ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.031

Human Action Recognition in Still Images Combining CBOW Language Processing and CNN

Quanzhi Gong [1], Yuxiang Xie [1+], Jie Yan[1] and Xidao Luan [2]

[1] College of Systems Engineering, National University of Defense Technology, Changsha, China
[2] College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, China

Abstract. In the field of computer vision, human action recognition task is a difficult problem, especially for static images. Traditional action recognition methods rely on local or global features designed manually, which are limited by people's visual cognition of image. Therefore, it is difficult to achieve the optimal recognition results. In recent years, the convolutional neural network (CNN) has become popular in the action recognition methods, which usually improve the performance by designing network structure. Most of the previous methods solve this task through the implicit image content association information contained in the image. Due to the lack of interpretability of convolution network, it is difficult to judge which part of image information plays an important role, so it is hard to optimize the network model by strengthening some image features. This paper proposes a method of action recognition which combines language processing technology with CNN. The proposed method uses the continuous bag-of-words (CBOW) model to assist CNN to complete the action recognition task by taking the cooccurrence information of the object pairs explicitly. The method is tested on two public datasets, namely Stanford 40 action and Pascal VOC Action 2012. The comparison result with the state-of-the-art methods shows that, as an exploration on the combination of language model and general CNN, the proposed method is satisfactory, whose accuracy rates reach 91.4% and 85.9% respectively.

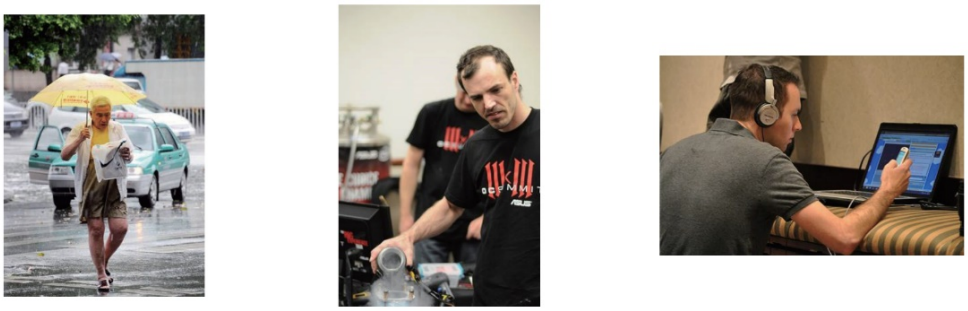
Keywords: action recognition, CBOW, DenseNet, deep learning.

1. Introduction

With the development of artificial intelligence technology, action recognition task has attracted more and more attention because of its rich applications in image retrieval, media tag generation, human-computer interaction and other fields. However, due to the lack of temporal features, the problem of action recognition is very challenging, especially in static images. In addition, the diversity of samples caused by viewpoint, shooting angle, light intensity, and resolution is also a tough problem. In this paper, we focus on the action recognition in static images.

Since the action recognition task was proposed, it has attracted extensive attention. In the early stage of research, action recognition methods mainly included traditional methods, such as methods based on global features and local features. These methods use artificial features and common classifiers for action recognition. Recently, many researchers have adopted convolutional neural networks (CNN), which can automatically extract features. The common method to improve the accuracy is to design a new network module. However, we changed the whole structure and designed a two-stream network, in which the visual-stream is composed of CNN and the language-stream is composed of the continuous bag-of-words (CBOW) model. It should be noted that for the CBOW model, it is important to give the object information of the action in the image in advance.

+ Corresponding author. Tel.: +86-13787100462
E-mail address: yxxie@nudt.edu.cn.



(a) (b) (c)

Fig. 1: Easily misjudged action images.

Unlike the object in object detection task, the concept of the action in action recognition task is abstract and cannot be accurately matched with the entity in the image. If the feature information in the image is directly used for recognition, it is easy to misjudge especially for complex images. As shown in fig.1, the action in the images is easily misjudged as the wrong action category due to the character pose and the shooting angle of the images. For example, due to the human posture, the action in Fig.1(a) is more likely to be recognized as reading rather than holding an umbrella, and the action in Fig.1(b) is difficult to be recognized as pouring water. For the action in Fig.1(c), it is hard to distinguish whether the correct result of action recognition is using a computer or sending a text message. However, if object information can be given, such as the umbrella, cup and mobile phone in fig.1, the error rate will be greatly reduced. Therefore, objects are very important in the task of action recognition. Understanding the relationship between objects and actions can improve the accuracy of action recognition.

