

Fast Walsh-Hadamard Transform and Smooth-Thresholding Based Binary Layers in Deep Neural Networks

Hongyi Pan

Diaa Badawi

Ahmet Enis Cetin

Department of Electrical and Computer Engineering

University of Illinois at Chicago

{hpan21, dbadaw2, aecyy}@uic.edu

Abstract

In this paper, we propose a novel layer based on fast Walsh-Hadamard transform (WHT) and smooth-thresholding to replace 1×1 convolution layers in deep neural networks. In the WHT domain, we denoise the transform domain coefficients using the new smooth-thresholding non-linearity, a smoothed version of the well-known soft-thresholding operator. We also introduce a family of multiplication-free operators from the basic 2×2 Hadamard transform to implement 3×3 depthwise separable convolution layers. Using these two types of layers, we replace the bottleneck layers in MobileNet-V2 to reduce the network’s number of parameters with a slight loss in accuracy. For example, by replacing the final third bottleneck layers, we reduce the number of parameters from 2.270M to 947K. This reduces the accuracy from 95.21% to 92.88% on the CIFAR-10 dataset. Our approach significantly improves the speed of data processing. The fast Walsh-Hadamard transform has a computational complexity of $O(m \log_2 m)$. As a result, it is computationally more efficient than the 1×1 convolution layer. The fast Walsh-Hadamard layer processes a tensor in $\mathbb{R}^{10 \times 32 \times 32 \times 1024}$ about 2 times faster than 1×1 convolution layer on NVIDIA Jetson Nano computer board.

1. Introduction

Deep convolution neural networks (CNN) are universally used in a wide range of applications including image classification [1, 2, 3, 4, 5, 6, 7, 8, 9], object detection [10, 11, 12, 13] and semantic segmentation [14, 15, 16, 17, 18]. On the other hand, implementing deep neural networks in real-time resource-constrained environments such as embedded devices is very difficult due to insufficient memory and limited computational capacity. For these reasons, there is an increasing demand to develop smaller and ef-

ficient neural networks as neural networks have become an instrumental technology for sensor data processing [19] and can be utilized in our modern society everywhere.

Efficient neural network models include compressing a large neural network using quantization [20, 21], hashing [22], pruning [23], vector quantization [24] and Huffman encoding [25]. Another approach is the SqueezeNet [26], which is designed as a small network with 1×1 filters during training. In [27] the neural network is slimmed by removing some layers from a well-developed model. Binary neural networks [28, 29, 30, 31, 32, 33, 34, 35, 36, 37] showed that binary weights can be used to slim and accelerate neural networks.

Although 1×1 convolutions reduce the computational load, they are still computationally expensive and time-consuming in regular deep neural networks. In this paper, we introduce a binary layer based on the fast Walsh-Hadamard transform to slim and speed up deep neural networks with 1×1 convolutions. Moreover, the recent literature [38, 39, 40, 41, 42] developed an energy-efficient neuron called multiplication-free (MF) kernel, which does not require any multiplications. We establish the relation between the MF operator and 2-by-2 Hadamard transform, and we fuse this idea to propose the depthwise separable multiplication-free convolution layer. We apply these two types of layers in bottleneck layers of MobileNet-V2 [43], and the new network is remarkably more slimmed and computationally efficient compared to the original structure according to our experiments. Similar results can also be obtained in other standard deep neural networks.

2. Methodology

In this section, we first review the Walsh-Hadamard (WH) transform. Then, we introduce the proposed fast Walsh-Hadamard transform layer. Finally, we describe the multiplication-free depthwise separable convolution layer, which turns out to be a combination of two binary layers.

2.1. The Walsh-Hadamard (WH) Transform

The WH transform is based on the so-called butterfly operation described by the following binary matrix:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad (1)$$

which is also the main building block of the fast Fourier transform (FFT). A 2-by-2 WH transform \mathbf{Y} of the vector $\mathbf{X} \in \mathbb{R}^2$ is

$$\mathbf{Y} = \mathbf{W}\mathbf{X} = \mathbf{H}\mathbf{X} \quad (2)$$

In general, the WH transform $\mathbf{Y} = \mathbf{W}_k\mathbf{X}$ of a vector $\mathbf{X} \in \mathbb{R}^m$ where $m = 2^k, k \in \mathbb{N}$ can be expressed via the orthogonal Walsh matrix $\mathbf{W}_k \in \mathbb{R}^{m \times m}$ which is generated using the Hadamard matrix which can be recursively constructed in two steps as follows [44]:

- First, we construct the Hadamard matrix \mathbf{H}_k :

$$\mathbf{H}_k = \begin{cases} 1, & k = 0, \\ \begin{bmatrix} \mathbf{H}_{k-1} & \mathbf{H}_{k-1} \\ \mathbf{H}_{k-1} & -\mathbf{H}_{k-1} \end{bmatrix}, & k > 0, \end{cases} \quad (3)$$

Alternatively, for $k > 1$, \mathbf{H}_k can also be computed as

$$\mathbf{H}_k = \mathbf{H}_1 \otimes \mathbf{H}_{k-1} \quad (4)$$

where \otimes denotes Kronecker product.

- Then, we shuffle the rows of \mathbf{H}_k to obtain \mathbf{W}_k by applying the bit-reversal permutation and the Gray-code permutation on row index. For example, when $k = 2, m = 4$, we have the Hadamard matrix

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \quad (5)$$

and the corresponding WH-transform is given by

$$\mathbf{W}_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \quad (6)$$

It is not difficult to verify that $\mathbf{X} = \frac{1}{m}\mathbf{W}_k\mathbf{Y}$, which implies the inverse Walsh-Hadamard transform is itself with normalization by m .

The fast Walsh-Hadamard transform (FWHT) algorithm is similar to the FFT algorithm, and, consequently, the complexity of FWHT is $O(m \log_2 m)$. The FWHT algorithm is completely based on the butterfly operations described in Eq. (1) in [45]. As a result, the implementation of FWHT can be realized by only addition and subtraction operations using butterflies. It was shown that the WH transform is the same as the block Haar wavelet transform [46].

