Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023408

Exploration of continual learning based on the comparison and analysis of LWF and EWC

Zihang Wang

Computer Science and Technology, SUSTC, Shenzhen, Guangdong, 518055, China

12012016@mail.sustech.edu.cn

Abstract. Nowadays, continual learning is seen as one of the important fields of strong artificial intelligence. People focus on it and propose a lot of methods in order to alleviate catastrophic forgetting problem which is serious and essential and cannot be avoided. Among all the methods, EWC and LWF are two typical methods and they have some similarities and differences. This paper adopts comparison method to research the effect of LWF and EWC applied on different datasets and also compares some similar methods like R-EWC and LFL. Meanwhile, some analysis and suggestions about these methods are also given. From the results, both methods improve memory ability. But LWF performs better in testing old knowledge. In contrast, EWC has a stronger ability to learn new knowledge. Both EWC and LWF have some limitations. There are some problems which cause errors for the principle of EWC while for LWF, the confirmation of hyper-parameters has great influence on the experimental results and the method itself has some errors.

Keywords: continual learning, regularization-based, catastrophic forgetting

1. Introduction

Human beings have the ability to reuse and expand knowledge from experience continuously. That is, people can not only apply the knowledge they have already gotten to the situation they never face, but also use them as the basis for future learning. However, for traditional deep learning models, they cannot have very complex and intelligent behaviors because they just learn from fixed datasets in one isolated task. Therefore, making models have such a kind of human-like ability is a hot and important topic for present researches in Artificial Intelligence (AI) fields. People want to create models which can have the ability of continuous learning and generalization of knowledge. And the agents can build a complex understanding of the world according to their own experience, so as to use more complex knowledge for independent and progressive training and improvement.

One of the research directions to solve such problems is continual learning, also known as life-long learning. Its purpose is to expand the adaptability of the model so that the model can learn different tasks at different times with little or less forgetting [1]. Data can be in different fields, different tasks, or simply different categories.

There are many challenges in the process of researching continual learning, and what the most concerned one is the high forgetting rate of old knowledge, which is called catastrophic forgetting [2]. Catastrophic forgetting means that after training on new datasets, the model almost completely or largely forgets what it learned from old datasets before, which can be reflected from the large value of loss

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/4/2023408

function and low accuracy when training on the old datasets again. And this serious problem is the major obstacle to continual learning.

People pay attention to the problem of catastrophic forgetting not only because of its seriousness, but also because of its obstinacy. The catastrophic forgetting problem is deeply rooted in genes of neural networks, which means it cannot be completely solved and can only be alleviated by various mechanisms and strategies to some extent [3]. That is its obstinacy. Neural network is composed of network structure and network parameters. But unfortunately, no matter the structure or parameters, catastrophic forgetting is hidden inside. From the view of the static structure, once the structure is determined, it is difficult to be adjusted during the process of training. The structure of network directly determines the capacity of learning models. The fixed structure neural network means that the capacity of the model is also limited. In the case of limited capacity, the neural network must erase the old knowledge in order to learn other new tasks. From the view of the dynamic changing, the neurons in the hidden layers are global, and the small changes of a single neuron can affect the output results of the whole network. In addition, the parameters of all feed-forward networks are connected to each dimension of the input. The new data may change all the parameters in the network. If the changed parameters include parameters that are highly relevant to historical knowledge, the final effect is that the new knowledge will cover the old knowledge, which will cause catastrophic forgetting. Therefore, how to design strategies and algorithms to alleviate catastrophic forgetting is an important topic. And so far, there have been a lot of methods to deal with this problem.

There are three different classes for the main state-of-the-art methods: replay-based methods, regularization-based methods and parameter isolation methods [4]. In the first class, the represented method is Incremental classifier and Representation Learning (iCaRL) [5], which achieves classification by using prototype network. In the second class, elastic weight consolidation (EWC), Learning without Forgetting (LWF), less-forgetting learning (LFL) are the represented methods. And this part is more concerned by the academic community than others. Many studies have been carried out around related aspects. In the third class, each task has a subset of parameters which is different from other tasks. And then, mask the parts corresponding to the previous tasks to avoid changes. The represented methods are including PathNet, Hard Attention to the Task (HAT) and so on.

