Pattern Recognition 108 (2020) 107348

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Few-shot activity recognition with cross-modal memory network

Lingling Zhang[a,b,*], Xiaojun Chang[d], Jun Liu[a,c], Minnan Luo[a,c], Mahesh Prakash[e], Alexander G. Hauptmann[f]

[a] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
[b] Ministry of Education Key Lab For Intelligent Networks and Network Security, Xi'an, China
[c] National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, Xi'an, China
[d] Faculty of Information Technology, Monash University, Australia
[e] Data61, CSIRO, Clayton, VIC, Australia
[f] School of Computer Science, Carnegie Mellon University, PA 15213, USA

ARTICLE INFO

Article history:
Received 13 August 2019
Revised 6 January 2020
Accepted 25 March 2020
Available online 30 July 2020

Keywords:
Few-shot learning
Activity recognition
Cross-modal memory

ABSTRACT

Deep learning based action recognition methods require large amount of labelled training data. However, labelling large-scale video data is time consuming and tedious. In this paper, we consider a more challenging few-shot action recognition problem where the training samples are few and rare. To solve this problem, memory network has been designed to use an external memory to remember the experience learned in training and then apply it to few-shot prediction during testing. However, existing memory-based methods just update the visual information with fixed label embeddings in the memory, which cannot adapt well to novel activities during testing. To alleviate the issue, we propose a novel end-to-end cross-modal memory network for few-shot activity recognition. Specifically, the proposed memory architecture stores the dynamic visual and textual semantics for some high-level attributes related to human activities. And the learned memory can provide effective multi-modal information for new activity recognition in the testing stage. Extensive experimental results on two video datasets, including HMDB51 and UCF101, indicate that our method could achieve significant improvements over other previous methods.

1. Introduction

The amount of video data is booming along with the rapid development of video-recording devices and digital media technology. Activity recognition is one of the most common studies in computer vision, which aims to predict the action category for correctly describing the content of video clips [1,2]. This task attracts increasing attention at present because it plays an important role in video behavior monitoring and analysis. Specifically, many researchers focus on constructing diverse deep networks to capture the excellent semantic representations of video clips. These deep learning methods such as [3–6] achieve great success on activity recognition, and they require large amount of labelled videos to optimize the massive parameters in deep networks. However, the annotation work is very time-consuming and some videos about certain activity classes are rare and unavailable in the real applications [7].

Different from deep learning, humans always learn a new activity from only one or few video clips successfully. For example, we can figure out that the long jump is a track and field event in which athletes leap as far as possible from a take off point according to one related video. Inspired from the human recognition procedure, few-shot learning, where just few labelled samples for each category are used for training, swarms into the community of machine learning. Notably, meta-learning paradigm [8] is proved to be an effective training pattern for few-shot learning, which learns the model from a large number of meta-tasks (also called episodes) and then testifies its ability to learn new meta-tasks. Taking three-way one-shot activity recognition for example in Fig. 1, each episode contains one unlabelled video (Query) and three labelled videos (Support Set) from three activity categories. In the training procedure, the model is trained on many episodes generated from base dataset by imitating the three-way one-shot activity recognition task. The learned model are evaluated on some new episodes that are generated from novel dataset in the testing stage. Note that there is no overlapping activity categories between base and novel dataset, but they are semantic related on some attributes.

The current few-shot models with meta-learning paradigm can be divided into three groups: metric-based methods [9–11],

[*] Corresponding author.
E-mail addresses: zhanglling@xjtu.edu.cn (L. Zhang), liukeen@mail.xjtu.edu.cn (J. Liu), minnluo@mail.xjtu.edu.cn (M. Luo), Mahesh.Prakash@data61.csiro.au (M. Prakash), alex@cs.cmu.edu (A.G. Hauptmann).

https://doi.org/10.1016/j.patcog.2020.107348
0031-3203/© 2020 Elsevier Ltd. All rights reserved.

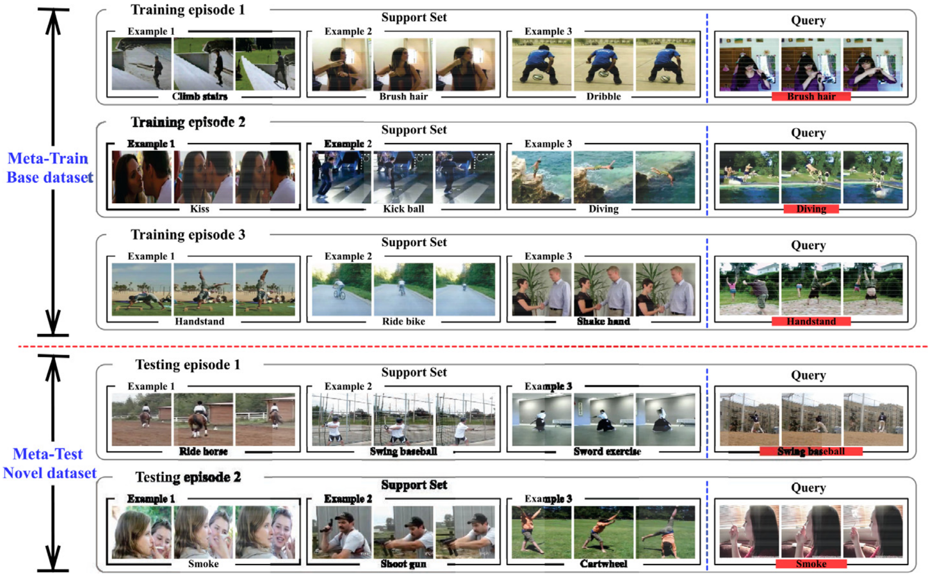


Fig. 1. The framework of meta-learning schema for three-way one-shot activity recognition.

optimization-based methods [12–14], and memory-based methods [15–18]. Note that memory-based methods are the focus of our paper. This type of methods generally designs an external memory to remember valuable information we have learned in training, and then applies the stored knowledge on few-shot prediction in testing. Key-value pair memory [19] is one type of popular memory structure for few-shot recognition, in which each memory slot is composed of a memory key and a memory value. The memory key denotes the visual representation of memory slot, and the memory value is an integer index or label embedding representing the class label of memory slot. With a given input example, two operations, namely reading and writing, are impacted on the memory content. The reading operation captures the related memory information to the given input, and the read content is often utilized to enrich the sample's original feature for better representing its characteristics. After that, memory content is updated by forgetting some old knowledge and strengthening specific fresh concepts in the writing stage. The goal of reading-writing mechanism is to control the memory to store the most prominent characteristics of the corresponding visual class. Particularly, current memory-based methods on few-shot learning just update the visual information with the fixed label embeddings as memory values. This single-modal memory has a great limitation on new few-shot prediction tasks because the visual categories in base and novel datasets are completely disjoint. In other words, the learned memory slot can reflect the intrinsic properties well of one training category, but is not directly related to the unseen visual categories in testing.

