A hybrid deep learning-based fruit classification using attention model and convolution autoencoder

Gang Xue¹ · Shifeng Liu¹ · Yicao Ma¹

Received: 1 July 2020 / Accepted: 18 August 2020
© The Author(s) 2020

Abstract

Image recognition supports several applications, for instance, facial recognition, image classification, and achieving accurate fruit and vegetable classification is very important in fresh supply chain, factories, supermarkets, and other fields. In this paper, we develop a hybrid deep learning-based fruit image classification framework, named attention-based densely connected convolutional networks with convolution autoencoder (CAE-ADN), which uses a convolution autoencoder to pre-train the images and uses an attention-based DenseNet to extract the features of image. In the first part of the framework, an unsupervised method with a set of images is applied to pre-train the greedy layer-wised CAE. We use CAE structure to initialize a set of weights and bias of ADN. In the second part of the framework, the supervised ADN with the ground truth is implemented. The final part of the framework makes a prediction of the category of fruits. We use two fruit datasets to test the effectiveness of the model, experimental results show the effectiveness of the framework, and the framework can improve the efficiency of fruit sorting, which can reduce costs of fresh supply chain, factories, supermarkets, etc.

Keywords Fruit classification · DenseNet · CBAM · Convolutional neural networks

Introduction

Nowadays, image classification method is very popular in a lot of fields, playing a pretty important role. Image recognition supports several applications, for instance, facial recognition, image classification, video analysis, and so on. Deep learning technology has been the core topic in machine learning and it has outstanding results in image identification [1]. Deep learning uses multilayer structure to process image features, which greatly enhance the performance of image identification [2]. Image recognition and deep learning are developing so fast and more and more fields benefit from them. Fresh supply chain, factories, supermarkets, etc. are the popular fields that are relying on image recognition and deep learning to obtain a better development. In other words, the application of image recognition and deep learning in logistics and supply chain field becomes a trend. For example, image recognition can help to guide the path of logistics and transportation, and it can solve the problem that several

automatic transportation vehicles make mistakes because of large path identification errors [3]. Another example is fruit and vegetable classification. Deep learning can extract image features effectively and then implement classification.

In the past, the fruit picking and processing is based on artificial methods, resulting in a large amount of waste of labor [4]. Recently, researchers tended to apply near-infrared imaging, gas sensor, and high-performance liquid chromatography devices to scan the fruit. Nevertheless, those methods need expensive devices (different types of sensors) and professional operators, and their overall accuracies are commonly lower than 85% [4]. With the fast development of 4G communication and widespread popularity of various mobile Internet devices, people have generated a great quantity of images, sounds, videos, and other information, and image recognition technology has been gradually mature. Image-based fruit identification has attracted the attention of researchers since their cheap device and excellent performance [4]. The intelligent identification of fruit can be used not only in the picking stage of the early fruit, but also in the picking and processing stage in the later stage. Fruit recognition technology based on deep learning can significantly improve the performance of fruit recognition, and has a positive effect on promoting the development of smart

✉ Gang Xue
19113032@bjtu.edu.cn

¹ School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China

agriculture. Comparing with artificial features + traditional machine learning combination method, deep learning can automatically extract features, and has better performance, which gradually becomes the mainstream method of smart identification [5]. For instance, Rocha et al. [6] presented a unified method that could combine features and classifier. Tu et al. [7] developed a machine vision method to detect passion fruit and identify maturity applying RGB-D images. Fruit and vegetable classification is challenging, because it is hard to give each kind of fruit an adequate definition. However, achieving accurate fruit and vegetable classification is very important in fresh supply chain for several reasons. First, fruit and vegetable automatic classification decreases labor cost, because factories do not need workers to do this classification task anymore. Labor cost is saved and this cost can be invested in other aspects. Second, accurate classification contributes to factory automatic fruit-packing and transportation [5]. In these fields, packing and transportation are two core parts. If any errors happen in these parts, it will cause a bad influence on the following processes. For instance, the errors will delay the time which the customers receive the fruits and vegetables. Thirdly, accurate classification can save the time and bring higher efficiency. After lots of training, automatic classification can reach a high standard, which can promise the accuracy of the classification result. In addition, unlike humans, automatic classification has memories and it can continue to repeat the classification process. It helps to save much time, enhancing the efficiency greatly [8].