It is meaningful to use the related logical patterns of objects and actions to assist action recognition. Because it can be well expressed in language, this paper proposes an action recognition model in still images that combines CBOW language model and CNN. The proposed method introduces the CBOW model to analyze the correlation logic of objects and actions in language, as an additional clue to improve the accuracy of recognition.

The contributions of this paper are as follows: (1) The performance of different general CNNs in the action recognition task is evaluated. (2) A CBOW language processing model is proposed for action recognition. (3) A framework combining CBOW model and DenseNet is proposed. (4) Experiments are carried out on two public datasets, Stanford 40 Action and Pascal VOC Action 2012, which proved the effectiveness of the proposed method.

The rest of the paper is organized as follows. First, the related work about this topic is reviewed in Section 2. Then, the method is introduced in Section 3. Section 4 presents the experiment results. In Section 5, further research directions are proposed.

2. Related work

For a long time, action recognition is an active research direction in the field of computer vision. Most of the existing work is about action recognition in video, which can use both spatial and temporal information. However, due to the lack of continuous information, action recognition in still images is a more challenging problem.

Unlike a large number of continuous images contained in the video, the still image has only one frame, in which each pixel is formed by the values red, green, and blue color channels. Early research mainly focused on traditional methods. Khan, F.S [1] uses the histogram of oriented gradient (HOG) features for classification. Sener, F [2] and Ali, K.H. [3] describe images based on local features, such as scale invariant feature transform (SIFT) feature. Many researchers also pay attention to the interaction, such as Zhang, Y. [4] and Yao, B. [5]. Specifically, Yao, B. [5] learned the action base and reconstruction coefficients of the image according to each action interaction, and constructed a special image representation. In addition, Khan, F.S. [6] paid more attention to posture information and built a semantic pyramid for image representation by combining the body region detection model and bag-of-words model for features.

Besides the RGB channels, some researchers introduced depth information and skeleton information as the clues for recognition. Depth information refers to the distance between the viewpoint and the surface of the objects in the scene. Xu, C. [7] used depth information to describe human's posture, while Bingbing Ni [8] obtained human motion features and different levels of context from the depth images. Skeleton information refers to the position of specific joints of the human body and related connected bones in the images. Batabyal, T. [9] uses the skeleton coordinates to describe the pose of the main character.

With the rapid development of deep learning network, researchers explore the application of various CNN in the field of action recognition actively. After Lavinia, Y. [10] used the basic CNN network, Gkioxari, G. [11] proposed to use R*CNN to extract action features in images. Based on the idea of hierarchical context, Zhu, H. [11] divided the image into different levels of parts to acquire diverse understandings from part to whole. Chao, Y. [12] proposed a three-stream structure to analyse the action from three perspectives: human pose, local context of objects, and the relationship between object pairs. Bhandari, B. [13] uses three sub-networks to extract three-level features of the image, including image stream, attention image stream and image partial stream. Chapariniya, M. [14] introduced a knowledge distillation framework to solve the problem of too many parameters and the demand for bounding boxes. Besides, acquiring missing temporal features is also a hot research direction to improve the recognition accuracy. Wang, Y. [15] extracts the temporal features by matching the images with related special video groups to realize "temporal hallucinating". Safaei, M. [16] obtains key pixels through domain mapping and predicts the optical flow for their temporal information.

In the past, these methods considered directly extracting and analysing features from images, and implicit key object information and relationship information in feature expressions. This unexplainability is not conducive to the improvement and development of the model. The method proposed in this paper considers introducing the language processing technology to explicitly model the relationship between objects, which is not only conducive to subsequent research, but also can explore the impact of the introduction of natural language related methods.

The method proposed in this paper is to combine the CBOW language model with DenseNet to form an action recognition framework for still images. It should be noted that the use of the CBOW language model is cross-domain. In addition, we verify the effectiveness of our method on two public datasets.

