Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.035377
Article

Tech Science Press

Check for updates

Identification of Key Links in Electric Power Operation Based-Spatiotemporal Mixing Convolution Neural Network

Lei Feng[1], Bo Wang[1,*], Fuqi Ma[1], Hengrui Ma[2] and Mohamed A. Mohamed[3]

[1]School of Electrical and Automation, Wuhan University, Wuhan, Hubei, 430072, China
[2]Tus-Institute for Renewable Energy, Qinghai University, Xining, Qinghai, 810016, China
[3]Electrical Engineering Department, Faculty of Engineering, Minia University, Minia, 61519, Egypt
*Corresponding Author: Bo Wang. Email: whwdwb@whu.edu.cn
Received: 18 August 2022; Accepted: 28 October 2022

Abstract: As the scale of the power system continues to expand, the environment for power operations becomes more and more complex. Existing risk management and control methods for power operations can only set the same risk detection standard and conduct the risk detection for any scenario indiscriminately. Therefore, more reliable and accurate security control methods are urgently needed. In order to improve the accuracy and reliability of the operation risk management and control method, this paper proposes a method for identifying the key links in the whole process of electric power operation based on the spatiotemporal hybrid convolutional neural network. To provide early warning and control of targeted risks, first, the video stream is framed adaptively according to the pixel changes in the video stream. Then, the optimized MobileNet is used to extract the feature map of the video stream, which contains both time-series and static spatial scene information. The feature maps are combined and non-linearly mapped to realize the identification of dynamic operating scenes. Finally, training samples and test samples are produced by using the whole process image of a power company in Xinjiang as a case study, and the proposed algorithm is compared with the unimproved MobileNet. The experimental results demonstrated that the method proposed in this paper can accurately identify the type and start and end time of each operation link in the whole process of electric power operation, and has good real-time performance. The average accuracy of the algorithm can reach 87.8%, and the frame rate is 61 frames/s, which is of great significance for improving the reliability and accuracy of security control methods.

Keywords: Security risk management; key links identifications; electric power operation; spatiotemporal mixing convolution neural network; MobileNet network

1 Introduction

With the continuous expansion of the power system, the environment for power operations becomes more complex, and at the same time, it faces challenges such as high-altitude and high-voltage operations [1]. In recent years, power production accidents occur frequently, and the operational risks at the power

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

site seriously threaten the safety of production personnel, at the same time, the production work will be stagnant, affecting the stable operation of the power system and safe production [2]. Therefore, it can be seen that it is of great significance to study the operational risk management and control of power sites.

The development and research of operational risk management and control in power sites have gone through a safety inspection mode based on manual supervision, a safety inspection mode based on information equipment such as fixed video surveillance, and an intelligent computer vision method [3]. With the development of high-definition video surveillance, high-definition cameras are widely used in power construction sets for all-round, large-scale, and long-term supervision. However, the viewing range of the human eye is limited, and long-term viewing can easily cause visual fatigue. With the rapid development of deep learning, deep learning algorithms can be used to perform hazard detection on the returned images. When risk information is detected, an alarm prompt is issued and then safety supervisors will check according to the video information, which greatly improves the safety of safety supervisors [4].

The current mainstream risk management and control algorithms focus on target detection algorithms. For example, Bai et al. [5] extracted the features of production site images by constructing a deep convolutional neural network, and then used the multi-scale classification network SSD (Single Shot Multibox Detector) to detect helmets. Xu et al. [6] used the improved YOLOv3 algorithm and adopted a multi-stage transfer learning strategy to achieve high-precision real-time detection of insulating clothing. However, the target detection algorithms proposed above can only set the same detection standard to detect any scene indiscriminately. For example, when detecting electric work scenes without seat belts, seat belts will be detected for all scenes, and if no seat belts are detected, it will be judged as illegal work. Therefore, this type of target detection method can only be applied to the detection that is common to all scenes. In fact, a complete power operation process includes multiple operation links, and the operation categories of each operation link have different operational requirements. For example, goggles are required for electric welding scenarios, and insulating gloves and insulating boots are required for high-voltage operation scenarios. That is to say, for different power operation links, it is necessary to detect different targets to determine whether they meet the requirements of safe operation, so as to achieve reliable safety management and control. It can be seen that, for the risk management and control of the whole process of electric power operation, realizing the identification and judgment of different operational links is particularly important to improve the reliability and accuracy of risk management and control of the whole process of electric power operation.

However, the current research on the identification of all aspects of power operations is still very rare, and it has not been fully integrated with the field of safety risk management. Hu et al. [7] exploratory put forward a model based on LRCN power on-site personnel operation scene recognition method, through the fusion of visible light and optical flow information to realize the operation scene, but this method needs adjusting, optical and optical flow, and to obtain implementation process is complicated, difficult to deploy to the job site. Popoola et al. [8] proposed an online detection method for personnel operation scene identification based on the elastic jump, and sliding window, which ensured real-time performance by skipping redundant data detection. However, this method could not identify different operation scenes and would miss a large number of reports, so its reliability was relatively low; Shi et al. [9] proposed an LSTM-based typical work behavior recognition method for power operation and maintenance operations. The on-site operators wear wrist sensors to collect motion information, and then use the LSTM model to identify and classify the motion information, and judge its behavior. Whether or not the operation is being performed, this method uses neural networks for the first time, and the accuracy is greatly improved, but it requires the operator to wear additional sensors; it can be seen that there is currently a lack of reliable and real-time identification algorithms for power operations.