The development of attention mechanism and autoencoder is constantly improving the performance of deep learning. The attention mechanism in the deep learning model is a model that can simulate the attention of the human brain [8]. Attention mechanisms were first presented in the field of image research and then were used in wider fields [9]. When people observe images, they only focus their attention on important parts instead of looking at every pixel of the image. As more and more people use attention mechanism, more cutting-edge models have been created and used in various studies. There are four categories: number of sequences, number of abstraction levels, number of positions, and number of representations [10]. Each category corresponds to different types. The number of sequences has distinctive, co-attention, and self-types. The number of abstraction levels has single-level and multi-level types. The number of positions has soft, hard, and local types. The number of representations has multi-representational and multi-dimensional types. As for autoencoder, it is popularly used as an effective feature extraction method. Autoencoder (AE) is a special artificial neural network used in semi-supervised learning and unsupervised learning and is composed of encoder and decoder [11]. Its function is to have representational learning from the input information using the input information as the learning target [12].

There are two main innovations and contributions in this paper: (1) scholars used manual features decided by experts, when there are many fruit subspecies, it is very difficult to identify specific fruit subspecies using artificial feature extraction. We construct a model which can identify specific fruit subspecies accurately. (2) An attention module and CAE are integrated into DenseNet, which refine the features of fruit images and improves the interpretability of the method.

In this paper, we develop a hybrid deep learning-based fruit recognition method, named attention-based densely connected convolutional networks with convolution autoencoder (CAE-ADN), which uses a convolution autoencoder to pre-train the images and uses an attention-based DenseNet to process the images. Experimental results illustrate the effectiveness of the method, and the model can improve the efficiency of related work.

The rest sections of this work are shown below. Section 2 delineates the related works of fruit classification. Section 3 delineates the detailed methodology. Section 4 describes the sets and results of experiment, and comparison with other studies. Finally, Sect. 5 makes a conclusion of this work.

Related work

With the rapid development of machine vision technology, automatic sorting method using machine vision has been used in production and processing fields. To resolve the problem of low efficiency and accuracy of traditional sorting method under the condition of huge fruit output, machine vision and deep learning technology can assuredly promote the efficiency and accuracy of sorting. However, under the actual conditions, the image will be affected by light, fruit reflection, and occlusion. For example, different shapes and colors of fruits make it difficult to identify and locate under different conditions (such as different light and noise). In addition, because the color and texture features of fruit image are related to the growth period, it also increases the complexity of fruit recognition.

In the previous studies, color, texture, and edge properties are considered to categorize fruits [13–15]. Garcia et al. [16] used artificial vision for fruit recognition and extracted shape, color chromaticity, and texture features. To distinguish large types of figures, features in low and middle levels are applied [17–19]. When referring to the product classification, the first attempt of a fruit recognition system in supermarket, which considered texture, color, and density, must be mentioned [20]. The accuracy could reach nearly 95% in some situations. Later, Jurie et al. and Agarwal et al. [21, 22] used the method of breaking down the classification problem to the recognition of different parts, that are features of each object class. These techniques were called bag-of-features and they

presented potential results, even though they did not model spatial constraints between features [23, 24].

Besides color, texture, and edge properties, many different methods are used in fruit and vegetable classification. For example, scholars use gas sensor, near-infrared, and high-performance liquid chromatography devices to scan the fruit [25–27]. Fei-Fei et al. [28] introduced prior knowledge into the estimation of the distribution, thus reducing the number of training examples to around ten images while preserving a good recognition rate. Even with this improvement, the problem of exponential growth with the number of parts persists, which makes it impractical for the problem presented in this paper, which requires speed for on-line operation. However, these methods, with expensive devices, do not bring good results, because the accuracies are lower than 85% [4]. To solve this problem, later scholars begin to pay attention to image-based fruit classification for its cheap device and wonderful performance. Wang et al. [29] applied backpropagation neural network (BPNN) and used fractional Fourier entropy (FRFE) as the features. Lu et al. [30] proposed an improved hybrid genetic algorithm (IHGA) to take the place of BPNN. Zhang et al. [31] presented a novel method called biogeography-based optimization and feed-forward neural network (BBO-FNN) which achieved higher accuracy. Zhang et al. [29] created a categorization model using fitness-scaled chaotic artificial bee colony (FSCABC) method to replace kSVM.

According to the past studies, two main problems must be solved to promote the accuracy of fruit recognition. One limitation is that some scholars used manual features decided by experts, when there are many fruit subspecies, it is very difficult to identify specific fruit subspecies using artificial feature extraction. And another limitation is that the accuracy of the existing models is not enough to support the recognition of dozens of fruits in related fields [32, 33].