To solve the problem, we propose a novel cross-modal memory network for few-shot activity recognition. This memory architecture remembers the visual and textual information of common attributes between the activity categories in base and novel datasets. To be specific, each memory slot in our framework is a visual-textual embedding pair describing the multi-modal semantics of one attribute related to person activity. These two embeddings are no longer distinguished as memory key and value, and they are treated equally in the training stage. With one labelled activity sample, we obtain two relevance probability vectors by comparing the video embedding with visual memory, and the label embedding with textual memory. After that, the visual correlation weights are applied to the read-write operations on textual memory, while the textual correlation weights are utilized in the read-write process of visual memory. This cross-modal read-write pattern guarantees that the visual and textual contents in each memory slot are corresponding to each other and they are related to one common activity attribute. After that, the learned cross-modal memory provides the meaningful attribute knowledge about person activities for few-shot prediction on novel dataset. Three contributions of this paper are summarized as follows:

- We propose an end-to-end framework for few-shot activity recognition, which consists deep embedding module, cross-modal memory module, and the few-shot activity recognition module.
- We design an innovative cross-modal memory structure, where each memory slot is an visual-textual embedding pair that stores the multi-modal semantic information for one activity attribute.
- We conduct extensive experiments on datasets HMDB51 and UCF101 to illustrate the effectiveness and superiority of the cross-modal memory network. The experimental results demonstrate that our method consistently outperforms other competitors.

The remainder of this paper is organized as follows. We give a brief introduction to the related work on few-shot learning and memory network in Section 2. In Section 3, a novel end-to-end deep cross-modal memory network is proposed for few-shot activity recognition. In Section 4, we conduct extensive experiments over two benchmark datasets to illustrate the effectiveness and superiority of the proposed method. Conclusions are given in Section 5.

2. Related work

In this section, we briefly review the related work on few-shot learning and memory network.

2.1. Few-shot learning

Few-shot learning refers to understanding new concepts with only a few labelled examples [20–22]. Earlier studies on few-shot learning focus on constructing probabilistic generative models without any auxiliary data [23,24]. These methods are only suitable for simple handwriting recognition, because the handwriting characters are composed of different strokes with the certain writing orders. After that, some researchers find that meta-learning paradigm is an effective training pattern for few-shot recognition. Thus they attempt to optimize some deep networks that has achieved great success on computer vision with the novel meta-learning paradigm. To be specific, they first collect a large-scale auxiliary base dataset and pre-train the deep models by imitating few-shot classification scenes, then they apply the learned models to recognize new classes with only few samples on target dataset. These studies make a great breakthrough on few-shot learning and this research direction attracts more and more attentions.

So far, few-shot deep methods with meta-learning schema are very diverse and they can be generally categorized into three groups, namely metric-based methods, optimization-based methods, and memory-based methods. The metric-based methods focus on learning a novel metric space [25,26] in which few-shot classification is particularly efficient. For example, Koch et al. [27] feed two examples into siamese networks that share the same network parameters to identify whether they belong to the same class. Vinyals et al. [9] designed a novel matching network that combines k-nearest neighbour (kNN) with deep representation network together. Snell et al. [10] utilized a prototype network to learn a metric space in which few-shot classification can be performed by computing distances to prototype representations of each class. Sung et al. [11] presented an end-to-end relation network to tune deep embedding and non-linear distance metric simultaneously for effective few-shot learning. Garcia et al.[28] regarded few-shot learning as a supervised message passing task which is trained end-to-end using a graph neural network. The optimization-based methods learn a meta-model on multiple few-shot learning tasks and then use it to predict the model weights for new few-shot classification tasks. For instance, Finn et al. [29] proposed a novel Model-Agnostic Meta-Learning (MAML) method to meta-learn an initial condition that is good for fine-tuning on new few-shot problem. Ravi et al. [12] and Andrychowicz et al. [13] designed one LSTM-based network (meta-learner model) to update the massive parameters of another network (learner network) such that the learner network effectively solves the new few-shot problem. Munkhdalai et al. [14] introduced a novel meta-network that learns a meta-level knowledge across tasks and shifts its inductive biases via fast parameterization for rapid generalization. The last memory-based methods design an external memory network to remember the experience we have learned before, and these memory contents are utilized to assist in new few-shot classification tasks. For example, Santoro et al. [15] constructed a memory-augmented neural network to rapidly assimilate new data, and leveraged this memory to make accurate predictions after only a few samples. Xu et al. [17] designed a key-value memory network to extract external knowledge from noisy web images for better accomplishing few-shot classification. Cai et al. [16] presented a memory matching network which writes the features of labelled samples into memory and reads the holistic stored knowledge when performing inference. Zhu et al. [18] proposed a novel key-value memory architecture, called compound memory network, for few-shot classification, in which each key memory involves multiple constituent keys.

2.2. Memory network

Memory networks are designed to alleviate the problem of learning long-term dependencies in sequential data. The seminal work for memory networks traces back to the beginning of 2014, when Weston et al. [30] designed an external memory component which can be read and written flexibly through read-write controllers. This novel work largely popularizes the memory networks into the machine learning community. After that, the memory network is developed by Sukhbaatar et al. [31], who achieved end-to-end training schema with less supervision and made memory network more realistic in common scenarios. Chandar et al. [32] explored a novel hierarchical memory network, which can be considered as a hybrid between hard and soft attention memory networks. Miller et al. [19] proposed the typical key-value memory architecture which stores prior knowledge in many key-value memory slots before reading them for prediction. Kumar et al. [33] presented a dynamic memory network that employs neural sequence models for input representation, attention, and response mechanisms. Ramachandran et al. [34] proposed an extension for dynamic memory network, which utilizes the popular neural tensor network [35] for original attention mechanism. Xiong et al. [36] improved dynamic memory network by designing input fusion layer to allow interactions between inputs and attention mechanism. Henaff et al. [37] utilized a fixed number of hidden blocks as memories and updated them in parallel with an independent gated recurrent neural network. At present, the memory networks have been widely used in many fields, such as question answering [36], cross-modal retrieval [38], and few-shot learning [16,18]. In this paper, we design a novel cross-modal memory architecture to store the learned visual-textual information, and then utilize the memory contents and few labelled examples to recognize new concept in the testing stage.

