Optics Communications 485 (2021) 126709

Contents lists available at ScienceDirect

Optics Communications

journal homepage: www.elsevier.com/locate/optcom

Express Wavenet: A lower parameter optical neural network with random shift wavelet pattern[☆]

Yingshi Chen [c], Naixing Feng [a,*], Binbin Hong [a], Mei Song Tong [b], Guo Ping Wang [a], Zhongzhu Liang [d]

[a] *Institute of Microscale Optoelectronics, Shenzhen University, Shenzhen 518060, China*
[b] *Key Laboratory of Embedded System and Service Computing of the Ministry of Education, College of Electronics and Information Engineering, Tongji University, Shanghai 200092, China*
[c] *Institute of Electromagnetics and Acoustics, and Department of Electronic Science, Xiamen University, Xiamen 361005, China*
[d] *State Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China*

ARTICLE INFO

Keywords:
Machine learning
Optical diffractive network
Random shift wavelet
Expressway network

ABSTRACT

Compared with other optical diffractive neural networks, the express wavenet (ExWave) is an improved and efficient network which adopts much fewer parameters. For advantages of the ExWave, firstly, the wavelet-like pattern is used to modulate the phase of optical waves; secondly, the number of parameters can be reduced from O(n^2) to O(n), for the input image with n^2 pixels; thirdly, the accuracy can be still high with only 1% of total parameters of the famous D^2 NN. Besides, in the modified national institute of standards and technology (MNIST) dataset, 93.5% accuracy can be achieved with only 2457 parameters, nevertheless, 250880 parameters are required to meet a similar accuracy for the standard optical network. Moreover, the random shift wavelets exhibit characteristics of optical network vividly, especially for the vanishing gradient phenomenon in the training process. We present the modified expressway structure for solving this problem. Afterward, simulation have been carried out to validate effects of both random shift wavelet and expressway structure. Finally, it can be shown that the optical diffractive network in this work utilizes much fewer parameters than other neural networks. The detailed source code can be available in the following link: https://github.com/closest-git/ONNet.

1. Introduction

The diffractive optical neural network (DNet) is a novel machine learning framework [1–4], which can ceaselessly learn from the changing of optical wave distributions. The DNet adopts stacker diffractive layers to modulate the amplitude or phase of optical wave propagation. In the output layer, different optical distributions correspond to different classifications. DNet is mainly composed of optical devices. The prediction (forward propagation) operation of DNet is carried out at the speed of light, which is much faster than that of electronic neural networks. Therefore, its main advantage is high speed and low power consumption. After Lin's innovative work [1], lots of papers appeared to study this framework from different perspectives. The accuracy of DNet was improved by the broadband structure [2]. The differential measurement technique was implemented to improve accuracy [3]. Afterward, Chen and Zhu presented the analysis to show some key differences between DNet and multiple layer perceptron (MLP) in detail [4]. Gao [5] exhibited how to hide and encrypt images by using DNet (or cascaded phase-only mask architecture). And the DNet was adopted to perform optical logic operations [6]. Mengu [7] showed how to integrate DNet with electronic neural networks to improve accuracy. Zheng et al. proved the orthogonality of DNet later [8]. [9] uses adjoint variable methods to derive the photonic analogue of the backpropagation algorithm. [10] is a comprehensive review on the photonic neural networks, which gives a concise explanation of the advantages of DNet over electrical network. In this paper, we show one more advantage of DNet which could use much fewer parameters than other types of neural networks.

As shown in Fig. 1, DNet use stacker diffractive layers to modulate the amplitude or phase of optical wave propagation. Each diffractive layer has many pixels, which is an independent modulation element.

[☆] This work was supported in part by the National Natural Science Foundation of China under Grant 61901274, by the Shenzhen Science and Technology Innovation Committee, China under Grant JCYJ20190808141818890, by the Guangdong Natural Science Foundation, China under Grant 2020A1515010467, and by the Open Fund of State Key Laboratory of Applied Optics under Grant SKLAO2020001A06.