2.2. The FWHT Layer

In the state-of-art deep convolution neural networks, 1×1 convolution layers are widely used as hidden layers [5, 43, 47, 48] to change the dimensions of channels. For instance, there are 17 bottleneck layers in MobileNetV2 [43]. The first bottleneck layer contains a 3×3 depthwise separable convolution layer and a 1×1 convolution layer. Other bottleneck layers are built using a 1×1 convolution layer named ‘‘conv_expand’’, a 3×3 depthwise separable convolution layer [49], and a 1×1 convolution layer named ‘‘conv_project’’. The layer ‘‘conv_expand’’ increases the number of channels 6 times. On the other hand, the layer ‘‘conv_project’’ decreases the number of channels by a factor of $\frac{1}{6}$. The depthwise separable convolution layer is the main component for feature extraction.

Notice that each 1×1 convolution layer contains as many parameters as the number of channels, and they are time-consuming during inference. We propose a novel FWHT layer to replace the 1×1 convolution layers. The FWHT layer is summarized in Algorithms 1 and 2. In general, an FWHT layer consists of an FWHT to change the tensor to the Hadamard domain, a smooth-thresholding operation as the non-linearity in the Hadamard domain, and an FWHT to change the tensor back to the feature-map domain. The FWHT is performed on the channel axis, which implies that completing FWHT on a tensor $\mathbf{X} \in \mathbb{R}^{n \times w \times h \times m}$ means performing $n \times w \times h$ m -length FWHTs in parallel.

In the WH transform domain, we denoise the parameters and eliminate coefficients with small amplitude values. Soft-thresholding (ST) is commonly used in wavelet domain denoising algorithms [50, 51] and defined as

$$y = S_T(x) = \text{sign}(x)(|x| - T)_+ = \begin{cases} x + T, & x < -T \\ 0, & |x| \leq T \\ x - T, & x > T \end{cases} \quad (7)$$

where T is the thresholding parameter and is trainable in our FWHT layer.

The denoising parameter T can be learned using the back-propagation algorithm during training. More specifically,

$$\frac{\partial(\text{sign}(x)(|x| - T)_+)}{\partial T} = \begin{cases} 1, & x < -T \\ 0, & |x| \leq T \\ -1, & x > T \end{cases} \quad (8)$$

However, the denoising parameter T in soft-thresholding can only be updated by $+1$ or -1 . We introduce smooth-thresholding by revising the definition of the soft-thresholding operator as follows

$$y = S'_T(x) = \tanh(x)(|x| - T)_+ \quad (9)$$

and the corresponding partial derivative is

$$\frac{\partial(\tanh(x)(|x| - T)_+)}{\partial T} = \begin{cases} -\tanh(x), & |x| > T \\ 0, & |x| \leq T \end{cases} \quad (10)$$

We observe that the derivative in Eq. (10) is the derivative in Eq. (8) multiplied by $\tanh(x)$. As a result, the convergence of the smooth-thresholding operator is smooth and steady in the back-propagation algorithm. We learn a different threshold T value for each WH transform domain coefficient. Other methods using the Hadamard transform include [52, 53] but they did not apply trainable thresholding in the WH-transform domain.

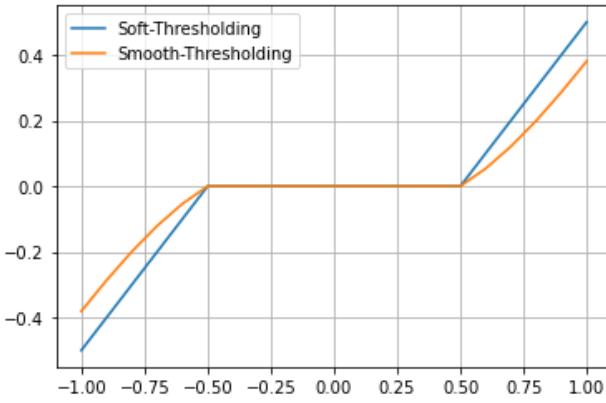


Figure 1: $y = S_T(x, 0.5)$ and $y = S'_T(x, 0.5)$

Figure 1 shows the soft-thresholding and smooth-thresholding operators for $T = 0.5$, respectively. The FWH transform coefficients can take both positive and negative values. Large positive and negative transform domain coefficients are equally important. Therefore, we cannot use the ReLU function in the transform domain. In fact, our experiments verify this observation, and soft-thresholding and smooth-thresholding improve the recognition accuracy compared to the ReLU.

There are $(2^d - 1)$ trainable thresholding parameters in the FWHT layer when there are 2^d channels. This means that if we perform smooth-thresholding on tensor \mathbf{Y} , each slice among the channel axis will share a common threshold value. The so-called DC channel $\mathbf{Y}[:, :, :, 0]$ usually contains essential information. Due to this reason, we do not perform smooth-thresholding on this channel.

The FWHT layer for channel expansion is described in Algorithm 1. We first compute the 2^d point WH transforms and perform smooth-thresholding in the transform domain. We pad $(2^d - c)$ zeros to the end of each input vector before the 2^d -by- 2^d WH transform to increase the dimension. After smooth-thresholding in the WH domain, we calculate the inverse WH transform.

Algorithm 1 The FWHT layer for channel expansion

Input: Input tensor $\mathbf{X} \in \mathbb{R}^{n \times w \times h \times c}$

Output: Output tensor $\mathbf{Z} \in \mathbb{R}^{n \times w \times h \times tc}$

- 1: Find minimum $d \in \mathbb{N}$, s.t. $2^d \geq tc$
- 2: $\hat{\mathbf{X}} = \text{pad}(\mathbf{X}, 2^d - c) \in \mathbb{R}^{n \times w \times h \times 2^d}$
- 3: $\mathbf{Y} = \text{FWHT}(\hat{\mathbf{X}}) \in \mathbb{R}^{n \times w \times h \times 2^d}$
- 4: $\hat{\mathbf{Y}} = \text{concat}(\mathbf{Y}[:, :, :, 0], \text{ST}(\mathbf{Y}[:, :, :, 1 :]))$
- 5: $\hat{\mathbf{Z}} = \text{FWHT}(\hat{\mathbf{Y}}) \in \mathbb{R}^{n \times w \times h \times 2^d}$
- 6: $\mathbf{Z} = \hat{\mathbf{Z}}[:, :, :, tc]$

7: **return** \mathbf{Z} .

