fourier Transform (FFT) to reduce the computation of convolution operations. After that, the FFT algorithm was refined by Vasilache (2014). Their two GPU implementations called

cuFFT and fbfft outperformed the convolution library made by NVIDIA at that time. Lavin (2016) exploited several element-level linear algebra methods to reduce the number of multiplications, including Winograd’s minimal filtering algorithm (Winograd 1980) and FFT. Nowadays, cuDNN (Chetlur et al. 2014), a state-of-art deep learning library, includes both the Winograd algorithm and the FFT algorithm as their convolution implementation.

Some researchers made their efforts to overcome the defects of the Winograd algorithm. Some of them tried to overcome the incompatibility of the Winograd algorithm and model sparsification. Li (2017) proposed a method to sparse native Winograd coefficients and obtains sparsity level beyond 90% with only 0.1% accuracy loss. Liu (2018) moved ReLU operation into the Winograd transformation and then pruned the weights in the Winograd domain. This approach reduced the number of multiplications by $10.8\times$ with loss of accuracy less than 0.1%. Choi (2018) proposed a prune principle to make the data keep sparse after the Winograd transformation. Some researchers attempted to solve the numerical accuracy and kernel size problem. Barabasz (2019) investigated a wider range of Winograd algorithms for DNNs, which significantly improve floating-point accuracy in many cases. In fp16, this approach has given up to 6.5 times better image recognition accuracy in one important case. Vincent (2017) decreased the numerical error of large tile Winograd algorithm by selecting the polynomial points. Meng and Brothers (2019) extended the Winograd algorithm to larger tiles by introducing complex numbers during the Winograd transformation. Other researchers focused on the hardware implementation of the Winograd algorithm. Due to the low memory ceiling of GPU hardware, the Winograd algorithm can be used to speed up CPU convolution operations and achieves 5 to 25-fold improvement in throughput compared to previous state-of-art implementations (Budden et al. 2017). Besides, for the mobile CPU acceleration, the Winograd algorithm achieves up to 60% performance improvements in the full network compared to im2col/im2row based optimization techniques (Maji et al. 2019).

Different from the researches mentioned above, our methods focus on the Winograd algorithm itself instead of the combination of the Winograd algorithm and other methods such as sparsification. By extending the Winograd algorithm to a much wider situation, we efficiently reduce the number of multiplications while keeping the calculation’s numerical accuracy stably high.

Preliminary on the Winograd Algorithm

The Winograd Algorithm

As an equivalent problem of multi-dimensional FIR filters problem, convolution can be implemented more efficiently using Winograd minimal filtering algorithm (Lavin and Gray 2016). Denoting the result of computing m outputs with an r -tap FIR filter as $F(m, r)$, the corresponding convolution algorithm for it requires $m + r - 1$ multiplications.

The original Winograd algorithm is derived from the relationship between polynomial multiplication and 1-D convolutions using the Chinese Remainder Theorem(CRT) (Wino-

grad 1980). For the fixed m and r , the whole algorithm contains three fixed transformation matrices: A , B and G .

Considering the original 1-D situation, the r element filter $g(x)$ and $l = m + r - 1$ element input signal $d(x)$ can be represented as polynomials ($0 < r < l$):

$$\begin{aligned} g(x) &= g_{r-1}x^{r-1} + g_{r-2}x^{r-2} + \dots + g_1x + g_0, \\ d(x) &= d_{l-1}x^{l-1} + d_{l-2}x^{l-2} + \dots + d_1x + d_0, \end{aligned} \quad (1)$$

then the result of convolution $g(x) * d(x)$ can be obtained by calculating the coefficients of polynomial multiplication

$$y(x) = g(x)d(x). \quad (2)$$

Applying CRT, we can get three transformation matrices A , B and G , and the process of doing convolution can be formulated as the following:

$$Y = A^T[(Gg) \odot (B^T d)]. \quad (3)$$

where Y denotes the convolution output and \odot denotes element-wise multiplication. For 2-D convolutions, we can nest the $F(m, r)$ with itself, and then get

$$Y = A^T[(GgG^T) \odot (B^T dB)]A. \quad (4)$$

From equation (4), we can derive the gradient of neuron (denoted as ∇d) and the gradient of weight (denoted as ∇g) of Winograd algorithm using the chain rule:

$$\begin{aligned} \nabla d &= B[(A\nabla Y A^T) \odot (GgG^T)]B^T, \\ \nabla g &= G^T[(A\nabla Y A^T) \odot (B^T dB)]G, \end{aligned} \quad (5)$$

where ∇Y is the gradient passed from the next layer.

