Research on computer vision technology based on BP-LSTM hybrid network

Qiaoling Yi¹, Shijia Ling†¹, Guoluan Chen¹, Liangfang Liu¹

¹ Zhongshan Polytechnic, Zhongshan, Guangdong 528400, China

Abstract

The computer vision direction in the field of artificial intelligence analyses the latest progress of computer vision technology from visual perception and visual generation, including but not limited to image recognition, target detection and image segmentation. First of all, for computer vision technology, this paper introduces the detailed application of image recognition technology, object detection technology and image segmentation technology. Then, we build a BP neural network combined with a deep LSTM neural network, use the BP network algorithm to select the input variables to reduce the dimension and complexity of the model, and use the selected variables as the input of the deep LSTM network. At the same time, deep LSTM is used to perform high-dimensional deep memory learning features on the selected variables. Finally, the model is separately experimented in computer vision. The experimental results show that the present model and other single models can be selected by BP neural network variables in computer vision applications, which can effectively reduce the complexity of the model and improve the generalisation ability of the model, so that it can be used in computer vision research.

Keywords: deep neural network, BP network, LSTM, computer vision

1 Introduction

The development of deep learning has not only solved many difficult visual problems but also accelerated the progress of computer vision and artificial intelligence related technologies [1]. In recent years, with the advancement of deep learning, computer vision is gradually infiltrating people’s lives [2], and influences several factors pertaining to everyday life that are within the reach of most common people, including municipal security, autonomous driving, film and television entertainment, fashion design, human–computer interaction, face authentication and photo album management; additionally, it is increasingly used in emerging applications in the field of e-commerce platforms and for other commercial purposes that are transacted in the online mode. Computer vision technology is continuing to find rapid penetration and acceptance in multiple fields [3].

† Corresponding author.
Email address: pennyling99@163.com

ISSN 2444-8656 doi:10.2478/amns.2021.2.00270
Computer vision based on deep learning technology can also have a profound impact on other disciplines [4], such as animation simulation and real-time rendering technology in computer graphics, microscopic image analysis technology in the field of materials, medical image analysis and processing technology, smart education for real-time assessment of teachers’ and students’ classroom performance and examination room behaviour, intelligent system for analysing athletes’ performance and technical statistics, and so on. Object detection algorithms have shifted from traditional algorithms such as HOG [5], SIFT [6] and LBP [7] based on handcrafted features to machine learning techniques based on deep neural networks. Among these, the target detection technology based on deep learning consists of RCNN [8], Fast RCNN [9] and other technologies. U-Net and fully convolutional network (FCN) are both networks used in language segmentation techniques [10,11]. For road and indoor scene understanding, SegNet [12] and VGG-16 [13] pixel-level image segmentation networks are designed in such a way as to be efficient in terms of memory and computation time. Image recognition techniques apply AlexNet [14] deep learning to image classification tasks. In order to solve a large amount of image data, VGGNet [15], ResNet [16] and DenseNet [17] deep learning networks are proposed to solve the problem of large data scale and multiple image categories.

Since the research work into computer vision technology of deep learning network uses a large amount of target data for training, strengthening and expanding the front-end data also constitutes an important research direction [18]. There are few researches on small target detection and video target detection. At the same time, since it introduces greater efficiency in the process of inference and deployment of models on mobile embedded devices, the lightweight nature of computer vision has always attracted the attention of the industry [19]. After collecting multi-modal information (such as text, images, point clouds, etc.), the means to improve the detection performance through better information fusion is also a key research direction in the future. Therefore, the demand for lightweight networks becomes more and more intense. The core of lightweight network design is to reduce the computational complexity, so that deep neural networks can be deployed on embedded edge devices with limited computing performance and storage space. In consideration of these concepts, to make the transition from academia to industry, this paper proposes a research on computer vision technology based on BP-LSTM hybrid network. The main innovative aspects of the research carried out in the present study are as follows:

1. Researching image recognition, target detection and image segmentation in computer vision technology.
2. Building a BP network [20] and a deep LSTM network [21], and combining them to form a BP-LSTM hybrid network [22] framework.
3. Applying the BP-LSTM hybrid network to computer vision technology, carrying out the experimental comparison and analysing the experimental results.

