Journal of Artificial Intelligence Practice (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2023.060105
ISSN 2371-8412 Vol. 6 Num. 1

A survey of Few-Shot Action Recognition

Congmin Wang[1], Yancong Zhou[2,*]

[1]School of Science, Tianjin University of Commerce, Tianjin, China
[2]School of Information Engineering, Tianjin University of Commerce, Tianjin, China
**Corresponding author*

Keywords: Few-Shot Learning, Action Recognition, Deep Learning

Abstract: In recent years, with the development of network technology, countless videos are produced every day. Many achievements have also been made in the field of action recognition in computer vision. Training action recognition models requires a large number of labeled samples, but in reality, the amount of data is scarce, and it is extremely difficult to obtain a large amount of data due to costs and other reasons. The few-shot learning aims to solve the problem of using several samples to learn new categories. This paper combs the relevant research in recent years of few-shot action recognition technology. According to the classification of training process, this paper summarizes the research progress and typical models of few-shot action recognition from the perspectives of data processing, feature embedding, feature augmentation, and metric learning; finally points out the challenges faced by current research and the future development directions.

1. Introduction

The training of efficient action recognition models often requires large-scale video data [1-4]. Due to the high cost and privacy problems in reality, many scenes cannot obtain a large amount of training data. In addition, some behaviors do not have a large amount of real data, which hinders the development of traditional action recognition models. In view of the achievements of few-shot learning in the field of text and image [5-7], it is introduced into action recognition. Few-shot learning refers to the relevant recognition learning with few training samples. That is, use one or several samples for learning. This method stems from the imitation of human learning ability: humans can learn to recognize a new class through a few samples or features. This is due to the accumulation of human prior knowledge. Through continuous learning, we have the ability to learn. When faced with new learning tasks, we can learn quickly. This is the goal of current deep learning, learning to learn.

At present, there are some research reviews related to few-shot learning. Yaqing Wang and others pointed out that the core problem of FSL is that the empirical risk minimization is unreliable [8]. Based on how to use prior knowledge to deal with this problem, they classified FSL methods from three perspectives: 1) Data, using prior knowledge to enhance supervision experience; 2) Algorithm, useing prior knowledge to change the search for the best hypothesis in the given hypothesis space; 3) Model, using prior knowledge to reduce the size of hypothesis space. Shengbiao An summarized the few-shot image classification methods under the three paradigms of supervised, semi-supervised and unsupervised, and summarized them from the perspectives of

metric learning, meta-learning, pseudo-labeling, and comparative learning according to different methods in various situations, they compared and analyzed the performance of these methods[9]. Different from the above methods, this paper summarizes the few-shot action recognition methods according to the training process, and introduces them from the perspectives of data processing, feature embedding, feature enhancement, measurement learning. The second section of this paper introduces the definition of few-shot action recognition symbols and related datasets, the third section introduces the relevant research of few-shot action recognition in detail, and the fourth section discusses the current challenges and future directions of few-shot action recognition.

2. Overview

2.1. Problem Definition

In few-shot action recognition, the video to be classified is called query video. One task of few-shot learning is to divide a query video without tags into one of several classes. Each class is composed of a small number of tagged samples that have not been seen in training, called support set. Each task is regarded as an N-way K-shot classification problem. N-way K-shot means that the support set contains N categories, and each category contains K tagged support set samples. The data of each task is composed of support set S and query set Q. In most studies, Q={q_1， ... , q_F} is a frame sample set uniformly sampled by F=8. S={S_c}, c=1..., N, is a support set containing N classes. The k video of class c is represented as Sc={ S^c_{k1},..., S^c_{kF} }. For the convenience of comparison, most studies have set N to 5 and K to 1 or 5.

2.2. Datasets

In this section, the common datasets for few-shot action recognition are introduced. The statistical data and common experimental settings of the datasets are listed below.

HMDB51 [10] contain 6849 videos, which are divided into 51 action categories, and each category contains at least 101 samples. Actions are mainly divided into five categories: general facial actions, general body actions, interaction with objects, and human actions. The recent research follows the experimental setup of [11], taking 31,10, and 10 categories, namely 4280,1194, and 1292 videos as training set, test set and verification set.