Comments: Function $\text{pad}(\mathbf{A}, b)$ pads b zeros on the channel axis of tensor \mathbf{A} . $\text{FWHT}(\cdot)$ is the normalized fast Walsh-Hadamard transform on the last axis. Function $\text{concat}(\cdot, \cdot)$ concatenates two tensors along the last axis. $\text{ST}(\cdot)$ performs smooth-thresholding. Index follows Python's rule.

Algorithm 2 The FWHT layer for channel projection

Input: Input tensor $\mathbf{X} \in \mathbb{R}^{n \times w \times h \times tc}$

Output: Output tensor $\mathbf{Z} \in \mathbb{R}^{n \times w \times h \times c}$

- 1: Find minimum $p, q \in \mathbb{N}$, s.t. $2^p \geq tc, 2^q \geq c$
- 2: $r = 2^{p-q}$
- 3: $\hat{\mathbf{X}} = \text{pad}(\mathbf{X}, 2^p - tc) \in \mathbb{R}^{n \times w \times h \times 2^p}$
- 4: $\mathbf{Y} = \text{FWHT}(\hat{\mathbf{X}}) \in \mathbb{R}^{n \times w \times h \times 2^p}$
- 5: $\hat{\mathbf{Y}} = \text{concat}(\mathbf{Y}[:, :, :, 0]/r, \text{avgpool}(\text{ST}(\mathbf{Y}[:, :, :, 1 : 2^p - r + 1]), r)) \in \mathbb{R}^{n \times w \times h \times 2^q}$
- 6: $\hat{\mathbf{Z}} = \text{FWHT}(\hat{\mathbf{Y}}) \in \mathbb{R}^{n \times w \times h \times 2^q}$
- 7: $\mathbf{Z} = \hat{\mathbf{Z}}[:, :, :, c]$

8: **return** \mathbf{Z} .

Comments: Function $\text{pad}(\mathbf{A}, b)$ pads b zeros on the channel axis of tensor \mathbf{A} . $\text{FWHT}(\cdot)$ is the normalized fast Walsh-Hadamard transform on the last axis. Function $\text{concat}(\cdot, \cdot)$ concatenates two tensors along the last axis. Function $\text{avgpool}(\mathbf{A}, b)$ is the average pooling on \mathbf{A} with pooling size and strides are b . $\text{ST}(\cdot)$ performs smooth-thresholding. Index follows Python's rule.

In FWHT layer for channel projection by a factor of $r = \frac{2^p}{2^q}$, we first compute the 2^p point WH transforms and perform smooth-thresholding in the transform domain as described in Algorithm 2. After this step, we compute the 2^q point WH transforms to reduce the dimension of the feature map. We divide the DC channel values by r to keep the energy at the same level as other channels after pooling. In Step 5 of Algorithm 2, we average pool the transform domain coefficients to reduce the dimension of the Walsh-Hadamard transform and discard the last $(r - 1)$ transform domain coefficients of \mathbf{Y} to make the dimension equal to 2^q . The last $(r - 1)$ coefficients are high-frequency coefficients, and usually, their amplitudes are negligible com-

pared to other WH coefficients.

Therefore, the dimension change operation from m dimensions to n dimensions can be summarized as follows:

$$\mathbf{Z} = \begin{cases} \frac{1}{2^q} \mathbf{U} \mathbf{W}_q \mathbf{S}'_{\mathbf{T}} \mathbf{W}_q \mathbf{P} \mathbf{X}, & m \leq n \\ \frac{1}{\sqrt{2^{p+q}}} \mathbf{U} \mathbf{W}_q \mathbf{A}_{\text{vg}} \mathbf{S}'_{\mathbf{T}} \mathbf{W}_p \mathbf{P} \mathbf{X}, & m > n \end{cases} \quad (11)$$

where p is the minimum integer such that $2^p \geq m$, q is the minimum integer such that $2^q \geq n$, \mathbf{P} describes zero-padding operation to make \mathbf{X} multipliable by \mathbf{W}_p or \mathbf{W}_q , $\mathbf{S}'_{\mathbf{T}}$ is the smooth-thresholding layer with DC channel excluded, \mathbf{U} is unpadding function to make the dimension the same as \mathbf{Z} , and \mathbf{A}_{vg} is average pooling on the channel axis with the DC channel of WH transform excluded.

Hence, the trainable number of parameters in FWHT layers is no more than the $(2^d - 1)$ where d is the minimum integer such that 2^d is no less than the number of input channels. The trainable parameters are only the threshold values of the smooth-thresholding. Therefore, it is clear that FWHT layer requires significantly fewer parameters than the regular 1×1 layer, which requires a different set of filter coefficients for each 1×1 convolution.

2.3. Multiplication-Free Depthwise Separable "Convolutions" (MF-DS-Conv)

Akbaş *et. al.* introduced multiplication-free (MF) kernel to replace regular convolution in CNNs [38, 39, 40, 41, 42]. The MF kernel requires no multiplication but only additions and sign operations. It is more energy-efficient compared to the standard convolution. In this section, we describe how we can generate new operators similar to the MF kernel from the 2×2 Walsh-Hadamard transform to implement depthwise separable convolution using binary operations, additions, and subtractions.