3. DenseNet Action Recognition Framework Combined with CBOW Language Model

This section outlines the proposed method, clarifies the composition and structure of the framework, describes the steps of training the framework, and introduces the method of applying CBOW language model to the action recognition in still images.

3.1. The Composition and Structure of the Framework

The CBOW model was originally a language processing technology. This method uses the co-occurrence configuration relationship between object pairs and actions to predict the actions in the image. Its advantage is that it has good recognition effect in most cases, but its disadvantage is that it cannot be well recognized when the co-occurrence configuration is not unique. The essential reason is that CBOW model does not make full use of other effective information in the image. However, in most cases, CNN can extract the global and local features of images, and has excellent versatility. Therefore, it is suggested to combine these two technologies to form an action recognition framework with DenseNet as the main body and CBOW language model as the supplement. This can not only improve the recognition accuracy of the network, but also ensure the robustness.

The DenseNet network used in the vision-stream was proposed by Huang, G. [17], which is mainly composed of dense blocks, the transmission layer and classification layer. The dense blocks of the network contain multiple convolution layers, and then the activation and feature map reduction operations are performed. The key point of this module is that the output of each convolution layer is not only the input of the next layer, but also the input of all convolution layers after this layer in the dense block. From the perspective of feature migration, DenseNet connects the input and image features obtained from each

convolution layer to subsequent convolution layer for processing. In this way, each layer contains the features of all previous layers, so each intermediate result in the extraction process will be represented in the final feature map. This feature multiplexing structure obviously improves the efficiency of feature use, improves the expressive ability of the entire network, and is conductive to action recognition.

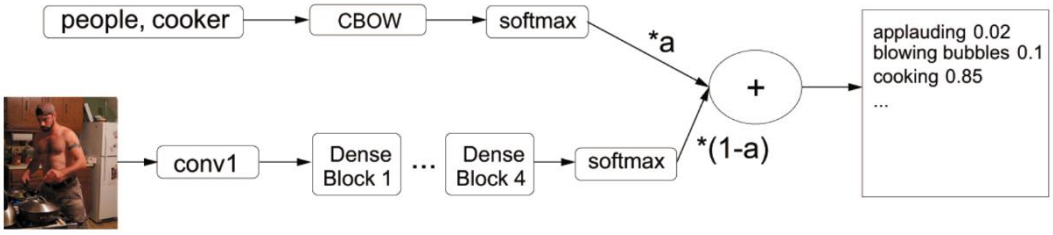


Fig. 2: DenseNet action recognition framework based on CBOW language model.

The structure of action recognition model is shown in Figure 2. In this model, the DenseNet module in vision-stream and the CBOW module in language-stream predict the results independently. Then, these two prediction distributions are fused to obtain the final recognition result. It should be noted that the coefficient *a* is a hyperparameter.

Because the training time gap between DenseNet and CBOW is large, and CBOW does not have a mature pre-training model, the following two-step training process is designed. We train the CBOW model firstly and then use the trained CBOW model to assist the training of DenseNet. We input the image into DenseNet to get the recognition result, and then input the given object pairs of the input image into the trained CBOW model to get other recognition results. Then, these two results are combined with the weight *a*, *(1-a)*. The final prediction is compared with the action label of the input image. After getting the loss, we use the stochastic gradient descent (SGD) algorithm to adjust the parameters of the model for training.

3.2. CBOW Language Model and Its Transplantation Method

During the training and testing process of the CBOW model used in this method, the object pairs corresponding to the action are given. Because there is almost no occlusion in the Stanford 40 Action dataset and Pascal VOC Action 2012 dataset, and the object pairs are located in the centre of the images, it can be easily recognized using existing advanced object detection methods. Therefore, in order to simplify the experiment, we assume that the object pairs corresponding to the action are known.

The two models included in the word2vec method were proposed by Mikolov T [18], in which the full name of CBOW is the continuous bag-of-words model. Its essence is to predict the target word through the background word without considering the word order. The schematic diagram of the model is shown in Figure 3. The four blocks *w(t-1)*, *w(t-2)*, *w(t+1)*, and *w(t+2)* in the INPUT layer represent the input words, which are usually one-hot encoded word vectors. The PROJECTION layer is the vector of the hidden layer, and the OUTPUT layer is the probability distribution vector of the predicted target word in all words.

