Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

A survey of Convolutional Neural Networks —From software to hardware and the applications in measurement

ARTICLE INFO

Keywords
Measurement
Convolutional neural network
Survey
Cross-research

ABSTRACT

The convolutional neural network is a subfield of artificial neural networks and has made great achievements in various domains over the past decade. The technique has been widely applied including computer vision, natural language processing, image analysis, etc. The field of measurement also achieves exciting progress with convolutional neural networks. This paper gives a survey of convolutional neural networks from software algorithms to hardware architectures perspective and the applications in measurement. The survey hopes to give contributions to the cross-research of artificial intelligent and measurement.

1. Introduction

The convolutional neural network (CNN) is one of the fundamental technologies in the field of artificial intelligence (AI), and it is a promising tool for solving the problem of pattern recognition. The technology has been the most general method to analyze visual imagery and has a wide range of research fields with superior performance such as image and video recognition, natural language processing, image analysis, etc [1,2]. The applications are extended to some special research fields, including cultural heritage protection by ancient characters recognition [3,4], environmental protection by fish species recognition [5], health care system by gender classification [6]. The nature of CNN makes it quite suitable for measurement and the technique has greatly promoted the development in the field of measurement [7].

As the emergence of CNNs can be traced back to 1980s, when was LeNet [8] devised. However, due to the constraints of hardware, it is not until AlexNet [9] in 2012 for CNNs to actually play their roles, when massive data and the high-performance computational hardware were affordable. Since AlexNet made great achievements in the field of computer vision, various CNN architectures have come to the fore over the past decade with a series of breakthroughs. And CNNs have made big impact on measurement.

Measurement refers to measure, monitor, and/or record physical phenomena for the purpose of advancing measurement science, methods, functionalities and applications. As AI transforming various domains, measurement is no exception. Measurement can be as simple as direct-reading just by thermometers. However, it also may be highly hard even impossible due to the complexity and lack of the exact information about the object. When it comes to such cases, conventional methods are unlikely to work well, yet the CNN technique provides a practical solution, which directly comes up to the result without fully aware of the system.

This paper firstly gives a short survey of CNNs from software algorithms to hardware architectures perspective, which are listed in Sections 2 and 3. Later, the CNN-based applications on measurement are surveyed in Section 4. Section 5 concludes this paper. The study hopes to give contributions to the cross-research of AI and measurement.

2. Survey on softwar algorithm

CNNs are classified into four categories based on software algorithms: early models, simple-deeper models, block models, and lightweight models. The models introduced in this section represent the state-of-the-art of the time.

2.1. Early models

Early models refer to the basic theories of CNNs. Fukushima proposed an embryonic model of CNN architecture, which consists of seven layers in 1980 [10]. Later, LeNet was put forward and represented one of the early CNN models. LeNet defines the basic CNN units which consist of convolutional layer, pooling layer, and fully connected layer [8]. The network was successfully applied in handwritten digit recognition provided by the U.S. Postal Service. Although CNN models nowadays are quite different from LeNet, all of the models are based on the architecture.

2.2. Simple-deeper models

In 2012, AlexNet was proposed for the normal object recognition, which consists of five convolutional layers and the followed three fully connected layers. The model has a similar architecture with LeNet except for deeper, with more filters per layer and stacked convolutional layers. It was the first time that CNNs achieved great successes in computer vision. The model can be seen as the milestone in making CNNs more widely-applicable.

As AlexNet achieves good performance by CNN operations, several simple-deeper models are proposed. The simple-deeper models mean the plain architectures with layers such as convolutional layers and activations stacked one by one to improve the accuracy.

https://doi.org/10.1016/j.measen.2021.100080

Available online 20 September 2021

VGG [11] is a typical simple-deeper model, which makes the improvement over AlexNet by replacing the large kernel-sized filters (11 and 5 that AlexNet adopts) with multiple 3*3 size kernel filters one after another, and pushing the model gets deeper with more layers. The network presents that the depth of CNNs with multiple stacked small size kernel filters is efficient for increasing the performance of CNNs, because the increase in depth with multiple non-linear layers enable it to learn more complex patterns.

However, the increasing in both depth and width of networks brings improving performance for neural networks as well as drawbacks: the enlarged network means a massive number of parameters and dramatically demand of computational resources.