Drawbacks

Large Kernel Size The benefit of the Winograd algorithm comes from the simplicity of transformation matrices. For example, applying the Winograd algorithm, the transformation matrices of $F(2, 3)$ are shown as follows:

$$\begin{aligned} B^T &= \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \\ G &= \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}, \\ A^T &= \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & -1 & -1 \end{bmatrix}. \end{aligned} \quad (6)$$

However, considering $F(2, 5)$ as an example, the transformation matrices becomes something like

$$\begin{aligned} B^T &= \begin{bmatrix} 4 & 0 & -5 & 0 & 1 & 0 \\ 0 & -4 & -4 & 1 & 1 & 0 \\ 0 & 4 & -4 & -1 & 1 & 0 \\ 0 & -2 & -1 & 2 & 1 & 0 \\ 0 & 2 & -1 & -2 & 1 & 0 \\ 0 & 4 & 0 & -5 & 0 & 1 \end{bmatrix}, \\ G &= \begin{bmatrix} 1/4 & 0 & 0 & 0 & 0 \\ -1/6 & -1/6 & -1/6 & -1/6 & -1/6 \\ -1/6 & 1/6 & -1/6 & 1/6 & -1/6 \\ 1/24 & 1/12 & 1/6 & 1/3 & 2/3 \\ 1/24 & -1/12 & 1/6 & -1/3 & 2/3 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \\ A^T &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & -1 & 2 & -2 & 1 \end{bmatrix}. \end{aligned} \quad (7)$$

The huge number of decimals in transformation matrices makes the Winograd transformation not only high consumption, but also less accurate.

Stride > 1 Another problem of the Winograd algorithm is that it cannot be applied to stride > 1 convolutions, for it is derived from polynomial multiplication which indicates stride 1 convolutions. Therefore, although the Winograd algorithm can implement convolutions much more efficiently, it is always used on 3×3 and stride 1 convolutions only.

The Decomposable Winograd Method

In this section, we propose a series of techniques called the Decomposable Winograd Method(DWM) to apply the Winograd algorithm mentioned above on more general cases like larger kernels and stride > 1.

Forward

Large Kernel Size As we denote before, $g(x)$ represents the 1-D convolution filter and thus r represents the size of convolution filter. Observing the derivation process above, we find that when r becomes larger, $g(x)$ in equation (1) can be split into several small polynomials, to which we can apply the original Winograd algorithm:

$$\begin{cases} g^{(0)}(x) = g_2x^2 + g_1x + g_0 \\ g^{(1)}(x) = (g_5x^2 + g_4x + g_3)x^3 \\ \vdots \\ g^{(\lfloor r/3 \rfloor)}(x) = \sum_{i=0}^{r-1 \bmod 3} g_{r-i-1}x^{(r-1 \bmod 3)-i}x^{3\lfloor r/3 \rfloor} \end{cases} \quad (8)$$

Then from $g(x) = \sum_{i=0}^{\lfloor r/3 \rfloor} g^{(i)}(x)$ we can get

$$y(x) = g(x)d(x) = \sum_{i=0}^{\lfloor r/3 \rfloor} [g^{(i)}(x)d(x)] = \sum_{i=0}^{\lfloor r/3 \rfloor} y^i(x), \quad (9)$$

We can apply $F(2, 2)$ or $F(2, 3)$ on each $y^i(x)$ separately, with half multiplication reduced. As for 2-D convolution,