Method

Model designing

To explore all fruit features contained in the image, we use an attention module to force the networks to learn the high-level feature. Attention not only illustrates where should be focused, but also enhances the representation of features. Woo et al. [34] proposed an effective attention module convolutional block attention module (CBAM) which can be widely used to boost the representation power of convolutional neural networks (CNNs). In addition, considering the complexity of fruit features, it is difficult for the traditional CNNs to train these images, densely connected convolutional networks (DenseNet) [35] is more suitable for the proposed problem due to the improvement of feature delivery. The

problem of gradient vanishment is resolved applying a direct path to all preceding layers to route residuals during back-propagation. We proposed an attention-based DenseNet to train the images.

An autoencoder (AE) is a special artificial neural network, and it is applied to unsupervised learning and efficient coding [36]. Hinton and PDP Group [37] first proposed AEs in the 1980s to solve the problem of “backpropagation without teachers” by applying training dataset as teachers [38]. Nowadays, AEs become more widely applied in the field of learning generative models [39], and convolutional autoencoder (CAE) also has a good performance of image identification [40].

In this work, an attention-based densely connected convolutional network with convolution autoencoder (CAE-ADN) framework is developed, which uses a convolution autoencoder to pre-train the attention-based densely connected convolutional networks. The whole framework of CAE-ADN is illustrated in Fig. 1. Details are as follows: in the first part of the framework, an unsupervised method with a set of images is applied to pre-train the greedy layer-wised CAE. We use CAE structure to initialize a set of weights and bias of ADN. In the second part of the framework, the supervised ADN with the ground truth is implemented. The final part of the framework makes a prediction of the category of fruits.

Attention-based DenseNet

Attention mechanism

CBAM represents the attention mechanism module of the convolution module. It is a kind of attention mechanism module which combines spatial and channel. Compared with channel only attention mechanism, it can achieve better results. An intermediate feature map $F \in \mathbb{R}^{C \times H \times W}$ is given as an input of the CBAM, and a 1D channel attention map $M_c \in \mathbb{R}^{C \times 1 \times 1}$ and a 2D spatial attention map $M_s \in \mathbb{R}^{1 \times H \times W}$ are computed consecutively, as shown in Fig. 2. The attention structure is concluded as:

$$F' = M_c(F) \otimes F \quad (1)$$

$$F'' = M_s(F') \otimes F', \quad (2)$$

where $\otimes$ represents an element-wise multiplication operation. F'' is the final refined output of CBAM. Details of each attention module are described as followed.

First, we aggregate the spatial feature of a feature map by utilizing average-pooling and max-pooling layers, and two different spatial context descriptors are generated: F_{avg}^c denotes average-pooled features and F_{max}^s denotes average-pooled features. Both F_{avg}^c and F_{max}^s are then inputted to a shared network to compute the channel attention map

Fig. 1 The structure of the CAE-ADN

Fig. 2 Structure of CBAM

$M_c \in \mathbb{R}^{C \times 1 \times 1}$. To avoid oversized parameters, the hidden activation size is set to $\mathbb{R}^{C/r \times 1 \times 1}$, where r denotes the reduction ratio. In a nutshell, we compute the channel attention as:

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\ = \sigma\left(W_1\left(W_0\left(F_{\text{avg}}^c\right)\right) + W_1\left(W_0\left(F_{\text{avg}}^c\right)\right)\right), \quad (3)$$

where σ represents the sigmoid function, $W_0 \in \mathbb{R}^{C/r \times C}$, and $W_1 \in \mathbb{R}^{C \times C/r}$ denote the weights of MLP, W_0 and W_1 , are employed for both inputs, and the ReLU activation function is followed by W_0 . It is proved that using pooling operations along the channel axis can highlight informative regions effectively [41–43]. On the concatenated feature descriptor, a convolution layer is applied to compute a spatial attention map $M_s(F) \in \mathbb{R}^{H \times W}$ which encodes where to focus on. Detailed operation is shown in the following [34].

The channel features of a feature map are aggregated by applying two pooling layers, computing two 2D maps: $F_{\text{avg}}^s \in \mathbb{R}^{1 \times H \times W}$ represents average-pooled features across the channel and $F_{\text{max}}^s \in \mathbb{R}^{1 \times H \times W}$ represents max-pooled features across the channel. In a nutshell, we compute the spatial attention as [34]:

$$M_s(F) = \sigma\left(f^{7 \times 7}([\text{AvgPool}(F)]; [\text{MaxPool}(F)])\right)$$

$$= \sigma\left(f^{7 \times 7}\left(F_{\text{avg}}^s; F_{\text{max}}^s\right)\right), \quad (4)$$

where σ represents the sigmoid function and $f^{7 \times 7}$ denotes a convolution operation with the filter size of 7×7 .

Attention-based dense block

Huang et al. [35] introduced direct connections from any layer to all subsequent layers. To refine the feature of each layer, we combine the CBAM and DenseNet to pay more attention to the scale features of targets.