UCF101 [12] is a real ground video action recognition data set. It provides great diversity in motion, and has great changes in camera motion, object appearance and posture, object scale, viewpoint, messy background, lighting conditions, etc. It contains 101 action categories and 13320 video clips. It can be divided into five categories: 1) human-object interaction, 2) simple body movements, 3) human-human interaction, 4) playing musical instruments, 5) sports. The recent research follows the experimental setup of [11]. The data set segmentation is divided into 70, 10 and 21 categories, namely 9154, 1421 and 2745 videos, which are used as train set, verification sets and test sets respectively.

SSV2 (Something-Something V2) [13] is a large collection of tagged video clips, which show that humans use everyday objects to perform predefined basic actions. It allows machine learning models to have a fine-grained understanding of the basic behaviors that occur in the physical world. The data set was created through large-scale crowd sourcing. It contains 220847 short video clips spanning 174 categories, each video clip lasts for 2-6s, which are marked as simple text descriptions. Recent studies often follow the experimental setup of [14,15], of which 168913 are in the training set, 24777 are in the verification set, and 27157 are in the test set.

Kinetics [16] is divided into Kinetics 100/400/600/700 according to the number of video

categories. Kinetics 100 and Kinetics 400 are more commonly used. Videos length is about 10s. The video is obtained from YouTube. These movements are human-centered and cover a wide range of categories, including human interactions with objects such as playing instruments and human-to-human interactions such as hand-shaking hands.

3. Methods

For the purpose of the action identification task of few-shot, the following four categories of existing action recognition methods based on few-shot are summarized according to the task process: 1) data processing; 2) feature embedding; 3) feature augmentation; 4) metric learning.

3.1. Data processing

Data processing-based methods include data augmentation and data feature augmentation.

Data augmentation is to enhance the diversity of the existing data. Improve the generalization ability of the model and reduce the overfitting risk [17]. Data augmentation is a method to generate new samples by performing basic image operations based on existing data. Common operations include cropping, flipping, rotating, random occlusion, and random color jitter. The advantage of this method is easier to implement and simple to operate, however it is limited as effective for improving the performance of the model, because no new semantic information is generated, and therefore it is usually used as an auxiliary technique for data preprocessing.

Data feature augmentation is different from data augmentation. The focus is not to change the data itself, but to enhance the data features after the input model, focusing on the extraction of semantic information. By learning the importance of video information, this method focuses on the information that affects the classification and reduces the attention of redundant information, which improves the efficiency of model recognition.

Specifically, G. Huang divided quiet and busy information by calculating the difference before and after each frame as a motion representation, and assigns different calculations to different information so as to reduce redundancy and improve computational efficiency [18]. AMeFu-Net introduces depth information to fuse the representation of the original RGB clip with multiple non-strictly corresponding depth clips sampled by the temporal asynchronous synchronous enhancement mechanism to synthesize new instances at the feature level [19]. OTAM proposed an ordered time alignment module that learns the depth distance measurement of a new class of agents of the query video on its alignment path [20]. The temporal ordering information in the video data is explicitly utilized by the ordered temporal alignment. This greatly improves the data efficiency of small-sample learning. TSN framework uses a sparse sampling scheme to extract short segments on long video sequences, where the samples are uniformly distributed along the temporal dimension. Therefore, a segmented structure is used to aggregate the information from the sampled fragments. That is, the timeline network is able to model the remote temporal structure over the entire video. Furthermore, this sparse sampling strategy retains relevant information at a significantly lower cost, thus enabling end-to-end learning of long video sequences with reasonable time and computational resource budgets [21]. TARN is a temporal attention relationship network. At the core of the network is a meta-learning method, learning to compare the variable length of time, that is, two different lengths of video or video and semantic representations such as word vector (in the case of zero action recognition) [22]. Specifically through the following implementation：a) using the attention mechanism to perform the temporal alignment, b) the deep distance measurements of learning the alignment representation at the video segment level.