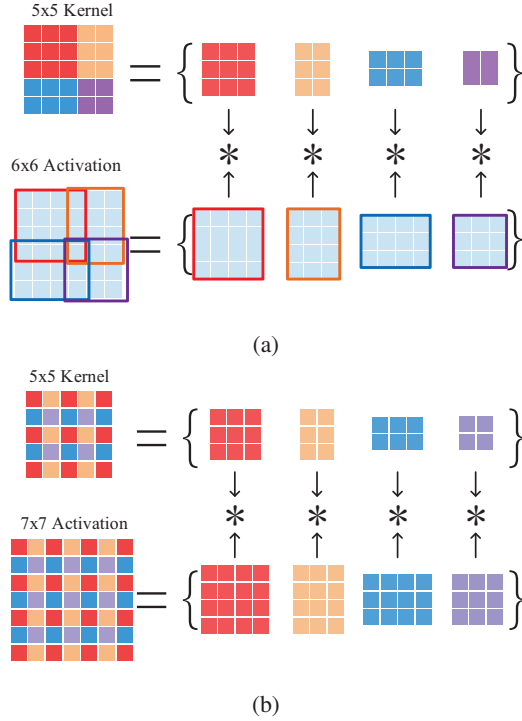


Figure 2: (a): Splitting a 5×5 and stride 1 convolution into four smaller convolutions; (b): Splitting a 5×5 and stride 2 convolution into four stride 1 convolutions. '*' denotes doing convolution with corresponding parts.

we can split the large kernel into small parts, and then apply Winograd algorithm to each part separately. The whole process is illustrated by Figure 3, which shows that we can process a common large kernel convolution in five steps:

- **Splitting.** Split the convolution kernel into several parts whose kernel size belows 3×3 , and then prepare the input signal by slicing the redundant edges. This method is shown in Figure 2a.
- **Transformation.** Apply corresponding Winograd transformation $B^T(\cdot)B$ and $G(\cdot)G^T$ (in $F(2, 2)$ or $F(2, 3)$) on each part.
- **Calculation.** Do element-wise multiplication and channel-wise summation.
- **Detransformation.** Do the $A^T(\cdot)A$ inverse transformation to change the intermediate results to spatial domain.
- **Aggregation.** Sum the calculation results of each part, which gives the final result that equivalent to the original convolution.

For example, when $r = 5$, we can split it into two parts

$$\begin{cases} g^{(0)}(x) = g_2x^2 + g_1x + g_0 \\ g^{(1)}(x) = (g_4x + g_3)x^3 \end{cases}, \quad (10)$$

and $g(x) = g^{(0)}(x) + g^{(1)}(x)$. Then we get:

$$y(x) = g(x)d(x) = g^{(0)}(x)d(x) + g^{(1)}(x)d(x). \quad (11)$$

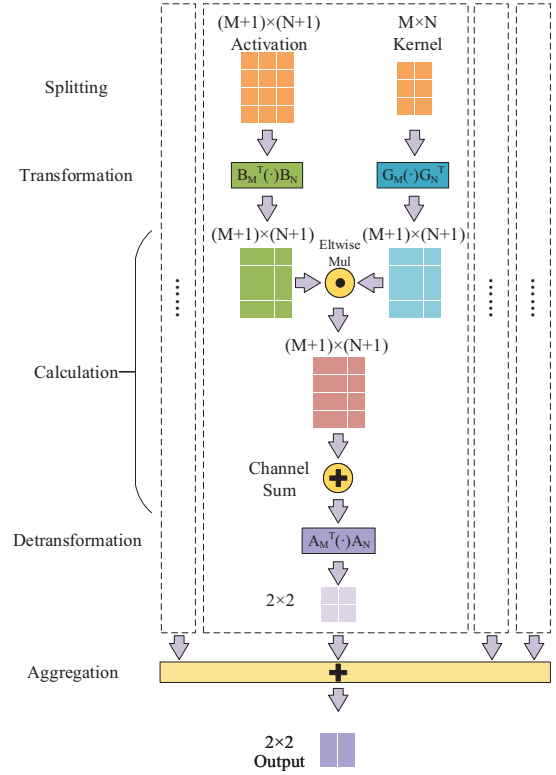


Figure 3: The process of doing convolution using Winograd algorithm. The four dotted frames denote convolution procedure of the four split parts, and each procedure can be divided into five steps: splitting, transformation, calculation, detransformation and aggregation.