Let w be a weight value and x be the corresponding input. In a regular neuron we multiply x by w to determine the contribution of the input value. In some binary networks and additive networks [38, 39, 40, 41, 42], the sign of the multiplication $w \times x$ is used as the basic operation of the neuron. In addition, we will scale the $\text{sign}(w \times x)$ by the ℓ_1 norm of the vector $y = [w + x \ w - x]^T$ which is the 2×2 WH transform of the vector $[w \ x]^T$ and define the following basic operation

$$\begin{aligned} w \odot x &= \text{sign}(w \times x)(|w + x| + |w - x|) \\ &= 2\text{sign}(w \times x) \max\{|w|, |x|\} \end{aligned} \quad (12)$$

which can be used to construct a vector "product". The operator can also replace the multiplication operator used in correlation or convolution calculations. Another related operator is

$$\text{sign}(w \times x)(|w + x| - |w - x|) = 2\text{sign}(w \times x) \min\{|w|, |x|\} \quad (13)$$

which can be also used in convolution and correlation operations. The above operations defined in Eq. (12) and (13) are very similar to the MF operation used in [38, 39, 40, 41, 42], which is

$$w \oplus x = \text{sign}(w \times x)(|w| + |x|) \quad (14)$$

What is common in all of the above operations is that the result of the operation $w \oplus x$ has the same sign as the multiplication operation, and the magnitude of the output is determined either by addition, min, or max operations. As a result, we can use all of the above three operations defined in Eq. (12)- (14) in vector product, correlation, and convolution operations as follows: Let $\mathbf{w} = [w_1 \dots w_D]^T \in \mathbb{R}^{D \times 1}$ and $\mathbf{x} = [x_1 \dots x_D]^T \in \mathbb{R}^{D \times 1}$ be two D -dimensional column vectors. Instead of performing the standard Euclidean inner product $\langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^T \mathbf{x} \triangleq \sum_{i=1}^D w_i x_i$, the MF dot product that is defined as

$$\mathbf{w}^T \oplus \mathbf{x} \triangleq \sum_{i=1}^D \text{sign}(w_i x_i)(|w_i| + |x_i|) \quad (15)$$

$$= \sum_{i=1}^D \text{sign}(w_i) x_i + w_i \text{sign}(x_i) \quad (16)$$

is used in neurons and convolutions. Vector "products" using Eq. (12) and (13) can be also defined in a similar manner.

Notice that the only multiplication operations that appear in Eq. (16) correspond to sign changes and can be implemented with very low complexity. For this reason, we do not count the sign changes towards multiplication operations and thus call Eq. (16) an MF dot product. It can easily be verified that the product in Eq. (16) and operators (12) and (13) induce a scaled version of ℓ_1 -norm as

$$\mathbf{x}^T \oplus \mathbf{x} = \sum_{i=1}^n |x_i| + |x_i| = 2\|\mathbf{x}\|_1 \quad (17)$$

Vector dot products described above can be extended to matrix multiplications as follows: Let $\mathbf{W} \in \mathbb{R}^{n \times m}$ and $\mathbf{X} \in \mathbb{R}^{n \times p}$ be arbitrary matrices. Then we define

$$\mathbf{W}^T \oplus \mathbf{X} \triangleq \begin{bmatrix} \mathbf{w}_1^T \oplus \mathbf{x}_1 & \mathbf{w}_1^T \oplus \mathbf{x}_2 & \dots & \mathbf{w}_1^T \oplus \mathbf{x}_p \\ \mathbf{w}_2^T \oplus \mathbf{x}_1 & \mathbf{w}_2^T \oplus \mathbf{x}_2 & \dots & \mathbf{w}_2^T \oplus \mathbf{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{w}_m^T \oplus \mathbf{x}_1 & \mathbf{w}_m^T \oplus \mathbf{x}_2 & \dots & \mathbf{w}_m^T \oplus \mathbf{x}_p \end{bmatrix} \quad (18)$$

where \mathbf{w}_i is the i th column of \mathbf{W} for $i = 1, 2, \dots, m$ and \mathbf{x}_j is the j th column of \mathbf{X} for $j = 1, 2, \dots, p$. In brief, the definition is similar to the matrix production $\mathbf{W}^T \mathbf{X}$ by only changing the element-wise product to element-wise

MF-product. By replacing product operation by MF product operation in depthwise separable convolution layer, we achieve MF-DS-Conv.

The standard back-propagation algorithm can be used for training the MF-DS-Conv layer with the need for small approximations. The partial scalar derivatives of the pre-activation response concerning are given as follows:

$$\frac{\partial(w \oplus x)}{\partial x} = \text{sign}(w) + 2w\delta(x) \quad (19)$$

$$\frac{\partial(w \oplus x)}{\partial w} = 2x\delta(w) + \text{sign}(x) \quad (20)$$

where $\delta(x)$ is the Dirac-delta function that directly results from the discontinuity of the signum function at $x = 0$. If we omit the delta function from the definitions of the partial derivatives, we end up with binary derivatives ($\text{sign}(w)$ and $\text{sign}(x)$). However, approximating the Dirac-delta function provides better convergence since we end up with smoother derivatives. In this regard, we approximate the derivative of the signum function to be that of a steep hyperbolic tangent, as follows:

$$\frac{d\text{sign}(x)}{dx} \approx \frac{dtanh(\alpha x)}{dx} = \alpha(1 - \tanh^2(\alpha x)) \quad (21)$$

for a scalar $\alpha \gg 1$. This is reasonable since $\text{sign}(x) = \lim_{\alpha \rightarrow \infty} \tanh(\alpha x)$. This way the terms associated with the delta function in the Eq. (19) and (20) will contribute to the partial derivatives when the arguments are close to zero. In our experiments, we choose $\alpha = 10$.

By replacing two 1×1 convolution layers by FWHT layers and the depthwise separable 3×3 convolution layer by a 3×3 MF-DS-Conv layer, we implement “multiplication-free” bottleneck layer. Although it is not wholly multiplication-free because there are batch-normalization layers between the convolution layers, it is much more efficient and slimmer than the original bottleneck layer. On the other hand, as it is shown in Table 1 because the input and the output of each layer have the same shape as the original layer, our layers are plug-and-play.

