Semi-supervised long short-term memory for human action recognition

Hong Liu¹, Chang Liu¹, Runwei Ding¹
¹Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Shenzhen, People’s Republic of China
E-mail: liuchang-0626@pku.edu.cn

Abstract: In real human action recognition task, it is a common phenomenon that there are many unlabelled samples and few labelled samples. How to make good use of unlabelled samples to improve the generalisation ability of models is the focus of semi-supervised learning research. In this study, the authors present two semi-supervised methods based on long short-term memory (LSTM) to learn discriminative hidden features. One is the LSTM ladder network, the other is the Symmetrical LSTM network. By them unlabelled samples can be used automatically to improve learning performance without relying on external interaction. Both on the NTU-RGB+D dataset and the Kinetics dataset, their methods achieve >10 and 5% improvements, separately.

1 Introduction

Human action recognition [1, 2] is an important research area in the field of computer vision. It has many applications in security monitoring, robot design, and intelligent home design [3]. The traditional methods [4] for human action recognition method are mainly based on RGB video, which is easy to be affected by background, illumination and other behavioural independent factors. It leads to low recognition accuracy. After the boom emergence of cheaper depth cameras such as Kinect, shortcomings of traditional RGB video-based method can be easily solved. Depth-based method is more robust to illumination and occlusion, so it has become a research hotspot in recent years. Finally, we adopt skeleton sequence as our input data modality.

Although the convolutional neural network (CNN [5]) model has achieved good performance in many fields, especially in image processing, it is not fully applicable to learning time series. CNN requires a variety of auxiliary processing, but the effect is not necessarily good. In face of the issues and tasks that are sensitive to time series, recurrent neural network (RNN) is usually more appropriate. Each learning result of RNN [6] is not only related to the data of the current moment, but also to the data of the previous time. The special structure of RNN can make full use of historical data, so it has obvious advantages in dealing with sequence problems. A researcher put forward a long short-term memory network (LSTM) [7], which is an optimal model of RNN. It inherits most of the advantages of RNN and solves the vanishing gradient problem caused by the gradual reduction. So we use LSTM as our baseline model. Generally speaking, there are four main ways of training deep learning network: supervised method, unsupervised method, semi-supervised method and reinforcement learning. In many practical applications of machine learning, it is easy to find a large number of samples without class tags, but it takes special equipment or a very expensive and very long experimental process to carry out manual marking to obtain samples with class labels. Then a very small number of samples with class labels and samples with excess class-free labels are generated. Therefore, people try to add a large number of samples without labels to a limited number of labelled data to train together, in the hope of improving learning performance. Thus semi-supervised learning is produced, as shown in Fig. 1. Semi-supervised method avoids the waste of data and resources, and solves the problems of weak generalisation ability of supervised learning models and imprecise of unsupervised methods.

In this paper, we study how to train a LSTM-based model with limited labelled samples and a large number of unlabelled data for human action recognition task. We present two semi-supervised method: the LSTM ladder network and the Symmetrical LSTM network. The contribution of our work is following: this is the first paper using LSTM-based semi-supervised methods for human action recognition task. Also our methods achieve great improvements on several mainstream datasets of human action recognition, which gains a lot from the unlabelled data.

2 Related work

2.1 Human action recognition

In early studies of human action recognition, researchers often use handcrafted features as input, which shows good performance. Using different machine learning models such as SVMs [8, 9], the hand-crafted features can be trained for human action recognition. However, it is not very easy to acquire domain knowledge with the aim to design handcrafted features. In recent years, with the improvement of GPU computing power, a large number of in-depth learning methods have emerged. The most widely used is the RNN and CNN. LSTM as a variant of RNN, which has a unique advantage in dealing with long time series data. Compared with traditional methods, LSTM achieved a better performance. Das et al. [10] used LSTM as encoder to recognise and predict skeleton sequence (see Fig. 2). LSTM uses gates to control the flow of long-time and short-time information, which can deal with long-time dependencies.