2.3. Block models

Block models refer to the architectures with specific function blocks such as Inception modules, Residual learning block, Dense block, and so on, which aim to eliminate problems within the module and improve the performance. These architectures have far more layers than simple-deeper models while less size and parameters. For example, DenseNet169 has about 10 times more layers than VGG19, with only tens of VGG19's size and parameters.

Inception [12–14]: As Neural Networks (NNs) with increasing depth and width, they might face the problem of overfitting, as well as much bigger size that demands more computing resources. The Inception block (Fig. 1 (a)) is proposed for overcoming the problem based on the Hebbian principle and the intuition of multi-scale processing. The model optimizes the computing resources inside the network: about four times more layers than VGG with only tenth of VGG's size and parameters.

Residual Net (ResNet): Although the depth is of crucial importance for NNs, the problem of the vanishing gradient has been raised as DNNs get deeper: accuracy gets saturated and then degrades rapidly in the training process. And naively adding layers leading to higher training error [15]. These problems indicate that DNNs are more difficult to train [16]. proposed the deep residual learning framework to address these problems: introduce the "shortcut connections" that skips one or more layers with an identity function. The building block is shown in Fig. 2.

Densely connected Convolutional Network (DenseNet) [17]: Instead of shorter connections with adjacent layers, the architecture strengthens the feature propagation by exploiting the potential of feature reuse: for each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers, as equation shows below:

$$x_l = H_l\,([x_0, x_1, \ldots, x_{l-1}]) \tag{1}$$

where x_l denotes the feature-maps of the *lth* layer, H_l is the function of operations such as Convolution, Batch Normalization, ReLU, Pooling. [x_0, x_1, …, x_{-1}] refers to the concatenation of preceding layers' outputs. The methods alleviate the problem of exploding gradients, substantially reduce the number of parameters and strengthen feature propagation.

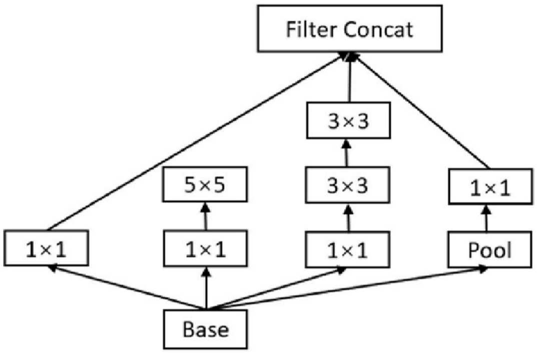

Fig. 1. An example of Inception block.

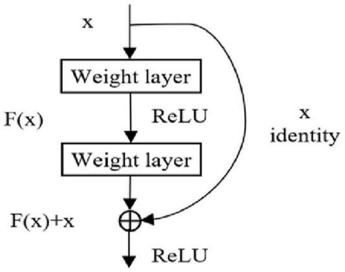

Fig. 2. An example of Residual block.

2.4. Lightweight models

As CNNs get deeper and larger with increase in hardware resources. Although Block models have optimized the size and parameters of CNNs with multiple algorithms, it is still difficult to deploy state-of-the-art CNN models on computation-constrained platforms such as embedded devices. And then, many works are focusing on designing high-efficiency architectures with small size for resource-limited devices. The representative light weight models are introduced in this section.

MobileNet [18] is based on a streamlined architecture by using depth-wise separable convolutions to reduce the model size and complexity. The method is to replace the computing expensive convolutional layers by a cheaper depthwise separable convolution, which is a 3*3 depthwise convolutional layer followed by a 1*1 convolutional layer.

ShuffleNet [19] uses pointwise group convolutions to reduce computation complexity of the costly dense 1*1 convolutions, with comparable accuracy and utilizes the channel shuffle operation to overcome the side effects brought by group convolutions.

MnasNet [20] proposes the automated mobile neural architecture search approach to find the best suitable architecture. The network incorporates platform-aware latency into the search process and utilizes a novel factorized hierarchical search space to get the resource-efficient result with good trade-offs between accuracy and latency.

3. Survey on hardware design

The successful application of CNNs in various fields with excellent performance is closely related to the support of computing capability of hardware. There are three major hardware solutions for the training and developing of CNNs: GPU, CPU, FPGA.