Aiming at the problem of insufficient reliability of power operation risk management and control methods due to the lack of scene identification methods, which can only detect common operation risks

in all scenes indiscriminately; this paper conducts a research on the identification method of key operation links in the whole process of electric power operation. Different operating scenarios are distinguished to provide a scenario basis for safety management and control methods, thereby improving the reliability and accuracy of safety management and control methods. The main research contents are as follows:

(1) Aim to the fact that the safety risk of power field operation is scene dependent, which means the safety risk of key links in the whole process of power operation is much greater than that of other links, a key link identification method of the whole process of power operation based on deep learning is proposed.

(2) According to the characteristics of obvious personnel action characteristics and long-time scale in power operation, a spatiotemporal mix convolution network based on deep separable convolution is constructed. In the convolution process, linear interpolation and feature fusion are carried out through the two-dimensional spatial features of each frame according to the time dimension of the operation link, to obtain the ability of time-series modeling. The network is very suitable for dynamic scene recognition of long-time scale, such as power operation.

(3) According to the features of power operation, such as located outdoors, changing locations and real-time security risk early warning, the acceleration and edge deployment strategy for this model is proposed. The model is quantified, pruned and operator fused, and the optimal operation strategy is adaptively matched according to the hardware layout and performance. The acceleration of the neural network is realized to ensure the real-time performance of the model, thereby, improving the reliability and practicability of the algorithm.

(4) According to the method and technology proposed in this paper, with the embedded chip of NVIDIA nano [10] at the edge end, typical personnel operation scenes can be detected. Such as the start and end time of pre-shift meetings, power inspection, and pulling the disconnector during the switching operation of the personnel of a power supply company. The accuracy is 87.8%.

The subsequent sections of this paper are arranged as follows: Section 2 introduces the relevant technical background; Section 3 introduces the model framework of this paper, Section 4 carries out comparative experiments and results analysis, and Section 5 introduces the conclusion and prospects.

2 Introduction of Technical Background

2.1 Convolution and Pseudo 3D convolution

3D convolution was first proposed by Ba et al. [11,12] and successfully applied to the field of motion recognition. *3D* convolution refers to the convolution neural network whose convolution kernel is three-dimensional, and its convolution kernel increases the time dimension on the basis of the original two-dimensional. Therefore, for *3D* data such as video data, *3D* convolution can extract the feature information of the data more completely than *2D* convolution. On the other hand, the amount of computation of *3D* convolution is also significantly larger than that of *2D* convolution. The research shows that the floating-point operation N_{3d} of *3D* convolution is $K \times T$ times that of *2D* convolution N_{2d}. Where K is the size of the convolution kernel and T is the time scale. The high amount of computation makes the deployment of *3D* convolution difficult, while the high amount of parameters makes *3D* convolution easy to overfit and fall into local optimal solutions during training, which greatly increases the difficulty of training. Therefore, many scholars are committed to finding a balance between the performance and the amount of computation of *3D* and *2D* convolution [13], so there are many Pseudo *3D* convolution structures.

Pseudo *3D* convolution structures refer to the *2D* convolution structure or mixed convolution structure constructed by imitating the idea of *3D* convolution to obtain spatiotemporal characteristics [14]. The feature extraction ability and parameter quantity of pseudo-*3D* convolution structure is between *2D* convolution and

3D convolution. At present, there are three main technical ideas: mix *2D* and *3D* convolution Architecture [15]; decompose *3D* convolution into *2D* convolution and *1D* convolution [16]; fuse *2D* convolution multiple channels [17]. Mixed *2D* and *3D* network architecture generally use *3D/2D* convolution modules in series, which can not only effectively increase the depth of *3D* CNN, but also strengthen the learning ability of *2D* airspace, to reduce the parameter size of the model and improve the efficiency of the model. The core idea of *3D* convolution decomposition is that each subset of *3D* convolution can be decomposed into *2D* convolution and *1D* convolution. After decomposition, the size of the convolution kernel can be reduced from $n_h \times n_w \times n_t$ to $n_h \times n_w + n_t$. Although this reduces the representational ability of *3D* convolution to a certain extent, it greatly reduces the number of parameters. The main concept of *2D* convolution fusion on multi-channel is to exchange *2D* convolution features of different frames artificially, to ensure a certain time modeling ability under the same amount of computation as *2D* convolution.

2.2 Power Vision Edge Intelligence

The concept of power vision edge intelligence was first put forward by a research team from Wuhan University, led by Wuhan University [18]: Power Vision edge intelligence is a new way that power depth vision and edge intelligence can empower each other. For the collected power vision image, it can complete the image analysis and calculation closer to the perception terminal. At present, there are two kinds of technical ideas of power edge intelligence, one is the model compression strategy through channel pruning and parameter reduction [19], and the other is the hardware-based model acceleration strategy [20]. The main methods of model compression strategy include network pruning, quantification, low-rank decomposition and knowledge distillation [21]. The basic idea of network pruning is to reduce the redundancy of the model by cutting the unimportant parts in the model. The parameter cutting can be written in the form of constrained optimization and transformed into a combinatorial optimization problem. Quantization is to convert the floating-point algorithm of the neural network into a fixed-point, which means replacing the float 16 or float 32 floating-point operations with 8 integer operations in the model inference stage. The basic idea of low-rank decomposition is to decompose the original large weight matrix into multiple small matrices, and use the low-rank matrix to approximate the original weight matrix. After decomposition, the sum of the calculation amount of all small matrices is smaller than that of the original large matrix, thus reducing the calculation amount. The basic idea of knowledge distillation is to use the knowledge learned from the large model to guide the training of the small model, so that the small model has the same performance as the large model, but the number of parameters is greatly reduced, in order to realize model compression and acceleration. The hardware-based model acceleration strategy is to complete the actual deployment of the deep learning model in the field of data center or edge computing through heterogeneous computing methods and a co-processing hardware engine.