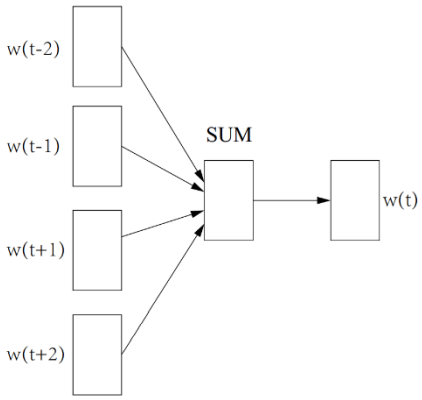


Fig. 3: Structure of the CBOW model.

In order to obtain the training data of the CBOW model, we extract the image content of text data from the image dataset. This process mainly uses the action label of the image and the main object pair to form a

sentence. When the interactive action involves two objects, the sentences can be designed as the pattern "object-action-object", such as "person is reading article". On the other hand, non-interactive action involves only one object, so the sentences can be "object-action-self", such as "person run self".

$$H_{1*n} = \frac{1}{2}\sum_{i=0}^{1}(m_i * W_{v*n}) \quad (1)$$

$$P_{1*v} = H_{1*n} * W'_{n*v} = [x_1, x_2, \dots, x_v]^T \quad (2)$$

$$P'_{1*v} = [\frac{e^{x_1}}{\sum_{j=1}^{v} e^{x_j}}, \frac{e^{x_2}}{\sum_{j=1}^{v} e^{x_j}}, \dots, \frac{e^{x_v}}{\sum_{j=1}^{v} e^{x_j}}]^T \quad (3)$$

The parameters of the model are two matrices W_{v*n} and W'_{n*v}, where *v* is the dimension of the one-hot word vector. In the other words, it represents the total number of words in the entire corpus. And *n* represents the dimension of the hidden layer. The CBOW module used in this article has two words in the input layer. The operation process of the model is not complicated, and the details are as follows. First, as shown in formula (1), the two one-hot word vectors m_1, m_2 are multiplied by the shared weight matrix W_{v*n}. Next, the products are averaged. The result can be regarded as a co-occurrence statement of background words. Then the hidden layer vector is multiplied by the second weight matrix W'_{n*v}, as shown in formula (2). Finally, according to formula (3), the probability distribution of the predicted target word is obtained as an output vector of size *1*v* after the softmax activation function. The word with the highest probability is the action word.

During training, the weight matricesW_{v*n} and W'_{n*v} need to be initialized, and the loss is calculated using the output vector and the one-hot encoding vector of the action word. The weight matrices W_{v*n} and W'_{n*v} are adjusted by gradient descent algorithm.

With this method, CBOW can predict the image action. In essence, the CBOW language model in this paper actually represents the semantic logic clues of the relationship between objects and actions. It is an innovative work to explicitly introduce the co-occurrence cues into computer vision tasks. At present, most action recognition models are implicit in global features. As entities that can be found in images, object pairs can be an effective recognition medium for abstract action. From the semantic logic of the relationship between objects and actions, language itself is a carrier of logical information. It is a simple and effective method to use text sequence to introduce relevant logic. Therefore, exploring the application of CBOW model to the field of action recognition in still images is valuable for research.

4. Experimental Results

In this section, we evaluate the performance of the proposed method on two public action recognition datasets: Stanford 40 Action dataset and Pascal VOC Action 2012 dataset. The former contains 40 action categories, each with 180-300 pictures, which has a total of 4000 training pictures and 5532 test pictures. The latter has 10 action categories and an "other" category, with a total of 9157 images. In our experiments, the object pairs corresponding to the action of these images are labelled in advance, which will be used in the CBOW model.

4.1. Performance of the Single Model

In the field of deep learning, the contrasts in the structure of different CNNs lead to different performance rankings in various application fields. Therefore, we conducted experiments on the performance of varied CNNs in action recognition tasks. Table 1 shows the action recognition accuracy of Shufflenet v2, Resnet34, Inception v3 and DenseNet121 on the two datasets, Stanford 40 Action and Pascal VOC Action 2012.