For $g^{(0)}(x)d(x)$, we can apply $F(2, 3)$ to it, and for $g^{(1)}(x)d(x)$, we can apply $F(2, 2)$. When it comes to a 2-D 5×5 convolution case, we can split the kernel into 4 parts: 3×3 , 3×2 , 2×3 and 2×2 , just as the method shown in Figure 2a. This method's advantage is that using $F(2, 2)$ and $F(2, 3)$ instead of larger ones not only reduces the multiplications of Winograd transformation efficiently but also keeps the Winograd algorithm's accuracy because of the few amounts of multiplications in transformation matrices. We will further illustrate this advantage in the experiment part.

Stride > 1 Normal polynomial multiplication indicates stride 1 convolution, but we can surmount this barrier by grouping the polynomial into several parts. Denoting convolution stride as s , 1-D convolution kernel $g(x)$ with r ele-

ments can be split into

$$\left\{ \begin{array}{l} g^{(0)}(x) = \sum_{i=0}^{\lfloor (r-1)/s \rfloor} g_{s*i} x^{s*i} \\ g^{(1)}(x) = \sum_{i=0}^{\lfloor (r-2)/s \rfloor} g_{s*i+1} x^{s*i+1} \\ \vdots \\ g^{(s-1)}(x) = \sum_{i=0}^{\lfloor (r-s-1)/s \rfloor} g_{s*i+s-1} x^{s*i+s-1} \end{array} \right. \quad (12)$$

The input signal $d(x)$ with l elements can be split into $d^{(0)}(x), d^{(1)}(x), \dots, d^{(s-1)}(x)$ similarly.

Then we get several stride 1 convolutions which can be represented by polynomials multiplications $y^{(i)}(x) = g^{(i)}(x)d^{(i)}(x)$ and $y(x) = \sum y^{(i)}(x)$. By applying Winograd algorithm to them, we have reduced the multiplications of 1-D convolutions with stride > 1 successfully. For 2-D convolutions, we nest the 1-D convolution methods. This process also contains five parts: splitting, transformation, calculation, detransformation and aggregation, which is similar to Figure 3. The splitting method is illustrated by Figure 2b.

For instance, when we are dealing with stride 2 convolutions, we can group the convolution kernel $g(x)$ and input signal $d(x)$ by their degree's parity, and then get

$$\left\{ \begin{array}{l} g^{(0)}(x) = \sum_{i=0}^{\lfloor (r-1)/2 \rfloor} g_{2i} x^{2i} \\ g^{(1)}(x) = \sum_{i=0}^{\lfloor (r-2)/2 \rfloor} g_{2i+1} x^{2i+1} \end{array} \right. \quad (13)$$

and

$$\left\{ \begin{array}{l} d^{(0)}(x) = \sum_{i=0}^{\lfloor (l-1)/2 \rfloor} d_{2i} x^{2i} \\ d^{(1)}(x) = \sum_{i=0}^{\lfloor (l-2)/2 \rfloor} d_{2i+1} x^{2i+1} \end{array} \right. \quad (14)$$

In a 2-D case, a 5×5 stride 2 convolution on 7×7 activation can be splitted into doing four stride 1 convolutions: 3×3 , 3×2 , 2×3 and 2×2 . Details are shown in Figure 2b.

Occasionally, we need to do a stride > 2 convolution with large kernel size, then we can combine the two techniques mentioned above. By applying these two techniques, we can optimize all kinds of convolutions with the Winograd algorithm, reducing the number of multiplications by half.

Backward

Using equation (5), we can apply DWM in the process of calculating gradients, and this will also reduce the number of multiplications by half. From SGD algorithm, we can derive that

$$\begin{aligned} g_{(t+1)}^{wino} &= G(g_{(t)} + \nabla g_{(t)})G^T \\ &= Gg_{(t)}G^T + G(G^T[(A\nabla Y_{(t)}A^T) \odot (B^T d_{(t)}B)])G^T, \end{aligned} \quad (15)$$

where W_{wino} denotes weight in Winograd form and t denotes iteration. We should notice that $GG^T \neq E$, so we keep weight W in the normal form (or called 'spatial domain') instead of Winograd form (or called 'frequency domain') during the training process in order to make DWM be the equivalent substitute of normal convolution. Considering DWM, the two techniques mentioned above, we can derive the corresponding back propagation rules. Denoting output as Y , the two techniques have the same form in the aggregation step:

$$Y = \sum_i Y^{(i)}. \quad (16)$$

From the derivative rules we know that for part j :

$$\begin{aligned} \nabla Y^{(j)} &= \nabla^{(j)}Y^{(j)} + \nabla^{(j)} \sum_{i \neq j} Y^{(i)} \\ &= \nabla^{(j)} \sum_i Y^{(i)} = \nabla^{(j)}Y. \end{aligned} \quad (17)$$