3. Approach

In the setting of few-shot activity recognition, we aim to train a deep cross-modal memory network that can correctly recognize new activities with only few labelled video samples. Like previous meta-learning methods, we optimize our cross-modal memory network on base dataset in episodic training regime, and then evaluate its few-shot classification performance on novel dataset. Note that the activity categories in base and novel video datasets are disjoint but semantic related. In this paper, we denote any one input episode as $(S, \hat{\mathbf{v}})$, where S is a small labelled video set (Support Set) and $\hat{\mathbf{v}}$ is the target video sample (Query) whose label needs to be predicted. Taking n-way one-shot learning for example, the support set S contains n classes and each class has one labelled activity video, *i.e.* $S = \{(\mathbf{v}_1, \ell_1, \mathbf{y}_1), (\mathbf{v}_2, \ell_2, \mathbf{y}_2), \cdots, (\mathbf{v}_n, \ell_n, \mathbf{y}_n)\}$. Each triple ($\mathbf{v}_i$, ℓ_i, $\mathbf{y}_i$) contains the visual content, textual content, and label indicator for the labelled activity sample:

- visual content $\mathbf{v}_i \in \mathbb{R}^{3\times L\times H\times W}$: a sequence of RGB frames, where L, H and W respectively denote the number of frames, the height and width of each frame.
- textual content $\ell_i \in \mathbb{R}^d$: a semantic representation for the corresponding activity label, which is obtained in an unsupervised manner such as from Word2Vec or Glove.
- label indicator $\mathbf{y}_i \in \{0, 1\}^n$: a one-zero indicator vector, where the j-th element in $\mathbf{y}_i$ equals to 1 if video $\mathbf{v}_i$ is in the j-th activity class, otherwise it equals to 0.

Note that the video query just contains the visual content $\hat{\mathbf{v}}$ with dimension $\mathbb{R}^{3\times L\times H\times W}$. During the training procedure, the deep network is fine-tuned via minimizing the following objective function:

$$\min_{\theta} \mathbb{E}(\theta, D_{base}) = \min_{\theta} \sum_{(S,\hat{\mathbf{v}})\sim D_{base}} \|\hat{\mathbf{y}}^p - \hat{\mathbf{y}}\|, \tag{1}$$

where D_{base} is the base video dataset used for training and the episode $(S, \hat{\mathbf{v}})$ is random sampled from the base dataset. θ indicates all trainable parameters of the proposed cross-modal memory network. The vector $\hat{\mathbf{y}} \in \{0, 1\}^n$ is the true label indicator of video query $\hat{\mathbf{v}}$. The vector $\hat{\mathbf{y}}^p \in \mathbb{R}^n$ is the predicted label probability distribution for video query $\hat{\mathbf{v}}$ generated by deep cross-modal

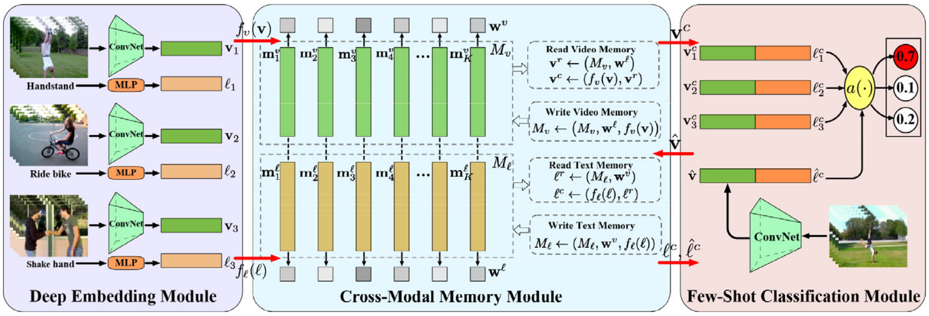

Handstand
Ride bike
Shake hand
ConvNet
MLP
$\mathbf{v}_1$
ℓ_1
$\mathbf{v}_2$
ℓ_2
$\mathbf{v}_3$
ℓ_3
$f_v(\mathbf{v})$
$f_\ell(\ell)$
$\mathbf{w}^v$
$\mathbf{w}^\ell$
$\mathbf{m}_1^v$ $\mathbf{m}_2^v$ $\mathbf{m}_3^v$ $\mathbf{m}_4^v$ … $\mathbf{m}_K^v$
M_v
$\mathbf{m}_1^\ell$ $\mathbf{m}_2^\ell$ $\mathbf{m}_3^\ell$ $\mathbf{m}_4^\ell$ … $\mathbf{m}_K^\ell$
M_ℓ
Read Video Memory
$\mathbf{v}^r \leftarrow (M_v, \mathbf{w}^\ell)$
$\mathbf{v}^c \leftarrow (f_v(\mathbf{v}), \mathbf{v}^r)$
Write Video Memory
$M_v \leftarrow (M_v, \mathbf{w}^\ell, f_v(\mathbf{v}))$
Read Text Memory
$\ell^r \leftarrow (M_\ell, \mathbf{w}^v)$
$\ell^c \leftarrow (f_\ell(\ell), \ell^r)$
Write Text Memory
$M_\ell \leftarrow (M_\ell, \mathbf{w}^v, f_\ell(\ell))$
$\mathbf{V}^C$
$\hat{\mathbf{V}}$
$\ell^c, \hat{\ell}^c$
$\mathbf{v}_1^c$ ℓ_1^c
$\mathbf{v}_2^c$ ℓ_2^c
$\mathbf{v}_3^c$ ℓ_3^c
$\hat{\mathbf{v}}$ $\hat{\ell}^c$
$a(\cdot)$
0.7
0.1
0.2
ConvNet
Deep Embedding Module
Cross-Modal Memory Module
Few-Shot Classification Module

Fig. 2. The framework of cross-modal memory network for few-shot activity recognition.

memory network. On the whole, the classification loss is measured with the Euclidean distance between the true label indicator $\hat{\mathbf{y}}$ and predicted label distribution $\hat{\mathbf{y}}^p$. The framework of deep cross-modal memory network for few-shot activity recognition is illustrated in Fig. 2. It includes three essential components, *i.e* deep embedding module (Left Part), cross-modal memory module (Center part), and few-shot classification module (Right part), whose specific details are introduced in Sections 3.1–3.3 respectively. And we briefly discuss the computational complexity of the proposed model in Section 3.4.

3.1. Deep embedding module

The deep embedding module is designed to extract visual and textual features for input videos and label embeddings. This module consists of two single-modal feature extraction networks, *i.e.* 3D convolutional network (3D ConvNet) for video modality and Multi-layer perceptron (MLP) for text modality. The architecture of 3D Convnet is similar to the proposed in [3] that has achieved significant success on video representation. It is composed of eight 3D convolutional layers (conv1*a*–conv5*b*), five max-pooling layers (pool1–pool5) and two fully connected layers (fc6–fc7). To be specific, the raw video $\mathbf{v}_i \in \mathbb{R}^{3\times L\times H\times W}$ first passes these convolutional and max-pooling layers to be represented as the deep feature map with dimension $\mathbb{R}^{512\times\frac{L}{16}\times\frac{H}{32}\times\frac{W}{32}}$, and then through two fully connected layers to be mapped to d_v-dimensional visual semantic space. The text MLP network is designed with two fully connected layers (fc1–fc2), which projects the original label embedding ℓ_i into a new d_ℓ-dimensional textual semantic space. For convenience, we denote the deep non-linear mapping functions for video and text networks as $f_v(\,\cdot\,)$ and $f_\ell(\,\cdot\,)$ respectively.