2.3 Deep Separable Convolution

Deep separable convolution was first proposed in [22]. Its core idea is to realize the low coupling of convolution structure and reduce the parameters of the model by extracting channel correlation and spatial correlation separately. Deep separable convolution is mainly divided into two steps, deep convolution and point-by-point convolution. Firstly, in the deep convolution layer, the number of convolution cores with the same number of channels as the input data is used for feature extraction to ensure that the number of channels of the output feature map is consistent. Then point by point convolution is carried out to make the characteristic information between channels interact and integrate. The point-to-point convolution multiple convolution cores of C_{11} size, where C is the number of channels of the input data. The feature map obtained by depth convolution is weighted and combined in the depth direction, to output a new feature map including channel correlation and spatial correlation. Deep separable convolution reduces the computational complexity of convolution calculation through deep

convolution and point-by-point convolution, and maintains the feature extraction effect of standard convolution, which has a positive influence on the lightweight of the network.

3 Introduction to Model Framework

3.1 Introduction to the Overall Framework

In this paper, a spatiotemporal hybrid convolution network module based on *2D* convolution is designed to linearly interpolate and fuse some features of continuous frames in a specific time order and mode, the mixed extraction of continuous frame timing information and static spatial information on the premise of far less than the amount of *3D* convolution calculation. Then the lightweight transformation is carried out to make it more accurate to identify the long-timescale scenes such as power operation and facilitate the edge deployment of the power field. The overall architecture of this method is shown in Fig. 1.

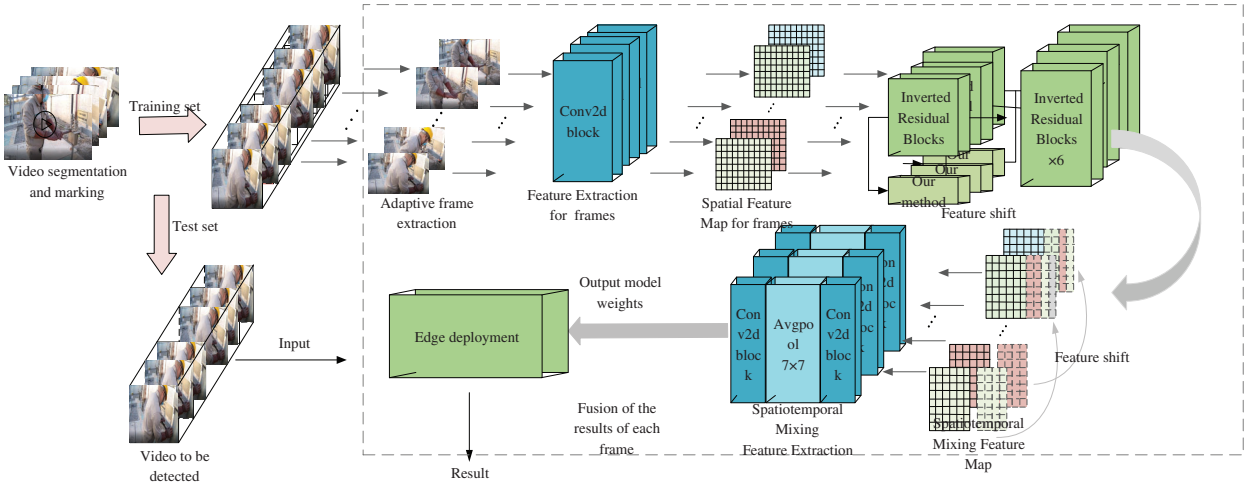

Figure 1: Model framework of the proposed method

When training, the collected actual monitoring video of electric work is divided into 1 to 3 min video segments according to the electric work scene it contains, and each video segment is labeled accordingly. Then, each video segment is differentiated between adjacent frames, and adaptive framing is performed according to the different results. In the training phase, take each video segment as the basic unit, input all images in the video segment into the feature extraction network in turn to obtain the spatial feature matrix, and then conduct temporal offset and fusion of the spatial feature matrix to obtain the Spatio-temporal hybrid feature matrix. Finally, input the Spatio-temporal hybrid matrix into the full connection layer for dimension reduction and normalization to obtain the final prediction result. The loss function iterates the model weight based on the predicted results and the actual results. After the best weight is obtained through iteration, the weight is quantized and compressed and deployed on a low-power computer.

3.2 Feature Extraction Basic Network

Considering the long-timescale characteristics of power operation scenarios and the dynamic real-time characteristics of power security risks, in order to avoid the explosion of calculation caused by the sudden increase of parameters in a long-timescale [23–25]. In this paper, the improved MobileNetv2 is used as the basic network for single-frame feature extraction [26]. MobileNetv2 is a lightweight convolution neural network based on deep separable convolution. The core structure is a separable convolution module and an inverted residual structure. Its network structure is shown in Table 1. In this table, conv2d

is a two-dimensional convolution operation, Bottleneck is a reverse residual block, *Avgpool* is a global pooling operation, it is the channel expansion factor, *C* is the number of output channels, n is the number of block repetitions, and S is the step size. MobileNetv2. The neural network can efficiently extract spatial features, but it does not have the modeling ability in any time dimension.

