Few–Shot Action Recognition

1Valeti Likhitha Chowdary, 2Nidadavolu Bhavya Sree Ratna, 3Ankitha Lunavath, 4Rajesh Kandakatla
1Student, 2Student, 3Student, 4Assistant Professor
BVRIT HYDERABAD College of Engineering for Women

Abstract - **The goal of few-shot action identification is to detect new action classes using only a few sets of training samples. The majority of available approaches use a meta-learning approach combined with episodic training. The few samples in a meta-training job are divided into support and query sets in each episode. The former is used to construct a classifier, which is subsequently assessed on the latter using a query-centered loss for updating the model. However, there are two key limitations, which are: a lack of data efficiency attributable to the query-centered only loss design and an inability to cope with outlying samples and inter-class distribution overlapping anomalies raised within the support set. We address these shortcomings in this study by introducing a new Prototype-centered Attentive Learning (PAL) model consisting of two revolutionary components. To make maximum use of the minimal training samples in each episode, a prototype-centered contrastive learning loss is introduced to enhance the traditional query-centered learning objective. Second, PAL incorporates a hybrid attentive learning method that can reduce the adverse effects of outliers while also promoting class separation.**

keywords - **FSL - Few-Shot Learning PAL - Prototype-centered Attentive Learning**

1. Introduction

In general, a large quantity of data is required for a system to classify a particular action. A large number of training samples may be required. For certain uncommon fine-grained action classes, collecting and categorizing such a vast amount of data is expensive, time-consuming, and even impractical. This is why few-shot action recognition has gained popularity since it seeks to build a video action classifier using only a few sets of training examples per class. The main objective of the "Few-shot action recognition" is to be able to identify a new action class within the limited number of the training dataset. The widespread use of smartphones has resulted in a substantial volume of films being posted on social media on social media sites, which are in desperate need of automated analysis. Video action identification has been extensively researched, with a new emphasis on fine-grained actions. Most modern action recognition algorithms use deep convolutional neural networks (CNNs), which are known to be data-hungry since they require a significant number of labeled examples to be trained for each action class.

2. Literature Survey

In this paper[1]**Few-shot action recognition with permutation-invariant attention**, To simulate short and long-range temporal patterns, this paper suggests a few-shot Action Recognition Network (ARN) that includes an encoder, comparator, and attention mechanism.Training and testing accuracy are 82% and 63%.In this paper[2]**Compound memory networks for few-shot video classification**,For few-shot video classification, the author proposes a compound memory network. This module keeps a record of matrix representations that may be quickly retrieved and modified. Training and testing accuracy are 78% and 60%.In this paper [3]**Few-shot video classification via temporal alignment**, the author proposed Ordered Temporal Alignment Module (OTAM), few-shot framework that can directly learn distance measure and representation in videos while ignoring non-linear temporal changes..Training and testing accuracy are 85.0% & 73.8%.In this paper [4] **Incremental few-shot object detection**, the proposed incremental few-shot object detection problem is a potential first step toward solving this challenge.Training and testing accuracy are 87.0% & 75.8%.In this paper [5]**Temporal-relational cross-transformers for few-shot action recognition**, In this by comparing the query to sub-sequences of all support set movies, TRX creates query-specific class prototypes. Training and testing accuracy are 75.0% & 60.8%.In this paper[6]**Optimization as a model for few-shot learning**, The parameter updates provided by gradient descent optimization algorithms prompted us to create an LSTM-based model for meta-learning. The learning updates of the parameters of a classifier are represented by the state of our LSTM meta-learner.Training and testing accuracy are 71% & 60.60%.In this paper[7]**Prototypical networks for few-shot learning**,says the challenge of few-shot classification, in which a classifier must generalize to new classes not observed in the training set with only a few samples of each new class..Training and testing accuracy are 68.20% & 66%.In this paper [8]**A closer look at few-shot classification**, the Baseline++ model is competitive to the state-of-the-art under standard conditions, and that the Baseline model achieves competitive performance with recent state-of-the-art meta-learning algorithms on both the CUB and mini-ImageNet benchmark datasets when using a deeper feature backbone.Training and testing accuracy are 76% & 62.04%.In this paper [9], **Two-stream convolutional networks for action recognition in videos** the video classification model based on ConvNets that has competitive performance and includes independent spatial and temporal recognition streams.Training and testing accuracy are 72.7% 59.9%.In this paper [10]**Few-shot image classification with differentiable earth mover's distance and structured classifiers**, They proposed cross-reference approach for determining node weights is critical in the EMD formulation and effectively reduces the detrimental impact of irrelevant regions. In the k-shot circumstances, the learnable structured fully connected layer can directly classify dense representations of images. Training and testing accuracy are 78.86% & 65%.