2 Computer vision and BP-LSTM framework design

2.1 Computer vision technology

There are many categories of computer vision technology. In the research in the field of computer vision, in addition to studying the novel visual processing problems such as face recognition, image segmentation, and image retrieval and classification, which are often mentioned, it is also worthwhile to conduct further exploration and ascertain the means for improvement in the technology as a whole, considered in its fundamentals, as indicated in the example in Figure 1.

2.1.1 Image identification

Image recognition, also known as image classification, is to input a picture, output the category of the image and let the computer recognise such information as people, traffic lights, animals, etc. This is image recognition in a broad sense [23]. In industry and academia, there is also recognition for specific targets. In order to deal
with the interference of factors such as illumination, occlusion, expression changes and pose estimation in face recognition pictures, feature selection has become a solution that researchers have been pursuing and exploring, thus finding hidden key structural information in face images.

2.1.2 Target detection

Another common application in computer vision tasks is object detection, which aims to output the location, class, etc. of a specific object in a given image [24]. It can be seen that target detection is a further development of target recognition. The computer not only outputs the position of the target in the image but also gives the category of the target. A common application of target detection is pedestrian detection. For example, at a traffic intersection, all pedestrians captured by the camera can be quickly detected, and the number of people can be estimated, thereby giving early warning of abnormal events. Object detection achieves object tracking by training a binary classifier on a set of labelled and unlabelled samples, and it uses positive and negative constraints on the labels of unlabelled data to guide the iterative learning process.

2.1.3 Image segmentation

Image segmentation refers to the task of semantic image segmentation, which requires classifying all pixels of the entire image into one of a number of predefined categories, and is also a technique and process. It is a key step from image processing to image analysis [25]. Since it is a pixel-level dense classification task, it is more difficult to programme and execute than image classification and object detection tasks. It is an important topic in image processing and computer vision. In recent years, a lot of work has been done to study image segmentation methods based on deep learning models.

2.2 BP-LSTM framework

This paper proposes a hybrid model based on BP model and deep LSTM model: one part is BP network and the other part is deep LSTM model. Figure 2 shows the BP-LSTM model flow.

2.2.1 BP network

The BP neural network algorithm has been continuously verified in practice, and its application algorithm structure has been developed to a considerable extent, as shown in Figure 3. The entire image shown in the
figure represents the basic function architecture of a BP neural network. The number of hidden layers in the architecture can be set to one layer or \( n \) layers. After many practical applications, it is shown that a single hidden layer BP network with enough hidden layer nodes can achieve a relatively ideal mapping effect. From the structural point of view, the BP network with a single hidden layer is mainly used to establish the function mapping relationship. The main steps are as follows:

1. First, after normalisation, data are imported into the input layer with the number of \( N \) nodes, and subsequently the data are imported from the input layer to the middle hidden layer with the number of \( M \) nodes in the mode of simulating the cell excitation.

2. Then, the features are processed by the hidden layer and thereafter transmitted to the output layer, with \( L \) number of nodes.

3. The final output layer obtains the output value according to the mode of cell excitation.

The output layer and hidden layer excitation function in the network architecture selects the \( \text{sigmoid} \) function, where the function formula is as follows:

\[
g(x) = f(x) = \frac{1}{1 + e^{-x}}
\]  

(1)

In relation to the function, the input formula for the input of \( q \) training samples is as follows:

\[(x_{q1}, x_{q2}, \ldots, x_{qn})\]  

(2)

The calculated output and target output formulas are as follows:

\[(y_{q1}, y_{q2}, \ldots, y_{ql})\]  

(3)

\[(t_{q1}, t_{q2}, \ldots, t_{ql})\]  