To this end, each triple ($\mathbf{v}_i$, ℓ_i, $\mathbf{y}_i$) in support set is encoded to ($f_v(\mathbf{v}_i)$, $f_\ell(\ell_i)$, $\mathbf{y}_i$) after the deep embedding module. For video query $\hat{\mathbf{v}}$ in the episode $(S, \hat{\mathbf{v}})$, we also represent it as d_v-dimensional deep features with 3D convNet, which is recorded as $f_v(\hat{\mathbf{v}})$. In this case, we simply denote the deep embeddings of episode $(S, \hat{\mathbf{v}})$ as $(S^e, \hat{\mathbf{v}}^e)$, namely $S^e = \{(f_v(\mathbf{v}_i), f_\ell(\ell_i), \mathbf{y}_i) : i = 1, 2, \cdots, n\}$ and $\hat{\mathbf{v}}^e = f_v(\hat{\mathbf{v}})$ in the following sections.

3.2. Cross-modal memory module

The cross-modal memory module contains a large external dynamic memory that can be read and written flexibly. It remembers the visual and textual semantic information about person activities that the model has learned. Specifically, we design K (memory size) number of cross-modal memory slots recorded as $M = \{(\mathbf{m}_1^v, \mathbf{m}_1^\ell), (\mathbf{m}_2^v, \mathbf{m}_2^\ell), \cdots, (\mathbf{m}_K^v, \mathbf{m}_K^\ell)\}$. Each memory slot $(\mathbf{m}_k^v, \mathbf{m}_k^\ell)$ $(k \le K)$ is a visual-textual embedding pair that describes the multimodal semantic characteristics of one high-level attribute associated with the person activity. Namely, it is composed of a d_v-dimensional visual vector $\mathbf{m}_k^v$ and a d_ℓ-dimensional textual vector $\mathbf{m}_k^\ell$. We respectively denote the contents of all visual memories and textual memories as matrices M_v and M_ℓ, *i.e.* $M_v = [\mathbf{m}_1^v, \mathbf{m}_2^v, \cdots, \mathbf{m}_K^v]$ and $M_\ell = [\mathbf{m}_1^\ell, \mathbf{m}_2^\ell, \cdots, \mathbf{m}_K^\ell]$.

At each time-stamp, the cross-modal memory module takes episode's deep embeddings $(S^e, \hat{\mathbf{v}}^e)$ as input, outputs the contextual embeddings conditioned on the current memory content, and updates the memory with the popular LSTM regime. Before reading and writing the cross-modal memory, for each triple in support set S^e, we first measure the visual and textual correlation weights between the deep embedding $f_v(\mathbf{v}_i)$, $f_\ell(\ell_i)$ and memory M_v, M_ℓ as follows:

$$\mathbf{w}^{v_i} = Softmax\big(M_v^\top f_v(\mathbf{v}_i)\big), \tag{2}$$

$$\mathbf{w}^{\ell_i} = Softmax\big(M_\ell^\top f_\ell(\ell_i)\big), \tag{3}$$

where the activation function $Softmax(\,\cdot\,)$ takes one vector as input and normalizes it into a probability distribution. The obtained visual weight vector $\mathbf{w}^{v_i}$ is utilized to read-write textual memory M_ℓ, while textual weight vector $\mathbf{w}^{\ell_i}$ is applied in the read-write procedure of visual memory M_v. With the cross-modal read-write pattern above, the visual memory $\mathbf{m}_k^v$ and textual memory $\mathbf{m}_k^\ell$ could interact with each other and represent semantic characteristics for one attribute in different modalities. Next, we will introduce the details of read-write operations on the cross-modal memory.

Read Memory: For any triple ($f_v(\mathbf{v}_i)$, $f_\ell(\ell_i)$, $\mathbf{y}_i$), we obtain the vector $\mathbf{v}_i^r$ by reading the visual memory M_v with textual weight vector $\mathbf{w}^{\ell_i}$:

$$\mathbf{v}_i^r = M_v \mathbf{w}^{\ell_i}. \tag{4}$$

Apparently, the $\mathbf{v}_i^r$ is the linear combination of the visual memory slot $\mathbf{m}_k^v$ in M_v. Similarly, we also read the textual memory M_ℓ with weight vector $\mathbf{w}^{v_i}$ as follows:

$$\ell_i^r = M_\ell \mathbf{w}^{v_i}. \tag{5}$$

The obtained vectors $\mathbf{v}_i^r$ and ℓ_i^r respectively are the summaries of previous visual and textual experiences related to the labelled sample ($\mathbf{v}_i$, ℓ_i, $\mathbf{y}_i$). We directly concatenate the sample's read contents and deep embeddings together for enriching the semantic representations. And then, the concatenated embeddings pass through the fully connected layer with a ReLU activation to get the final contextual embeddings:

$$\mathbf{v}_i^c = Relu(W_v^\top [f_v(\mathbf{v}_i), \mathbf{v}_i^r] + \mathbf{b}_v), \tag{6}$$

$$\ell_i^c = Relu(W_\ell^\top [f_\ell(\ell_i), \ell_i^r] + \mathbf{b}_\ell), \tag{7}$$

where parameters $\{W_v, \mathbf{b}_v\}$ and $\{W_\ell, \mathbf{b}_\ell\}$ denote the weight matrices and biases of fully connected layers. The activation function $Relu(\,\cdot\,)$ limits all the elements in vectors $\mathbf{v}_i^c$ and ℓ_i^c to be positive numbers. In the end, each triple ($\mathbf{v}_i$, ℓ_i, $\mathbf{y}_i$) is encoded to $(\mathbf{v}_i^c, \ell_i^c, \mathbf{y}_i)$ after the cross-modal memory module. Thus the contextual features of support set S is recorded as $S^m = \{(\mathbf{v}_1^c, \ell_1^c, \mathbf{y}_1), (\mathbf{v}_2^c, \ell_2^c, \mathbf{y}_2), \cdots, (\mathbf{v}_n^c, \ell_n^c, \mathbf{y}_n)\}$. At the same time, we also obtain the textual contextual embedding $\hat{\ell}^c$ for video query $\hat{\mathbf{v}}$ according to Equations (2) (5) and (7).