3.2. Feature Embedding

Embedding refers to the data coding for characteristic part of the network, most models adopt the classic pre-training deep neural network as a feature embedded network, such as resnet50 [23], VIT [24], using the same embedded network is beneficial to the fairness of the model performance comparison, in addition, the use of mature pre-training classical network can reduce the training time, make the model focus on the performance of other modules. But the use of fixed embedded networks ignores the innovation of this part. Therefore, some studies have improved the feature embedding section. The FEAT [25] introduces an adaptive step to learn the task-specific embedding. We propose an embedding adaptive method based on few-shot models to adjust the instance embedding model from visible classes. This model-based embedding adaptation requires a set-to-set function mapping: obtains all instances from few-shot support sets, and outputs adaptive support instance embedding sets, and the elements in the set adapt to each other. These outputs were then embedded and assembled as prototypes for each visual category and used as a nearest neighbor classifier.

3.3. Feature Augmentation

Feature augmentation refers to the information enhancement between feature embedding and metric learning, common such as statistical method clustering: Haddad M obtain a precise representation of human action by clustering the results given by GF-OF using the vector-quantified K-means method. Then, a Gaussian mixture model continues to represent the result of one instance of each action. In addition, the Kullback-Leibler divergence (kl-divergence) was used to find the similarity between the trained movements and the action in the test video [26]. Deep learning methods such as spatio-temporal information enhancement: ARN establishes a C3D encoder that captures short-range action patterns through spatio-temporal video blocks that are aggregated by permuted invariant pools, subsequently, representations are combined into simple relational descriptors, the metric network compares relational descriptors to output the similarity of videos [27]. TRX references CrossTransformers [29] to action recognition to build the class prototype, which looks at all subsequences of supporting videos to thereby enhance temporal information, rather than using a class average or single best match [28]. This allows video comparison of action subsequences at different speed and time offsets. STRM introduces spatio-temporal enrichment modules based on TRX, which aggregate spatio-temporal context to enrich sub-modules through dedicated local patch-level and global frame-level features [30]. A query class similarity classifier is further introduced on patch-level enriched features to enhance class-specific feature recognition by strengthening feature learning at different stages of this framework. That is to say, on the basis of rich time information, the spatial information is enriched.

3.4. Metric Learning

Metric learning is one of the most common and effective methods to solve few-shot classification. Metric learning also refers to similarity learning. Metric learning can be interpreted as a method of spatial mapping, able to learn some kind of feature space. In the few-shot classification, it is understood to translate the data into feature vectors [9]. It measures the similarity or distance of two target features or multiple in the embedded space. The same class in the feature space are closer, while different classes are far apart. Feature extraction is achieved through convolution neural network and recurrent neural network. The metric classifier can use a fixed metric of Euclidean distance, Mahalanobis distance and cosine distance based on Bregman divergence or a learnable metric [31] based on deep neural networks. To design a feature extractor

with strong expression ability and match the extracted features with the requirements of the classifier, is very important to improve the classification performance of the network.

Current classical measures such as 1) The prototype network [32] aims to classify by learning a new metric space. In the new metric space, the mean of the embedded eigenvector of each class is used as the Euclidean distance between the embedded eigenvector and the idea of nearest neighbor is used to divide the query sample into the correct category. 2) The relational network [33], different from the traditional measurement learning method, introduces a neural network to learn the distance function in measuring the similarity between samples for the first time. Relationship network consists of two parts: embedded module and relationship module. The model first obtains the embedding space eigenvector of the support set label sample and the query set sample, and then splicing the two feature vectors. Finally, the metric learning network in the relational module is used to compare the similarity between samples and give the relational score. Matching network [34] is a novel neural network structure that combines metric learning and memory enhancement neural networks. It provides a network framework that can map a small number of data sets and unlabeled instances to their own labels, and avoid adapting to new classes by fine-tuning the trained model [35]. There is also the class prototype-centered metric PAL, the lack of data efficiency for query-centered loss design, and the inability to handle the problem of outlier samples and inter class distribution overlap in support sets [36]. Both limitations are overcome by proposing a new prototype-centered attentive learning (PAL) model consisting of two new components. It combines the prototype-centered loss function to improve data efficiency.