(4)

The calculated output value of the constructed BP neural network is:

\[
y_{ql} = g \left[ \sum_{i=1}^{M} w_{ij} f \left( \sum_{j=1}^{N} w_{ij} x_{qj} + b_i \right) + b_l \right]
\]  

(5)

where \( y_{ql} \) represents the output value of the network calculation; \( q \) represents the input data serial number and the sequence \( 1, 2, \ldots, Q \) represents samples taken; \( w_{ij}, w_{ij} \) is the weight of the hidden layer, the input layer and the hidden layer node; and \( b_i, b_l \) represent the hidden layer and the output node neuron threshold, respectively.
The error between the calculated output value of the BP network and the corresponding target output value can have different processing forms, and there are also many training algorithms. In this paper, the batch processing algorithm for reducing the total sample error is adopted, which is relatively fast in the training sample calculation process. The formula for the average error between the output value and the sample learned by training is:

\[
E_m = \frac{1}{2Q} \sum_{q=1}^{Q} \sum_{l=1}^{L} (y_{ql} - y_{ql})^2
\]  

(6)

2.2.2 LSTM network

The LSTM model mainly introduces memory cells to solve long-term dependency problems, that is, gradient explosion and gradient disappearance. Each LSTM unit mainly includes one or more cells, as shown in Figure 4. The main steps can be summarised as:

1. First, the feature input forget gate (forget gate, \( f_t \)) determines the features to be eliminated from the memory unit.

\[
f_t = \sigma(b_f + W_f x_t + U_f h_{t-1})
\]  

(7)

where \( \sigma \) represents the sigmoid activation function, and the feature flow weight is set to a value between 0 and 1 (0 means to delete the feature completely, 1 means to keep all the features); \( x_t \) is the current input vector; \( h_t \) is the current hidden layer vector; and \( b_f, W_f, U_f \) represents the bias, the input weight and the loop weight of the forget gate.

2. Then, the state in the memory cell is updated.

\[
g_t = \sigma(b_g + W_g x_t + U_g h_{t-1})
\]  

(8)

where \( g_t \) is an external input gate between 0 and 1 controlled by the sigmoid activation function.

Then, based on \( C_{t-1} \), the state \( C_t \) in the memory unit will continue to be updated.

\[
C_t = f_t \ast C_{t-1} + g_t \ast \tanh(b_c + W_c x_t + U_c h_{t-1})
\]  

(9)
3. Finally, the output is controlled by the output gate.

\[ h_t = o_t \ast \tanh(C_t) \]  

(10)

With reference to Eq. (10), it also needs to be mentioned that the sigmoid activation function controls the output gate.

\[ o_t = \sigma(b_o + W_o x_t + U_o h_{t-1}) \]  

(11)

4. The LSTM neural network requires input of feature vector values on successive time steps for training. Assuming that each feature sequence is input with a total of \( N \) time steps, and \( S_N \) represents the feature vector of the \( N \)th time step, the \( M \)th sequence can be expressed as:

\[ S = \{ s_M, s_{M+1}, \ldots, s_{M+N+1} \} \]  

(12)

5. We suppose that the input feature vector is structured as:

\[ \{ 1, 2, \ldots, N, N+1 \} - \{ s_1, s_2, \ldots, s_N, s_{N+1} \} \]  

(13)

where \( s_1 \sim s_n \) is the feature sequence 1; \( s_2 \sim s_{n+1} \) is the feature sequence 2; and so on up to sequence \( n \).

2.2.3 Deep LSTM network

A deep LSTM neural network is composed of multiple one-layer LSTM modules. Mainly, the local features of the input sample data are established in one LSTM layer and integrated in the higher LSTM layer. The learning ability of the deep LSTM neural network structure is stronger, and the extracted features are more complete.