Although there are so many methods, because one method can have different performances when it is used in different situations, there is not a universal method to deal with all situations. This paper tries to research the methods EWC and LWF and compare these two methods from principle and results of training on different datasets. Also analyse their shortcomings and put forward suggestions and understanding for their improvement directions.

2. The EWC method and its improvement

The EWC method by James et al. [6], known as elastic weight consolidation, is one of famous algorithms to solve catastrophic forgetting problem and achieve continual learning. This method belongs to regularization-based methods and it can be classified as prior-focused class [7].

2.1. Idea of EWC

EWC is an algorithm which constrains changing of the parameters. Because not all neurons can make a difference to the result, just reduce the changing on those relatively important neurons during the next training and then the model can achieve continual learning to some extent. In other words, for some certain weights, the learning speed can be decreased based on the importance they have to previous tasks, which is the core idea of EWC.

2.2. Principle

With regularization-based methods, how to calculate and evaluate the importance of each parameter for previous tasks is what EWC focuses on. If the model can learn parameters according to the Bayesian Method, catastrophic forgetting will not happen. Suppose that the prior probability of the parameters is $P(\theta)$, then posterior distribution of parameters is: (D is the given data, D_A represents task A and D_B represents task B)

DOI: 10.54254/2755-2721/4/2023408

$$P(\theta|D) = \frac{P(D_B|\theta)P(\theta|D_A)}{P(D_B)} \quad (1)$$

Because task A and B are independent of each other, it can be written as:

$$\log P(\theta|D) = \log P(D_B|\theta) + \log P(\theta|D_A) - \log P(D_B) \quad (2)$$

The first term is the minus of loss function on task B: $-L_B(\theta)$, and the third term is a constant, so the optimization goal is:

$$\theta = \arg\max \log P(\theta|D) = \arg\min L_B(\theta) - \log P(\theta|D_A) \quad (3)$$

Dealing with $P(\theta|D_A)$ with Gaussian distribution approximately:

$$P(\theta|D_A) = N(\sigma, \mu) \quad (4)$$

$$\log P(\theta|D_A) = \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(\theta-\mu)^2}{2\sigma^2} \quad (5)$$

Suppose the parameters on task A after training are θ_A^*. And suppose function:

$$f(\theta) = \log P(\theta|D_A) \quad (6)$$

Do Taylor expansion on $\theta = \theta_A^*$:

$$f(\theta) \approx f(\theta_A^*) + \frac{1}{2}(\theta - \theta_A^*)^2 f''(\theta_A^*) \quad (7)$$

Using Fisher Information Matrix to get the value of $f''(\theta_A^*)$. And the final loss function is:

$$L(\theta) = L_B(\theta) + \frac{\lambda}{2}\sum_i F_i(\theta_i - \theta_{A,i}^*)^2 \quad (8)$$

Meanwhile, Fisher Information Matrix also reflects uncertainty for evaluating parameters. The larger the second derivative is, the more certain estimation of parameters is. And the larger Fisher Information Matrix is, the larger the penalty term is. Therefore, the more important parameters are, the smaller change they have in the subsequent tasks.

2.3. Experiment and Results

Apply EWC on SplitCIFAR10 data-set which contains 10 classes different colorful pictures and is split as five tasks while compare the accuracy under EWC and no EWC these two cases. The size of all pictures is 32×32. The learning rate is 0.01 and λ in EWC is 0.5, which represents the hyper-parameter to weigh the penalty inside the total loss. The larger for λ, the larger for regularization. The results are as follows.

Table 1. The accuracy of EWC and No EWC these two methods applied on SplitCIFAR10 dataset.

Task Index	EWC	No EWC
1	0.5885	0.3764
2	0.6397	0.4160
3	0.5605	0.2931
4	0.4540	0.2651
5	0.8200	0.6891

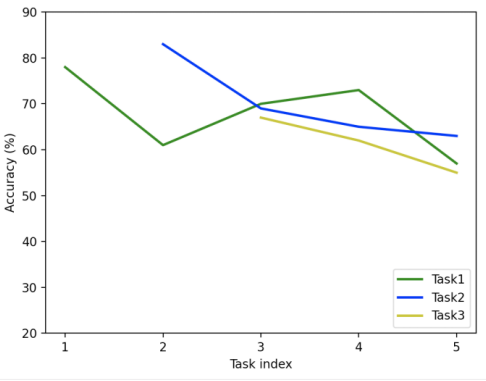


Figure 1. The accuracy of task1-3 after each task trained.