Table 1: Framework of MobileNetV2

Input	Operation	t	c	n	s
$224^2 \times 3$	Conv2d	–	32	1	2
$112^2 \times 32$	Bottleneck	1	16	2	1
$112^2 \times 16$	Bottleneck	6	24	3	2
$56^2 \times 24$	Bottleneck	6	32	4	2
$28^2 \times 32$	Bottleneck	6	64	3	2
$14^2 \times 64$	Bottleneck	6	96	3	1
$14^2 \times 96$	Bottleneck	6	160	1	2
$7^2 \times 160$	Bottleneck	6	320	1	1
$7^2 \times 320$	Conv2d	–	1280	1	1
$7^2 \times 280$	Avgpool 7×7	–	–	1	–
$1^2 \times 280$	Conv2d	–	–	–	–

In order to solve the above shortcomings, this paper improves the MobileNet neural network as follows: As shown in Fig. 2, this paper retains the separable convolution module as the spatial feature extraction branch, and embeds the feature exchange module proposed in this paper into the skip connection branch of the structure of the inverted residual as the sequential action feature extraction branch. For the improved feature extraction network, on the one hand, the backbone and framework backbone is not damaged, and its spatial feature extraction ability is almost identical to the original. On the other hand, the embedded timing feature offset module is connected in series and parallel with the structure of the inverted residual and carries out feature exchange in turn, which realizes the deep exchange and fusion of adjacent frame features, so it has the ability of efficient timing action feature extraction [27].

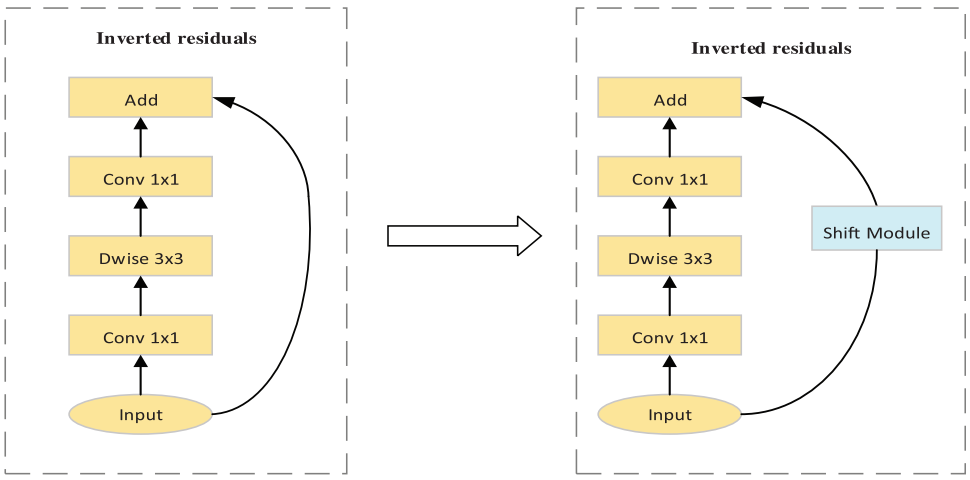

Figure 2: Improved backbone

3.3 Spatiotemporal Mix Feature Extraction Network

The spatiotemporal mix feature extraction network is the core of this model. It imitates the working mode of *3D* convolution. By interleaving and fusing the features of adjacent frames of the video stream in the time dimension, the low parameter *2D* convolution neural network MobileNet can obtain the ability of temporal action modeling. It is an efficient, lightweight behavior recognition network for power personnel. Spatiotemporal mix feature extraction network is based on the idea of time-series offset, namely exchanging features between different frames along the time dimension, which can not only promote the information exchange between adjacent frames, but also bring no additional computation. The mathematical essence behind the temporal shift network is the separability and linear interpolation of shift operation and multiplication operation in convolution operation, which means the linear interpolation of the characteristic graph during the shift operation, will not affect the subsequent multiplication operation and lead to the failure of the convolution operation. If the convolution kernel is $W = (w_1, w_2, w_3)$, and the input is a one-dimensional vector X, the convolution operation can be written as follows:

$$Y_i = w_1 X_{i-1} + w_2 X_i + w_3 X_{i+1} \tag{1}$$

At the same time, the convolution operation can also be divided into two steps. First, the shift operation is carried out, and the original input is X_i^0. The input is moved forward and backward by one bit respectively to obtain the sum $X_i^{-1}$and X_i^{+1}. Then the multiplication operation is carried out, which can be expressed as follows:

$$Y_i = w_1 X_i^{-1} + w_2 X_i^0 + w_3 X_i^{+1} \tag{2}$$

The separability of shift operation and multiplication operation in convolution operation ensures the effectiveness of convolution after the feature exchange. In order to ensure the lowest parameters and light weight of the model, the space-time mix feature extraction network directly embeds the temporal shift operation into the feature extraction network MobileNet, and changes the original spatial extraction network in a space-time extraction network. The key is how to take the feature extraction ability of time and space into account. Too much time-series offset network intrusion will destroy the spatial modeling ability of the backbone and seriously consume memory resources. Too little deployment will lead to insufficient time modeling ability. To solve the above problems, the temporal shift operation proposed in this paper is shown in Fig. 3, which is divided into the feature interleaving part and feature fusion partitions. After many experiments, this paper believes that deploying the feature exchange module on the skip connection branch of the inverted residuals structure in the MobileNetv2 network can better consider the spatial modeling ability of the backbone itself and the sequential action extraction ability of adjacent frames. In the figure, *T*, *C*, *W*, *H* respectively represent the time dimension, number of channels, width and height of the characteristic diagram of the video stream at the skip connection of the inverted residuals structure. The first step is to segment the feature map according to the number of channels along the time dimension. The segmentation proportion is $K(0 \leq K \leq 1)$. In this model, K is related to the time dimension T and frame extraction frequency f, which is:

$$K = k \cdot \frac{1}{T \cdot f} \tag{3}$$

where, k is the scale coefficient. After segmentation, the following result can be obtained:

$$\begin{cases} M_1 = K \times C \times T \times WH \\ M_2 = (1 - K) \times C \times T \times WH \end{cases} \tag{4}$$