Figure 3 illustrates the structure of attention-based dense block. Finally, the ℓ th layer connects the feature maps of all preceding layers, $x_0, x_1, \dots, x_{\ell-1}$, as input:

$$x_\ell = H_\ell([x_0, x_1, \dots, x_{\ell-1}]), \quad (5)$$

where $[x_0, x_1, \dots, x_{\ell-1}]$ denotes a concatenate operation of the $x_0, x_1, \dots, x_{\ell-1}$ generated in $0, \dots, \ell-1$. And we concatenate the multiple inputs of $H_\ell(\cdot)$ in Eq. (5) into a single tensor. Inspired by Huang et al. [35], $H_\ell(\cdot)$ is defined as a composite function of four sequential operations: batch normalization (BN) $\rightarrow$ rectified linear unit (ReLU) $\rightarrow 3 \times 3$ convolutional layer (Conv) $\rightarrow$ CBAM. And other settings of attention-based dense block include growth rate, Bottleneck layers, compression, etc. are the same as DenseNet.

Attention-based DenseNet

The input of attention-based DenseNet is the image followed by a convolution layer with a 7×7 filter, represented by D_0^1 :

$$D_i^q = \odot(D_{i-1}^p, W_i^d), \quad (6)$$

Fig. 3 Structure of attention-based dense block

where D_i^q denotes the i th feature map of attention-based DenseNet with the channel size of q , $\odot$ denotes a set of operations: attention-based dense block followed by a 1×1 convolution layer and a 2×2 average-pooling layer, stride 2, and W_i^d denotes to a set of parameters of the i th attention-based DenseNet.

Stacked convolutional autoencoders

The AEs is a well-known learning algorithm, derived from the idea of sparse coding, which has a huge advantage in data feature extraction. The traditional AEs consist an encoder and a decoder, using a backpropagation algorithm to find the optimum solution to make the prediction result equal to the ground truth. Traditional AEs ignore the 2D image structure. The convolutional autoencoder uses a convolutional layer instead of a fully connected layer. The principle is the same as that of an autoencoder. It downsamples the input symbol to provide a smaller representation of the latent dimension, and forces the autoencoder to learn a compressed version of the symbol [44, 45]. When processing high-dimensional data like images, the traditional AEs will cause the network to generate many redundant parameters, especially for three-channel color images. And due to the network layer parameters of traditional AEs which are global, so traditional AEs cannot retain the spatial locality, and slow down the network learning rate.

The structure of a CAE is shown in Fig. 4, which has N attention-based dense blocks, two blocks are connected by a convolution operation and a pooling operation; the output is the same as input image, which can pre-train the structure using supervision training method, and there is a convolution operation between the input or output.

Classification and parameter learning

The structure of an ADN is shown in Fig. 5, the i th attention-based dense block in ADN is the same as CAE, and ADN remains first half attention-based dense blocks of CAE. If N is an odd number, we round up the number, and ADN has $\text{Ceiling}(N/2)$ attention-based dense blocks. Finally, a SoftMax function is applied to make the final prediction. And we train the networks by minimizing logarithmic loss between the prediction results and the labels:

$$\theta = \arg \min_{\theta} \left\{ - \sum_{i=1}^M g_i \log(p_i) \right\}, \quad (7)$$

where M is the number of images in the training set, g_i is the ground truth of the image, and p_i is the output of the neural network after SoftMax. And θ represent a set of parameters of the framework and we learn θ using an Adam [44] optimizer with backpropagation algorithm.

Experimental result

Datasets

We use two fruit datasets to test the effectiveness of the model. Mureşan et al. [45] collected a fruit dataset including 26 labels (Fruit 26), they displayed a fruit with the original background in the image, and after removing the background, the image was downsampling to 100×100 pixels. The Fruit 26 dataset included 124,212 fruit images spread across 26 labels. 85,260 images are for training and 38,952 images are for testing. Figure 6 shows some samples of Fruit 26.

Fig. 4 The structure of CAE section

Fig. 5 The structure of ADN section

Fig. 6 Samples of Fruit 26

Hussain et al. [46] collected a fruit dataset including 15 labels (Fruit 15). There are 15 different kind of fruits consisting of 44,406 images. All the figures are captured on a clear background with resolution of 320×258 pixels. Figure 7 shows some samples of Fruit 15.

Experimental implementation

The loss is defined as the cross-entropy loss between the predicted result and the ground truth, which is defined as:

$$L_i = - \sum_{i=1}^M g_i \log(p_i), \quad (8)$$

Fig. 7 Samples of Fruit 15

where M represents the number of images of training dataset, g_i denotes ground truth of the image, and p_i is the output of the neural network after SoftMax.

Accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (9)$$

where TP and FP represent True Positive and False Positive; TP and FN represent True Positive and False Negative.

Considering the GPU memory, we set batch size to 10. The learning rate is set to 0.0001 and reduced to half at epochs 50, 100, and 150, and the network is fine-tuned with a learning rate of 0.000001 at the last 10 epochs. And we conduct 200 epochs to train the network.

In this paper, we evaluate ADN-q of CAE-ADN with $q \in [121, 169, 201]$. The q in ADN-q denotes the number of layers in ADN. Table 1 shows the ADN-q architectures. The CAE-p is generated by the corresponding ADN-q to pre-train the structure. All experiments are conducted on one GPU card (NVIDIA GeForce RTX 2080) with 8 GB RAM using Tensorflow.

Results

Comparison with baselines

We compare CAE-ADN with two baselines: ResNet-50 [47], and DenseNet-169 [35] (these two states of the art methods have the similar parameter numbers as CAE-ADN). For ResNet-50, we use the kernel size of 3×3 at each convolutional layer, and for DenseNet-169, k value sets 4, k represents the growth rate of network, and for convolutional layers with kernel size 3×3 , each side of the inputs is zero-padded by one pixel to keep the feature-map size fixed. A 1×1 convolution followed by 2×2 average pooling is applied as transition layers between two contiguous dense blocks. The training configurations of ATP-DenseNet, ResNet-50, and DenseNet-169 are the same for a fair comparison. Table 2 illustrates the Top-1 and Top-5 accuracy of fruit classifica-

tion with several networks, ADN without pre-train perform better than ResNet-50 and DenseNet-169, and ADN with pre-train has higher accuracy than ADN without pre-train. Results show that AND-169 with pre-train is the best structure of CAE-ADN, whose Top-1 accuracy is 95.86% and 93.78%, Top-5 accuracy is 99.98%, and 98.78% of Fruit 26 and Fruit 15 respectively.

Performance of each class

The precision and recall of the testing dataset is presented in Tables 3 and 4. The results show that the model has good performance in all kinds of fruits. And the model has a good ability of fruit color recognition, which can identify different varieties of the same kind of fruit. For instance, the Precisions of Apple Red 1, Apple Red 2, and Apple Red 3 are 96.07%, 95.36%, and 95.28%, respectively, and the Recalls of Apple Red 1, Apple Red 2, and Apple Red 3 are 95.67%, 95.58%, and 94.89%, respectively. Due to the difference in shooting scenes, the accuracy of Fruit 26 is generally higher than Fruit 15.

Comparison with other studies

We compared our framework with five up-to-date methods: PCA + FSCABC [48], WE + BBO [49], FRFE + BPNN [50], FRFE + IHGA [51], and 13-layer CNN [52]. Some of them used traditional machine learning methods with feature engineering, and others used simple BPNN and CNN structures. The overall accuracies of these methods are shown in Table 5, and the accuracies of our method reach 95.86% and 93.78% for Fruit 26 and Fruit 15, which illustrate that CAE-ADN performs better than state-of-the-art approaches. It can be summarized that the attention model and pre-training using CAE can improve the performance of CNN model in fruit classification problem. Compared with the traditional recognition method using sample features, the deep learning algorithm based on convolution neural networks have stronger adaptability, better robustness,

Table 1 ADN of CAE-ADN architectures

Layers	ADN-121	ADN-169	ADN-201
Convolution	7×7 conv, stride 2		
Pooling	3×3 Max pool, stride 2		
A-Dense block (1)	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 6$
Transition layer (1)	1×1 conv 2×2 Average pool, stride 2		
A-Dense block (2)	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 12$
Transition layer (2)	1×1 conv 2×2 average pool, stride 2		
A-Dense block (3)	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 48$
Transition layer (3)	1×1 conv 2×2 average pool, stride 2		
A-Dense block (4)	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ BRCC} \end{bmatrix} \times 32$
Classification layer	7×7 global average pool 1000D fully connected, SoftMax		

The growth rate of ADN is set to $k = 4$. And each “conv” operation denotes to the sequence BN-ReLU-Conv; each “BRCC” operation denotes to sequence BN-ReLU-Conv-CBAM

Table 2 Performance of fruit classification with different networks

Net	Fruit 26		Fruit 15	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
ResNet-50	93.59	98.35	91.44	96.41
DenseNet-169	93.87	98.85	91.46	96.57
ADN-121 (without pre-train)	94.21	98.42	92.07	97.52
ADN-169 (without pre-train)	94.65	98.97	92.52	97.89
ADN-201 (without pre-train)	94.26	98.24	92.12	97.38
ADN-121 (pre-train)	95.15	99.58	93.03	98.29
ADN-169 (pre-train)	95.86	99.98	93.78	98.78
ADN-201 (pre-train)	95.64	99.32	93.49	98.64

and higher recognition accuracy. Image recognition based on deep learning is learned by the abstract features of the algorithm, which avoids the difficulty of generating specific features for specific tasks and makes the whole recognition process more intelligent. Due to its strong learning ability, it can be transplanted to other tasks well, only need to retrain the convolutional neural network. Considering the development of the technology of IoT [34, 53], it is meaningful to establish a decision-making system [54–56], and we could use the proposed method in practice.