4. Challenges and Future Directions

Despite the development of few-shot action recognition rapidly, there is still a lot of space for progress and many challenges.1) data set. Most of the existing few-shot action recognition models use pre-trained embedding models for feature extraction, or they are now pre-trained on large-scale datasets. The commonly used embedding module resnet50 all uses ImageNet [37] as a pre-training dataset. The action recognition data set used for the formal training of the model only has HMDB51, UCF101, SSV2, kinetics and other datasets, while some data such as kinetics, the video has been damaged or lost or deleted.2) Few-shot learning hypothesis challenge. The first is the separation class hypothesis, which is difficult to maintain because the video clips of the target class may appear in the video of the source class. Secondly, the current few-shot action recognition methods require pre-training on large-scale non-target datasets, that is to say, the model still needs a large amount of annotated data for training, which violates the definition of few-shot learning.

Through the summary of current action recognition studies on few-shot, the future direction is 1) the construction of relevant data sets. The action subjects of the existing data sets are almost all young people, the scenes are mostly common indoor and outdoor movements. According to the situation of few existing data sets and single scenes, more action video data for special groups under different application scenarios can be constructed in the future, such as the action scenes for the elderly: the nursing home, the scene for family living alone; action scenes for children: kindergarten and amusement park; action scenes for the disabled, etc. In addition, in order to ensure the stability of the data, try to provide the existing video format data as far as possible, instead of video links like kinetics. 2) Solve the problem of few-shot learning assumptions. The first is to solve the problem of separation class hypothesis. We need to maintain accurate annotation when labeling the data, and pay attention to the similarities and differences between the action class in the source domain target domain, to avoid overlapping or high similarity between the training set and the test set and the validation set. Secondly, the model does not rely on pre-training model or large-scale pre-training data, we could use other prior knowledge for recognition.

5. Conclusions

The goal of few-shot learning (FSL) is to narrow the gap between AI and human learning. Combined with prior knowledge, we learn new action categories that contain only a few video samples. Few-shot action recognition will help reduce the burden of collecting large-scale supervised samples in the application. This paper makes a systematic introduction to the study of few-shot action recognition, first introduces the relevant research background, problem definition, and common data sets, and then classifies the training process, namely data, embedding, feature enhancement, and measurement; finally, summarizes the challenges and future direction of few-shot action recognition. At present, there is still much room for progress in few-shot action recognition, which still needs continuous exploration by scholars.

References

[1] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, 28.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advancesin neural information processing systems, 2012, 25.

[3] Yan L, Zheng Y, Cao J. Few-shot learning for short text classification. Multimedia Tools and Applications, 2018, 77(22): 29799-29810.

[4] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press, 2016.

[5] Qin T, Li W, Shi Y, et al. Diversity helps: Unsupervised few-shot learning via distribution shift-based data augmentation. arXiv:2004.05805, 2020.

[6] Xu H, Wang J, Li H, et al. Unsupervised meta-learning for few-shot learning. Pattern Recognition, 2021, 116: 107951.

[7] Zhang H, Zhan T, Davidson I. A self-supervised deep learning framework for unsupervised few-shot learning and clustering. Pattern Recognition Letters, 2021, 148: 75-81.

[8] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a Few Examples: A Survey on Few-shot Learning. ACM Computing Surveys, 2020, 53(3):1-34.

[9] An Shengbiao, Guo Yuqi, Bai Yu, Wang Tengbo. Summary of image classification studies in small samples. And Computer Science and Exploration. 2022. 1-22.

[10] H Kuehne, T Serre, H Jhuang, E Garrote, T Poggio, and T Serre. HMDB: A large video database for human motion recognition. In International Conference on Computer Vision, nov 2011. 2, 4, 10.

[11] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H S Torr, and Piotr Koniusz. Few-shot Action Recognition with Permutation-invariant Attention. In European Conference on Computer Vision, 2020. 1, 2, 3, 4, 5, 10,11

[12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah.UCF101: A Dataset of 101 Human Actions Classes from Videos in The Wild. arXiv, 2012. 2, 4, 10.