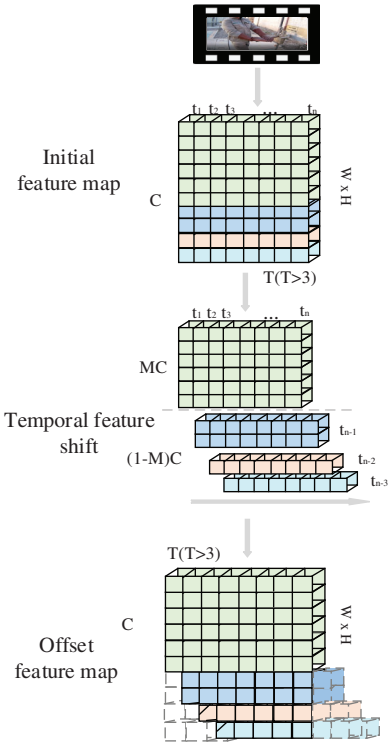

Figure 3: Temporal shift operation

The second step is to offset the feature map M_1 to be exchanged, whose mathematical essence is linear interpolation. As shown in the figure, first divide the number of channels of M_1 by 2:1:1. M_{11}, M_{12}, M_{13} backward by +1, +2 and +4 steps along the time dimension. For the characteristic image of the frame $(t \geq 4)$, before the offset, the following is obtained:

$$M^t = M_1^t + M_2^t \tag{5}$$

After the offset is completed, the result can be written as:

$$M^t = M_{11}^{t-1} + M_{12}^{t-2} + M_{13}^{t-4} + M_2^t \tag{6}$$

Thus, the feature map of the t-th frame at this time is composed of the features of the current frame, the previous frame, the first two frames and the first four frames. For each feature exchange, the temporal receptive field will be expanded by 4. This operation mode is like running the convolution with a kernel size of 5 in the time dimension. Finally, our model has a very large time receptive field for highly complex time modeling.

In step three, the reasoning results of each frame of the whole video segment are fused to give the final result. In this paper, the attention mechanism is used to fuse the results of each frame, and its calculation formula is as follows:

$$C_i = \sum_{j=1}^{L_x} a_{ij} h_j \tag{7}$$

3.4 Model Acceleration and Edge Deployment

Achieving the simplification of the model and high hardware efficiency is an important step in the actual deployment at the edge. The model proposed in this paper is a lightweight model architecture, which has

natural advantages in the deployment at the edge. However, the running environment of the server side is very different from that on the edge side. If the model of the server-side is transplanted directly to the edge side, the execution efficiency of the algorithm is often poor, resulting in the reduction of the practicability and real-time performance of the algorithm. To solve this problem, this paper proposes an edge deployment strategy based on the action recognition model, including model acceleration strategy and hardware acceleration strategy. As shown in Fig. 4, first implement the model acceleration strategy. Since the server-side model pays more attention to the flexibility of model construction and the edge-side model pays more attention to the efficiency of model execution, it is necessary to convert the server-side model.

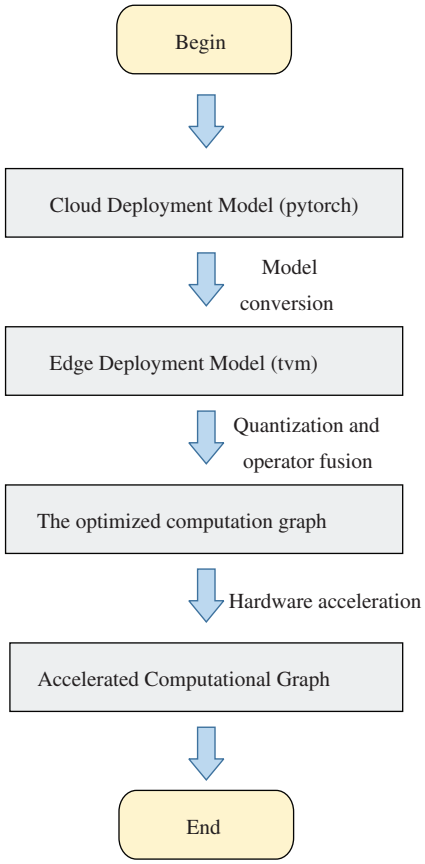

Figure 4: Model acceleration and deployment on the edge

This paper uses the open-source architecture TVM architecture to transform the PyTorch model suitable for server-side into the open-source architecture TVM model, and generates an efficient model calculation diagram suitable for edge deployment. Then the model calculation diagram is simplified and optimized, including int8 quantization and operator fusion to generate the optimized calculation diagram. Among them, operator fusion refers to combining multiple operators into the same core without saving the intermediate results to the global memory, to reduce the time required for execution; Int8 quantization refers to reducing the saving type of model weight from float16 to int8, to reduce the memory occupied by the weight. Secondly, implement the hardware acceleration strategy, deploy the optimized calculation diagram to the hardware terminal, use the hardware automatic search strategy to optimize the memory usage of the calculation diagram and finally complete the deployment.