Conclusion

In this work, we develop a hybrid deep learning-based fruit image classification approach, which uses a convolution autoencoder to pre-train the images and uses an attention-based DenseNet to extract the features of image. In the first

part of the framework, an unsupervised method with a set of images is applied to pre-train the greedy layer-wised CAE. We use the CAE structure to initialize a set of weights and bias of ADN. In the second part of the framework, the supervised ADN with the ground truth is implemented. The final part of the framework makes a prediction of the category of fruit. We use two fruit datasets to test the effectiveness of the model, the overall accuracies of these methods are shown in Table 5, and the accuracies of our method reach 95.86% and 93.78% for Fruit 26 and Fruit 15, which illustrate that CAE-ADN performs better than other approaches. It can be summarized that the attention model and pre-training using CAE can promote the performance of CNN algorithm in fruit classification problem. In addition, compared with the traditional algorithm, deep learning is a method to simulate human visual perception through neural network. Through abstracting the local features of the lowest layer, each layer receives the input of the higher layer. It can automatically learn the

Table 3 Performance of each class (Fruit 26)

Label	Precision (%)	Recall (%)	Number of test images
Not fruits	96.07	95.67	15,750
Apple red 1	95.36	95.58	984
Apple red 2	95.28	94.89	984
Apple red 3	95.90	96.20	864
Apple red yellow	95.50	95.02	984
Apricot	95.85	96.64	984
Avocado	95.47	96.21	858
Braeburn (apple)	95.66	95.64	984
Cherry	95.61	96.11	984
Apple golden 1	95.23	95.82	984
Apple golden 2	95.40	94.83	984
Apple golden 3	96.00	95.99	966
Granny Smith (Apple)	96.33	96.69	984
Grape	95.62	95.42	984
Grapefruit	95.92	95.97	984
Kiwi	95.87	96.54	936
Lemon	95.57	95.68	492
Nectarine	95.94	96.35	984
Orange	95.57	95.45	960
Papaya	95.58	95.48	984
Peach	96.40	96.20	984
Peach flat	95.12	95.45	984
Pear	96.25	95.67	984
Plum	95.50	95.88	906
Pomegranate	95.25	95.78	492
Strawberry	95.26	95.77	984

Table 4 Performance of each class (Fruit 15)

Label	Precision (%)	Recall (%)	Number of test images
Apple	93.70	93.35	1507
Banana	94.29	93.34	908
Carambola	93.33	94.34	624
Guava	94.61	94.13	1202
Kiwi	93.32	93.45	1252
Mango	94.33	93.61	1246
Orange	93.91	94.19	904
Peach	93.27	93.75	792
Pear	93.85	93.23	904
Persimmon	94.04	93.85	622
Pitaya	93.40	93.27	750
Plum	93.33	94.64	689
Pomegranate	93.64	93.71	650
Tomatoes	94.02	92.96	651
muskmelon	93.63	94.45	623

Table 5 Comparison with other studies

Method	Fruit 26 (accuracy) (%)	Fruit 15 (accuracy) (%)
PCA + FSCABC [48]	89.18	87.75
WE + BBO [49]	89.85	87.17
FRFE + BPNN [50]	88.78	87.52
FRFE + IHGA [51]	89.96	88.85
13-Layer CNN [52]	94.59	92.38
CAE-ADN	95.86	93.78

hidden features of the data. At last, it can obtain the perception of the whole target.

Our method now stays in the offline experiment. In the future, we will build an application to experiment in the actual supply chain scenario. In addition, we will expand the data set, so that the model can adapt to more complex classification tasks. The proposed model cannot be installed in some embedded systems of machine vision, and it is still in the research of two-dimensional image, which cannot really realize the location of spatial points. In the future, we need to study the transformation from 2-D coordinates to 3-D coordinates, which can be combined with Kinect sensor for three-dimensional positioning.

Acknowledgements This work was supported by Beijing Social Science Foundation Grant 19JDGLA002 and was partially supported by Beijing Logistics Informatics Research Base. We are very grateful for their support.

