Research and Discussions on the Method of Video Tracking on In-depth Learning

Xiaomin Gao [1 a]

[1.]Zhengzhou University of Industry Technology, Department of Information Engineering, Zhengzhou, China

[a.]Houyanyang521@sina.com

Keywords: In-depth learning; Video tracking; Convolution neural network; Recurrent neural network; Self-encoder

Abstract. In recent years, in-depth learning has developed rapidly, which overturns the idea of algorithm design in the fields of speech recognition, image classification and text comprehension. Because of its strong feature extraction ability, in-depth learning is especially outstanding in the field of image recognition. However, there is not much combination between in-depth learning and video monitoring. Because the depth model has multi-layer network structure and the complexity of algorithm, it is time-consuming to train and update the model, it is difficult to meet the real-time requirement. This paper reviews the development history of in-depth learning, introduces the main models of in-depth learning at home and abroad in recent 10 years, discusses the target tracking algorithm based on in-depth learning, points out the advantages and disadvantages of each algorithm, and finally summarizes the existing problems and prospects in this field.

Introduction

In October 2015, AlphaGo, a computer program based on in-depth learning technology, beat the European Go-game champion Fan Hui in five consecutive games. In March 2016, AlphaGo beat the fourth-ranked Korean professional player Lee Shek at 4-1, and beat the top Chinese professional player Ke Jie at 3-0 in May 2005.Thanks to the powerful feature extraction ability of in-depth learning and the great success in computer vision[1-2], speech recognition[3] and big data[4], people are gradually turning their attention to video tracking. The combination of in-depth learning and video tracking has become a hot topic in the field.

Overview of In-depth Learning

In 1986, the scientists led by Rumelhart and McClelland proposed a multi-layer feed forward neural network, also called BP neural network[5], which is the most widely used neural model at present. In-depth learning is a branch of artificial neural network, whose concept is also derived from the study of artificial neural network.

In 2006, by the University of Toronto's G. E. Hinton et al.[6] proposed the in-depth learning theory for the first time, that is, the machine learning process based on sample data to obtain an in-depth network structure model with multi-levels through a specific training method. It is easy to converge to the local minimum due to the random initialization of network weight parameters by traditional neural network. In order to solve this problem, Hinton proposes a method of layer-by-layer pre-training with unsupervised restricted boltzmannmachine (RBM) algorithm to realize the efficient training of multi-layer neural networks. Since then, it has begun the prologue of in-depth learning.

In 2011 year, Microsoft applied in-depth learning to speech recognition for the first time, reducing recognition error rate by nearly 30%[7] compared with previous algorithms, making it the biggest breakthrough in speech recognition in more than a decade. In March of 2014, Facebook's DeepFace project brought the accuracy of face recognition technology to 97.25%, nearly comparable to that of humans. In 2015, Geoffrey Hinton, Joshua Bengio and Yann LeCun[8] co-authored an article for the first time in "Nature" to mark the 60th anniversary of artificial

  DOI: 10.25236/icmcs.2019.045

intelligence, giving a brief introduction to the basic principles and core strengths of in-depth learning. In 2016, W Liu and others proposed SSD algorithm. Compared with other object detection methods, SSD advantage lies in higher precision, faster. In October of 2017, the Google DeepMind team launched AlphaGo Zero[9], a self-reinforcing learning algorithm that is not based on human experience. After three days of training, it beat the AlphaGo version of Li Shishi at 100-0. So far, the development of in-depth learning has become more and more fierce. Because of its great achievements in transportation, medical and military fields, it has been widely used in science and industry.

The Main Models of In-Depth Learning

In recent years, with the rise of in-depth learning, more and more models have been proposed. This paper briefly introduces the following four widely used mainstream models, namely convolution neural network, self-encoder, recurrent neural network and generation antagonistic network.

Convolutional Neural Network.

Convolutional neural network (CNN) is a kind of feed forward neural network which is popular in recent years and has high recognition ability. It is composed of convolution layer, pool layer and full connection layer. Convolution operation, sparse connection and weight sharing are three prominent features of convolution neural network.

In 2012 year, Krizhevsky et al. proposed a CNN-based AlexNet model, which is essentially a multi-layer artificial neural network. The model has won the title in the Image Net large-scale image recognition competition with great advantages, and its excellent classification performance has attracted wide attention from the scientific community. Some of the better convolutional neural network models, such as Oxford University's VGGNet[10], Google's Google LeNet[11] and Microsoft's ResNet[12], were proposed after AlexNe.

Recurrent Neural Network.

Recurrent neural network (RNN)has a fixed weight and internal state, usually used to describe the dynamic time behavior sequence, is a kind of neural network which can process information of any length sequence. Because RNN is susceptible to gradient explosion and gradient disappearance, Schmidhuber et al.[13] proposed the long short term memory (LSTM) model in 1997, in which "forget valve" and "renewal valve" were added outside the RNN structure. Experiments show that the model can effectively deal with the problems caused by gradient disappearance or gradient explosion. Subsequently, a number of popular LSTM variants have been proposed. In 2014 year, Cho et al. proposed the gate cycle unit (GRU), which combines the forget gate and the input gate into a "renewal gate", and combines the neuron state and the hidden layer state. The model is simpler than the standard LSTM model.