This means that we don't need to store each output $Y^{(i)}$ for the backward. Furthermore, it is easy to derive that in the splitting step, the backward acts as follows:

$$\nabla d = \sum \nabla d^{(i)}, \nabla g = \sum \nabla g^{(i)}, \quad (18)$$

where i denotes different parts produced by DWM.

Comparison and Discussion

We are not the first one who tries to solve the large kernel size problem that appeared during applying the Winograd algorithm. Lu et al. (Lu et al. 2017) tried to implement the Winograd algorithm on FPGAs, and they solved large kernel problems by padding the kernel. However, the padded elements will be filled with non-zero values after the Winograd transformation, which means that paddings will bring lots of extra calculations. For DWM, we find a way that precisely separates the convolution operations without padding. This method avoids any redundant floating-point multiplications during Winograd transformation, and thus achieves the best acceleration without any numerical accuracy loss.

Furthermore, with the rise of neural architecture search (NAS), convolutions with larger kernel sizes such as 5×5 or 7×7 have become more popular than ever. Networks like ProxylessNAS (Cai, Zhu, and Han 2018) show that different computation environments will breed different neural architectures. For example, when we want to implement our neural networks on GPUs rather than on mobile CPUs, large kernel sizes may be more suitable choices than small ones because of the computation parallelism of GPUs. From beginning to the end, the architecture of neural networks follows the most popular hardware, not the other way round. Hence, we believe that reducing the computation overhead of neural networks is worthy in any case.

Experiments

Setup

All the results were tested on NVIDIA V100 GPU. We have implemented DWM on TensorFlow (Abadi et al. 2016) and

Table 1: Mean squared error (MSE) between different acceleration algorithms and the FP64 result by running a forward convolution, i.e FP64 convolution result is the baseline. H/W means the size of featur map, and FP32/FP16 indicates doing convolution in FP32/FP16 format.

Kernel Size	H/W	Channel	Filters	FP32	FP16	Winograd FP32	Winograd FP16	DWM FP32	DWM FP16
3x3	14	256	256	2.21E-08	2.71E-04	5.24E-10	1.44E-03	5.32E-10	3.42E-02
3x3	28	128	128	1.11E-10	1.41E-04	1.43E-10	7.48E-04	1.47E-10	9.08E-03
5x5	14	256	256	1.05E-02	6.93E-04	7.14E-08	1.07E-01	1.47E-09	9.72E-02
5x5	28	128	128	3.15E-10	3.80E-04	2.00E-08	5.78E-02	4.33E-10	2.83E-02
7x7	14	256	256	6.13E-10	1.25E-03	1.47E-03	NaN	2.97E-09	1.97E-01
7x7	28	128	128	5.61E-10	7.16E-04	4.24E-04	NaN	8.86E-10	5.88E-02
9x9	14	256	256	9.90E-10	1.90E-03	5.31E-03	NaN	3.67E-09	2.36E-01
9x9	28	128	128	8.52E-10	1.14E-03	1.62E-03	NaN	1.18E-09	7.33E-02
11x11	14	256	256	1.47E-09	2.60E-03	1.48E-02	NaN	5.30E-09	3.46E-01
11x11	28	128	128	1.15E-09	1.63E-03	4.35E-03	NaN	1.81E-09	1.15E-01

Table 2: Top-1 accuracy, FLOPs and speedup of several acclerating algorithm on different networks. Origin means the original top-1 accuracy and FLOPs of networks.