[13] Raghav Goyal, Vincent Michalski, Joanna Materzy, Susanne Westphal, Heuna Kim, Valentin Haenel, Peter Yianilos, Moritz Mueller-freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In International Conference on Computer Vision, 2017. 1, 2, 4, 10.

[14] Linchao Zhu and Yi Yang. Compound Memory Networks for Few-Shot Video Classification. In European Conference on Computer Vision, 2018. 1, 2, 4, 5, 10, 11.

[15] Linchao Zhu and Yi Yang. Label Independent Memory for Semi-Supervised Few-shot Video Classification. Transactions on Pattern Analysis and Machine Intelligence, 14(8), 2020. 1, 2, 4, 5, 7, 8, 10, 11.

[16] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Computer Vision and Pattern Recognition, 2017. 1, 2, 4, 10.

[17] Lawrence S, Giles C L, Tsoi A C. Lessons in Neural Network Training: Overfitting May be Harder than Expected. Fourteenth National Conference on Artificial Intelligence & Ninth Innovative Applications of Artificial Intelligence Conference. AAAI Press, 1997.

[18] G. Huang, A. G. Bors. Busy-Quiet Video Disentangling for Video Classification.2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, 756-765.

[19] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, Yu-Gang Jiang. Depth Guided Adaptive Meta-Fusion Network for Few-shot Video Recognition. accepted by ACM Multimedia 2020.

[20] K. Cao, J. Ji, Z. Cao, C. -Y. Chang and J. C. Niebles. Few-Shot Video Classification via Temporal Alignment.2020

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, 10615-10624.
[21] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision, pages 20–36. Springer, 2016. 1, 2, 4, 6.
[22] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition. In British Machine Vision Conference, 2019. 1, 2, 4, 5, 8, 10, 11.
[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In Computer Vision and Pattern Recognition, 2016. 5.
[24] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. 2021.
[25] H.-J. Ye, H. Hu, D.-C. Zhan and F. Sha. Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions.2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, 8805-8814.
[26] Haddad, M., Ghassab, V.K., Najar, F. et al. A statistical framework for few-shot action recognition. Multimed Tools Appl 80, 2021，24303–24318.
[27] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H S Torr, and Piotr Koniusz. Few-shot Action Recognition with Permutation-invariant Attention. In European Conference on Computer Vision, 2020. 1, 2, 3, 4, 5, 10, 11
[28] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In CVPR, 2021. 1, 2, 3, 6, 7, 8, 12.
[29] Carl Doersch, Ankush Gupta, and Andrew Zisserman.CrossTransformers: Spatially-Aware Few-Shot Transfer. In Advances in Nerual Information Processing Systems, 2020.1, 2, 3, 5.
[30] Thatipelli, A., Narayan, S., Khan, S.H., Anwer, R.M., Khan, F.S.& Ghanem,B. Spatio-temporal Relation Modeling for Few-shot Action Recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 19926-19935.
[31] Xing E, Jordan M, Russell S J, et al. Distance metric learning with application to clustering with side-information. Advances in neural information processing systems,2002, 15.
[32] SNELLJ, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning. In Advances in Neural Information Processing Systems, Long Beach: MIT Press, 2017.4077-4087.
[33] SUNG F, YANG YX, ZHANG L, XIANG T, TORR P H, Hospedales T M. Learning to compare: relation network for few-shot learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, USA: IEEE, 2018. 1199-1208.
[34] VINYALS O, BLUNDELL C, LILLICRAP T, KORAY K. Matching networks for one shot learning. Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: MITPress, 2016. 3630-3638.
[35] Wei Shihong, Liu Hongmei, Tang Hong, Zhu Longjiao. Small-sample learning of multilevel metric networks. Computer Engineering and Application, 2023,59 (02): 94-101.
[36] Zhu X,Toisoul A,Prez-Ra J M,et al. Few-shot Action Recognition with Prototype-centered Attentive Learning. 2021.
[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In Computer Vision and Pattern Recognition, 2009. 5.