In the deep LSTM network structure shown in Figure 5, \( h^l_t \) represents the output feature vector of the first layer of the LSTM neural network at time \( t \). The input feature vector is \( X_t \) at time \( t+1 \) and the output vector is \( h^l_t \) at time \( t \); after passing through the first layer of LSTM, the output vector is \( h^{l+1}_t \), which is used as the output of the second layer of LSTM and the input of the first layer of LSTM at time \( t+2 \), at which point the iterative layer ends.

![Deep LSTM network structure](image)

**Fig. 5** Deep LSTM network structure

3 Experimental results and analysis

3.1 Experimental data

This paper seeks to verify that BP-LSTM can be applied in computer vision technology. The image data uses public datasets including: ImageNet dataset for image classification, Harmful Garbage dataset for target detection (4000 images in training samples and test samples, 21 different categories + background noise) and image segmentation using Select Network Vegetation dataset to perform segmentation processing (a total of 8000 images for training samples and test samples).
3.2 Lab environment

The experiments are based on the deep learning framework of Tensorflow and Keras, and the operation is performed on a platform having the following configuration: Windows 10 as the OS, Intel(R) Core(TM) i5-3470 CPU @ 3.20 GHz as the processor and with a 32.00 GB memory.

3.3 Experimental evaluation index

3.3.1 Object detection evaluation index

Average precision (AP) is a common evaluation index for object detection, which is used to calculate the average detection accuracy and measure the performance of the detector in each category. The AP calculation formula is:

\[ AP = \frac{TP}{TP + FP} \]  \hspace{1cm} (14)

where \( TP \) is the number of positive classes that are predicted to be positive, and \( FP \) is the number of false positives that are predicted to be negative classes.

In addition to the \( AP \) rate, the average recall (AR) rate is also proposed to indicate the proportion of detected positive samples to the total number of actual positive samples, and its calculation formula is:

\[ AR = \frac{TP}{TP + FN} \]  \hspace{1cm} (15)

where \( FN \) is the number of negative classes predicted as negative classes.

The detection speed is evaluated using frames per second (FPS), and the larger the value, the better the real-time performance.

3.3.2 Image segmentation evaluation index

For the semantic segmentation method, the mean intersection over union (mIoU) is used as the accuracy measurement index. In the problem of image segmentation, these two sets are the set of true values and predicted values, respectively.

\[ mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \]  \hspace{1cm} (16)

where \( k \) represents the number of categories of the target, and there are \( k+1 \) categories (including the target and the background); \( i, j \) both represent the category number; \( p_{ii} \) represents the correctly classified pixel; and \( p_{ij} \) and \( p_{ji} \) both represent the wrongly classified pixel.

For instance, in segmentation methods, pixel accuracy (PA) is used as a measure to indicate the proportion of correctly classified pixels to total pixels. The specific calculation formula is the following:

\[ PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \]  \hspace{1cm} (17)

3.4 Experimental verification method

3.4.1 Experiment 1: Comparison of image recognition and traditional deep learning models

In order to verify the application based on the BP-LSTM hybrid network method in image classification, the preprocessed images were individually input into the AlexNet, VGG-16, Xception model and BP-LSTM method pretrained by ImageNet, and the training and testing results are shown in Table 1. The prediction is correct only when the class with the highest probability in the prediction result is the correct class.
Table 1 Comparison of classification accuracy of different models

<table>
<thead>
<tr>
<th>Input</th>
<th>Models</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet Data</td>
<td>AlexNet</td>
<td>66.91</td>
</tr>
<tr>
<td></td>
<td>VGG-16</td>
<td>67.32</td>
</tr>
<tr>
<td></td>
<td>Xception</td>
<td>85.56</td>
</tr>
<tr>
<td></td>
<td>BP-LSTM</td>
<td>89.94</td>
</tr>
</tbody>
</table>

From Table 1, we infer that the classification accuracy of BP-LSTM is higher than those of AlexNet, VGG-16 and Xception models. The main reason is that these models use the features of the fully connected layer for classification, and there is much background noise in the image, which affects the classification accuracy. On the contrary, the method in this paper uses the BP network selection feature, which contains the detailed information of the salient area, removes the interference of background noise features on the classification, and improves the accuracy of image classification to a certain extent.