3. Proposed System

Existing few-shot action detection systems, however, have two major drawbacks. The first drawback is a lack of data efficiency as a result of the previous approach's query-centered learning purpose. As a result, this query-centered learning aim does not fully use the limited training data in each episode. The second drawback is their inability to solve two core FSL challenges: outlying samples and overlapping interclass distributions in the support set. Because of the considerable spatiotemporal variability in video data, these difficulties are generally more severe than in images.

To address the constraints, we apply the Prototype-centered Attentive Learning (PAL) approach. To make use of the most limited training samples in each episode, a prototype-centered contrastive learning loss is added first. Second, PAL includes a hybrid attentive learning approach that can help to mitigate the negative effects of outliers while simultaneously increasing class separation.A subset of few-shot learning is few-shot action recognition (FSL). Most FSL approaches adhere to a meta-learning paradigm typified by episodic training, which seeks to learn a model or optimizer from a collection of base/seen tasks in order to generalize well to new tasks with a few labeled training samples/shots. To sample a large number of training episodes, a meta-training set with numerous training samples per seen/base class is utilized. The training data is divided into two sets in each episode: support set with N classes and K samples per class to simulate the setup of target meta-test tasks, and a query set with the same N classes.

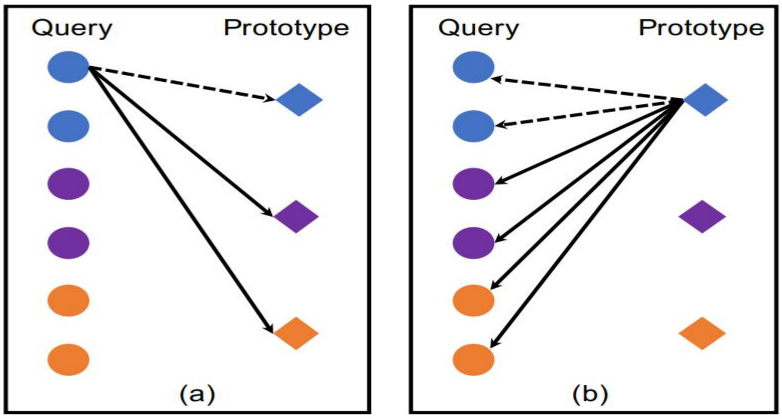


Figure 1
a)Query-centered Learning b) Prototype-Centered Learning

4. Implementation

4.1 Dataset

The Something-Something dataset (version 2) is a collection of 220,847 labeled video clips of humans performing pre-defined, basic actions with everyday objects. It is designed to train machine learning models in fine-grained understanding of human hand gestures like putting something into something, turning something upside down and covering something with something.

4.2 Data Preprocessing

Data analysis is very essential to get a deeper insight into the dataset. Considering that the data was collected from an online website the data cannot be in a proper state and can be biased. After a detailed analysis of the dataset, we have implemented some pre-processing steps. The first step includes the removal of noise using Gaussian blur smoothing. We applied Gaussian blur to enhance the image. We cropped out uninformative areas from the image and kept only the necessary parts. The second step consists of thresholding. It is the method of segmenting images. It separates objects from foreground to background pixels. The third step includes resizing the image. The fourth step is Data normalization which makes sure that each value has a similar distribution. In our project, the images are normalized by dividing each pixel value by 255.0. It replaces the range of pixel values from (0,255) to (0,1).

5. Algorithms

5.1 Implementation

As the backbone network, we employed an ImageNet pretrained ResNet-50. The original fully linked layer was replaced with a new cosine similarity-based classification layer. For the first training stage, we used SGD to optimize the TSN feature embedding, starting at 0.001 and declining by 0.1 every 30 epochs for a total of 70 epochs. We performed end-to-end meta-training of both the TSN feature backbone and our PAL model in the second stage. On Sth-Sth-100, we trained a total of 35 epochs with fading epochs at 15 and 30, each epoch containing 200 episodes, starting with a learning rate of 0.0001. We discovered that training 10 epochs was sufficient for the other datasets, with declining points at 5, 7, and 8.