The prevalent machine learning frameworks such as Caffe, TensorFlow and Pytorch support both NVIDIA GPUs and CPUs, and the frameworks have been optimized for the two architectures with high-performance libraries (such as cuDNN for GPUs and MKL-DNN for CPUs). As to FPGA, various toolflows are also available to optimize and deploy trained deep learning networks with high-level programming languages, such as Intel's Deep Learning Acceleration (DLA) Suite and Xilinx's AI inference development stack-Vitis AI.

3.1. GPU and CPU

GPUs were originally designed for the graphic process. However, the architecture was started to be used as a general-purpose massive parallel processor. With the support of software frameworks such as CUDA to easily program for various algorithms executed on the hardware [21], GPUs have been one of the best efficient platforms for CNNs [22].

CPU is the primary processor for general computation and has been widespread deployment on various occasions. In terms of data-level parallelism, CPU performs the same operation simultaneously on multiple data points to speed up the data process with the support of the single-instruction-multiple-data (SIMD) [23] technique. Intel has released high-performance library of building blocks (MKL-DNN) to optimize the operators in CNN models, including convolution, pooling, batch normalization and activation. For now, performing DNNs on CPU is the preferred choice, especially for the simple network and the inference phase.

3.2. FPGA

FPGA is a highly flexible programmable device with the nature of reconfigurable and parallel processing ability, achieving low latency, low power consumption, high throughput and so on. These characteristics lend FPGA well suit to CNN applications and make up the shortcoming (Sparsity and floating-point operation vs approximate computing in CNNs) of GPU and CPU.

Sparsity brings redundant computation in CNN [24]. However, to fully optimize these meaningless operations, the activation sparsity is irregular and should be dynamically checked during the inference process of the model, which is instruction-consuming with large branch misses for CPU and GPU. In terms of the irregularity of sparse models, FPGA provides unsubstitutable solutions [25]: proposed Crane architecture, a hardware level load-balancing method, to mitigate the under-utilization problem caused by all kinds of sparsity irregularities. The method achieves 1.27x–1.88x speedup while saving 16%–48% energy consumption compared with the state-of-the-art accelerator baselines.

Approximate computing based on FPGA can significantly improve the efficiency of CNNs by balancing the tradeoff between accuracy, speed and power in limited hardware resources [26]. As “Approximate Multiply-Accumulate Array” proposed in [27], which offers a range of possible parametrizations: number of approximate multiply iterations and data bit width, improves the efficiency by 280% compared with general conventional multiply-accumulate operations.

In sum, each hardware has its own characteristics that serves specific needs with excellent performance. It is hard to say which is the best platform for DNNs tasks and it is up to concrete applications.

4. CNN applications in measurement

The field of Measurement has been greatly impacted by CNN applications. Techniques with CNN applied in the field of measurement not only make the improvement on the accuracy, robustness and stability, but also offer solutions for complex problems that traditional techniques lost their ways. One of the crucial strengths of CNNs is the capability of feature extraction. With sufficient datasets of objects and a pretrained model, CNNs are capable of extracting key patterns efficiently from large amounts and complex data. According to this, there has been multiple successful applications in measurement.

4.1. Vision-Based Measurement

For the outstanding performance of CNN on image recognition and the importance of Fault Detection (FD) and the general method of surface defect detection, Vision-Based Measurement (VBM) is the most widely used approach based on CNN for its nature in feature extraction [28]. For example: FD is crucial for the normal state of industrial equipment, and the surface defect detection is a general means [29]. proposed the deep neural network DEA_RetinaNet that based on RetinaNet with difference channel attention and adaptively spatial feature fusion for detecting the surface defects of steel. The method achieves 79.13 mean average precision for steel surface defect detection based on ResNet152 [30]; proposed an end-to-end combining prior knowledge with CNN to detect the weak scratch of optical components, with the pixel accuracy of 92.48% on the test data set [31]; proposed a method by combing symmetrize dot pattern (SDP) representation with squeeze-and-excitation-enabled convolutional neural network (SE-CNN) for bearing fault diagnostic. The method adopts CNN to implement the feature extraction of SDP images with high-efficiency so that the diagnostic can devote to major features without redundancy. The model achieves a classification rate of over 99% with better generalization ability and stability [31,32] proposed a Visual odometry (VO) system named DL_Hybrid for mobile robot Simultaneous Localization and Mapping (SLAM). The system incorporates CNN for image processing and geometric theory-based pose localization methods for SLAM. The experiment proves the better accuracy and robustness than traditional VO systems.