DOI: 10.54254/2755-2721/4/2023408

From the results, it can be seen obviously that EWC can alleviate the problem of catastrophic forgetting to some extent. And with the increase of learning tasks, the accuracy of previously learned tasks shows a downward trend.

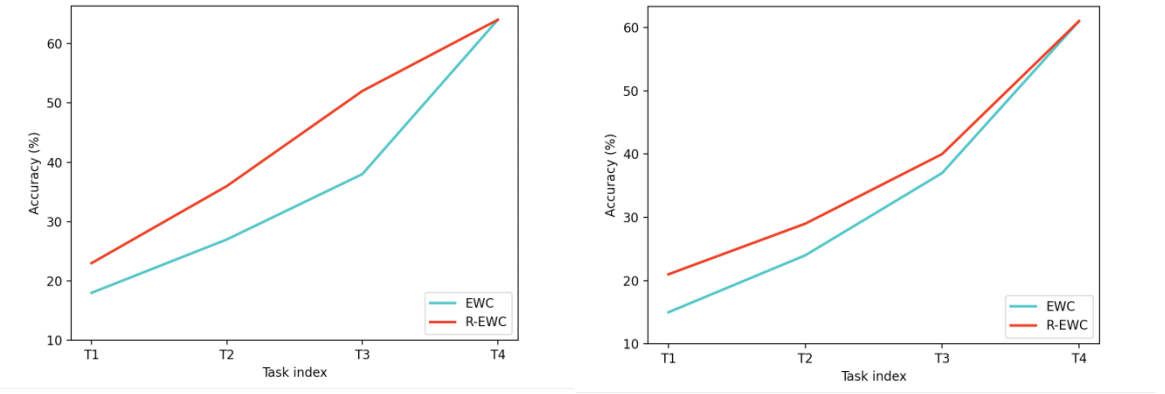
2.4. The improvement of EWC

One model optimized for EWC is called Rotated EWC (R-EWC), proposed by Xialei Liu et al [8].

2.4.1. The problem R-EWC overcomes. For EWC, the correlations of different weights are neglected when diagonal approximation is carried out, which may cause some forgetting in Fisher Information Matrix. That is, catastrophic forgetting may happen because EWC ignores the relativity of the distribution of parameters. And R-EWC is a way to solve this problem.

2.4.2. Idea of R-EWC. The main idea of R-EWC is to rotate the parameter space and let the diagonal approximation of the Fisher Information Matrix be a good approximation of the full Fisher Information Matrix, then to learn the next task in the rotated space. To achieve the operation of rotating, R-EWC applies rotations layer by layer. And it approximates each rotation with additional layers that will rotate the input and the output of that layer.

2.4.3. Experiment and results. Apply EWC and R-EWC on CIFAR-100 and CUB-200 datasets, then compare the accuracy of each task of them. CIFAR-100 contains 100 different classes of colorful pictures and CUB-200 contains 200 different classes of pictures of birds. Both of the datasets are divided into four tasks. The results are as follows:



(a) (b)

Figure 2. The figure of the results on two datasets. (a) The accuracy of EWC and R-EWC training on CIFAR-100 dataset. (b) The accuracy of EWC and R-EWC training on CUB-200 dataset.

From the results, the performance of R-EWC is better than EWC for each single previous task, that means it can avoid forgetting better than EWC.

3. The LWF method

LWF is proposed by Zhizhong Li et al [9]. It is data-focused and also belongs to regularization-based methods.

3.1. Background

New tasks can share the parameters of old tasks for learning without catastrophic forgetting. It is a simple idea for people. Based on this idea, let θ_s be a set of shared parameters for CNN model, θ_o be specific parameters on previously learned tasks and θ_n be specific parameters randomly initialized for new tasks. There are three ways for this framework.