Write Memory: For each labelled sample ($\mathbf{v}_i$, ℓ_i, $\mathbf{y}_i$) in support set S, we encode its deep embeddings $f_v(\mathbf{v}_i)$, $f_\ell(\ell_i)$ into cross-modal memory for storing the meaningful characteristics related to person activities. Inspired by the input and forget gates in LSTM, we first erase the useless semantic information in current memory and then add new valuable knowledge into them. For visual memory M_v, the erase vector $\mathbf{e}_{v_i}$ and add vector $\mathbf{a}_{v_i}$ are computed as follows under given deep visual embedding $f_v(\mathbf{v}_i)$:

$$\mathbf{e}_{v_i} = Sigmoid(E_v^\top f_v(\mathbf{v}_i) + B_v^e) \tag{8}$$

$$\mathbf{a}_{v_i} = Tanh(A_v^\top f_v(\mathbf{v}_i) + B_v^a) \tag{9}$$

where the transformation matrices E_v and A_v are of shape $d_v \times d_v$, biases B_v^e and B_v^a are column vectors with d_v elements. Apparently, all elements in erase vector $\mathbf{e}_v$ lie in the ranges (0,1) with *Sigmoid*(·) activation function, meanwhile the elements in add vector $\mathbf{a}_v$ are in $(-1, 1)$ owing to *Tanh*(·) activation function. In this case, each visual memory slot $\mathbf{m}_k^v$ is modified as follows with the textual weight vector $\mathbf{w}^{\ell_i}$ obtained by Eq. (3):

$$\mathbf{m}_k^v = \mathbf{m}_k^v \odot \left(1 - \mathbf{w}_k^{\ell_i} \mathbf{e}_{v_i}\right) + \mathbf{w}_k^{\ell_i} \mathbf{a}_{v_i}, \tag{10}$$

where the real number $\mathbf{w}_k^{\ell_i}$ denotes the k-th element in textual weight vector $\mathbf{w}^\ell$. The update procedure of textual memory M_ℓ is similar to that of visual memory M_v. Specifically, with the deep textual embedding $f_\ell(\ell_i)$, the erase vector $\mathbf{e}_{\ell_i}$ and add vector $\mathbf{a}_{\ell_i}$ for textual memory M_ℓ is obtained by:

$$\mathbf{e}_{\ell_i} = Sigmoid(E_\ell^\top f_\ell(\ell_i) + B_\ell^e). \tag{11}$$

$$\mathbf{a}_{\ell_i} = Tanh(A_\ell^\top f_\ell(\ell_i) + B_\ell^a). \tag{12}$$

The E_ℓ and A_ℓ are two transformation matrices with dimension $d_\ell \times d_\ell$, and the B_ℓ^e and B_ℓ^a are two d_ℓ-dimensional biases vectors. With the visual weight vector $\mathbf{w}^{v_i}$ in Eq. (2), the value of each text memory $\mathbf{m}_\ell^k$ is updated by:

$$\mathbf{m}_k^\ell = \mathbf{m}_k^\ell \odot \left(1 - \mathbf{w}_k^{v_i} \mathbf{e}_{\ell_i}\right) + \mathbf{w}_k^{v_i} \mathbf{a}_{\ell_i} \tag{13}$$

The erase-add updating mechanism allows to forget useless information and strengthen the memory of new knowledge. In the end, the cross-modal memory stories rich cross-modal semantics about the attributes of person activities.

3.3. Few-shot classification module

With contextual embeddings of support set $S^m = \{(\mathbf{v}_1^c, \ell_1^c, \mathbf{y}_1), (\mathbf{v}_2^c, \ell_2^c, \mathbf{y}_2), \cdots, (\mathbf{v}_n^c, \ell_n^c, \mathbf{y}_n)\}$, the few-shot classification module aims to classify the video query $\hat{\mathbf{v}}$ along with its textual contextual embedding $\hat{\ell}^c$ to one activity class. To be specific, for labelled video in support set, we concatenate its visual and textual contextual embeddings as $\mathbf{s}_i = [\mathbf{v}_i^c, \ell_i^c]$, and then project it into a d_f-dimensional multi-modal fusion space by weight matrix W_s. The concatenated embedding $\hat{\mathbf{q}} = [f_v(\hat{\mathbf{v}}), \hat{\ell}^c]$ for video query is also mapped to the fusion space by another weight matrix W_q. Under the condition, we obtain the predicted label distribution $\hat{\mathbf{y}}^p$ for video query $\hat{\mathbf{v}}$ by the following attention mechanism:

$$\hat{\mathbf{y}}^p = \sum_{i=1}^{n} a\left(W_q^\top \hat{\mathbf{q}}, W_s^\top \mathbf{s}_i, S^m\right) \mathbf{y}_i. \tag{14}$$

Apparently, vector $\hat{\mathbf{y}}^p$ is a linear combination of the label indicators $\{\mathbf{y}_i; i = 1, 2, \cdots, n\}$ in the support set. The attention kernel function $a\left(W_q^\top \hat{\mathbf{q}}, W_s^\top \mathbf{s}_i, S^m\right)$ could be computed by the following softmax function:

$$a\left(W_q^\top \hat{\mathbf{q}}, W_s^\top \mathbf{s}_i, S^m\right) = \frac{e^{\hat{\mathbf{q}}^\top W_q W_s^\top \mathbf{s}_i}}{\sum_{j=1}^{n} e^{\hat{\mathbf{q}}^\top W_q W_s^\top \mathbf{s}_j}} \tag{15}$$

Finally, the predicted label distribution $\hat{\mathbf{y}}^p$ is applied to compute the loss function (1) in the training stage, and to evaluate the model's performance in testing.

3.4. Computational complexity discussion

The trainable parameters in the proposed deep model are optimized with the typical back-propagation strategy. To be specifc, the deep embedding module contains the typical 3D ConvNet and MLP network, which are widely used in many research fields. The remain two modules are designed by ourselves, so we just analyze the time complexity of their forward procedure to testify the efficiency in theory.