As is shown above, CNNs present their efficiency and powerful with multiple applications in VBM system.

4.2. Vibration signal detection system

CNNs are also widely used for fault detection according to vibration signals. For example [33], proposed the adaptive densely connected convolutional auto-encoder (ADCAE) for feature extraction from 1-D vibration signals: The method takes adaptive attention mechanism (AAM) for feature filtering, proposes a multiscale convolution based on AAM for fusion of multiscale information, and develops a new unsupervised-learning network, densely connected convolutional auto-encoder to improve information flow between encoder and decoder. The method performs quite better on gearbox fault diagnosis than typical CNNs [34]. put forward a fault diagnosis method for the high-speed train (HST) by virtue of the improved complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN) and 1-D convolutional neural network (1-D CNN), with the prediction accuracy of the proposed method is no less than 98%. As is shown that CNNs perform very well in terms of vibration signals.

4.3. Multi-signal detection system

Compared with the single signal diagnosis for FD, multi-signal detection system offers better accuracy, robustness and little overfitting problems [35]. presents a DCNN-based multi-signal fault diagnosis framework for induction motors. In the method, CNN is adopted to learning discriminative representations from the Time-Frequency Distribution (TFD) image, which is converted from multiple types of sensor signals generated simultaneously. Based on CNN, the system learns by itself and extracts specified features that contribute to accurate fault diagnosis, which are powerless for traditional techniques.

As can be seen above, from VBM to vibration signal, from single signal to multiple signals, CNN exhibits its superiority in the field of measurement. The research of CNN on measurement is still on the way, more and more solutions will be put forward for unresolved issues and make significant performance improvements.

4.4. Discussion about CNNs selection in measurement

First, many kinds of CNN architectures have been designed as introduced in Section 2. Each of them has its own specific features. As surveyed above, applications in measurement based on a certain CNN have shown CNNs’ advantages. However, there is not enough research on the CNN architectures that which are best suitable for the applications. Optimization of CNNs on measurement awaits further study.

Second, due to the hardware resource constraints, large models such as VGG with the size of more than 500MB are difficult to deploy for measurement, and light weight models would be better choices. Research would be committed to designing the suitable light weight models for various measurement applications.

It can be seen that although CNNs have achieved remarkable advances in measurement, there is still much room for improvement.

5. Conclusions

This paper investigates the CNN-based techniques applied in measurement in detail. Multiple representative CNN models are analyzed as well as the hardware architectures that support CNNs. In the field of measurement, CNN-based methods provide the solutions with high efficiency, accuracy and better stability, especially for the complex systems that without enough understanding of the measure objects, such as the principles of parameters. In terms of such problems, conventional techniques are ineffective, while CNNs offer the straightforward solutions with feature extraction and recognition by self-learning, all that is needed is enough data. As CNN algorithms develop with high efficiency and accuracy, CNNs have great potential in measurement and will continue to promote the development of measurement greatly.