DOI: 10.54254/2755-2721/4/2023408

The first is feature extraction. θ_s and θ_o do not change. Build a new parameter layer to deal with new tasks.

The second is fine-tuning. θ_o is fixed while θ_s and θ_n are updated with a small learning rate.

The third is joint training. All parameters are jointly optimized and cross sampled from each task, which can be considered as the upper limit of lifelong learning.

These three ways have corresponding disadvantages. Intuitively, feature extraction will only train the parameters of the new task, while the convolution layer and most connection layers are fixed, and the effect on the new task will be poor. For fine-tuning, the shared parameters are changed, and catastrophic forgetting will happen on the old tasks. Joint training deviates from the request (the old task will not appear).

To solve these problems, LWF is proposed. And it is very similar to joint training, but it does not need to use the data of old tasks.

3.2. Idea of LWF

LWF adopts distillation network and also uses the idea of fine tuning. The core of the latter is to find a local minimum through a low learning rate, combined with the existing network and related data information processing new tasks. Distillation is an ingenious way to learn and get proper parameters in an easy and simple network in order that the output set produced by the network is the same as that produced by the original large network.

The difference of LWF is that the trained parameters can deal with the tasks better, whether old or new, also supervise the learning process of new tasks with the same data and provide unsupervised output guidance for old tasks.

The loss function for new task is:

$$L_{new}(\mathrm{Y}_n, \hat{\mathrm{Y}}_n) = -\mathrm{Y}_n \log \hat{\mathrm{Y}}_n \tag{9}$$

$\hat{\mathrm{Y}}_n$ is the output of network and Y_n is true label. For old task, the loss function is: (ℓ is the number of labels; $y_o^{'(i)}$and $\hat{\mathrm{y}}_o^{'(i)}$ are the modified value of recording value $y_o^{(i)}$ and current value $\hat{\mathrm{y}}_o^{(i)}$)

$$L_{old}(\mathrm{Y}_o, \hat{\mathrm{Y}}_o) = -\sum_{i=1}^{\ell} y_o^{'(i)} \log \hat{\mathrm{y}}_o^{'(i)} \tag{10}$$

$y_o^{'(i)}$ and $\hat{\mathrm{y}}_o^{'(i)}$ can be defined as follows:

$$y_o^{'(i)} = \frac{\left(y_o^{(i)}\right)^{1/T}}{\Sigma_j \left(y_o^{(j)}\right)^{1/T}} \tag{11}$$

$$\hat{\mathrm{y}}_o^{'(i)} = \frac{\left(\hat{\mathrm{y}}_o^{(i)}\right)^{1/T}}{\Sigma_j \left(\hat{\mathrm{y}}_o^{(j)}\right)^{1/T}} \tag{12}$$

3.3. Advantages of LWF

In terms of classification performance, LWF is better than feature extraction. At the same time, fine-tuning is performed on new tasks, while using fine-tuning parameters on old tasks can also achieve good results, that is, it can learn without forgetting.

When it comes to the efficiency of computation, the time spent on training is faster than joint training, close to fine-tuning. In addition, the time spent on testing is faster than using multiple fine-tuning networks to perform different tasks.

Furthermore, it has simple deployment. That means, once the task is understood, there is no need to retain or reapply training data to maintain the performance of the adaptive network.

3.4. Contrast and results

LFL is proposed by Heechul Jung et al [10]. It is the same kind of method as LWF, regularization-based and data-focused. Apply them on PermutedMNIST dataset which is split as four tasks. α is 0.4, which represents distillation hyper-parameter. And T is 2, which is the softmax temperature for distillation. Learning rate is 0.01 and weight-decay is 0.0005. The results are as follows:

DOI: 10.54254/2755-2721/4/2023408

Table 2. The accuracy of LWF and LFL these two methods applied on PermutedMNIST dataset.

Task Index	LWF	LFL
1	0.8979	0.8532
2	0.9188	0.8707
3	0.9256	0.9062
4	0.9347	0.9192

From the results, LWF has better accuracy and performance than LFL, which also shows the superiority of LWF in continual learning field.

4. Comparison of EWC and LWF

Both EWC and LWF are regularization-based methods. However, they have different performance on the same data-set and both of them have their limitations. Here is some analysis.