Network	Origin		Winograd			DWM		
	Acc	GFLOPs	Acc	GFLOPs	speedup	Acc	GFLOPs	speedup
AlexNet (Krizhevsky, Sutskever, and Hinton 2012)	56.52	0.71	56.51	0.56	1.28	56.51	0.45	1.57
GoogLeNet (Szegedy et al. 2015)	69.79	1.51	69.79	0.97	1.55	69.77	0.92	1.65
Inception-V3 (Szegedy et al. 2016)	69.54	2.86	69.47	2.34	1.22	69.46	1.92	1.49
ResNet-152 (He et al. 2016)	78.31	11.62	78.31	8.78	1.32	78.31	8.60	1.35
DenseNet-161 (Huang et al. 2017)	77.14	7.88	77.13	6.32	1.25	77.12	6.23	1.26
ProxylessGPU (Cai, Zhu, and Han 2018)	75.08	0.49	74.77	0.49	1.01	75.06	0.47	1.05
ProxylessMobile (Cai, Zhu, and Han 2018)	74.59	0.35	74.47	0.34	1.01	74.57	0.33	1.06

PyTorch (Paszke et al. 2017). On both platforms, the implemented DWM performs like a plug-and-play operator, which makes it convenient to use during inference and training. We tested the numerical accuracy of the algorithms by doing convolution with the standard normal distribution random numbers on one single layer, and then calculate the mean squared error (MSE) with FP64 convolution results. The numerical accuracy on a single layer was tested on both two platforms, and the results were the same. For all single layer tests, the batch size is set to 256 and the layers are the same padded. We measured the accuracy and FLOPs of networks based on PyTorch. The traditional model architectures are gained from TorchVision, and the ProxylessNAS (Cai, Zhu, and Han 2018) analysis is based on the official PyTorch implementation. The Networks' accuracy was measured on ImageNet 2012 (Russakovsky et al. 2015).

Numerical Accuracy of Single Layer

We estimated the mean squared error (MSE) between several methods and the FP64 results by doing a forward convolution. The input signal and convolution weights are random numbers generated by standard normal distribution using Numpy, set seed 11. We assumed two application situations: larger feature map (set 28×28) with fewer channels (set 128) and smaller feature map (set 14×14) with more channels (set 256), which is consistent with the reality. As shown in Table 1, (1) DWM has better numerical accuracy than the traditional Winograd algorithm almost in all situations. It is obvious that the traditional Winograd algorithm faced with a serious numerical accuracy problem as

the kernel size grows up. When the kernel size is 7×7 or larger, the error of the FP32 Winograd algorithm approaches FP16's, which may cause accuracy problems. By contrast, DWM's numerical error stays at a low level, which is close to the result of FP32, meaning that it can be applied to all kinds of convolution operations without any problem. (2) When using FP16, the traditional Winograd algorithm may get an overflow. This may be caused by the intermediate results of the Winograd transformation. When kernel size becomes larger, the transformation matrices will be filled with large numbers, and these numbers may make the results of matrix multiplication become too large to be represented in FP16. This also shows the advantages of using DWM instead of the traditional Winograd algorithm. Furthermore, the FP16 DWM is implemented without any optimization, which means the numerical accuracy of it still can be improved. (3) In some situations, DWM FP32 is more accurate than FP32 ones. This is reasonable because DWM consumes fewer multiplications, which may make it has better numerical accuracy.

FLOPs Estimation on Single Layer

We calculated the FLOPs of convolutions with different kernel sizes and 2 kinds of stride. Due to the decimals in the transformation matrix in the traditional Winograd algorithm, the FLOPs caused by transformation cannot be ignored, and thus we only eliminate the influence of $0, \pm 1, \pm 2^n$ and $\pm \frac{1}{2^n}$ which can be easily implemented by shifting. As shown in Table 3, we can get the following results: (1) DWM cost less computation than the traditional Winograd algo-

Table 3: The speedup of several accelerating algorithms on different kinds of convolutions. We assume that the output size is fixed to 14×14 .