A cosine-based classifier model was used. The feature and latent space dimensions were set at d = da = 2024. During training, we enlarged each video frame to 256 256 pixels, from which a random 224 224 region was selected.cropped to create the input Three coarse-grained grains During training, we applied a random horizontal flip to the datasets. Many classes in the fine-grained Sth-Sth-100 range, however,are direction-sensitive (for example, pulling something from the left to the right)horizontal (drawing anything from the right to the left), right .As a result, no flip was used. During the test, we solely used the center crop to predict test data and model performance.

5.2 Prototype-centered Attentive Learning

PAL was created specifically to address the problem of inter-class boundary ambiguity and outlier support-set samples that are inherent to new tasks.In a new 5-way 5-shot challenge, we showed the change in feature representations and per-class prototypes of support samples to better understand the internal process of our model. The feature distributions with and without our PAL model were compared to show how the feature space is enhanced. Figure 4 shows that PAL is effective.

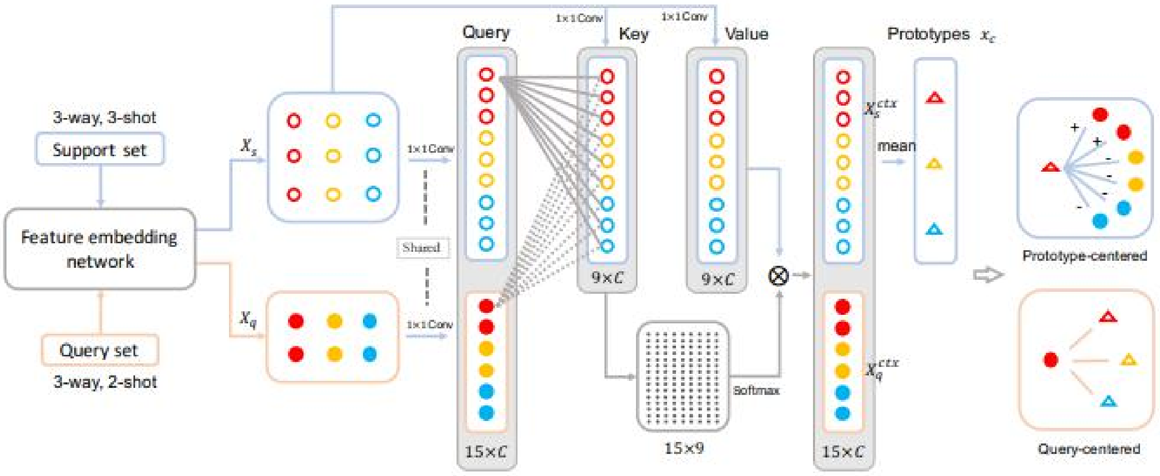
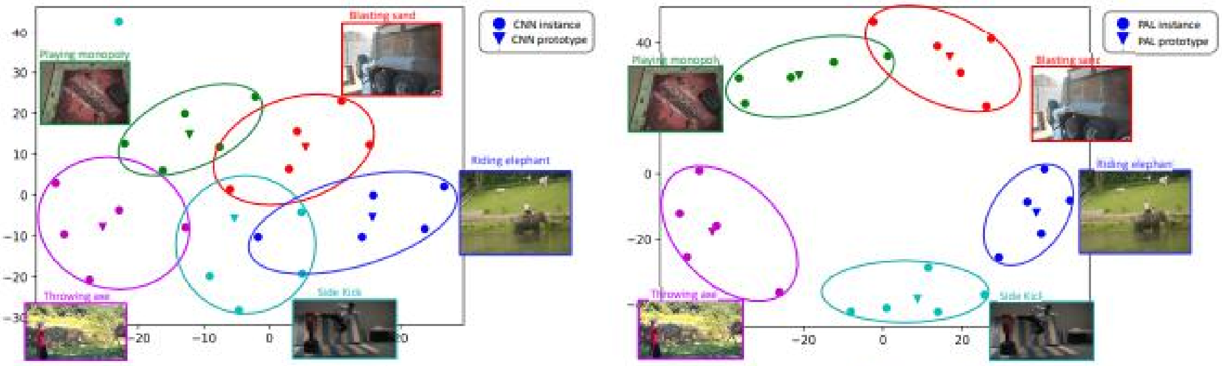


Figure 2

Figure(2) : Prototype-centered Attentive Learning is depicted schematically (PAL). A CNN feature embedding network is used to extract the feature vectors Xs and Xq from an episode of training video data that includes a support set (3-way 3-shot) and a query set (2 samples per class). The feature vectors Xs of all support samples are then picked and linearly transformed to generate the Key and Value. The transformation of Xs is set as the Query and required to perform attentive learning using Key and Value for support-set self-attention learning.For support-set self-attention learning, the transformation of Xs is set as the Query and is utilized to perform attentive learning using Key and Value. Xq is instead utilized as the input of Query for query-to-support cross-attention learning.