3.4.2 Experiment 2: Comparison of performance indicators between various target detection mainstream algorithms

In order to compare the performance indicators of various target detection mainstream algorithms, Table 2 lists the specific results of various deep learning target detection algorithms. On the whole, in the Harmful Garbage dataset, the performance of the algorithm based on BP-LSTM is better. From the perspective of target detection accuracy, the AP value of the BP-LSTM algorithm is greatly improved compared to other algorithms, the average detection accuracy reaches 69.98% and the overall detection effect is better than those of other algorithms.

<table>
<thead>
<tr>
<th>Models</th>
<th>AP (%)</th>
<th>AR (%)</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN</td>
<td>34.57</td>
<td>36.57</td>
<td>16</td>
</tr>
<tr>
<td>YOLOv3</td>
<td>45.62</td>
<td>47.68</td>
<td>23</td>
</tr>
<tr>
<td>CenterNet</td>
<td>49.89</td>
<td>51.09</td>
<td>15</td>
</tr>
<tr>
<td>YOLOv4</td>
<td>57.34</td>
<td>55.12</td>
<td>37</td>
</tr>
<tr>
<td>BP-LSTM</td>
<td>69.98</td>
<td>65.23</td>
<td>49</td>
</tr>
</tbody>
</table>

AP, average precision; AR, average recall; FPS, frames per second

From Table 2, we infer that the algorithm has good real-time detection performance, which is far superior to that of the two-stage target detector (Faster-RCNN), and the detection speed is close to that of YOLOv4, which is about 68% higher than that of YOLOv3. The studied algorithm mainly uses the BP network selection feature, which contains detailed information pertaining to the salient area, and also removes the interference of the background noise feature on the classification.

3.4.3 Experiment 3: Comparison of performance indicators between PSPNe, FCN, CRF-RNN and DPN

In order to verify the application of the image segmentation of the BP-LSTM hybrid network, the performance indicators of the image segmentation methods of PSPNe, FCN, CRF-RNN and DPN are compared quantitatively, as shown in Table 3.

As can be seen from Table 3, compared with other image segmentation models, the proposed BP-LSTM model can achieve the highest accuracy on the dataset, amounting to 37.23%, whereas the PA can reach 36.34%.
Table 3 Performance of different semantic segmentation methods

<table>
<thead>
<tr>
<th>Models</th>
<th>mIoU (%)</th>
<th>PA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSPNe</td>
<td>22.76</td>
<td>33.12</td>
</tr>
<tr>
<td>FCN</td>
<td>31.77</td>
<td>36.34</td>
</tr>
<tr>
<td>CRF-RNN</td>
<td>36.36</td>
<td>34.72</td>
</tr>
<tr>
<td>DPN</td>
<td>35.21</td>
<td>31.32</td>
</tr>
<tr>
<td>BP-LSTM</td>
<td>37.23</td>
<td>36.34</td>
</tr>
</tbody>
</table>

FCN, fully convolutional network; PA, pixel accuracy

4 Conclusion

In order to verify the use of BP-LSTM hybrid network in small data sets to solve a wide range of technical problems, through the research and exploration of several key technologies of deep learning, this paper makes some suggestions for the development of image interpretation technology. The following research is conducted. First, the several prevalent mainstream directions of computer vision technology are studied individually, including image recognition, target detection and image segmentation. Then, a BP-LSTM hybrid network is built by combining the BP network and the deep LSTM network. Finally, the BP-LSTM hybrid network is applied in image recognition, object detection and image segmentation computer vision technology. The accuracy of the experimental results is not adequate for achieving industrial applications. In future work, features should be extracted from front-end data, transfer learning ideas should be introduced and the network should be deepened to improve the accuracy.

References