Kernel Size	Stride	Direct FLOPs	Winograd		DWM	
			FLOPs	speedup	FLOPs	speedup
3x3	1	1.76E+03	784	2.25	784	2.25
5x5	1	4.90E+03	1.48E+04	0.33	2.40E+03	2.04
7x7	1	9.60E+03	9.72E+04	0.10	4.90E+03	1.96
9x9	1	1.59E+04	3.16E+05	0.05	7.06E+03	2.25
11x11	1	2.37E+04	1.07E+06	0.02	1.10E+04	2.15
3x3	2	1.76E+03	N/A	N/A	1.23E+03	1.44
5x5	2	4.90E+03	N/A	N/A	2.40E+03	2.04
7x7	2	9.60E+03	N/A	N/A	4.90E+03	1.96
9x9	2	1.59E+04	N/A	N/A	8.28E+03	1.92
11x11	2	2.37E+04	N/A	N/A	1.10E+04	2.15

Table 4: The actual runtime of several accelerating algorithms on different kinds of convolutions tested by nvprof. The batch size, the channels and the filters are 256. The input size is fixed to 14×14 .

Kernel Size	DWM(ms)	Winograd(ms)	cuDNN(ms)	Wino/DWM	cuDNN/DWM
3x3	3.67	3.35	2.80	0.91	0.76
5x5	11.37	69.70	11.26	6.13	0.99
7x7	22.83	133.34	24.51	5.84	1.07
9x9	30.67	248.71	50.92	8.11	1.66
11x11	48.29	349.33	94.63	7.23	1.96

rithm in all situations. As the kernel size becomes larger, the FLOPs of traditional Winograd algorithm increases heavily, and most FLOPs concentrates on the Winograd transformation because the transformation process becomes a non-sparsity matrix multiplication. On the contrary, the speedup of DWM keeps steady for its simple transformation matrices. (2) DWM speeds up stride 2 convolutions by $2 \times$, which cannot be achieved by the traditional method. Not surprisingly, due to the splitting method of DWM, the speedup of stride 2 convolution still holds stably. These advantages lead to stably speedup on almost all kinds of convolutions.

We also tested the actual runtime of convolution operations based on naive implementations of different kinds of convolutions. The result was tested by nvprof, an NVIDIA profiling tool. According to Table 4, we can conclude that: (1) DWM outperforms the original Winograd algorithm, especially in large kernel situation. (2) DWM is faster than cuDNN in some situations. Actually, when the size of the feature map increases to around 100, cuDNN performs better than DWM (not presented due to the space left). The instability of cuDNN may be caused by some other accelerating algorithms. Hence, both DWM and cuDNN have their advantages. Furthermore, the naive DWM implementation can be optimized by kernel fusion, soft pipeline and some other optimization techniques unlike the simple traditional 3×3 Winograd algorithm. Since we mainly focuses on speed up Winograd through the algorithm aspect in this paper, the implementation optimization of DWM is our future work.

Total Analysis on Networks

Finally, we analyzed several representative networks. This analysis includes the comparison of total FLOPs of the network and the top-1 accuracy of inference on the ImageNet

2012 dataset. The top-1 accuracy of the two accelerating algorithms is tested using FP16. As shown in Table 2, we can conclude that: (1) the top-1 acc is very close. This result is not surprising, because Table 1 shows that only large kernel size will make DWM and the traditional Winograd algorithm different. Thus, the networks most of which consist of 3×3 kernel convolutions like will get similar inference results. (2) When there are more large kernels in the network such as 11×11 in AlexNet and $1 \times 5/7$ or $5/7 \times 1$ in GoogLeNet and Inception-v3, DWM performs better.

However, in some cases such as ResNet-152, most of the computation is produced by 3×3 kernels. Worse, when it comes to some modern NAS architectures, the calculation is concentrated on 1×1 kernels because of the separable convolution. Over 90% amount of calculation of convolutional neural networks is caused by convolution operations. Hence, most of the architectures are designed or searched based on separable convolutions to reduce the amount of calculation. Although separable convolutions can cut the FLOPs down effectively, it reduces the representation of the original neural networks. When the computing power grows up and more fast convolution algorithms are invented, FLOPs will not be the main consideration of architecture designing.