4.1. Regularization-based methods

Regularization-based methods are a kind of methods which focus on how to design a good and clever regularization loss term in the continual learning field. Due to privacy policies and other reasons, it is often impossible to store previous task information. In this scenario, a clever regularization loss can be designed to limit the forgetting of old knowledge when learning new data (such as limiting the range of changes of model parameters). These methods can be further divided into data-focused methods and prior-focused methods.

Methods like EWC belong to prior-focused methods while methods like LWF, LFL belong to data-focused methods. In the data-focused methods, people can weight the input data and select the data with migration ability. While in the prior-focused methods, people use the estimated distribution of model parameters as a priori when learning from new data and impose loss penalties when changing parameters that are important for previous tasks.

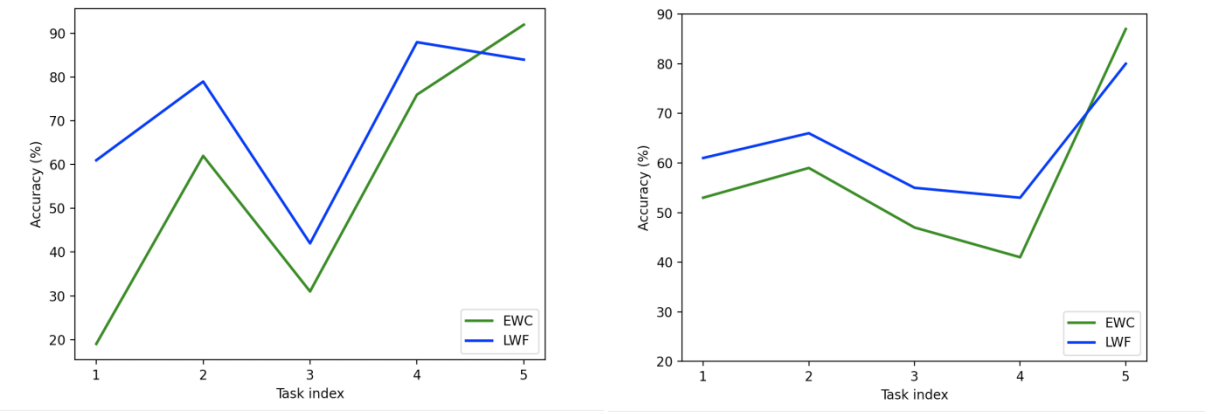
4.2. Experiment and results

4.2.1. Datasets. The Apply EWC and LWF on Split-MNIST and Split-CIFAR10 datasets. Both of these two datasets are split into 5 tasks. Observe the accuracy of them.

4.2.2. Hyper-parameter selection. Using a single-head multiple-layer perceptron with two hidden layers for the Split-MNIST dataset and a Convolutional Neural Network (CNN) with 4 convolution layers and 2 dense layers with dropouts for the Split-CIFAR10 dataset. For EWC, the learning rate is 0.001 and λ in EWC is 0.3. For LWF, the learning rate is 0.001, α is 0.3 and T is 2.

4.2.3. Results. The tasks learned before have different degrees of forgetting. And because of the respective algorithm characteristics, under this hyper-parameter selection, EWC has better learning ability for new tasks but has a higher degree of forgetting for old tasks than LWF.

DOI: 10.54254/2755-2721/4/2023408



(a) (b)

Figure 3. The figure of the results on two datasets. (a) The accuracy of EWC and LWF training on Split-MNIST dataset. (b) The accuracy of EWC and LWF training on Split-CIFAR10 dataset.

4.3. Problems and limitations

From the principle of EWC, it uses Taylor expansion to deal with $\log P(\theta|D_A)$. And it adopts approximate treatment, that is, it ignores higher order terms. This error is small when the tasks are similar, but when there is a big difference between them or the quantity of tasks is large, this error will be enlarged. And this is one of the problems of EWC. In addition, diagonal approximation in Fisher Information Matrix ignores the relativity of parameters distribution, which may cause catastrophic forgetting in some cases.

For LWF, it has many hyper-parameters and it is inconvenient and complex for parameter adjustment. For tasks not similar, the performance for old tasks will be decreased sharply. Furthermore, the LWF needs additional networks to assist in the construction of new models, it will occupy more space when the quantity of tasks is large. The above shortcomings may be addressed in future research using CNN based networks [11-15] with a pyramidal structure for feature extraction.