Input	Operator	Output
$h \times w \times k$	FWHT expansion, ReLU6	$h \times w \times tk$
$h \times w \times tk$	3×3 MF-DS-Conv, ReLU6	$\frac{h}{s} \times \frac{w}{s} \times tk$
$\frac{h}{s} \times \frac{w}{s} \times tk$	FWHT projection, ReLU6	$\frac{h}{s} \times \frac{w}{s} \times k'$

Table 1: Revised bottleneck residual block transforming from k to k' channels, with stride s , and expansion factor t . Batch normalization is used after each layer. Input and output are same as Table 1 in [43].

3. Experimental Results

In our experiments, training and accuracy tests proceed on a laptop with Intel Core i7-7700HQ CPU, NVIDIA GTX-1060 GPU with Max-Q design, and 16GB DDR4 RAM. Code is written in TensorFlow-Keras in Python 3. In the following sections, we will compare the accuracy of neural networks on the Fashion MNIST dataset and CIFAR-10 dataset, then compare the speed of neural networks by feeding a large tensor to a single layer.

In the accuracy test, we use the ImageNet-pretrained MobileNet-V2 model with 1.0 depth. As it is shown in Table 2, We replace layers after the last global average pooling layer by a dropout layer and a dense layer with 10 outputs as TensorFlow official transfer learning demo [54]. This is how we build the fine-tuned MobileNet-V2. In case the minimum input of MobileNet-V2 is $96 \times 96 \times 3$, we interpolate the images to this resolution, as shown in Figures 2 and 3. Because images in Fashion MNIST dataset are in gray-scale, after interpolating, we copy the values 2 times to make them $96 \times 96 \times 3$.

Input	Operator	t	c	n	s
$96^2 \times 3$	Conv2D	-	32	1	2
$48^2 \times 32$	Bottleneck	1	16	1	1
$48^2 \times 16$	Bottleneck	6	24	2	2
$24^2 \times 24$	Bottleneck	6	32	3	2
$12^2 \times 32$	Bottleneck	6	64	4	2
$6^2 \times 64$	Bottleneck	6	96	3	1
$6^2 \times 96$	Bottleneck	6	160	3	2
$3^2 \times 160$	Bottleneck	6	320	1	1
$3^2 \times 320$	Conv2D	-	1280	1	1
$3^2 \times 1280$	AvgPool	-	-	1	-
1280	Dropout (rate=0.2)	-	-	1	-
1280	Dense (units=10)	-	-	1	-

t , c , n and s represent expansion factor, channel, repeat time and stride [43]. Initial weights before Dropout are from ImageNet checkpoint float_v2_1.0_96 in [55].

Table 2: Structure of fine-tuned MobileNet-V2 (baseline)

We modify the MobileNet-V2 with methods described in Section 2. Detaillly, in Tables 3 and 4, we try to only change 1×1 convolution layers for channel projection with name “MobileNet-V2 with projection conv. changed” because 1×1 convolution layers for channel expansion plays a more important role than 1×1 convolution layers for channel projection in bottleneck layers for feature extraction, and we try to change 1×1 convolution layers for both channel expansion and projection with name “MobileNet-V2 with 1×1 conv. changed”, and change 1×1 convolution layers and 3×3 depthwise convolution layers with the name “MobileNet-V2 with bottleneck changed”. Note that

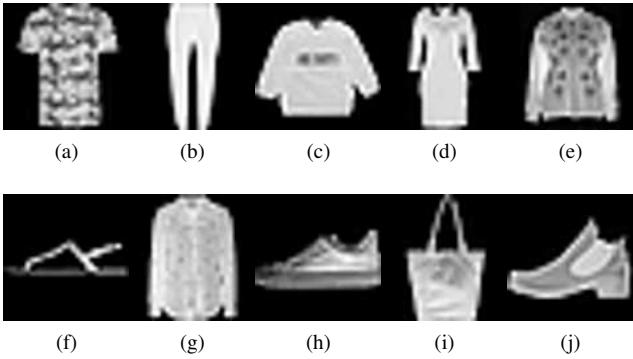


Figure 2: Samples of Fashion MNIST: (a) T-shirt/top, (b) trouser, (c) pullover, (d) dress, (e) coat, (f) sandal, (g) shirt, (h) sneaker, (i) bag, (j) ankle boot.

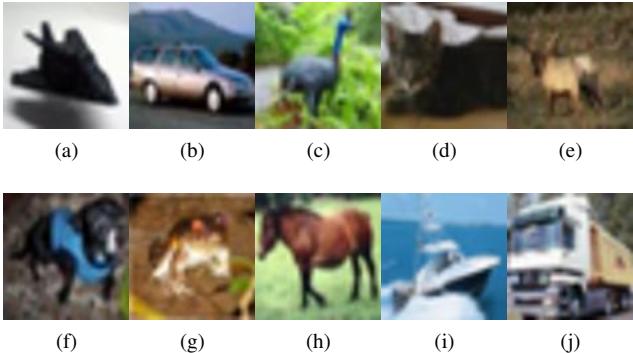


Figure 3: Samples of CIFAR-10: (a) airplane, (b) automobile, (c) bird, (d) cat, (e) deer, (f) dog, (g) frog, (h) horse, (i) ship, (j) truck.

early layers contain very few parameters, but they are much more important in feature extraction than the latter layers, especially for early 1×1 convolution layers for channel expansion, so we also try to keep some early layers, as changing these layer cannot save too many parameters but will bring a lot of accuracy loss. For example, if we only change 1×1 convolution layers for both channel expansion and projection in the latter half of bottleneck layers, we call it as ‘‘MobileNet-V2 with 1/2 1×1 conv. changed’’.