[*] Corresponding author.
E-mail address: fengnaixing@gmail.com (N. Feng).

https://doi.org/10.1016/j.optcom.2020.126709
Received 2 August 2020; Received in revised form 6 December 2020; Accepted 17 December 2020
Available online 31 December 2020

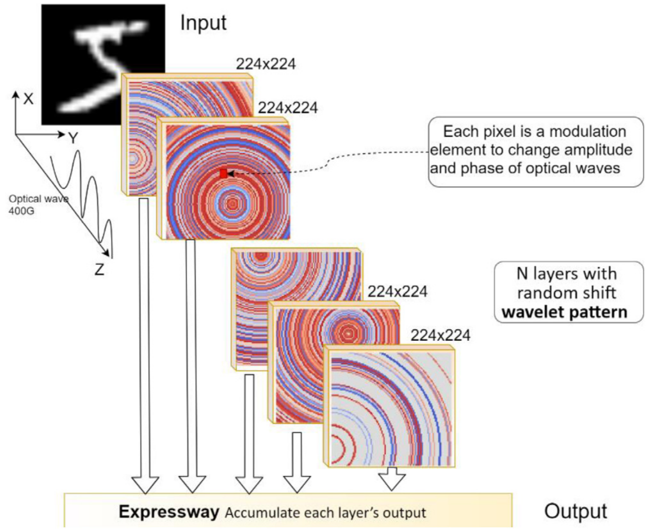

Input
224x224
224x224
X
Y
Z
Optical wave
400G
Each pixel is a modulation element to change amplitude and phase of optical waves
N layers with random shift wavelet pattern
224x224
224x224
Expressway Accumulate each layer's output
Output

Fig. 1. The structure of Express Wavenet: the shift wavelet which translates the wavelet randomly.

Eq. (1) gives the modulation of each pixel p:

$$modulation_p = a_p exp(j\phi_p) \tag{1}$$

where a_p is denoted as changing the amplitude of optical wave and ϕ_p is adopted to change the phase. As described in [4], the modulation is actually an activation function on optical transformation. In this work, we here only discuss the modulation of phase. Only phase modulator ϕ_p would be learned by the network and a_p would always be 1. Each pixel is treated as an independent parameter [1–3]. To the best of our knowledge, we find these pixels could be arranged in a wavelet pattern. As shown in Fig. 1, pixels at the same wave circle have same ϕ_p. So for the layer with n^2 pixels, the wavelet pattern would reduce the number of parameters from O(n^2) to O(n). For the case with a 10-layer network, the side length for each layer is 224, the number of parameters would be reduced from 250880 to 2457. Only 1% parameters are adopted in our work, compared to the famous D^{2}NN [1–3].

As far as we know, the wavelet pattern not only reduces the number of parameters greatly, but also reveals more intrinsic characteristics of optical network. By observing the changing process of wavelet in the training, we can find the obvious “vanishing gradient” phenomenon — the gradients of earlier layers become smaller and smaller [11]. This would make it really hard to learn and tune the parameters of the earlier layers. In the case of the deep convolutional neural networks (CNN) [12,13], the vanishing gradient problem is a long-standing problem and has been studied deeply. There are many powerful techniques to overcome this problem. However, in the much simpler DNet, there are limited techniques. As Fig. 4 shows, the wavelets in earlier layers are almost the same as the initial pattern and only the last few layers have significant changes. Compare to the random patterns used in other groups [1–4], the wavelet pattern would show the change of parameters in each layer much more clearly. So wavelet pattern would help optical network researchers to study the “vanishing gradient” phenomenon more effectively.

In this paper, two efficient and creative techniques are proposed to train a DNet with quite few parameters. As reflected in Fig. 1, the first technique we present is the shift wavelet which translates the wavelet randomly. And the second one is the expressway structure. In the proposals, the output of each layer not only plays a role of the input of the next one, but also directly accumulates with the weighting coefficient to the final layer. In this work, the improved network with these two features is referred as the express wavenet (ExWave), and its effectiveness has been verified by software simulation. Furthermore, in the modified national institute of standards and technology (MNIST) dataset, only two thousand parameters are required to achieve 93.5% accuracy, leading to much more efficient than the classical deep CNN. To our knowledge, it should be the minimal DNet for the MNIST classification problem.