Author contributions Conceptualization, SL; methodology, GX; software, GX; writing—original draft preparation, GX and YM.

Funding Beijing Social Science Foundation Grant 19JDGLA002.

Availability of data and material The authors declare availability of data and material.

Code availability The authors declare code availability.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the

permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Pak M, Kim S (2017) A review of deep learning in image recognition. In: 2017 4th International conference on computer applications and information processing technology (CAIPT)
- Zhai H (2016) Research on image recognition based on deep learning technology. In: 2016 4th International conference on advanced materials and information technology processing (AMITP 2016)
- Jiang L, Fan Y, Sheng Q, Feng X, Wang W (2018) Research on path guidance of logistics transport vehicle based on image recognition and image processing in port area. EURASIP J Image Video Process
- Liu F, Snetkov L, Lima D (2017) Summary on fruit identification methods: a literature review. Adv Soc Sci Educ Hum Res 119:1629–1633
- Getahun S, Ambaw A, Delele M, Meyer CJ, Opara UL (2017) Analysis of airflow and heat transfer inside fruit packed refrigerated shipping container: Part I—model development and validation. J Food Eng 203:58–68
- Rocha A, Hauagge DC, Wainer J, Goldenstein S (2010) Automatic fruit and vegetable classification from images. Comput Electron Agric 70(1):96–104. <https://doi.org/10.1016/j.compag.2009.09.002>
- Tu S, Xue Y, Zheng C, Qi Y, Wan H, Mao L (2018) Detection of passion fruits and maturity classification using red-green-blue depth images. Biosyst Eng 175:156–167. <https://doi.org/10.1016/j.biosystemseng.2018.09.004>
- Wang C, Han D, Liu Q, Luo S (2018) A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. IEEE Access 7:1–1
- Mnih V, Heess N, Graves A, Kavukcuoglu K (2014) Recurrent models of visual attention. In: Advances in neural information processing systems
- Chaudhari S, Polatkan G, Ramanath R, Mithal V (2019) An attentive survey of attention models
- Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. Machine learning. In: Proceedings of the twenty-fifth international conference (ICML 2008), Helsinki, Finland, June 5–9, 2008. ACM
- Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
- Unser M (1986) Sum and difference histograms for texture classification. IEEE TPAMI 8(1):118–125
- Pass G, Zabih R, Miller J (1997) Comparing images using color coherence vectors. In: ACMMM, pp 1–14
- Stehling R, Nascimento M, Falcao A (2002) A compact and efficient image retrieval approach based on border/interior pixel classification. In: CIKM, pp 102–109
- Garcia F, Cervantes J, Lopez A, Alvarado M (2016) Fruit classification by extracting color chromaticity, shape and texture features: towards an application for supermarkets. IEEE Lat Am Trans 14(7):3434–3443
- Serrano N, Savakis A, Luo J (2004) A computationally efficient approach to indoor/outdoor scene classification. In: ICPR, pp 146–149
- Lyu S, Farid H (2005) How realistic is photorealistic? IEEE Trans Signal Process (TSP) 53(2):845–850
- Rocha A, Goldenstein S (2007) PR: more than meets the eye. In: ICCV, pp 1–8
- Bolle R, Connell J, Haas N, Mohan R, Taubin G (1996) Veggievision: a produce recognition system. WACV, Sarasota, pp 1–8
- Jurie F, Triggs B (2005) Creating efficient code books for visual recognition. ICCV 1:604–610
- Agarwal S, Awan A, Roth D (2004) Learning to detect objects in images via a sparse, part-based representation. TPAMI 26(11):1475–1490
- Marszalek M, Schmid C (2006) Spatial weighting for bag-of-features. In: CVPR, pp 2118–2125
- Sivic J, Russell B, Efros A, Zisserman A, Freeman W (2005) Discovering objects and their location in images. In: ICCV, pp 370–377
- Pardo-Mates N, Vera A, Barbosa S, Hidalgo-Serrano M, Núñez O, Saurina J et al (2017) Characterization, classification and authentication of fruit-based extracts by means of HPLC-UV chromatographic fingerprints, polyphenolic profiles and chemometric methods. Food Chem 221:29
- Shao W, Li Y, Diao S, Jiang J, Dong R (2017) Rapid classification of chinese quince (*Chaenomeles speciosa nakai*) fruit provenance by near-infrared spectroscopy and multivariate calibration. Anal Bioanal Chem 409(1):115–120
- Radi CS, Litananda WS et al (2016) Electronic nose based on partition column integrated with gas sensor for fruit identification and classification. Comput Electron Agric 121:429–435
- Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE TPAMI 33(3):239–253
- Zhang Y, Phillips P, Wang S, Ji G, Yang J, Wu J (2016) Fruit classification by biogeography-based optimization and feedforward neural network. Expert Syst 33(3):239–253
- Wang S, Lu Z, Yang J, Zhang Y, Dong Z (2016) Fractional Fourier entropy increases the recognition rate of fruit type detection. BMC Plant Biol 16(S2):85
- Lu Z, Lu S, Wang S, Li Y, Lu H (2017) A fruit sensing and classification system by fractional Fourier entropy and improved hybrid genetic algorithm. In: International conference on industrial application engineering 2017
- Zhang Y, Wang S, Ji G, Phillips P (2014) Fruit classification using computer vision and feedforward neural network. J Food Eng 143:167–177
- Kuo Y-H, Yeh Y-T, Pan S-Y, Hsieh S-C (2019) Identification and structural elucidation of anti-inflammatory compounds from Chinese olive (*Canarium album* L.) fruit extracts. Foods 8(10):441. <https://doi.org/10.3390/foods8100441>
- Zhang Y, Dong Z, Chen X, Jia W, Du S, Muhammad K et al (2017) Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. Multimed Tools Appl 78:3613
- Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: convolutional block attention module. Springer, New York
- Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: IEEE Conference on computer vision and pattern recognition (CVPR), Honolulu, HI, 2017, pp 2261–2269
- Liou CY, Cheng WC, Liou JW, Liou DR (2014) Autoencoder for words. Neurocomputing 139:84–96
- Rumelhart DE (1986) Learning internal representations by error propagation, parallel distributed processing. Explorations in the microstructure of cognition. MIT Press, Cambridge
- Baldi P (2012) Autoencoders, unsupervised learning, and deep architectures. ICML Unsuperv Transf Learn 27:37–50
- Kingma DP, Welling M (2013) Auto-encoding variational Bayes
- Masci J, Meier U, Cireşan D, Schmidhuber J (2011) Stacked convolutional auto-encoders for hierarchical feature extraction. Artif Neural Netw Mach Learn ICANN 89:52–59. https://doi.org/10.1007/978-3-642-21735-7_7