References

[1] A. Khan, et al., A survey of the recent architectures of deep convolutional neural networks, Artif. Intell. Rev. 53 (2020) 5455–5516.
[2] S. Kiranyaz, et al., 1D convolutional neural networks and applications, A surveyMechanical Systems and Signal Processing 151 (No) (2021) 107398.
[3] L. Meng, et al., Oracle Bone Inscription Detector Based on SSD, ICIAP2019, 2019.
[4] B. Lyu, et al., Frame Detection and Text Line Segmentation for Early Japanese Books Understanding, ICPRAM2020, 2020, pp. 600–606.
[5] L. Meng, et al., Underwater-drone with panoramic camera for automatic fish recognition based on deep learning, IEEE ACCESS 6 (No.1) (2018) pp17880–17886.
[6] Z. Wang, et al., Deep Learning-Based Elderly Gender Classification Using Doppler Radar, Personal and Ubiquitous Computing, 2021.
[7] M. Khanafer, et al., Applied AI in instrumentation and measurement: the deep learning revolution, IEEE Instrum. Meas. Mag. 23 (No.6) (2020) 10–17.
[8] Y. LeCun, et al., Backpropagation applied to handwritten zip code recognition, Neural Comput. (1989) 541–551.
[9] A. Krizhevsky, et al., Imagenet classification with deep convo-lutional neural networks, NIPS2012 (2012) 1097–1105.
[10] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (1980) 193–202.
[11] S. Karen, et al., Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015, p. ICLR2015.
[12] C. Szegedy, et al., Going Deeper with Convolutions, CVPR2015, 2015, pp. 1–6.
[13] S. Ioffe, et al., Batch normalization: accelerating deep network training by reducing internal covariate shift, the 32nd, Int. Conf. on Machine Learning (2015) 448–456.
[14] C. Szegedy, et al., Rethinking the Inception Architecture for Computer Vision, CVPR2016, 2016, pp. 2818–2826.
[15] K. He, J. Sun, Convolutional Neural Networks at Constrained Time Cost, CVPR2015, 2015, pp. 5353–5360, 2.
[16] K. He, et al., Deep Residual Learning for Image Recognition, CVPR2016, 2016, pp. 770–778.
[17] G. Huang, et al., Densely Connected Convolutional Networks. CVPR2016, 2017, pp. 2261–2269.
[18] A.G. Howard, et al., Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications, CORR, 2017.
[19] X. Zhang, et al., ShuffleNet: an Extremely Efficient Convolu-Tional Neural Network for Mobile Devices, CVPR2018, 2018.
[20] M. Tan, et al., MnasNet: Platform-Aware Neural Architecture Search for Mobile, CVPR2019, 2019, pp. 2815–2823.
[21] D. Strigl, et al., Performance and scalability of GPU-based convolutional neural networks, in: The 18th Euromicro Conf. On PDP, 2010, pp. 317–324.
[22] G. Stoica, et al., Speeding-up image processing in reaction-diffusion cellular neural networks using CUDA-enabled GPU platforms, in: The 6th Int. Conf. On ECAI, 2014, pp. 39–42.
[23] Y. He, et al., A configurable SIMD architecture with explicit datapath for intelligent learning, in: 2016 Int. Conf. On SAMOS, 2016, pp. 156–163.
[24] J. Albericio, et al., Ineffectual-Neuron-Free Deep Neural Network Computing, ISCA2016, 2016.
[25] Y. Guan, et al., Crane: mitigating accelerator under-utilization caused by sparsity irregularities in CNNs, Trans. on Computers 69 (7) (2020) 931–943.
[26] S. Venkataramani, et al., Efficient AI system design with cross-layer approximate computing, Proc. IEEE 108 (12) (2020) 2232–2250.
[27] Z. Wang, et al., Approxiate multiply-accumulate Array for convolutional neural ntworks on FPGA, in: The 14th Int. Symp. On ReCoSoC, 2019, pp. 35–42.
[28] S. Shirmohammadi, et al., Camera as the instrument: the rising trend of vision based measurement, *IEEE* Instrum. Meas. Mag. 17 (3) (2014) 41–47.
[29] X. Cheng, et al., RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection, *IEEE trans*. instrum. meas. 70 (2021) 1–11.
[30] W. Hou, et al., Combining prior knowledge with CNN for weak scratch inspection of optical components, *IEEE trans*. instrum. meas. 70 (2021) 1–11.
[31] H. Wang, et al., A new intelligent bearing fault diagnosis method using SDP representation and SE-CNN, *IEEE trans*. instrum. meas. 69 (5) (2020) 2377–2389.
[32] X. Ban, et al., Monocular visual odometry based on depth and optical flow using deep learning, *IEEE trans*. instrum. meas. 70 (2021) 1–19.
[33] M. Miao, et al., Adaptive densely connected convolutional auto-encoder-based feature learning of gearbox vibration signals, IEEE trans. instrum. meas. 70 (2021) 1–11.
[34] D. Huang, et al., fault diagnosis of high-speed train bogie based on the improved-CEEMDAN and 1-D CNN algori-thms, *IEEE trans*. instrum. meas. 70 (2021) 1–11.
[35] S. Shao, et al., DCNN-based multi-signal induction motor fault diagnosis, *IEEE trans*. instrum. meas. 69 (6) (2020) 2658–2669.

Hengyi Li
Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan

Xuebin Yue
Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan

Zhichen Wang
Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan

Wenwen Wang
Department of Computer Science, University of Georgia, Athens, GA, USA

Hiroyuki Tomiyama
Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan

Lin Meng*
Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan

* Corresponding author.
E-mail address: menglin@fc.ritsumei.ac.jp (L. Meng).