Table 1: Performance comparison of general CNNs on Stanford 40 Action dataset and Pascal VOC Action 2012 dataset

Mean AP(%)	Stanford 40 Action	Pascal VOC Action 2012
Shufflenet v2	20.0	40.7
Resnet34	87.8	79.2
Inception v3	87.0	76.5
DenseNet121	**88.8**	**83.6**

The results of the experiment basically met expectations. As a lightweight network, Shufflenet v2 has the advantages of short training time, small memory footprint, and low resource requirements. But its recognition accuracy is bad. The result of this lightweight network proves that the dataset used in this article is difficult. The other three networks are well-known models for recognition accuracy in various fields of computer vision. And the results in the field of action recognition are also relatively satisfactory. On the Stanford 40 Action dataset, the performance gap between Resnet34, Inception v3 and DenseNet121 is not large, and the last one is better than the other two. On the Pascal VOC Action 2012 dataset, DenseNet121 has obvious advantages. The experiment results prove that DenseNet does have better applicability in action recognition tasks. Its high efficiency in feature use and its ability to avoid over-fitting are both important in action recognition tasks.

We conducted experiments on the Pascal VOC Action 2012 dataset to verify the performance of the proposed CBOW model transplantation method in the action recognition task. The results in each action category are shown in Table 2. Except for “Running” and “Walking”, the rest action classes are recognized correctly. This abnormal situation stems from the special settings of the model. Specifically, the text corpus of the CBOW model used in this article is composed of object pairs and action extracted from image datasets. In this case, the mapping between the CBOW model and the corpus is “background words - object pairs” and “target words - action”. Since the “jumping”, “running” and “walking” action in the Pascal VOC Action 2012 dataset have the same object pairs, they will be recognized incorrectly. In the original image dataset, the number of samples for each action class is not exactly the same. Therefore, when the object pair is “person” and “self”, the model is more inclined to judge the action as the action with more samples in the training dataset, “Jumping”. So, the recognition accuracy of this action is 100%, and the accuracy of the other two action is lower.

Table 2: The CBOW model's recognition accuracy of each action class on the Pascal VOC Action 2012 dataset

Action type	AP(%)
Jumping	100
Phoning	100
Playinginstrument	100
Reading	100
Ridingbike	100
Ridinghorse	100
Running	50
Takingphoto	100
Usingcomputer	100
Walking	33.3
Others	100

From these results, it can be inferred that, the results of the CBOW model will be terrible in the datasets where the most action is with the same object pairs. Therefore, it is not suitable to take the CBOW model as the main body of the action recognition framework. Because the CBOW model has a great effect in most time, we choose to take this model to assist the general CNN, DenseNet. In most cases, proposed method can combine the superior versatility of the general CNN and the high performance of the language processing model to obtain good recognition results.

4.2. Performance of the Overall Model

We also tested the model performance of the overall model on the Stanford 40 Action dataset and Pascal VOC Action 2012 dataset. Table 3 shows the performance (mAP) of four models. By analysing the data in the table, it is not hard to find that the overall models have a certain improvement in performance with the assistance of the CBOW model, whether the backbone network is ResNet or DenseNet. Therefore, the introduction of the CBOW model is effective. In the other words, it is feasible to use language processing methods to assist CNN to complete the action recognition task.

Table 3: Comparison of the overall model performance

Mean AP(%)	Stanford 40 Action	Pascal VOC Action 2012
Resnet34	87.8	79.2
Resnet34 + CBOW	91.1	85.2
DenseNet121	88.8	83.6
DenseNet121 + CBOW	**91.4**	**85.9**