For simulation applications of the ExWave, we have developed the ONNet. It is an open-source Python package on the framework of PyTorch [14,15], and then codes can be available and downloaded at https://github.com/closest-git/ONNet. The ONNet can provide many tools to users for studying optical neural networks.

2. Express wavenet with random shift wavelet

To describe the problem more concisely, we give the following formulas and symbols. Let complex function $f(x;\theta)$ is the target function parameterized by θ in complex space C. Given training data $D = \{(x,y)\}$ where x denotes the input image and y the class number. The DNet tries to learn the function $f(x;\theta)$ by stacked L diffractive layers. For a pixel i in layer l, z_i^l is its input and h_i^l is its activation (output). The z_i^l is a weighted summary of all activation h_k^{l-1} from the previous layer.

$$z_i^l = \sum_k \omega_{k,i}^{l-1} h_k^{l-1} \tag{2}$$

$\omega_{k,i}^l$ is the weight coefficient for each pixel k in the previous layer:

$$\omega_{k,i}^l = \frac{(i-k)}{r^2}\left(\frac{1}{2\pi r}+\frac{1}{j\lambda}\right) exp\left(\frac{j2\pi r}{\lambda}\right) \tag{3}$$

where r is the distance between pixel k and i, λ is the wavelength of optical wave. For the detailed derivation process of this coefficient of $\omega_{k,i}^l$, please see [1–3].

The diffractive layer would change the input z_i^l by the modulation of its amplitude and phase. Let a_i^l is the change of amplitude, and ϕ_i^l is the change of phase, then the modulation coefficient t_i^l is:

$$t_i^l = a_i^l exp(j\phi_i^l) \tag{4}$$

Finally, the activation h_i^l is the production of modulation coefficient t_i^l and input z_i^l:

$$h_i^l = t_i^l z_i^l \tag{5}$$

The key components of ExWave are respectively random shift wavelet and expressway network structure which are described in detail as follow.

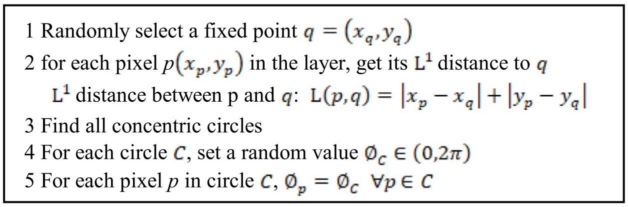
2.1. Random shift wavelet

In the first place, let n is the side length of each layer, each pixel p is represented by its coordinate as $(x,y), x,y \in \{1,2,\ldots,n\}$. In the previous work [1–3], all the n^2 pixels are independent variables as the formula (5). They take random initial value between $(0,2\pi)$. Then network would adjust the $\emptyset_p$ to learn a better model. So the total number of parameters is n^2.

$$\emptyset_p = \text{random}(0,2\pi) \text{ for each } p(x,y)\ x,y \in \{1,2,\ldots,n\} \tag{6}$$

We here present a wavelet-like pattern to decrease the number of parameters. At first, we select a fixed point q randomly, then set $\emptyset_p$ on the distance between p and q. The detailed algorithm is as follows.

Algorithm 1 Random shift wavelet algorithm to set the value $\emptyset_p$



```
1 Randomly select a fixed point $q = (x_q, y_q)$
2 for each pixel $p(x_p, y_p)$ in the layer, get its $L^1$ distance to $q$
    $L^1$ distance between p and $q$: $L(p,q) = |x_p - x_q| + |y_p - y_q|$
3 Find all concentric circles
4 For each circle $C$, set a random value $\emptyset_C \in (0,2\pi)$
5 For each pixel $p$ in circle $C$, $\emptyset_p = \emptyset_C \ \ \forall p \in C$
```

In view of some typical pattern in each layer shown in Fig. 1, it is clear that many concentric circle pixels just like wavelets on the water surface, therefore, these patterns are called as the wavelet. Every layer contains at most $\sqrt{2n}$ concentric circles, and hence total parameters are reduced to O(n).

To our knowledge, if we adopt the wavelet pattern without random translation, accuracies will descend due to the fact that all fixed points q are the center of layer ($q = (n/2, n/2)$) Moreover, simulation from different datasets have validated this point. As the training in the deep CNN, the training with randomness can further improve the accuracy so that the learned model could have better generality.