However, the traditional RNN (LSTM) ignores the spatial information in the skeleton data, that is, the relative position of the skeleton points. Wang and Liang [11] consider the spatial structure of human skeleton, the data were divided into five parts: trunk and extremities, and five bidirectional circulatory neural networks (bidirectional recurrent neural network, BRNN) were used to extract the features, respectively. Then the features are merged layer by layer to the next BRNN for training. After four layers of BRNN, the spatial relationship between the various parts of the
human body is modelled from the local to the whole. Finally, the whole model is completed. The features of the volume are sent to the classifier for classification. To some extent, this hierarchical RNN model excavates the spatial features of skeleton data. The disadvantage is that because the model is too large and the number of parameters is too large, only the last layer uses the bidirectional LSTM, but the former layers using the normal bi-directional RNN, which greatly reduce the performance of the model.

In order to make full use of the spatial information of skeleton data, a spatio-temporal LSTM model was proposed in [12]. Traditional LSTM only considers the use of cellular state to store long-term information in the time dimension. For one input at any time, forget gates, input gates and output gates are used to discard or increase information. While in space-time LSTM, the state of the current skeleton point is not only related to the state of the previous moment, but also to the state of the previous skeleton point. It uses two forgotten gate groups controls the current state of time and space, respectively. In this way, both spatial and temporal features can be excavated at the same time. In addition, the skeleton points are not only arranged according to the traditional order. Considering that the human action is usually determined by some adjacent skeleton points, it proposes a circular traversal tree structure to further mine the spatial information of the skeleton points.

In human action recognition, different skeleton points at different times are of different amount of information provided by recognition, so attention model is also widely used in this field. Song et al. [13] use two networks to train the spatial attention model and the time domain attention model, respectively. The spatial attention model acts on the input skeleton of the network, and the time domain attention model acts on the output characteristics of the main network. The information of different time series and different skeleton points are weighted and end-to-end action recognition is realised. It shows that different time domain attention models can give more weight to the frame with more discriminant power, and have a greater correlation to the action. Skeleton points will also give greater weight to the overall and human perception consistent.

2.2 Semi-supervised learning

As we all know, machine learning is divided into three kinds: supervised learning, unsupervised learning and semi-supervised learning. Among them, supervised learning and unsupervised learning are relatively common, semi-supervised learning is less in contact with the former two methods. Semi-supervised learning is a combination of supervised learning and unsupervised learning. Its main idea is to use a small amount of labelled data and a large amount of unlabelled data to train a model with better results.

The research of semi-supervised learning originated in the middle of 1980s [14]. Since then, some scholars have carried out extensive research on semi-supervised learning, including semi-supervised regression [15, 16], semi-supervised feature extraction [17] and semi-supervised data manifold analysis. With the further research on semi-supervised learning, scholars have also developed semi-supervised learning methods. For example, from the early popular semi-supervised learning with mixed models, the generative models, to the relevant collaborative filtering methods based on data features [18, 19], then to the improved semi-supervised methods that introduced new mathematical methods. For example, the minimal cut of graphs [20] of Gaussian random field [21], graph theory, graph-based semi-supervised learning and so on.

With the improvement of the research system of semi-supervised learning method and the advantage of using unlabelled data to learn, the theoretical results of semi-supervised learning are also applied to practical problems. One of the typical fields which is widely used is the related application of natural language processing. Provoost and Moens [22] attempt to apply semi-supervised learning theory to the practical problem of word sense disambiguation. In processing, the need for manual tagging data is greatly reduced. Grira [23] proposed an active semi-supervised system based on cooperative training, syntactic analysis, which can effectively reduce the workload of about half of the manual marking. Another important application area of semi-supervised learning is content-based image retrieval, which can effectively improve the performance of image retrieval.

3 Methodology

We adopt the LSTM since it shows strengths in modelling the complex dynamics of human actions as time series data, and achieved good performance for human action recognition task. So we use the LSTM-based structure for both our supervised and semi-supervised methods.

3.1 LSTM for supervised learning

In this section, we briefly introduce the LSTM to make the paper self-contained. LSTM is an advanced RNN, whose neuron contains a memory cell \( C_t \) which has self-connected recurrent edge of weight 1. As shown in Fig. 3, at each time step \( t \), the neuron can choose to write, reset and read the memory cell governed by the input gate \( i_t \), forget gate \( f_t \) and output gate \( o_t \).