3.1. Fashion MNIST Experiments

We train the networks with the same parameter setting (SGD with the learning rate = 0.005, momentum = 0.9 on 100 epochs with batch size = 64) for a fair comparison and record the best accuracy training phase. We take categorical cross-entropy without label smoothing as the loss function. We horizontally flip the training images to enrich the

dataset. The results of this experiment are summarized in Table 3. By replacing about a half of 1×1 convolution layers with FWHT layers, we can save 67.82% parameters with 0.98% accuracy loss. Besides, by also replacing about half of 3×3 depthwise separable convolution layers by MF-DS-Conv, we only lose an additional 0.05% accuracy.

Moreover, to show the advantage of smooth-thresholding over the soft-thresholding, ReLU ($y = \text{ReLU}(x - T)$) and the linear identity function ($y = x$), we performed a set of experiments in the FWHT layers. Accuracy history is shown in Figure 4. We only changed the non-linearity in the WH transform domain, and everything else (even the random seed) is identical. If we do not use any function, the network has a very rough converge behavior, as shown in the green curve in Figure 4. With ReLU, MobileNet-V2 with the half of 1×1 convolutions changed reaches an accuracy of 0.24% less than the soft-thresholding-based approach. This is because ReLU is only the non-negative part of the soft-thresholding operator. The positive values are as important as negative values in the WH transform domain. Because of these reasons, soft- and smooth-thresholding can retain more information than ReLU in FWHT layers. On the other hand, with smooth-thresholding, MobileNet-V2 with half of the bottleneck changed reaches a 0.18% higher accuracy than the soft-thresholding due to the fact that the derivative of the soft-thresholding operator can only be either 1, 0, or -1.

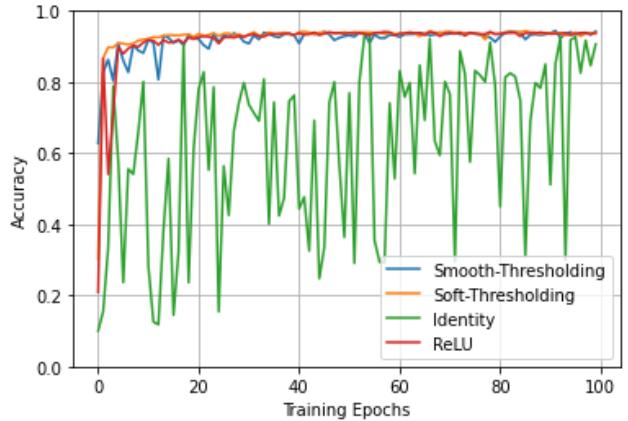


Figure 4: Accuracy history of networks with different threshold functions in the FWHT layers.

3.2. CIFAR-10 Experiments

We train the networks with the same parameter setting (SGD with learning rate = 0.005, momentum = 0.9, on 150 epochs with the batch size = 128) for a fair comparison and record the best accuracy in the training phase. We take categorical cross-entropy without label smoothing as the

Method	Threshold	Parameters	Saving Rate	Accuracy
Fine-tuned MobileNet-V2 (baseline)	-	2,270,794	-	95.37%
MobileNet-V2 with all projection conv. changed	Soft	1,317,126	42.00%	94.46%
MobileNet-V2 with all 1×1 conv. changed	Soft	574,838	74.69%	91.49%
MobileNet-V2 with 2/3 1×1 conv. changed	Soft	616,449	72.85%	93.90%
MobileNet-V2 with 1/2 1×1 conv. changed	Identity ^a	716,362	68.45%	93.45%
MobileNet-V2 with 1/2 1×1 conv. changed	ReLU ^b	730,648	67.82%	94.14%
MobileNet-V2 with 1/2 1×1 conv. changed	Soft	730,648	67.82%	94.38%
MobileNet-V2 with 1/2 1×1 conv. changed	Smooth	730,648	67.82%	94.49%
MobileNet-V2 with 1/2 bottleneck changed	Soft	730,648	67.82%	94.26%
MobileNet-V2 with 1/2 bottleneck changed	Smooth	730,648	67.82%	94.44%

^a is achieved without using any function at where smooth-thresholding is in FWHT layers. So there is no trainable variable. Although it reaches 93.45%, its accuracy plot does not converge like other cases.

^b is achieved with ReLU instead of smooth-thresholding in FWHT layers. Threshold values are also trainable.

Table 3: Results on Fashion MNIST Dataset

Method	Threshold	Parameters	Saving Rate	Accuracy
Baseline MobileNet-V2 model ^a in [27]	-	2.2378M	-	94.3%
RMNv2 ^b [27]	-	1.0691M	52.22%	92.4%
Fine-tuned MobileNet-V2 (baseline)	-	2,270,794	-	95.21%
MobileNet-V2 with all projection conv. changed	Soft	1,317,126	42.00%	87.22%
MobileNet-V2 with 1/2 projection conv. changed	Soft	1,399,328	38.38%	92.46%
MobileNet-V2 with 1/3 projection conv. changed	Soft	1,514,036	33.33%	94.28%
MobileNet-V2 with 1/2 1×1 conv. changed	Soft	730,648	67.82%	90.16%
MobileNet-V2 with 1/2 1×1 conv. changed	Smooth	730,648	67.82%	90.91%
MobileNet-V2 with 1/3 1×1 conv. changed	Soft	947,759	58.26%	93.21%
MobileNet-V2 with 1/3 1×1 conv. changed	Smooth	947,759	58.26%	93.23%
MobileNet-V2 with 1/2 bottleneck changed	Soft	730,648	67.82%	90.11%
MobileNet-V2 with 1/2 bottleneck changed	Smooth	730,648	67.82%	90.77%
MobileNet-V2 with 1/3 bottleneck changed	Soft	947,759	58.26%	92.47%
MobileNet-V2 with 1/3 bottleneck changed	Smooth	947,759	58.26%	92.88%

Parameters and accuracy of baseline model^a and RMNv2^b are from Table 3 in [27].