2.2. Expressway network structure

"Vanishing gradient" is a common problem in the training of neural networks, especially for deep networks with many layers. Similarly, this kind of phenomenon also exists in the optical diffractive networks. In the case of wavelet network, intuitive and vivid displays are given, as shown in Fig. 4. It can be concluded that only the wavelet in the last few layers has significant changes, however, changes are much not obvious in the early layers.

To conquer this problem, several techniques have been shown in the classical deep learning [16,17]. In the highway networks [16], the forward propagation is defined in Eq. (7):

$$z^l = f(z^{l-1}, W^{l-1})T^l + z^{l-1}(1 - T^l) \tag{7}$$

Where z^l is the output at l layer, W^{l-1} is the weight matrix. The gating function T^l in [14] is just the sigmoid function. That is $T^l(x) = 1/(1+e^{-x})$. So when at some layer $l, T^1 \approx 0$, the output z^l is nearly the input z^{l-1}. The layer l will directly utilize the outputs of previous layers, which can create an information highway to pass gradient information in the back propagation. Therefore, highway networks will train out very deep networks with hundreds of layers.

But for the optical network, there is no easy way to implement the sigmoid gating function. The output of each diffractive layer is actually optical waves at speed of light. For some diffractive layers in the middle, how to accumulate outputs from previous layers? It means that the transmission of light in the middle layers is stopped to merge outputs of previous layers. To merge outputs of middle layers into the output of final layer, we here develop a modified version of the highway network instead of merging operations in the middle layers. The output is just accumulated in the final layer, defined as the following equation.

$$z_{output} = \sum_l z_l w_l \tag{8}$$

In the same way, we still observe similar phenomenon in the testing dataset. In a conclusion, these expressway structures will send gradient information to the previous layer, leading to improving the accuracy.

3. Results and discussions

In this section, we test the ExWave on two datasets including both MNIST [18] and fashion-MNIST [19]. The MNIST dataset is a commonly used dataset of handwritten digits, which contains 60000 training and 10000 testing images, respectively. The fashion-MNIST consists of a training set consisting of 60000 examples belonging to ten different classes and a test set of 10000 examples.

As reflected in Fig. 2, it shows some contents of both MNIST and fashion-MNIST datasets. As indicated in [4], the optical diffractive network is still in its infancy. So, these two simple testing datasets can still be used for testing. Afterward, we will plan to test larger and harder datasets in the future work.

Fig. 2. MNIST and fashion-MNIST datasets show some contents of MNIST and fashion-MNIST datasets (10 classes).

Table 1
Accuracy on the different algorithm combinations.

	MINST	Fashion-MNIST
Express Wavenet	93.5%	83.5%
Only shift wavelet	90.8%	78.7%
Only Expressway	89.3%	77.4%
No shift wavelet, no Expressway	80.3%	71.7%
Accuracy in [1]	93.39%	81.13%

3.1. Accuracy and model parameters

Now, we train a 10-layer ExWave with the side length of each layer is 224. This network only requires 2457 parameters whose distribution for each layer in number is set as below (260, 238, 224, 286, 253, 209, 282, 190, 219, 286). It is shown in Fig. 3 that learning curves are plotted for these two testing datasets. The classification accuracy is 93.5% for the MNIST and 83.5% for the fashion-MNIST.

As depicted above, ExWave has two unique techniques, including shift wavelet technique and expressway structure. As shown in Table 1, therefore, four combinations are listed. It is clear that both techniques can significantly improve accuracy. (1) Compared to fixpoint wavelet, the random translation can improve accuracies from 80.3% to 90.8% for the MNIST, and from 71.7% to 78.7% for the fashion-MNIST. (2) With expressway network structure, the accuracy increased from 90.8% to 93.5% for the MNIST, and from 78.7% to 83.5% for the fashion-MNIST. This significant improvement does verify the effectiveness of expressway structure.

Table 1 also lists the accuracy of famous D^{2}NN in [1]. Their model needs 0.2 million parameters to reach 93.39% in MINST and 81.13% in Fashion-MNIST. And our model only needs 2000 parameters! So we only use 1% of the parameters to achieve comparable accuracy. This comparison also verified the effectiveness of shift wavelet and expressway structures.