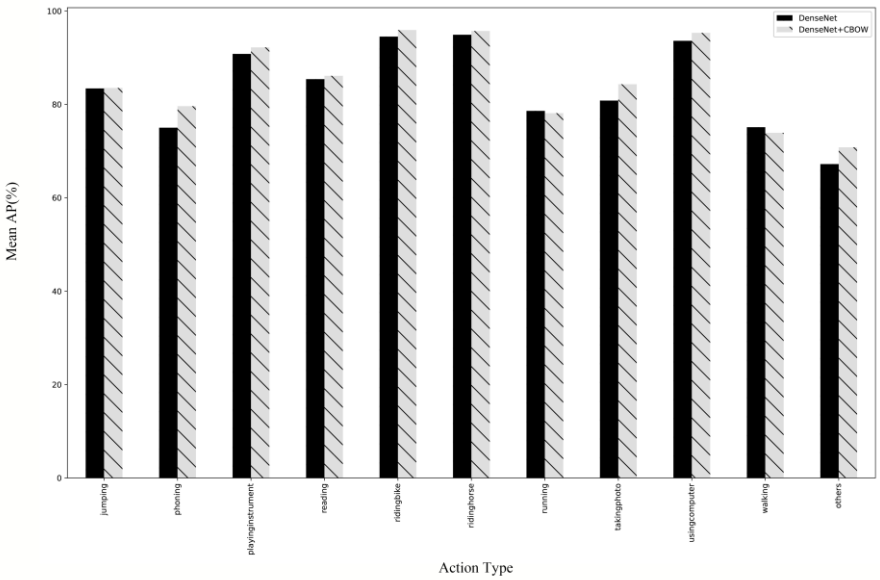


Fig. 4: The performance of the proposed model on each action class on the Pascal VOC Action 2012 dataset.

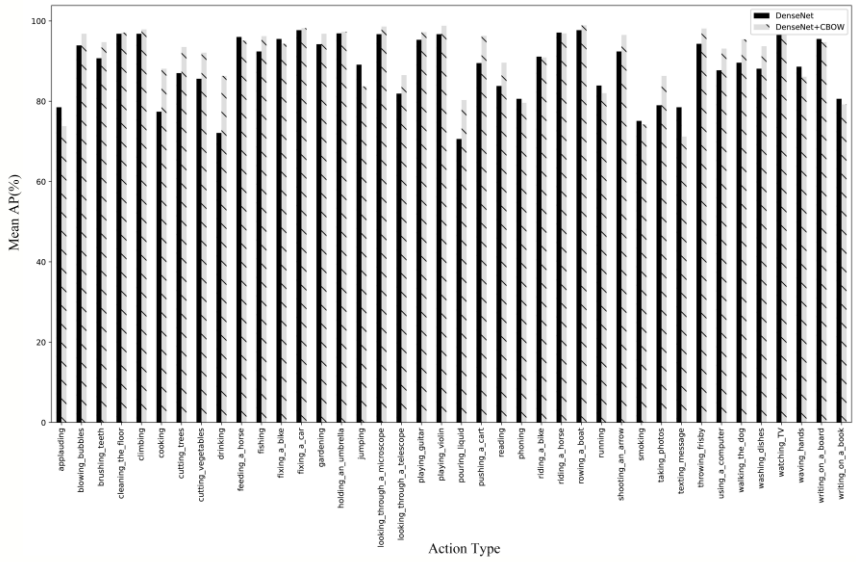


Fig. 5: The performance of the proposed model on each action class on the Stanford 40 Action dataset.

In table 2, the recognition results show that the CBOW model has obvious defects in special cases. In order to investigate whether this shortcoming will be left in the overall model, we use both DenseNet and DenseNet + CBOW model to test each action class on two datasets. The experimental results are shown in Figure 4 and Figure 5, which prove that this weakness will not have a significant negative impact on the performance of the proposed model. The results of the experiment on the Pascal VOC Action 2012 dataset in Figure 4 show that the recognition performance of the DenseNet + CBOW model in specific action classes is slightly worse than that of DenseNet, but the accuracy of other action classes is significantly improved. This shows that the disadvantage of the CBOW model is left in the model, but it did not lead to an obvious drop in the performance. Overall, the assistance of the CBOW model has greatly improved the performance of the DenseNet model. Similar conclusions can be obtained from the results of the Stanford 40 Action dataset. We

randomly selected the positive and negative samples of the proposed method from the four types of images in the Stanford 40 action dataset.

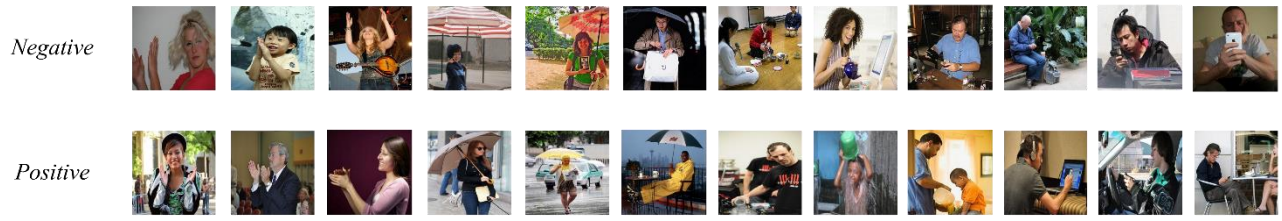


Fig. 6: The positive and negative samples of the proposed method from the images in the Stanford 40 Action dataset.