5. Conclusion

The EWC method requires a prior probability hypothesis and adds regular terms to the loss function to limit the change of past knowledge, while the essence of LWF is to balance the network parameters trained in the new task through the output of the old network, so as to obtain better performance on both the old and new task networks. From the results running on Split-MNIST and Split-CIFAR10 datasets, both of EWC and LWF provide strategies to deal with forgetting problem. LWF has a better performance in memory knowledge and EWC seems to have stronger ability to learn new knowledge under the parameters and datasets used.

When there is a big difference between tasks or there are plenty of tasks, the effect of EWC will be poor. Also, EWC neglects the correlation of parameter distribution to deal with Fisher matrix. These are the main problems for EWC and the optimization for Fisher Information matrix processing is a direction that can be considered to improve the performance of EWC like R-EWC and so on. LWF keeps the output of the old task while trains the new network to approach the optimal results, which can keep the old knowledge a lot but influence the learning for new knowledge. In addition, the LWF has a lot of parameters needed to confirm, so suitable parameters can also improve its performance.

The research in this paper provides the reference of possible optimization directions and suggestions for EWC and LWF, also points out the differences and shortcomings of the two methods and tries to analyse them from a theoretical perspective. It is helpful for people to further study and optimize the methods of continual learning. However, there are also some limitations to this research. It only focuses

DOI: 10.54254/2755-2721/4/2023408

on LWF and EWC methods, more methods need to be tested. And the optimal hyper-parameters are difficult to confirm, which is also a lack in the experiment.

References

[1] Nguyen, C. V., Li, Y., Bui, T. D. and Turner, R. E. 2017. Variational continual learning. arXiv preprint arXiv:1710.10628.

[2] Ratcliff, R. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. Psychological review, 97(2), 285.

[3] Toneva, M., Sordoni, A., Combes, R. T. D., Trischler, A., Bengio, Y. and Gordon, G. J. 2018. An empirical study of example forgetting during deep neural network learning. arXiv preprint arXiv:1812.05159.

[4] Van de Ven, G. M. and Tolias, A. S. 2019. Three scenarios for continual learning. arXiv preprint arXiv:1904.07734.

[5] Rebuffi, S. A., Kolesnikov, A., Sperl, G. and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 2001-2010).

[6] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... andHadsell, R. 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526.

[7] Han, X. and Guo, Y. 2021, September. Continual Learning with Dual Regularizations. In JointEuropean Conference on Machine Learning and Knowledge Discovery in Databases (pp.619-634). Springer, Cham.

[8] Liu, X., Masana, M., Herranz, L., Van de Weijer, J., Lopez, A. M. and Bagdanov, A. D. 2018, August. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 2262-2268). IEEE.

[9] Li, Z. and Hoiem, D. 2017. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12), 2935-2947.

[10] Jung, H., Ju, J., Jung, M. and Kim, J. 2016. Less-forgetting learning in deep neural networks. arXiv preprint arXiv:1607.00122.

[11] Beini, Z., Xuee, C., Bo, L., and Weijia, W. 2021. A new few-shot learning method of digital PCR image detection. IEEE Access, 9, 74446-74453.

[12] Kim, S. W., Kook, H. K., Sun, J. Y., Kang, M. C., and Ko, S. J. 2018. Parallel feature pyramid network for object detection. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 234-250).

[13] Seferbekov, S., Iglovikov, V., Buslaev, A., and Shvets, A. 2018. Feature pyramid network for multi-class land segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 272-275).

[14] Lyu, Y., Yang, Z., Liang, H., Zhang, B., Ge, M., Liu, R., ... and Yang, H. 2022. Artificial intelligence‐assisted fatigue fracture recognition based on morphing and fully convolutional networks. Fatigue & Fracture of Engineering Materials & Structures, 45(6), 1690-1702.

[15] Zhang, B., Luo, X., Lyu, Y., Wu, X., and Wen, W. 2022. A defect detection method for topological phononic materials based on few-shot learning. New Journal of Physics, 24(8), 083012.