Self Encoder.

Self-encoder is a kind of neural network based on unsupervised learning, which is composed of encoder and decoder. It is usually used for feature learning or data dimension reduction. The encoder encodes the input data into a potential variable, and the decoder reconstructs the potential variable into the original data. Since self-encoder can reduce dimension of data and filter redundant information effectively, it is widely used in image recognition and pedestrian detection.

Target Tracking Based on In-depth Learning

In recent years, due to its excellent feature modeling ability, in-depth learning has achieved good results in video tracking field. From the first depth network to extract the adaptive features of objects, then fusion of other tracking methods to achieve the target tracking, and now has been able to train the end-to-end depth neural network model to directly predict the target location. The researchers are not satisfied with the first in-depth learning method to extract adaptive features and apply them to the problem of target tracking. Nowadays, in-depth learning technology has obvious improvement in the expression of target features, the accuracy of predicting target position and the speed of image frame processing. More and more in-depth neural network models, such as

Recurrent Neural Network (RNN), Automatic Encoding Machine (ADE), Convolutional Neural Network (CNN), have been applied to the field of target tracking, and achieved good results. Of course, the combination of in-depth learning and target tracking is not long, and there are still many problems to be solved, such as: (1) Video target tracking is strict in real-time, and the large-scale in-depth neural network is difficult to meet the real-time requirements. This requires a comprehensive consideration of network structure and running speed. (2) In video target tracking, strictly speaking, only the first frame of data is the real annotation data. In the subsequent online tracking process, the number of positive and negative samples is only a few hundred. Therefore, video target tracking is a typical small sample on-line learning problem, which makes it difficult to develop the advantage of in-depth learning technology in processing big data. In spite of the difficulties mentioned above, because of the strong advantages of in-depth learning in apparent modeling and feature extraction, researchers still propose some target tracking algorithms based on depth learning in different ways, combining with the characteristics of target tracking tasks. This paper mainly introduces two target tracking methods based on self-encoder and recurrent neural network.

Target Tracking Based on Self-Encoder.

Since the self-encoder can be effectively compressed, many video tracking models use it for video drop-down and generation. In multi-class self-encoder, the de-noising stack self-encoder is first applied to the target tracking field of online video without specific target because of its excellent characteristic learning ability and anti-noise performance.

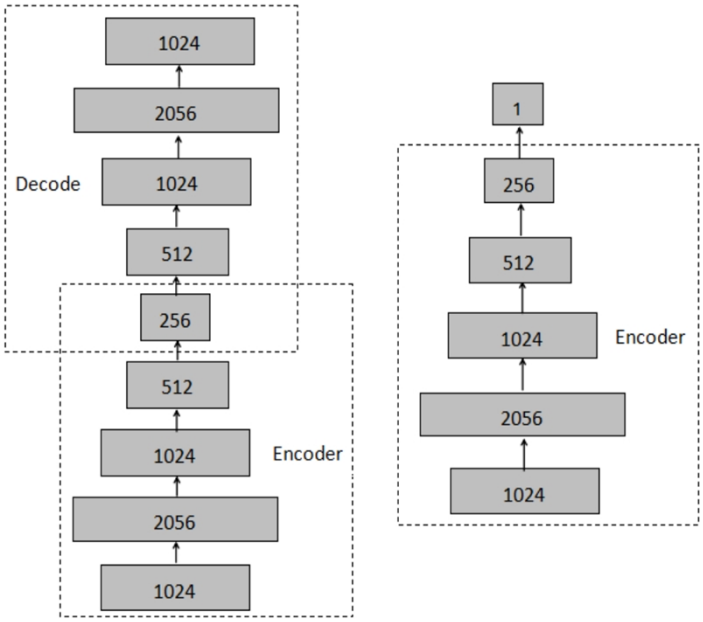

Figure. 1 The de-noising self-encoder architecture for target tracking

Wang et al.[14] first trained a stack denoising self-coding machine offline in a large scale small-scale image sample data set[15].The structure of its depth network model is shown in Figure 13 on the left. The trained multi-layer network is then used to extract the apparent feature of the target when tracking. To make use of on-line tagged information, add a logical regression binary classifier at the top of the network. The "1" represents the target and the "0" represents the background, as shown in the right figure in Figure 1. When initializing, the network parameters are fine-tuned by using the given annotation information of the first frame. When tracking the target online, we continue to fine-tune the depth network model through real-time sampling, so as to adapt to the apparent change of the target. The tracking model is based on the particle filter framework. In order to reduce the computation, the system update is not about every frame, but only once at a certain number of frames or the system confidence is less than the set threshold. The experimental

results show that the tracking effect is better than that of some methods based on traditional feature expression. It is a typical method of in-depth learning and tracking in "Off-line learning + on-line fine tuning" mode, and its architecture is exemplary. The network structure is relatively simple and easy to train. But there are some shortcomings.