4 Comparative Experiment

The software and hardware platform used in this experiment is configured as follows: Intel Core i5 on the server-side- 10400F@2.90 GHz × 6 CPUs, NVIDIA Geforce RTX 2060, operating system: Ubuntu 16.04lts, deep learning framework: PyTorch; The edge end is a small embedded device with low power consumption, NVIDIA Jetson nano, quad-core arm A57@1.43 GHz × 4 CPUs, 128-core Maxwell GPUs.

4.1 Data Set

The experimental data used in this paper are from the on-site images recorded by the safety recorder during the switching operation of a power supply company in Xinjiang in recent three years. The safety recorder completely records the whole operation process; the duration is generally 60 to 180 min. According to the safety regulations of the State Grid, the switching operation includes six types of typical personnel operation scenarios, including a pre-shift meeting, an inspection of insulating gloves, an inspection of the electric rod, operation of the disconnector, and power inspection and five prevention simulations. As shown in Fig. 5, the safety regulations are also different in different operational scenarios. Operators must wear insulating gloves in specific operation scenarios such as power inspection and closing disconnects, but they do not need to wear them at other times. This paper first clips and filters the video into 1–3 min video segments. Each video segment contains at least one kind of typical operational behavior, and labels the video segments with labeling tools; including action category and action start time. The final data set includes 235 videos. The dataset details are shown in Table 2.

(a) Pre shift meeting (b) Check Insulated Gloves (c) Electricity check

(d) Operate the Isolator (e) Check the electroscope (f) Five Electric Preventions

Figure 5: Key links in the whole process of electric power operation

Table 2: Description of dataset

Key links category	Pre-shift meeting	Check insulated gloves	Check the electroscope	Operate the isolator	Electricity check	Five electric preventions
Number of videos	51	49	32	37	31	35

4.2 Model Training and Parameter Setting

It can be seen from Table 2, the characteristic of the formal data set in this paper is the number of data samples is small, but the action time is generally much longer than that of ordinary actions. Taking the power test as an example, it generally takes 60 s from the beginning of wearing insulating gloves to the end of the power test, and 1800 frames are included at the frame rate of 30 frames/second. In view of the above characteristics, this paper adopts the following training strategies:

(1) Transfer learning. This paper adopts the idea of transfer learning. Firstly, the training iteration is 10000 times in the public auction data set somethingv2, so that the model can be fully trained and convergent. After obtaining the optimal weight of pre-training, the training can be officially started in the data set.

(2) Random sparse sampling to optimize the data set. In this paper, the original video segment of the data set is divided into 30 segments according to the time length, and then 3~5 pictures are randomly sparse sampled in each segment, and finally spliced into a new video segment. If the original video segment is repeated many times, multiple new videos can be obtained. On the one hand, random sparse sampling ensures that there is little loss of video level features, while greatly reducing the redundancy of the video stream [1], to speed up the convergence speed of the model. On the other hand, repeated multiple sampling can effectively expand the data set. After sampling, the video segment of the data set has expanded from 235 to 940 segments.

In the training process, the videos in the video segment are uniformly adjusted to 320 * 320. After the K-mean clustering of the data set, the adaptive sampling frame rate f is adopted, the time dimension resolution T is set to 64, and the offset coefficient K is set to 12. The random gradient descent method is used to process one video segment at a time, the initial learning rate is 0.001, and the cosine annealing learning rate is used to carry out 3000 iterations. After full convergence, the best model is obtained.

4.3 Experimental Result

The experimental part of this paper is divided into server-side performance comparison experiment and edge-side efficiency comparison experiment. Since the model proposed in this paper is based on the MobileNet network and the parameters are almost consistent with MobileNet, the comparative experiment is carried out on the server side with this model and the MobileNet model with zero offsets. The loss convergence curve of the loss value function is shown in Fig. 6. The red line represents the MobileNet result curve with zero offsets, the blue line is the resulting curve of the improved method in this paper, the abscissa represents the number of iterations of the network model, and the ordinate represents the loss value in the training process. It can be seen from the figure that the convergence speed of the two is basically the same. However, the loss value of the model proposed in this paper is finally stable at 0.017, which is lower than the 0.035 of MobileNet, indicating that the training difficulty of the model proposed in this paper is almost the same as that of the 2D convolutional neural network MobileNet. But the learning ability of video data is obviously stronger than that of 2D convolutional neural networks without

time modeling ability. The mean average precision (map) curvatures and F1 curve of this model and the zero offset MobileNet model are shown in Fig. 7.