Considered a dataset with \( N \) labelled examples, the goal of supervised method is to find the cost function using a lot of labelled data. Here we use a basic structure contains three LSTM layers and a fully connected layer prior a top-level softmax classifier. Then the supervised LSTM cost function is

\[
C_S = - \frac{1}{N} \sum_{i=1}^{N} \log P(y^* = y_i | x_i)
\]
y* here is the prediction of the model.

3.2 LSTM ladder

Considered a dataset with N labelled examples \((x_i, y_i), \ldots, (x_N, y_N)\) and M unlabelled examples \((x_{M+N}, x_{M+N+1}, \ldots, x_{M+N+N})\), where \(M \gg N\). The goal of semi-supervised method is to find the cost function by using both the labelled examples and the large number of unlabelled examples. Inspired by the ladder network, we design the LSTM ladder network, which consists of three sub-networks: a noisy encoder, a clean encoder and a decoder as shown in Fig. 4. This structure involves both vertical connection and lateral connection.

The whole algorithm is divided into three parts. The first part, the input data pass through both clean encoder and noisy encoder. The two coding processes are basically the same; the only difference is that adding Gaussian random noise to each layer of encoder-decoder pair corresponding to each layer.

The second part is to decode the unlabelled data using the result of noisy encoder–decoder pair. The supervised loss is obtained by adding noise to the supervised measurement of loss function. The loss functions of this LSTM ladder network are two: unsupervised loss and supervised loss. The supervised cost \(C_S\) is the averaged cross-entropy between the classification result and the actual label. Note that the classification task is performed by the output of clean encoder at test time. As shown in formulation (3), the unsupervised cost is the averaged square error between the reconstruction output \(\hat{z}_i\) and the variable \(z_i\) of every layer.

\[
C = C_S + \sum_{i=1}^{L} \beta \lambda C_P = -\frac{1}{N} \log P(y^* = y_i | x_i) + \frac{1}{M} \sum_{i=N+1}^{N+M} \sum_{l=1}^{L} \lambda \| \hat{z}_i - z_i \|^2_2
\]

where \(x_i\) is the given input, \(y^*\) is the output of noisy encoder and \(y_i\) is the target label. The supervised cost \(C_S\) is the averaged cross entropy of \(y_i\) and \(y^*\).

3.3 Symmetrical LSTM (Sym-LSTM)

Inspired by the Mean Teacher, we designed a simple LSTM-based method for semi-supervised learning. As shown in Fig. 5, it consists of the following steps: (i) Take a supervised architecture and make a copy of it. Let us call the original model A and the new one B. (ii) At each training step, use the same minibatch as inputs to both A model and B model, but add random augmentation or noise to the inputs separately. (iii) Add an additional consistency cost between the A model and B model outputs (after softmax). (iv) Let the optimiser update the A model weights normally. (v) Let the B model weights be an exponential moving average (EMA) of the student weights. That is, after each training step, update the B model weights a little bit towards the A model weights. When compared with other methods using shared parameters between the two models, this one is more accurate and applicable to large datasets and human action recognition task.

In detail, we define the consistency cost \(J\) as the expected distance between the prediction of the A model and the expected prediction of the B model. The definition of \(J\) is

\[
J(\theta) = E_{x, y^*} \left[ \| f(x, \theta') - f(x, \theta, \eta') \| \right]
\]

where \(\theta\) and \(\eta\) are the weights and noise of model A, \(\theta'\) and \(\eta'\) are the weights and noise of model B, respectively. With regard to optimisation, we treat the B model parameters as constants.

The parameters are initialised randomly in the initial training stage of the network, so most of the classification of categories is incorrect. It is normal for the two models to output different labels. In the beginning, the value of \(a\) should not be too large, it should start from 0. The value of \(a\) begins to increase after the B network reaches a certain accuracy rate, \(\theta'\) finally increasing to 0.99. We define \(\theta\) at training step \(t\) as the EMA of successive weights:

\[
J_{\text{Eng}}, 2020, \text{Vol. } 2020 \text{ Iss. } 13, \text{pp. } 373-378
\]

This is an open access article published by the IET under the Creative Commons Attribution -NonCommercial License (http://creativecommons.org/licenses/by-nc/3.0/)

![Fig. 4 Structure of LSTM ladder](Image)
\[ \theta'_i = \alpha \theta'_{i-1} + (1 - \alpha) \theta_i \]  

(5)

where $\alpha$ is a smoothing coefficient hyper-parameter. In addition, for the same input samples, we use extended transform, add noise and so on to generate the data feed into network. The input of the two networks may be different due to the addition of noise, so the trained network not only has the ability of anti-noise, but also has higher accuracy.