Mainly lies in:

(1) The image data used in off-line pre-training are 32 * 32 images with low resolution. The network can access to some general pictures image features, but for tracking tasks, focus on the effective expression of target features rather than on the entire image. Therefore, there is no theoretical guarantee that enough strong representation can be learned to maximize the distinction between goals and backgrounds.

(2) The back end of the network is a binary classifier, i. e., only the tracking is considered as a binary classification problem. When marking on-line samples, samples close to the current target position are selected as positive samples and those far away are selected as negative samples. Since both "near" and "far" need to set specific thresholds, it is easy to introduce error samples to cause template drift.

Target Tracking Based on Recurrent Neural Network.

In recent years, the recurrent neural network (RNN), especially the long-time memory network (LSTM) with gate structure, has shown outstanding performance in temporal tasks. Many scholars began to explore how to use recurrent neural network to deal with the problems existing in existing tracking tasks.

At present, most trackers based on convolutional neural network regard tracking as a classification problem. Because their models focus on inter-class classification, they are sensitive to target similarity and tend to track drift. To solve this problem, Fan et al.[16] designed SANet, using the structure information of the target itself to distinguish from the interference. Specifically, the recurrent neural network (RNN) is used to model the target structure and combine it with convolutional neural network to improve its robustness in the face of similar disturbance. Considering that convolution layers at different levels represent objects from different angles, the document[51] uses multiple recurrent neural networks to model target structures at different levels and provides a jump connection strategy to fuse the features of convolution neural networks and recurrent neural networks. Thus, it can provide more rich information for the next layer, and thus improving the tracking performance.

Summary and Prospect

In reality, video target tracking is a complex and difficult research topic, because there are too many factors that can interfere with the tracking process. After decades of efforts, the tracking model can handle some simple scenarios well. However, the tracking effect is still not ideal in complex environments. The emergence of the theory of in-depth learning provides the possibility for the establishment of a more robust target apparent model. However, the combination of in-depth learning and target tracking is still short, and a lot of research work needs to be done. At present, the research focus and development trend are mainly focused on the following points: (1) the integration of in-depth learning and online learning. (2) To establish an in-depth network suitable for target tracking. (3) Tracking the creation of the data platform. In this paper, the development history of in-depth learning is introduced. And then the main models of in-depth learning are described. The research on target tracking algorithm based on depth learning in the last 10 years is reviewed. And the advantages and disadvantages of each algorithm are summarized. Then, the characteristics, problems and difficulties of in-depth learning method applied in target tracking are analyzed and summarized. It is believed that in the near future, there will be a better in-depth learning model.

References

[1] Kumar S, Singh S K, Singh R, et al. Deep Learning Framework for Recognition of Cattle Using Muzzle Point Image Pattern[J]. Measurement, 2017, 116.

[2] Y.W. Hou, H.L. Zhao. Handwritten Digit Recognition Based on Depth Neural Network[C]. ICIIBMS 2017 - 2nd International Conference on Intelligent Informatics and Biomedical Sciences , Nov.24-26, 2017, Okinawa, Japan. P35-38.

[3] Tian C, Liu J, Peng Z M. Acceleration Strategies for Speech Recognition Based on Deep Neural Networks[J]. Applied Mechanics & Materials, 2014, 556-562：5181-5185.

[4] Q. Zhang, L.T. Yang and Z. Chen, et al. A survey on deep learning for big data[J]. Information Fusion, 2018, 42:146-157.

[5] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors.[J]. Nature,1986,323(6088):399-421.

[6] Hinton G E, Osindero S, The Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7):1527-1554.

[7] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8):1798-1828.

[8] Lecun Y,Bengio Y, Hinton G. Deep learning[J].Nature,2015:436-444.

[9] T.T. Zhen, K. Shao and B.Z. Dong, et al. Recent progress of deep reinforcement learning:from AlphaGo to AlphaGo Zero[J]. Control Theory & Applications, 2017.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014.

[11] C. Szegedy, W. Liu, Y. Jia, et al.. Going deeper with convolutions[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015:1-8.

[12] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 770-778.

[13] Hochreiter S, Schmidhuber J. Long Short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[14] Wang N, Yeung D Y. Learning a deep compact image representation for visual tracking[C].International Conference on Neural Information Processing Systems. Curran Associates Inc. 2013:809-817.

[15] Torralba A, Fergus R, Freeman W T, et al. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1958-1970.

[16] Fan H, Ling H. SANet: Structure-Aware Network for Visual Tracking[C]. Computer Vision and Pattern Recognition Workshops. IEEE, 2017:2217-2224.