Table 4: Results on the CIFAR-10 dataset

loss function. We horizontally flip, rotate, shift, sheer and zoom the training images to enrich the dataset. By replacing about one-third of 1×1 convolution layers with FWHT layers, we can save 58.26% parameters with 1.98% accuracy loss. Our network with 947,759 parameters is more accurate (93.23%) than the comparable slim RMNv2 network (accuracy=92.4%) described in [27].

In addition, by also replacing about one-third of 3×3 depthwise separable convolution layers with MF-DS-Conv, we lose an additional 0.35% accuracy.

3.3. The Speed Test

We implemented the proposed network on an NVIDIA computer board to show the efficiency of the FWHT layer. As pointed in Section 2.2, the FWHT is an $O(m \log_2 m)$

operation. We compared the 1×1 convolution layer and the FWHT layer by processing a random tensor from $\mathbb{R}^{10 \times 32 \times 32 \times 1024}$ to $\mathbb{R}^{10 \times 32 \times 32 \times 1024}$. We also implemented the Walsh-Hadamard transform (WHT) using ordinary matrix multiplication. The only difference between the WHT layer and the FWHT layer is that the former employs matrix-vector-product-based Walsh-Hadamard transform instead of the $O(m \log_2 m)$ fast Walsh-Hadamard transform. The model uploaded to [56] only contains one layer (FWHT, WHT or 1×1 convolution). The speed tests are performed on the same laptop computer and an NVIDIA Jetson Nano (4GB RAM version). We run the TensorFlow PB model on the laptop, and we convert the PB file to TFLite format for NVIDIA Jetson Nano because it is an ARM-based edge device. The test result is stated in Ta-

ble 5. There is no doubt that the WHT layer runs the slowest because its complexity is $O(m^2)$. Although the FWHT layer’s speed advantage to 1×1 convolution layer is not obvious on the laptop computer, the FWHT layer runs about 2 times faster than 1×1 convolution layer on the NVIDIA Jetson Nano board. This is because the TFLite model is optimized by TensorFlow, while the PB model is not optimized as much as the TFLite model.

Layer	On Laptop	On NVIDIA Jetson Nano
1×1 conv.	0.0975s	2.3524s
WHT	0.1064s	5.1580s
FWHT	0.0952s	1.1973s

Table 5: Speed test, code is available at [56]

4. Conclusion

In this paper, we propose a binary layer called FWHT layer to replace 1×1 convolution layer based on the fast Walsh-Hadamard (WH) transform and smooth-thresholding to change the channel size in deep neural networks. The FWHT layer using an $m \times m$ WH matrix requires only N trainable parameters compared to 1×1 convolution layers because the only trainable parameters are the threshold values of the smooth-threshold, which is a tanh-smoothed version of the well-known soft-thresholding operator. The number of parameters m is equal to the channel number or a power of 2 integers larger than the number of channels, and the same threshold values can be used throughout the layer. The WH transform is implemented in $O(m \log_2 m)$ arithmetic. As a result, the FWHT layer runs about 2 times faster than the 1×1 convolution layer on NVIDIA Jetson Nano with TFLite model format.

We also establish the relationship between the MF operator and the 2-by-2 WH transform and proposed a depthwise separable multiplication-free convolution layer, a combination of two binary layers. By using the architecture of MobileNet-V2 with FWHT and MF layers, we reduce the number of parameters of MobileNet-V2 by 67.82% with 0.93% accuracy loss on the Fashion MNIST dataset and 58.26% savings with 2.33% accuracy loss on the CIFAR-10 dataset. Although we perform experiments using MobileNet-V2 to compare our results with [27], similar results can be obtained using other deep neural networks.

References

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. [1](#)
- [2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [1](#)
- [3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. [1](#)
- [4] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. [1](#)
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [1, 2](#)
- [6] Diaa Badawi, Hongyi Pan, Sinan Cem Cetin, and A Enis Çetin. Computationally efficient spatio-temporal dynamic texture recognition for volatile organic compound (voc) leakage detection in industrial plants. *IEEE Journal of Selected Topics in Signal Processing*, 14(4):676–687, 2020. [1](#)
- [7] Chirag Agarwal, Shahin Khobahi, Dan Schonfeld, and Mojtaba Soltanalian. Coronet: a deep network architecture for enhanced identification of covid-19 from chest x-ray images. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, page 1159722. International Society for Optics and Photonics, 2021. [1](#)
- [8] Harris Partaourides, Kostantinos Papadamou, Nicolas Kourtellis, Ilias Leontiades, and Sotirios Chatzis. A self-attentive emotion recognition network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7199–7203. IEEE, 2020. [1](#)
- [9] Dimitrios Stamoulis, Ting-Wu Chin, Anand Krishnan Prakash, Haocheng Fang, Sribhuvan Sajja, Mitchell Bognar, and Diana Marculescu. Designing adaptive neural networks for energy-constrained image classification. In *Proceedings of the International Conference on Computer-Aided Design*, pages 1–8, 2018. [1](#)
- [10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. [1](#)
- [11] Süleyman Aslan, Uğur Güdükbay, B Uğur Töreyin, and A Enis Çetin. Deep convolutional generative adversarial networks for flame detection in video. In *International Conference on Computational Collective Intelligence*, pages 807–815. Springer, 2020. [1](#)
- [12] Guglielmo Menchetti, Zhanli Chen, Diana J Wilkie, Rashid Ansari, Yasemin Yardimci, and A Enis Çetin. Pain detection from facial videos using two-stage deep learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019. [1](#)

[13] Süleyman Aslan, Uğur Güdükbay, B Uğur Töreyin, and A Enis Çetin. Early wildfire smoke detection based on motion-based geometric image transformation and deep convolutional generative adversarial networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8315–8319. IEEE, 2019. 1

[14] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 1

[15] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019. 1

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[17] Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*, 2019. 1

[18] Yinli Jin, Wenbang Hao, Ping Wang, and Jun Wang. Fast detection of traffic congestion from ultra-low frame rate image based on semantic segmentation. In *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 528–532. IEEE, 2019. 1