Taking n-way one-shot learning for example, each input episode $(S, \hat{\mathbf{v}})$ first passes the deep embedding module to be encoded to $(S^e, \hat{\mathbf{v}}^e)$, where $S^e = \{(f_v(\mathbf{v}_i), f_\ell(\ell_i), \mathbf{y}_i) : i = 1, 2, \cdots, n\}$ and $\hat{\mathbf{v}}^e = f_v(\hat{\mathbf{v}})$. After that, for each triple ($f_v(\mathbf{v}_i)$, $f_\ell(\ell_i)$, $\mathbf{y}_i$) in support set S^e, we need to compute the visual and textual correlation weights, read and update memory in the cross-modal memory module. The calculation of correlation weights needs $\mathcal{O}(d_v K + d_\ell K)$. The time complexity of reading memory is $\mathcal{O}(d_v K + d_\ell K + d_v d_v^c + d_\ell d_\ell^c)$, where d_v^c and d_ℓ^c are the dimensions of contextual features $\mathbf{v}_i^c$ and ℓ_i^c. The time complexity of writing memory is $\mathcal{O}(d_v d_v + d_\ell d_\ell)$. Generally, the value of memory size K is smaller than d_v and d_ℓ. since the triple ($f_v(\mathbf{v}_i)$, $f_\ell(\ell_i)$, $\mathbf{y}_i$) consumes the following time when it passes through the cross-modal memory module:

$$\begin{aligned} &\mathcal{O}(d_v K + d_\ell K) + \mathcal{O}(d_v K + d_\ell K + d_v d_v^c + d_\ell d_\ell^c) + \mathcal{O}(d_v d_v + d_\ell d_\ell) \\ &\quad \approx \mathcal{O}(d_v d_v^c + d_\ell d_\ell^c + d_v d_v + d_\ell d_\ell) \end{aligned} \tag{16}$$

According to the mentioned above, the support set S^e containing n triples passes the cross-modal memory module with the time complexity $\mathcal{O}(n d_v d_v^c + n d_\ell d_\ell^c + n d_v d_v + n d_\ell d_\ell)$. For the video query $\hat{\mathbf{v}}^e$, we just compute its corresponding visual weights and read the textual memory with the time complexity $\mathcal{O}(d_v K + d_\ell K + d_\ell d_\ell^c)$. Therefore, for the cross-modal module, the time complexity with the input episode $(S^e, \hat{\mathbf{v}}^e)$ is as follows:

$$\begin{aligned} &\mathcal{O}(n d_v d_v^c + n d_\ell d_\ell^c + n d_v d_v + n d_\ell d_\ell) + \mathcal{O}(d_v K + d_\ell K + d_\ell d_\ell^c) \\ &\quad \approx \mathcal{O}(n d_v d_v^c + n d_\ell d_\ell^c + n d_v d_v + n d_\ell d_\ell). \end{aligned} \tag{17}$$

Finally, given the contextual embedding of episode $(S^e, \hat{\mathbf{v}}^e)$, the few-shot classification module classify the video query to one activity class. This stage is implemented by an attention mechanism with the time complexity $\mathcal{O}(n d_f d_v^c + n d_f d_\ell^c)$. In practice, the values of d_v, d_ℓ, d_v^c, d_ℓ^c, and d_f are in the same magnitude, and d_v is slightly larger than d_ℓ, d_v^c, d_ℓ^c, and d_f. Therefore, the forward time complexity of cross-modal memory module and few-shot classification module is about $\mathcal{O}(n d_v d_v)$, *i.e* linear in the number of triples n of support set S.

4. Experiment

In this section, extensive experiments are conducted on two datasets to validate the effectiveness and superiority of our cross-modal memory network.

4.1. Datasets

We perform experiments on two video datasets HMDB51 and UCF101. Their details are as follows:

Fig. 3. Examples from two video datasets. Top: HMDB51; Bottom: UCF101.

- **HMDB51** [39]: The dataset is a widely used benchmark dataset for activity recognition. It contains 6849 video clips divided into 51 activity classes, and each class includes a minimum of 101 clips. In our experiment, we select 35 activity classes for base dataset, and the rest 16 classes for novel dataset.
- **UCF101** [40]: The dataset is a well-known action recognition dataset in which the video data are collected from YouTube. It consists of 101 action classes and 13,320 clips over 27 hours. We respectively choose 61 classes for base dataset and 40 classes for novel dataset in the experiment.

Some examples of video clips in HMDB51 and UCF101 are shown in Fig. 3. Note that the label embeddings of activities in base and novel datasets are required for the proposed cross-modal memory network. In HMDB51 and UCF101 datasets, the activity labels are generally made up of one or two noun words with specific meanings, we thus directly utilize the average of 300-dimensional Word2Vec features as their semantic representations.

4.2. Competitors

We compare the proposed cross-modal memory network with one baseline designed by ourself and five existing state-of-the-art methods. Their details are introduced as follows.

- **Baseline Classifier**: We train the 3D ConvNet on base dataset in traditional classification regime. The learned network is utilized to extract deep features for video samples in novel dataset. In the case, few-shot classification is directly achieved by measuring the feature similarity between video query and labelled videos in support set.
- **Siamese Network** [27]: This method performs few-shot classification by learning a binary classifier to distinguish whether two samples are from same/different class. The binary classifier is constructed by a deep siamese network that has the same architecture with our 3D ConvNet.
- **Matching Network** [9]: This method is the pioneer with meta-learning regime to train deep network for few-shot recognition. It maps the target query and a small support set into a shared embedding space. It also adds a full context embedding module after the feature-extracting network for adjusting sample features in the condition of support set. Note that the architecture of the full context embedding module is a bidirectional LSTM.
- **Prototype Network** [10]: This method learns a metric space in which few-shot classification is performed by computing distances to prototype representations of each class. The prototype network reflects a simpler inductive bias that is beneficial in the limited-data regime, thus it achieves excellent results on few-shot recognition.
- **Relation Network** [11]: This method learns the deep feature embedding and non-linear distance metric simultaneously for measuring the similarity between query and sample items. This method is far simpler and it produces state-of-the-art results on few-shot recognition task.
- **Memory Network** [16]: This method designs a key-value memory module to preserve the visual knowledge learned before. The feature embeddings of samples in the support set are contextually augmented with the holistic knowledge in memory slots.

4.3. Implementation

We implement our model based on the open-source PyTorch framework. In deep embedding module, the input of 3D ConvNet is a sequence of resized RGB video frames with dimension 3 × 16 × 112 × 112 because of the limited GPU memory, and the number of neurons for fully connected layers (fc6-fc7) are set to 4096 and 512 respectively. Meanwhile, the input of text MLP network is the 300-dimensional Word2Vec, and the output dimensions of fully connected layers (fc1-fc2) are set to 300 and 256. In cross-modal memory module, we respectively assign the dimensions of visual and textual contextual embeddings as 512 and 256. Except for the layers in memory writing stage, other network layers in proposed end-to-end network adopt the ReLU function as the activation function. We also add the dropout layers with probability 0.1 and apply batch normalization for fully connected layers in deep embedding module and few-shot classification module to improve the network stability.

Under above settings, we utilize the Stochastic Gradient Descent (SGD) algorithm with the momentum 0.9 and weight decay 0.001 to optimize the cross-modal memory network. Notably, the parameters of convolutional layers (conv1*a* to conv 5*b*) in 3D ConvNet are pre-trained on large-scale video dataset Sports-1M to avoid overfitting. The learning rates of parameters in three modules are assigned to different values. The learning rate of deep embedding module and few-shot classification module is initialized as 0.1 and then decreased by 5% after each 1000 iterations; the learning rate for cross-modal memory is assigned to 0.5 at first and reduced by 10% every 500 iterations. Lastly, we set the mini-batch with 16 video clips and run the experiment on four TITAN GPUs, each with 12G memory.