3.2. Wavelet pattern

The wavelet pattern of ExWave provides a vivid way to reveal the training process in detail. Fig. 4 shows the wavelet pattern at different epochs. These images in Fig. 4 actually exhibit the sine transformation

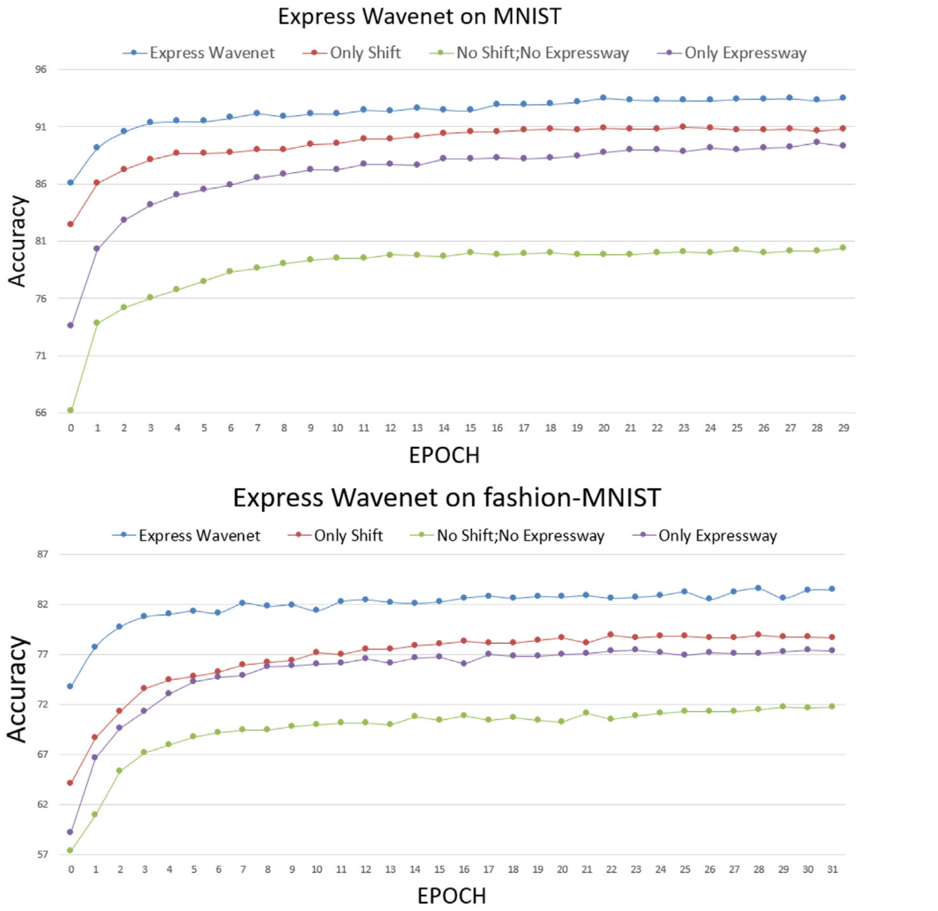


Fig. 3. Comparison of Express Wavenet on MNIST and fashion-MNIST. Four curves correspond to four algorithms. The blue curve corresponds to "Express Wavenet". The red curve corresponds to network with only shift wavelet. The violet curve corresponds to the network with only expressway. The cyan curve corresponds to an trivial network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