[19] Holger Schwenk and Yoshua Bengio. Boosting neural networks. *Neural computation*, 12(8):1869–1887, 2000. 1

[20] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016. 1

[21] Usama Muneeb, Erdem Koyuncu, Yasaman Keshtkar-jahromi, Hulya Seferoglu, Mehmet Fatih Erden, and A Enis Cetin. Robust and computationally-efficient anomaly detection using powers-of-two networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2992–2996. IEEE, 2020. 1

[22] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. In *International conference on machine learning*, pages 2285–2294. PMLR, 2015. 1

[23] Hongyi Pan, Diaa Badawi, and Ahmet Enis Cetin. Computationally efficient wildfire detection method using a deep convolutional network pruned via fourier analysis. *Sensors*, 20(10):2891, 2020. 1

[24] Mingchao Yu, Zhifeng Lin, Krishna Narra, Songze Li, Youjie Li, Nam Sung Kim, Alexander Schwing, Murali Annavaram, and Salman Avestimehr. Gradiveq: Vector quantization for bandwidth-efficient gradient aggregation in distributed cnn training. *arXiv preprint arXiv:1811.03617*, 2018. 1

[25] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arxiv* 2015. *arXiv preprint arXiv:1510.00149*, 2019. 1

[26] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 1

[27] Maneesh Ayi and Mohamed El-Sharkawy. Rmnv2: Reduced mobilenet v2 for cifar10. In *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0287–0292. IEEE, 2020. 1, 7, 8

[28] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*, 2016. 1

[29] Adrian Bulat and Georgios Tzimiropoulos. Hierarchical binary cnns for landmark localization with limited resources. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):343–356, 2018. 1

[30] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016. 1

[31] Zhiqiang Shen, Zechun Liu, Jie Qin, Lei Huang, Kwang-Ting Cheng, and Marios Savvides. S2-bnn: Bridging the gap between self-supervised real and 1-bit neural networks via guided distribution calibration. *arXiv preprint arXiv:2102.08946*, 2021. 1

[32] Zechun Liu, Zhiqiang Shen, Marios Savvides, and Kwang-Ting Cheng. Reactnet: Towards precise binary neural network with generalized activation functions. In *European Conference on Computer Vision*, pages 143–159. Springer, 2020. 1

[33] Brais Martinez, Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Training binary neural networks with real-to-binary convolutions. *arXiv preprint arXiv:2003.11535*, 2020. 1

[34] Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Bats: Binary architecture search. *arXiv preprint arXiv:2003.01711*, 2020. 1

[35] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4114–4122, 2016. 1

[36] Milad Alizadeh, Javier Fernández-Marqués, Nicholas D Lane, and Yarin Gal. An empirical study of binary neural networks’ optimisation. In *International Conference on Learning Representations*, 2018. 1

[37] Tom Bannink, Adam Hillier, Lukas Geiger, Tim de Bruin, Leon Overweel, Jelmer Neeven, and Koen Helwegen. Larq compute engine: Design, benchmark, and deploy state-of-the-art binarized neural networks. *arXiv preprint arXiv:2011.09398*, 2020. 1

[38] Cem Emre Akbaş, Alican Bozkurt, A Enis Çetin, Rengül Çetin-Atalay, and Ayşegül Üner. Multiplication-free neural networks. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pages 2416–2418. IEEE, 2015. 1, 4

[39] Arman Afrasiyabi, Baris Nasir, Ozan Yildiz, Fatos T Yarman Vural, and A Enis Çetin. An energy efficient additive neural network. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2017. 1, 4

[40] Arman Afrasiyabi, Diaa Badawi, Baris Nasir, Ozan Yildi, Fatos T Yarman Vural, and A Enis Çetin. Non-euclidean vector product for neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6862–6866. IEEE, 2018. 1, 4

[41] Diaa Badawi, Ece Akhan, Ma'en Mallah, Ayşegül Üner, Rengül Çetin-Atalay, and A Enis Çetin. Multiplication free neural network for cancer stem cell detection in h-and-e stained liver images. In *Compressive Sensing VI: From Diverse Modalities to Big Data Analytics*, volume 10211, page 102110C. International Society for Optics and Photonics, 2017. 1, 4

[42] Hongyi Pan, Diaa Badawi, Xi Zhang, and Ahmet Enis Cetin. Additive neural network for forest fire detection. *Signal, Image and Video Processing*, pages 1–8, 2019. 1, 4

[43] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1, 2, 5

[44] Joseph L Walsh. A closed set of normal orthogonal functions. *American Journal of Mathematics*, 45(1):5–24, 1923. 2

[45] Bernard J. Fino and V. Ralph Algazi. Unified matrix treatment of the fast walsh-hadamard transform. *IEEE Transactions on Computers*, 25(11):1142–1146, 1976. 2

[46] A Enis Cetin, Omer N Gerek, and Sennur Ulukus. Block wavelet transforms for image coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 3(6):433–435, 1993. 2

[47] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2

[48] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2

[49] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2

[50] PM Agante and JP Marques De Sá. Eeg noise filtering using wavelets with soft-thresholding methods. In *Computers in Cardiology 1999. Vol. 26 (Cat. No. 99CH37004)*, pages 535–538. IEEE, 1999. 2

[51] David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995. 2

[52] T Ceren Deveci, Serdar Cakir, and A Enis Cetin. Energy efficient hadamard neural networks. *arXiv preprint arXiv:1805.05421*, 2018. 3

[53] Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Christopher De Sa, and Zhiru Zhang. Building efficient deep neural networks with unitary group convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11303–11312, 2019. 3

[54] Transfer learning and fine-tuning. https://www.tensorflow.org/tutorials/images/transfer_learning. Accessed: 2021-03-01. 5

[55] Mobilenet in tensorflow’s official github. <https://github.com/tensorflow/models/tree/master/research/slim/nets/mobilenet>. Accessed: 2021-03-01. 5

[56] Hadamard network speed test code. <https://github.com/phy710/Hadamard-Network>. 7, 8