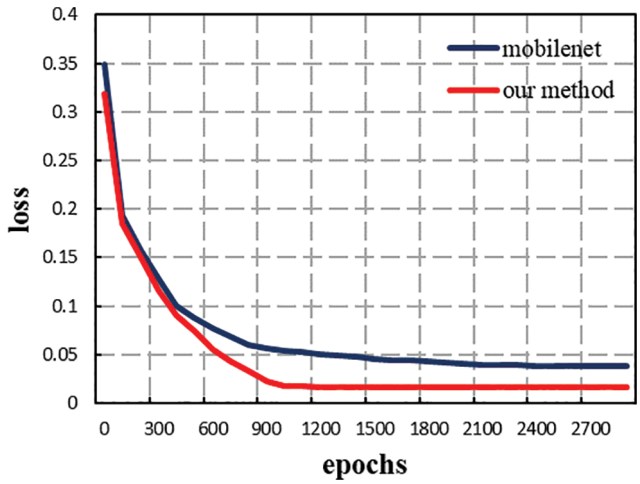

Figure 6: Loss value function curvatures

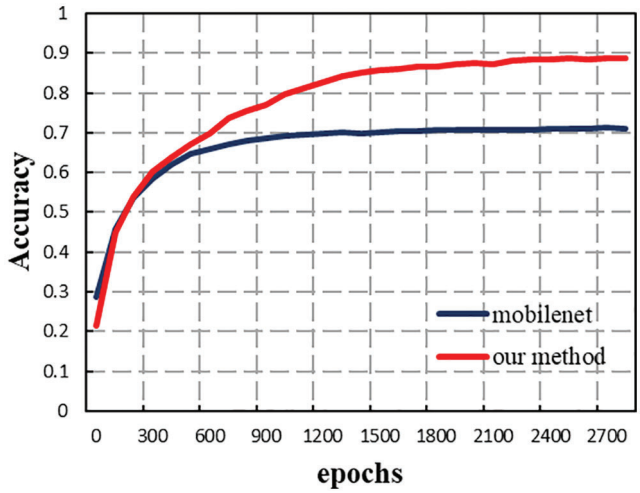

Figure 7: Accuracy function curvatures

As can be seen from Fig. 7, the accuracy of the model in this paper rose to 0.6 after iterating about 30 times, and finally stabilized at about 0.88, achieving high video-level detection accuracy, while MobileNet finally stabilized at about 0.7, which fully illustrates the effectiveness of the spatiotemporal mixing network. The spatiotemporal mixing networks can indeed enable 2D convolutional neural networks to gain temporal modeling capabilities. Table 3 shows the model output.

Table 4 shows the recognition accuracy of each person's work scene. It can be seen from the table that the prediction accuracy of pre-shift meetings is up to 94%, and the accuracy of complex actions such as operating disconnects is also up to 86%.

Table 3: Model output

Video segment	Predict result	Confidence	Begin frame	Finish frame
1	Check insulated gloves	0.67	33	112
2	Background	0.91	130	154
3	Background	0.78	158	202
…	…	…	…	…
16	Check the electroscope	0.83	1197	1276

Table 4: MAP of the proposed method

Key links category	Key links category	Pre-shift meeting	Check Insulated Gloves	Check the electroscope	Operate the Isolator	Electricity check	total
mAP/%	94.1	89.3	87.3	86.7	85.5	85.1	88.1

In order to further verify the advantages of this algorithm, this algorithm is compared with the mainstream action recognition network, 3D convolutional neural network and slow-fast model. As shown in Table 5, the 3D convolutional neural network and slow-fast model are slightly higher than the model in this paper, but the reasoning time is much longer than the model in this paper.

Table 5: MAP and time cost comparison of different methods

Method	mAP/%	Time cost/s
Slow/fast	92.6	0.575
3D CNN	88.3	0.330
	69.1	0.017
Proposed method	87.8	0.021

The comparison between the method in this paper and MobileNetV2 model shows that the timing offset module proposed by us is meaningful and successful for the transformation of MobileNetV2 network. A spatiotemporal hybrid convolution network module based on 2D convolution is designed to linearly interpolate and fuse some features of continuous frames in a specific time order and mode, the mixed extraction of continuous frame timing information and static spatial information on the premise of far less than the amount of 3D convolution calculation. At the same time, the 3D consistent neural network and slow fast model are slightly higher than the model in this paper, but the reasoning time is much longer than the model in this paper This shows that the method in this paper achieves the same effect as other complex models for power operation scenarios (in which only people are dynamic but environment and equipment are static), and has the advantage of extremely fast speed, which is very practical.

On the edge side, this paper adopts two deployment strategies. One is to deploy the model built on the server-side based on PyTorch framework directly, and the other is to deploy it through model transformation and acceleration. The test results under the two strategies are shown in Table 6.

Table 6: Comparison between unaccelerated model and accelerated model

	Model size(M)	Speed(frame/s)	mAP/%
Unaccelerated model	28.8	23	87.8
Accelerated model	18.9	61	85.6

It can be seen from the table that the speed of the original model is significantly reduced by 6.2 times. At the same time, the accuracy is not significantly reduced compared with the original model, which shows that the acceleration strategy in this paper is effective. The accelerated model takes into account the accuracy and timeliness, and has high practical value.

5 Conclusion

With regard to the entire process of power operation, this paper constructs a key link recognition model based on a spatiotemporal mix convolution neural network, and introduces power vision edge intelligence technology to accelerate and deploy the model proposed in this paper, which improves the reliability and real-time performance of the algorithm. By identifying the key links in the whole process of switching operation of a power company in Xinjiang, this algorithm identifies six typical links such as pre-shift meeting, power inspection and pulling the disconnector, with an accuracy of 87.8% and a recognition speed of 61 frames/s. It is proved that the proposed algorithm can accurately identify the key operation links in the real power operation process, and has high real-time performance. This is of great significance for improving the reliability of risk management and control methods in electric power operations, which will help reduce the occurrence of electric power accidents and protect the personal safety of operators.

Acknowledgement: This paper is supported by the Science and technology projects of Yunnan Province (Grant No. 202202AD080004).

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. A. Mohamed, T. Chen, W. Su and T. Jin, "Proactive resilience of power systems against natural disasters: A literature review," *IEEE Access*, vol. 7, pp. 163778–163795, 2019.

[2] J. Qian, M. Zhu, Y. Zhao and X. He, "Short-term wind speed prediction with a two-layer attention-based lstm," *Computer Systems Science and Engineering*, vol. 39, no. 2, pp. 197–209, 2021.