42. Zagoruyko S, Komodakis N (2017) Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: ICLR
43. Lowe D (1999) Object recognition from local scale-invariant features. *Proc Seventh IEEE Int Conf Comput Vis* 2:1150–1157
44. Serre T, Wolf L, Poggio T (2007) Object recognition with features inspired by visual cortex. In: *Proceedings of computer vision and pattern recognition conference* (2007)
45. Kingma D, Ba J (2014) ADAM: a method for stochastic optimization. *Comput Sci*
46. Mureşan H, Oltean M (2017) Fruit recognition from images using deep learning
47. Israr H, Qianhua H, Zhuliang C, Wei X (2018) Fruit recognition dataset (version V 1.0). Zenodo. <https://doi.org/10.5281/zenodo.1310165>
48. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *IEEE Conference on computer vision and pattern recognition*. IEEE computer society
49. Ji G (2014) Fruit classification using computer vision and feedforward neural network. *J Food Eng* 143:167–177
50. Wei L (2015) Fruit classification by wavelet-entropy and feedforward neural network trained by fitness scaled chaotic ABC and biogeography-based optimization. *Entropy* 17(8):5711–5728
51. Lu Z (2016) Fractional Fourier entropy increases the recognition rate of fruit type detection. *BMC Plant Biol* 16(S2):10
52. Lu Z, Li Y (2017) A fruit sensing and classification system by fractional fourier entropy and improved hybrid genetic algorithm. In: *5th International conference on industrial application engineering (IIAE)*. Kitakyushu, Institute of Industrial Applications Engineers, Japan, pp 293–299
53. Brahmachary TK, Ahmed S, Mia MS (2018) Health, safety and quality management practices in construction sector: a case study. *J Syst Manag Sci* 8(2):47–64
54. Hai L, Fan Chunxiao W, Yuexin LJ, Lilin R (2014) Research of LDAP-based IOT object information management scheme. *J Logist Inform Serv Sci* 1(1):51–60
55. Zhao PX, Gao WQ, Han X, Luo WH (2019) Bi-objective collaborative scheduling optimization of airport ferry vehicle and tractor. *Int J Simul Model* 18(2):355–365. [https://doi.org/10.2507/IJSIMM18\(2\)CO9](https://doi.org/10.2507/IJSIMM18(2)CO9)
56. Xu W, Yin Y (2018) Functional objectives decision-making of discrete manufacturing system based on integrated ant colony optimization and particle swarm optimization approach. *Adv Prod Eng Manag* 13(4):389–404. <https://doi.org/10.14743/apem2018.4.298>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.