4 Results

We validate our methods on two datasets, the NTU-RGB + D dataset and the Kinetics dataset. We compare our method to the traditional supervised model, LSTM, which is the baseline model of the architecture. Then, in order to understand the generality of our method, we do experiments with varying number of labelled and unlabelled data. The models we use here are: baseline LSTM network, LSTM ladder network, Symmetrical LSTM network, which are all performed on 4 NVIDIA 1080ti GPUs and 64 GB memory (see Figs. 6 and 7).

4.1 Datasets

The NTU-RGB + D dataset [24]. This dataset is the largest human action recognition dataset so far, which is established in 2016. All clips in this dataset are shot by Kinetic v2 and contain 60 action classes. Compared with other datasets, the actions in this dataset are more complex, which include the gesture action of single person (clapping, skipping), the interaction action between people and objects (brush teeth, put on a shoe), and the interaction action between people (hug). In order to test the experimental results, it is divided into two benchmarks: Cross-Subject dataset and Cross-View dataset.

4.2 Comparison

Kinetics-400 dataset [12]. All the videos are extracted from YouTube, with a total of 600 categories, each category contains at least 600 videos and each video lasts $\sim 10$ s. There are three main categories: interaction between people and things, such as playing musical instruments; interaction between people, such as shaking hands, hugs, sports and so on. In other words, it is person, person–person and person–object. This dataset has 400 categories, each action has 400–1150 video clips. The current version has 306,245 videos, divided into three parts. During training, each class using 250–1000 videos.

In order to extract skeleton sequence from the RGB video, we use the tool named OpenPose [3] which can get the 2D coordinates of 18 skeleton joint of human body. We compared the performance of three models on two datasets. The results are shown in Table 1. As for the supervised method, we use the baseline LSTM as model with 20% labelled samples, the rest are unlabelled samples. As for the semi-supervised method, we use LSTM ladder and Symmetrical LSTM to do experiments. Both on NTU-RGB + D dataset and Kinetics dataset, the LSTM ladder and Sym-LSTM show better performance than LSTM. In particular, LSTM ladder
achieves 15.15%, 6.75% improvements on NTU-RGB + D dataset and Kinetics dataset, separately. While Sym-LSTM achieves 19.64%, 8.21% improvements on the two datasets. Also the Sym-LSTM outperforms LSTM at any time. Those results suggest that LSTM ladder and Sym-LSTM can effectively process the unlabelled samples, which improves the accuracy rapidly.

4.3 Varying amount of labeled data

Here we do experiments about the performance of our model with varying number of labelled data. As shown in Figs. 8 and 9, we evaluate the accuracy of supervised LSTM, LSTM ladder and Symmetrical LSTM trained on 1%, 5%, 10%, 15%, 20% labelled samples. The rest examples are all unlabelled data. Fig. 8 shows the results on NTU-RGB + D dataset. We can see that the accuracy of all the three models improve when there are more labelled samples. With a certain number of labelled samples, LSTM ladder and Symmetrical LSTM show better performance than baseline LSTM. It is notable that when our models are learned from 20% labelled samples, its accuracy is pretty competitive with the supervised LSTM learned from the whole dataset, which proves the powerful capabilities of our model. Fig. 9 shows the results on Kinetics dataset, which has the similar trend with the results on NTU-RGB + D dataset.

4.4 Varying amount of unlabelled data

In this section, we do experiments about the performance of our model with varying number of unlabelled data. As shown in Figs. 10 and 11, we evaluate the accuracy of supervised LSTM, LSTM ladder and Symmetrical LSTM trained on 1%, 5%, 25%, 45%, 70% unlabelled examples. Fig. 10 shows the results on NTU-RGB + D dataset. It can be seen that with the increasing number of unlabelled samples, the accuracy of both LSTM ladder and Symmetrical LSTM gains a lot. The phenomenon suggests that better latent features in our models can be trained with more unlabelled examples. The accuracy improves due to the more unlabelled samples help adjust the latent LSTM features. In the meanwhile, Fig. 11 shows the results on Kinetics dataset.