[3] R. Francis and B. Bekera, "A metric and frameworks for resilience analysis of engineered and infrastructure systems," *Reliability Engineering & System Safety*, vol. 121, no. 4, pp. 90–103, 2014.

[4] F. Q. Ma, B. Wang, X. Z. Dong, H. G. Wang, P. Luo *et al.,* "Power vision edge intelligence: Power depth vision acceleration technology driven by edge computing," *Power System Technology*, vol. 44, no. 6, pp. 2020–2029, 2020.

[5] X. T. Bai, D. D. Sun, X. C. Zhang and B. C. Sun, "Intelligent safety monitoring system for nuclear power plant based on the convolution neural network," in *Int. Symp. on Software Reliability, Industrial Safety, Cyber Security and Physical Protection for Nuclear Power Plant*, Singapore, Springer, pp. 696–705, 2020.

[6] Q. Xu, H. Huang, C. Zhou and X. Zhang, "Research on real-time infrared image fault detection of substation high-voltage lead connectors based on improved YOLOv3 network," *Electronics*, vol. 10, no. 5, pp. 544, 2021.

[7] K. Hu, J. Jin, F. Zheng, L. Weng and Y. Ding, "Overview of behavior recognition based on deep learning," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 1–33, 2022.

[8] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition-A review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, 2012.

[9] Z. Shi and T. -K. Kim, "Learning and refining of privileged information-based RNNs for action recognition from depth sequences," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4684–4693, 2017.

[10] F. Ma, M. Li, X. Dong, B. Wang, Y. Zhou *et al.,* "Thinking and prospect of power chip specificity," *International Journal of Photoenergy*, vol. 2021, pp. 14, Article ID 1512629, 2021. https://doi.org/10.1155/2021/1512629.

[11] Y. Ba, "Power dynamics and corporate power in governance processes: Evidence from US environmental governance systems," *The American Review of Public Administration*, vol. 52, no. 3, pp. 206–220, 2022.

[12] X. Li, Y. Wang, Z. Zhou and Y. Qiao, "SmallBigNet: Integrating core and contextual views for video classification," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 1089–1098, 2020.

[13] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun *et al.,* "A closer look at spatiotemporal convolutions for action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, pp. 6450–6459, 2018.

[14] C. H. Pham, A. Ducournau, R. Fablet and F. Rousseau, "Brain MRI super-resolution using deep 3D convolutional networks," in *2017 IEEE 14th Int. Symp. on Biomedical Imaging (ISBI 2017)*, Melbourne, VIC, Australia, pp. 197–200, 2017.

[15] S. Sudhakaran, S. Escalera and O. Lanz, "Gate-shift networks for video action recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 1102–1111, 2020.

[16] Z. Zhao, W. Zou and J. Wang, "Action recognition based on c3d network and adaptive keyframe extraction," in *2020 IEEE 6th Int. Con. on Computer and Communications (ICCC)*, Chengdu, China, pp. 2441–2447, 2020.

[17] J. Lin, C. Gan and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Korea, pp. 7083–7093, 2019.

[18] Y. Li, B. Wang, H. Wang, F. Ma, H. Ma *et al.,* "An effective node-to-edge interdependent network and vulnerability analysis for digital coupled power grids," *International Transactions on Electrical Energy Systems*, vol. 2022, no. 6, pp. 1–13, Article ID 5820126, 2022. https://doi.org/10.1155/2022/5820126.

[19] Y. He, X. Zhang and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proc. of the IEEE Int. Conf. On Computer Vision*, Venice, Italy, pp. 1389–1397, 2017.

[20] R. Kim, G. Kim, H. Kim, G. Yoon and H. Yoo, "A method for optimizing deep learning object detection in edge computing," in *2020 Int. Conf. on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, pp. 1164–1167, 2020.

[21] X. Chen, G. Liu, J. Shi, J. Xu and B. Xu, "Distilled binary neural network for monaural speech separation," in *2018 Int. Joint Conf. on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, pp. 1–8, 2018.

[22] D. Sinha and M. El-Sharkawy, "Thin mobilenet: An enhanced mobilenet architecture," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conf. (UEMCON)*, Columbia, USA, pp. 280–285, 2019.

[23] F. Ma, B. Wang, J. Zhou, R. Jia, P. Luo *et al.,* "An effective risk identification method for power fence operation based on neighborhood correlation network and vector calculation," *Energy Reports*, vol. 7, pp. 6995–7003, 2021.

[24] J. Liu, R. Jia, W. Li, F. Ma, H. M. Abdullah *et al.,* "High precision detection algorithm based on improved RetinaNet for defect recognition of transmission lines," *Energy Reports*, vol. 6, pp. 2430–2440, 2020.

[25] H. Ma, Z. Liu, M. Li, B. Wang, Y. Si *et al.,* "A two-stage optimal scheduling method for active distribution networks considering uncertainty risk," *Energy Reports*, vol. 7, pp. 4633–4641, 2021.

[26] H. Wang, F. Lu, X. Tong, X. Gao, L. Wang *et al.,* "A model for detecting safety hazards in key electrical sites based on hybrid attention mechanisms and lightweight Mobilenet," *Energy Reports*, vol. 7, pp. 716–724, 2021.

[27] H. Wang, B. Wang, P. Luo, F. Ma, Y. Zhou *et al.,* "State evaluation based-feature identification of measurement data for resilient power system," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 4, pp. 983–992, 2021.