4.4. Performance comparison

For our method, we set the memory size K for datasets HMDB51 and UCF101 to 24 and 40 respectively, while the dimension of fusion space in few-shot classification module is assigned to 320. For six competitors, including Baseline Classifier, Siamese Net, Matching Net, Prototype Net, Relation Net, and Memory Net, their video feature-extracting network is set to be the same as that of our 3D ConvNet for fair comparison. With two videos as input, the Siamese Net computes the ℓ_1 component-wise distance between two 512-dimensional video vectors and then add a fully connected

Table 1
Few-shot classification performance comparison in terms of ACC over HMDB51.

Models	3-way		5-way		10-way	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Base Classifier	65.30 ± 1.60	76.03 ± 1.23	52.94 ± 2.03	66.12 ± 1.13	41.04 ± 1.23	54.26 ± 1.12
Siamese Net	66.45 ± 1.54	77.21 ± 0.86	54.03 ± 0.98	67.34 ± 1.12	42.23 ± 1.32	55.38 ± 1.17
Matching Net	67.71 ± 0.98	78.92 ± 0.36	56.78 ± 1.21	68.79 ± 1.32	43.76 ± 1.21	57.21 ± 1.07
Prototype Net	68.03 ± 0.62	79.32 ± 0.88	57.21 ± 1.67	69.03 ± 0.99	44.21 ± 1.21	57.89 ± 0.97
Relation Net	68.23 ± 0.65	79.78 ± 0.54	57.09 ± 1.04	69.13 ± 0.36	44.15 ± 1.01	57.48 ± 1.34
Memory Net	68.12 ± 0.78	79.34 ± 0.88	57.26 ± 0.76	69.21 ± 0.72	44.38 ± 1.11	57.81 ± 1.26
Our model	**69.13 ± 0.93** (↑0.90)	**80.84 ± 0.67** (↑1.06)	**58.92 ± 0.72** (↑1.66)	**70.45 ± 0.66** (↑1.24)	**46.56 ± 1.45** (↑2.18)	**59.10 ± 0.86** (↑1.21)

Table 2
Few-shot classification performance comparison in terms of ACC over UCF101.

Models	3-way		5-way		10-way	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Base Classifier	84.05 ± 0.23	90.13 ± 0.32	75.37 ± 0.87	83.73 ± 0.64	64.89 ± 0.76	76.25 ± 0.78
Siamese Net	85.36 ± 0.12	91.32 ± 0.14	76.72 ± 0.87	85.23 ± 0.43	66.36 ± 1.21	77.89 ± 0.97
Matching Net	86.92 ± 0.24	91.98 ± 0.21	77.58 ± 1.01	86.12 ± 0.87	67.21 ± 0.76	78.68 ± 0.68
Prototype Net	87.32 ± 0.24	92.57 ± 0.14	78.13 ± 0.63	86.46 ± 0.17	67.52 ± 0.36	79.12 ± 0.72
Relation Net	87.26 ± 0.18	92.89 ± 0.10	77.94 ± 1.04	86.37 ± 0.26	67.28 ± 0.89	79.02 ± 0.88
Memory Net	87.42 ± 0.19	92.68 ± 0.13	78.08 ± 0.87	86.51 ± 0.46	67.56 ± 0.12	79.20 ± 0.65
Our model	**89.28 ± 0.36** (↑1.86)	**93.72 ± 0.11** (↑0.83)	**78.92 ± 0.54** (↑0.79)	**87.72 ± 0.24** (↑1.21)	**68.62 ± 0.79** (↑1.06)	**80.12 ± 0.38** (↑0.92)

layer with Sigmoid activation function to obtain the similarity for estimating whether they are from the same class. For Matching Net, the number of unrolling steps in the LSTM context embedding module is set to 32. In Relation Net, we concatenate the 512-dimensional feature of any labelled video in support set with the feature vector of video query together, and then pass two fully connected layers with 128 and 1 hidden units to generate the relation scores. The memory size in Memory Net is set the same as our proposed model.

Under the condition, the few-shot classification accuracy on datasets HMDB51 and UCF101 are reported in Table 1 and 2 respectively. Note that the value of way is assigned in {3, 5, 10} and the value of shot is set in {1, 5}, to adequately verify the effectiveness of proposed cross-modal memory network. The symbol (↑) denotes the percentage increase in our method compared to the competitor that achieves the best result. By comparing the results of six baselines and our model, we have the following observations:

- Four few-shot methods with meta-learning schema consistently perform better than competitors Baseline Classifier and Siamese Network. This result proves that meta-learning is indeed an effective training pattern for few-shot recognition.
- Compared to four meta-learning methods, our cross-modal memory network consistently achieves the optimal accuracy of few-shot activity recognition. It indicates the learned multi-modal attribute knowledge for person activities is useful for few-shot prediction in the testing stage.
- The performance of each method becomes better when the value of shot increases and the way decreases on two datasets. The situation is reasonable because more video samples for each activity class and less activity classes could decrease the difficulty of few-shot recognition.

In addition, we show the loss and performance curves of our cross-modal memory network for three-way one-shot learning in Fig. 4. It demonstrates that the model converges after about 15,000 and 25,000 iterations for datasets HMDB51 and UCF101 respectively. In Fig. 5, we show some results of three-way one-shot recognition in testing stage, where the predicted label appears with red check mark.

Table 3
The influence of visual/textual contextual embeddings for three-way few-shot activity recognition.