4.3. Comparison with State-of-the-Art Methods

We compared the performance of the proposed method and the state-of-the-art methods on the Stanford 40 Action and Pascal VOC Action 2012 datasets. The experimental results are shown in table 4. From the results on the Stanford 40 Action dataset, it can be seen that our method is relatively satisfactory. The other comparison results show that the effect of our method is relatively poor when the number of action classes with the same object pairs accounts for a large proportion of the total number of action classes in the dataset. This also shows that in the further work, the defects in this area need to be eliminated to improve the accuracy.

Table 4: Comparison of our method with other methods on the Stanford 40 Action and Pascal VOC Action 2012 datasets. The value in bold font indicates the best result. The symbol "-" indicates that the experimental results have not been announced.

Mean AP(%)	Stanford 40 Action	Pascal VOC Action 2012
Safaei, M. [16]	81.76	-
Safaei, M. [20]	82.9	-
Batabyal, T. [9]	84.8	-
Wu, W. [19]	88.73	-
Zhu, H. [11]	-	86.4
Xin, M. [22]	-	90.6
Xin, M. [21]	-	**92.1**
Yan, S. [23]	90.7	90.2
Ours	**91.4**	85.9

Purely from the perspective of recognition accuracy, the accuracy of the proposed method on the Stanford 40 Action dataset is only equal to that of the state-of-the-art methods. However, the model structure in this paper is much simpler than them, which is only composed of the basic general CNN and CBOW network with a few parameters. The main work of this paper is to explore the application of language model in image action recognition task. The experiment results show that this application is effective. Our proposed method is better than the recognition result of simple general CNN, which is in line with the analysis of CBOW model in advance. From the perspective of research, the value of this paper is the explicit application of the cooccurrence information of object pairs, as well as the semantic logic of object pairs and action. This work is a pioneering exploration, which has been proved to be effective. Adjusting the selection of specific visual model and language model will certainly improve the recognition accuracy, which is the research direction in our future work.

5. Conclusion

In this paper, a model framework combining CBOW language model and DenseNet is proposed. Experiments on two datasets, Stanford 40 Action and Pascal VOC Action 2012, show the effectiveness of the proposed model. The main contribution of this paper is to explore the impact of introducing language model into action recognition task. The essence of this method is to use language model to explicitly model the co-occurrence information and semantic logic of object pairs and actions. This method can effectively

improve the recognition accuracy of general CNN, and achieve the same results as the state-of-the-art methods without significantly increasing the calculation cost and time.

A limitation of this paper is that the overall framework needs object pair information of each image during training and testing, which is a demanding requirement. Therefore, the main direction of future work is to use appropriate object detection models and filtering algorithms to automatically extract object pairs in the images. In addition, the CBOW language model has poor performance in action recognition of the same object pair. Therefore, it is proposed to enrich the information in the corpus by introducing other clues (such as the relative position of the objects) to improve the recognition accuracy. In future research, it is also an effective research direction to use more mature action recognition CNN instead of DenseNet.

6. References

[1] F. S. Khan, J. Xu, J. van de Weijer, A. D. Bagdanov, R. M. Anwer, and A. M. Lopez. Recognizing Actions Through Action-Specific Person Detection. *IEEE Trans. on Image Process.* 2015, **24**(11): 4422–4432.

[2] F. Sener, C. Bas, and N. Ikizler-Cinbis. On Recognizing Actions in Still Images via Multiple Features. In: *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Heidelberg: Springer Berlin Heidelberg, 2012, pp. 263–272.

[3] K. H. Ali and T. Wang. Recognition of Human Action and Identification Based on SIFT and Watermark," In: *Intelligent Computing Methodologies*. Cham: Springer International Publishing, 2014, pp. 298–309.