Conclusion

In this paper, we propose a novel DWM to extend Winograd’s minimal filtering algorithm to a wide and general convolutions. Winograd’s minimal filtering algorithm has been widely used to reduce the number of multiplications for faster processing. However, it has the drawbacks of suffering from significantly increased FLOPs and numerical accuracy problem for kernel size larger than 3×3 and failing on convolution with stride larger than 1, so it is only effec-

tive on convolutions with kernel size as 3x3 and stride as 1. To solve this problems, we propose DWM to break through the limitation of original Winograd’s minimal filtering algorithm on convolutions of large kernel and large stride. DWM decomposes kernels with large size or large stride to several small kernels with stride as 1 for further applying Winograd method, so that DWM can reduce the number of multiplications while keeping the numerical accuracy. Experimental results show that the proposed DWM is able to support all kinds of convolutions with a speedup of ~ 2 , without affecting the numerical accuracy. These good properties of DWM enables the fast exploring of larger kernel size and larger stride value in CNNs for high performance, accuracy and even the potential for new CNNs.

Acknowledgments

This work is partially supported by the National Key Research and Development Program of China (under Grant 2017YFA0700902), the NSF of China (under Grants 61432016, 61532016, 61672491, 61602441, 61602446, 61732002, 61702478, 61732007, 61732020 and 61906179), Beijing Natural Science Foundation (JQ18013), the 973 Program of China (under Grant 2015CB358800), National Science and Technology Major Project (2018ZX01031102), the Transformation and Transfer of Scientific and Technological Achievements of Chinese Academy of Sciences (KFJ-HGZX-013), Key Research Projects in Frontier Science of Chinese Academy of Sciences (QYZDB-SSW-JSC001), Strategic Priority Research Program of Chinese Academy of Science (XDB32050200, XDC01020000) and Standardization Research Project of Chinese Academy of Sciences (BZ201800001).

References

- Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283.
- Barabasz, B., and Gregg, D. 2019. Winograd convolution for dnns: Beyond linear polinomials. *arXiv preprint arXiv:1905.05233*.
- Budden, D.; Matveev, A.; Santurkar, S.; Chaudhuri, S. R.; and Shavit, N. 2017. Deep tensor convolution on multicores. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 615–624. JMLR. org.
- Cai, H.; Zhu, L.; and Han, S. 2018. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*.
- Chetlur, S.; Woolley, C.; Vandermersch, P.; Cohen, J.; Tran, J.; Catanzaro, B.; and Shelhamer, E. 2014. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*.
- Choi, Y.; El-Khany, M.; and Lee, J. 2018. Compression of deep convolutional neural networks under joint sparsity constraints. *arXiv preprint arXiv:1805.08303*.
- Cong, J., and Xiao, B. 2014. Minimizing computation in convolutional neural networks. In *International conference on artificial neural networks*, 281–290. Springer.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.
- Lavin, A., and Gray, S. 2016. Fast algorithms for convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4013–4021.
- Li, S.; Park, J.; and Tang, P. T. P. 2017. Enabling sparse winograd convolution by native pruning. *arXiv preprint arXiv:1702.08597*.
- Liu, X.; Pool, J.; Han, S.; and Dally, W. J. 2018. Efficient sparse-winograd convolutional neural networks. *arXiv preprint arXiv:1802.06367*.
- Lu, L.; Liang, Y.; Xiao, Q.; and Yan, S. 2017. Evaluating fast algorithms for convolutional neural networks on fpgas. In *2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 101–108. IEEE.
- Madisetti, V. 1997. *The digital signal processing handbook*. CRC press.
- Maji, P.; Mundy, A.; Dasika, G.; Beu, J.; Mattina, M.; and Mullins, R. 2019. Efficient winograd or cook-toom convolution kernel implementation on widely used mobile cpus. *arXiv preprint arXiv:1903.01521*.
- Meng, L., and Brothers, J. 2019. Efficient winograd convolution via integer arithmetic. *arXiv preprint arXiv:1901.01965*.
- Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.
- Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.
- Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.
- Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.
- Vasilache, N.; Johnson, J.; Mathieu, M.; Chintala, S.; Piantino, S.; and LeCun, Y. 2014. Fast convolutional nets with fbfft: A gpu performance evaluation. *arXiv preprint arXiv:1412.7580*.
- Vincent, K.; Stephano, K.; Frumkin, M.; Ginsburg, B.; and Demouth, J. 2017. On improving the numerical stability of winograd convolutions.
- Winograd, S. 1980. *Arithmetic complexity of computations*, volume 33. Siam.