5 Conclusion

In this paper, we design the LSTM ladder and Symmetrical LSTM for semi-supervised human action recognition. The experiment results performed on the two datasets: NTU-RGB + D dataset and Kinetics dataset, shows that our method can achieve significant improvements than traditional supervised models with less labelled data. Also it is helpful to use unlabelled data to better learn latent features of the models.

6 Acknowledgments

This work is supported by National Natural Science Foundation of China (NSFC, U1613209), Specialised Research Fund for Strategic and Prospective Industrial Development of Shenzhen City (No. ZLZBCXLIJZI20160729020003), Scientific Research Project of Shenzhen City (JCYJ20170306164738129), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (ZDSYS201703031405).

7 References

[1] Poppe, R.: ‘A survey on vision-based human action recognition’, Image Vis. Comput., 2010, 28 (6), pp. 976–990
[2] Tao, H., Zhu, W., Guo, X., et al.: ‘Human action recognition based on scene semantics’, Multimed. Tools Appl., 2018, 2, pp. 1–22
[3] Zhang, S., Guo, C., Zhang, J., et al.: ‘Discriminative part selection for human action recognition’, IEEE Trans. Multimed., 2018, 20 (4), pp. 769–780
[4] Zhang, Z., Ma, X., Song, R., et al.: ‘Deep learning based human action recognition: a survey’. Chinese Automation Congress, Jinan, People’s Republic of China, 2018
[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ‘Imagenet classification with deep convolutional neural networks’. Int. Conf. Neural Information Processing Systems, Lake Tahoe, NV, USA, 2012
[6] IEEE: Proceedings of 1993 IEEE International Conference on Neural Networks (ICNN ’93). IEEE Int. Conf. Neural Networks, New Orleans, LA, USA, 2002

Table 1 Comparison

|                  | Supervised LSTM (baseline) | Semi-supervised LSTM ladder | Sym-LSTM | Improvement LSTM-ladder | Sym-LSTM |
|------------------|----------------------------|----------------------------|----------|-------------------------|----------|
| NTU-RGB + D      | 40.13                      | 55.28                      | 59.77    | 15.15                   | 19.64    |
| kinetics         | 16.01                      | 22.76                      | 24.22    | 6.75                    | 8.21     |

Fig. 8 Varying labelled data on NTU-RGB + D dataset

Fig. 9 Varying labelled data on Kinetics dataset

Fig. 10 Varying unlabelled data on NTU-RGB + D dataset

Fig. 11 Varying unlabelled data on Kinetics dataset
7] Gerx, F.A., Schmidhuber, J., Cummins, F., et al.: ‘Learning to forget: continual prediction with LSTM’, Neural Comput., 2000, 12, (10), pp. 2451–2471
8] Chang, C.C., Lin, C.J.: ‘LIBSVM: a library for support vector machines’. ACM Transactions on Intelligent Systems and Technology, 2011, 27
9] Hsu, C.W., Lin, C.J.: ‘A comparison of methods for multiclass support vector machines’. IEEE Trans. Neural Netw., 2002, 13, (4), p. 1026
10] Das, S., Kopierski, M., Bremond, F., et al.: ‘A fusion of appearance based CNNs and temporal evolution of skeleton with LSTM for daily living action recognition’. 2018, arXiv:1802.00421 [cs.CV]
11] Wang, H., Liang, W.: ‘Learning content and style: joint action recognition and person identification from human skeletons’. Pattern Recognit., 2018, 81, pp. 23–85, 50031320318301195
12] Yan, S., Xiong, Y., Lin, D.: ‘Spatial temporal graph convolutional networks for skeleton-based action recognition’. AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2018
13] Song, S., Lan, C., Xing, J., et al.: ‘An end-to-end spatio-temporal attention model for human action recognition from skeleton data’. AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 2016
14] Chapelle, O., Schlkopf, B.S., Zien, A.: ‘Semi-supervised learning’ in Chapelle, O., Schlkopf, B., Zien, A. (Eds.): ‘Handbook on neural information processing’ (The MIT Press, USA, 2013), pp. 13–16
15] Zhou, Z.H., Li, M.: ‘Semi-supervised regression with co-training’. Int. Joint Conf. Artificial Intelligence, San Francisco, CA, USA, 2005