[4] Y. Zhang, L. Cheng, J. Wu, J. Cai, M.N. Do, J. Lu. Action Recognition in Still Images With Minimum Annotation Efforts. *IEEE Trans. on Image Process*. 2016, **25**(11): 5479–5490.

[5] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In: *International Conference on Computer Vision*. Spain: IEEE Computer Society, 2011, pp. 1331–1338.

[6] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta. Semantic Pyramids for Gender and Action Recognition. *IEEE Trans. on Image Process*. 2014, **23**(8): 3633–3645

[7] C. Xu, L. N. Govindarajan, Y. Zhang, and L. Cheng. Lie-X: Depth Image Based Articulated Object Pose Estimation, Tracking, and Action Recognition on Lie Groups. *International Journal of Computer Vision*. 2017, **123**(3): 454–478.

[8] Bingbing Ni, Yong Pei, Z. Liang, Liang Lin, and P. Moulin. Integrating multi-stage depth-induced contextual information for human action recognition and localization. In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. China: IEEE Computer Society. 2013, pp. 1–8.

[9] T. Batabyal, T. Chattopadhyay, and D. P. Mukherjee. Action recognition using joint coordinates of 3D skeleton data. In: *2015 IEEE International Conference on Image Processing*. Canada: IEEE Computer Society. 2015, pp. 4107–4111.

[10] Y. Lavinia, H. H. Vo, and A. Verma. Fusion Based Deep CNN for Improved Large-Scale Image Action Recognition. In: *2016 IEEE International Symposium on Multimedia*. USA: IEEE Computer Society. 2016, pp. 609–614.

[11] H. Zhu, J.-F. Hu, and W.-S. Zheng. Learning Hierarchical Context for Action Recognition in Still Images. In: *Advances in Multimedia Information Processing*. Cham: Springer International Publishing. 2018, pp. 67–77.

[12] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to Detect Human-Object Interactions. In: *2018 IEEE Winter Conference on Applications of Computer Vision*. NV: IEEE Computer Society. 2018, pp. 381–389.

[13] B. Bhandari, G. Lee, and J. Cho. Body-Part-Aware and Multitask-Aware Single-Image-Based Action Recognition. *Applied Sciences*. 2020, **10**(4): 1531-1549.

[14] M. chapariniya, S. S. Ashrafi, and S. B. Shokouhi. Knowledge Distillation Framework for Action Recognition in Still Images. In: *10th International Conference on Computer and Knowledge Engineering*. Iran: IEEE Computer Society. 2020, pp. 274–277.

[15] Y. Wang, L. Zhou, and Y. Qiao. Temporal Hallucinating for Action Recognition with Few Still Images. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. UT: IEEE Computer Society. 2018, pp. 5314–5322.

[16] M. Safaei and H. Foroosh. Still Image Action Recognition by Predicting Spatial-Temporal Pixel Evolution. In: *2019 IEEE Winter Conference on Applications of Computer Vision*. USA: IEEE Computer Society. 2019, pp. 111–120.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition*. HI: IEEE Computer Society. 2017, pp. 2261–2269.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs], 2013.

[19] W. Wu and J. Yu. A Part Fusion Model for Action Recognition in Still Images. *Neural Information Processing - 27th International Conference*. Cham: Springer International Publishing. 2020, pp. 101–112.

[20] M. Safaei and H. Foroosh. A Zero-Shot Architecture for Action Recognition in Still Images. In: *25th IEEE International Conference on Image Processing*. Athens: IEEE Computer Society. 2018, pp. 460–464.

[21] M. Xin, S. Wang, and J. Cheng. Entanglement Loss for Context-Based Still Image Action Recognition. In: *2019 IEEE International Conference on Multimedia and Expo*. China: IEEE Computer Society. 2019, pp. 1042–1047.

[22] M. Xin, H. Zhang, D. Yuan, and M. Sun. Learning discriminative action and context representations for action recognition in still images. In: *2017 IEEE International Conference on Multimedia and Expo*. Hong Kong: IEEE Computer Society. 2017, pp. 757–762.

[23] S. Yan, J. S. Smith, W. Lu, and B. Zhang. Multibranch Attention Networks for Action Recognition in Still Images. *IEEE Transactions on Cognitive and Developmental Systems*. 2018, **10**(4): 1116–1125.