Method	**Sth-Sth-100**	
	1-shot	5-shot
ProtoNet++	33.6	43.0
TRN++*	38.6	48.9
CMN++*	34.4	43.8
OTAM	42.8	52.3
FEAT	45.3	61.2
PAL	**46.4**	**62.6**

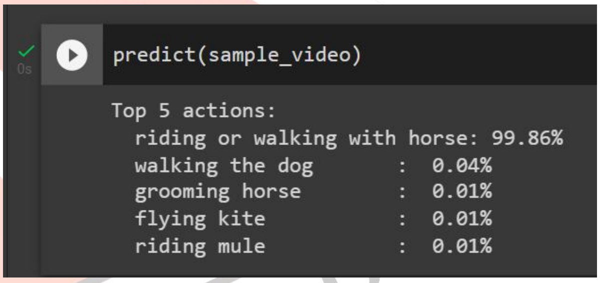


(a) Before PAL (b) After PAL

Figure 3

Figure(3) : t-SNE features projection of support samples and per-class prototypes in a 5-way 5-shot task on Kinetics-100 (a) before and (b) after the suggested PAL. Each class's variation and border are indicated by a circle. Each class has its own color scheme. As can be seen, each class's boundary becomes closer, and the overlap between classes is significantly decreased.reduce intra-class variation by decreasing the distracting effect of outlier samples in building class decision border, and (2) minimize interclass overlap by executing query-centered and prototype centered discriminative learning concurrently.

6. Output



```
predict(sample_video)

Top 5 actions:
  riding or walking with horse: 99.86%
  walking the dog       :  0.04%
  grooming horse        :  0.01%
  flying kite           :  0.01%
  riding mule           :  0.01%
```

7. Comparative analysis

When compared to the traditional FSL approaches, the suggested PAL clearly outperforms them in both scenarios and across all datasets.While more powerful action recognition models improve the outcomes, they still fall short of our model. Surprisingly, the first FSL action model, CMN, is revealed to be inferior to both TRN++ and TARN, implying that the quality of its memory network is less important than improved temporal structure modeling.The cutting-edge OTAM boosts performance even more by employing a pairwise temporal alignment technique. Nonetheless, it is outperformed by our PAL, especially on the more difficult Sth-Sth-100 dataset. This is not surprising given that OTAM's temporal warping is functionally sensitive to outlier (less consistent) support-set films, which are common in fine-grained activities involving human-object interactions. Furthermore, due to the lack of prototype-centered discrimination, OTAM performs poorly when using a few shots per class during meta-learning when compared to our model. On the simpler Kinetics-100 dataset, where class overlapping is less of an issue, our PAL model outperforms OTAM by a narrower margin, indicating that tackling outlier samples is consistently a more successful technique.

8. Conclusion and Future scope

We propose a novel Prototype-centered Attentive Learning method for few-shot action recognition in this paper. It is specifically designed to address existing methods' data efficiency and inability to deal with outliers and class overlapping issues. To that end, two complementary components are created: prototype-centered contrastive learning, which allows for more efficient use of a few shots per class, and hybrid attention learning, which aims to mitigate the negative effects of outlier support samples as well class overlapping. They can be combined into a single framework and trained from start to finish to achieve maximum complementarity.Extensive studies show that the proposed PAL achieves new state-of-the-art results on four action benchmarks, with the most convincing improvement on the more difficult fine-grained action recognition benchmark.

9. Reference

[1] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li,Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In ECCV, 2020.

[2] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In ECCV, 2018.

[3] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles.Few-shot video classification via temporal alignment. In CVPR, 2020.

[4] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang.Incremental few-shot object detection. In CVPR, 2020.

[5] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational cross-transformers for few-shot action recognition. In CVPR, 2021.

[6] .Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In ICLR, 2017.

[7] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In NeurIPS, 2017 .

[8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In ICLR, 2019.

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In NeurIPS, 2014.Gottigundala, Tanay. (2020).Predicting Personality Type from Writing Style.

[10] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot age classification with differentiable earth mover’s distance and structured classifiers. In CVPR,2020.