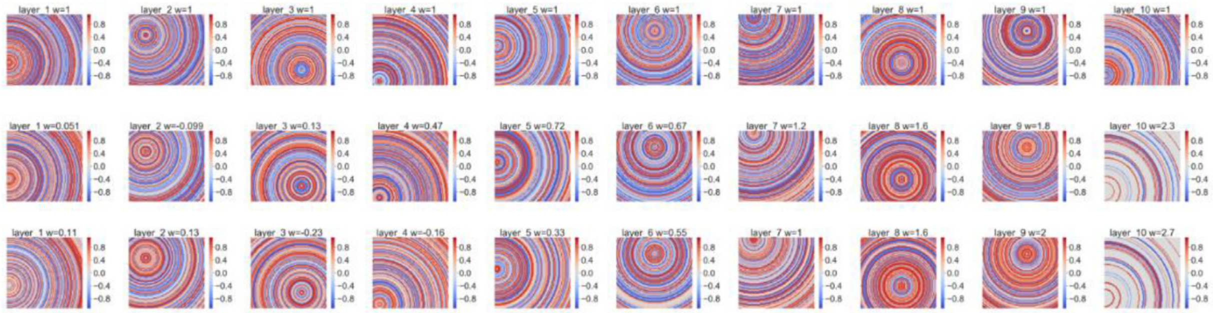


Fig. 4. The phase maps of 10 layers at different epochs. The first row shows wavelet pattern at epoch 0. The middle row shows wavelet pattern at epoch 10. And the last row shows wavelet pattern at epoch 20.

of $\emptyset_p$. Regardless of the value of $\emptyset_p, \sin(\emptyset_p)$ will always be in the range of $[-1, 1]$. So the range of colorbars in these figures are -1 to 1.

In Fig. 4, the first row is the initial random pattern, containing dense concentric circles with many mutations. As the training goes on, then changes are obviously observed in the last few layers, at which the wavelet becomes sparse and more regular. The pattern in the early layers is also changing, but it is much fainter. Therefore, it is a clear sign of "vanishing gradient" phenomenon.

4. Conclusions

In this paper, we have developed the ExWave, which reveals more characteristics of optical diffractive networks. In the first place, our optical diffractive network with random shift wavelet only needs one percent parameters to achieve comparable accuracy of classical DNet. In addition, the vanishing gradient phenomenon in the training process is hard to be overcome by the simple structure in the classical DNet. To relieve and improve this problem, an expressway structure is adopted.

Based on this technology, we can further improve the accuracy. We have developed an open-source package ONNet, which would help researchers in this area.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] X. Lin, et al., All-optical machine learning using diffractive deep neural networks, Science 361 (6406) (2018) 1004–1008.

[2] Y. Luo, et al., Design of task-specific optical systems using broadband diffractive neural networks, Light Sci. Appl. 8 (1) (2019) 1–14.

[3] J. Li, et al., Class-specific differential detection in diffractive optical neural networks improves inference accuracy, 2019, arXiv preprint arXiv:1906.03417.

[4] Y. Chen, J. Zhu, An optical diffractive deep neural network with multiple frequency-channels, 2019, arXiv preprint arXiv:1912.10730.

[5] Y. Gao, et al., Multiple-image encryption and hiding with an optical diffractive neural network, Opt. Commun. 463 (2020) 125476.

[6] C. Qian, et al., Performing optical logic operations by a diffractive neural network, Light Sci. Appl. 9 (1) (2020) 1–7.

[7] D. Mengu, et al., Analysis of diffractive optical neural networks and their integration with electronic neural networks, IEEE J. Sel. Top. Quantum Electron. 26 (1) (2019) 3700114.

[8] S. Zheng, et al., Orthogonality of diffractive deep neural networks, 2018, arXiv preprint arXiv:1811.03370.

[9] Tyler W. Hughes, et al., Training of photonic neural networks through in situ backpropagation and gradient measurement, Optica 5 (7) (2018) 864–871.

[10] Lorenzo De Marinis, et al., Photonic neural networks: A survey, IEEE Access 7 (2019) 175827-175841.

[11] S. Hochreiter, et al., Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies, Wiley-IEEE Press, 2001, pp. 237–243.

[12] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[13] Léon Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT' 2010, Physica-Verlag HD, 2010, pp. 177–186.

[14] Adam Paszke, et al., Automatic Differentiation in Pytorch, 2017.

[15] Adam Paszke, et al., PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019.

[16] Rupesh Kumar Srivastava, Klaus Greff, Jürgen Schmidhuber, Highway networks, 2015, arXiv preprint arXiv:1505.00387.

[17] K. He, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[18] Li Deng, The mnist database of handwritten digit images for machine learning research [best of the web], IEEE Signal Process. Mag. 29 (6) (2012) 141–142.

[19] Han Xiao, Kashif Rasul, Roland Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, 2017, arXiv preprint arXiv:1708.07747.