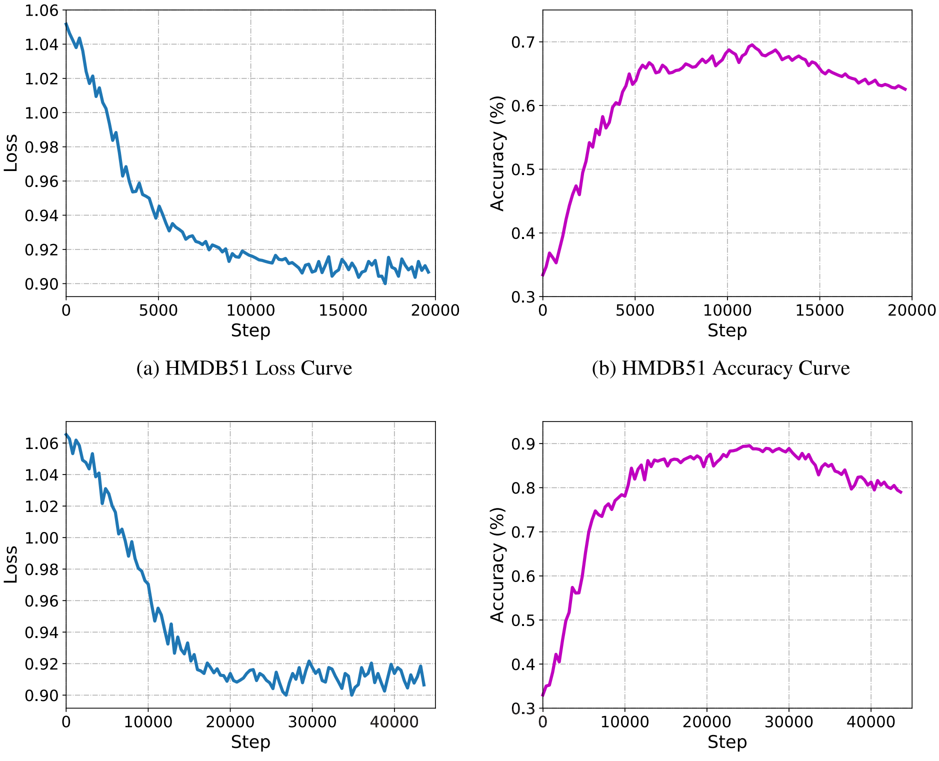
Model	HMDB51		UCF101	
	1-shot	5-shot	1-shot	5-shot
Our Model^V	67.17 ± 0.86	78.85 ± 0.74	86.94 ± 1.01	92.08 ± 0.87
Our Model^T	56.91 ± 1.24	69.92 ± 1.09	74.21 ± 1.54	80.63 ± 1.08
Our Model	**69.13 ± 0.93**	**80.84 ± 0.67**	**89.28 ± 0.36**	**93.72 ± 0.11**

4.5. Impact of visual/textual contextual embedding

In the proposed model, for any triple ($\mathbf{v}_i$, ℓ_i, $\mathbf{y}_i$) we utilize the cross-modal read mechanism to obtain the textual and visual contextual embeddings from cross-modal memory module, and then concatenate them together for few-shot classification. To further testify the contribution of textual and visual contextual embeddings, we compare the proposed model with two variants that only contains single-modal contextual embedding for few-shot classification. Their details are as follows:

- Our Model: Concatenating textual and visual contextual embeddings together for few-shot classification. Namely, video query is represented as $[\hat{\mathbf{v}}, \hat{\ell}^c]$, and the sample in support set is encoded to $[\mathbf{v}_i^c, \ell_i^c]$.
- Our Model^V: Just using the visual contextual embedding for few-shot classification. Namely, video query is represented as $\hat{\mathbf{v}}$, and the sample in support set is encoded to $\mathbf{v}_i^c$.
- Our Model^T: Just using the textual contextual embedding for few-shot classification. Namely, video query is represented as $\hat{\ell}^c$, and the sample in support set is encoded to ℓ_i^c.

As shown in Table 3, for three-way few-shot learning, if just using the visual contextual embedding, the variant "Our Model^V" can get the general level of performance on activity recognition. The performance of "Our Model" is improved by considering the textual contextual embedding into "Our Model^V". Therefore, both visual and textual contextual embeddings are helpful to few-shot activity recognition, especially the visual contextual embedding.



(a) HMDB51 Loss Curve

(b) HMDB51 Accuracy Curve

(c) UCF101 Loss Curve

(d) UCF101 Accuracy Curve

Fig. 4. The loss and performance curves for three-way one-shot activity recognition.

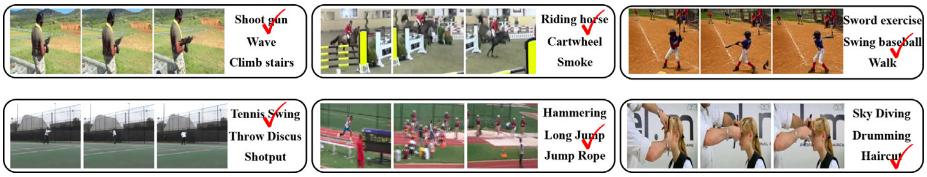
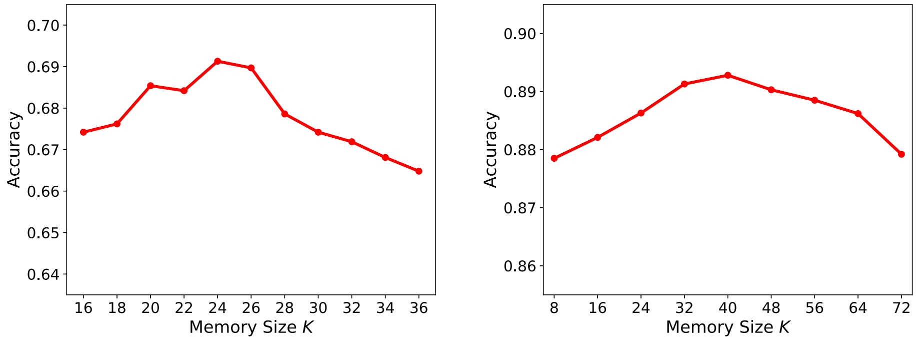


Fig. 5. Some testing results of three-way one-shot classification. Top: HMDB51; Bottom: UCF101.



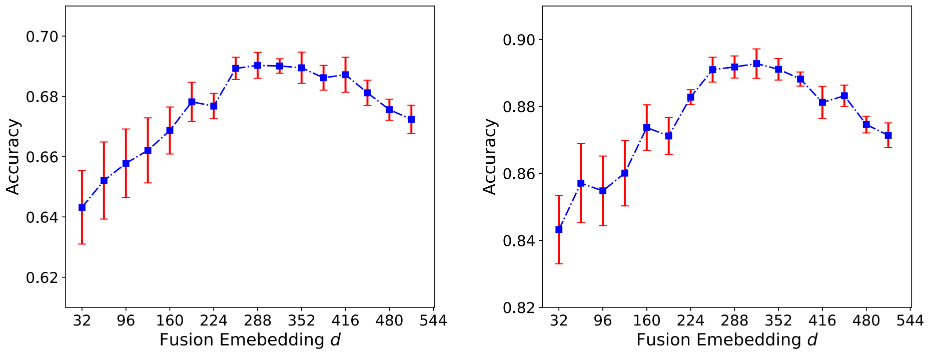
(a) HMDB51

(b) UCF101

Fig. 6. The influence of memory size for three-way one-shot activity recognition.

4.6. Impact of memory size

The memory size represents the number of attributes associated with human activity that are mined during the training process. In this section, we study the influence of memory size on three-way one-shot activity recognition. For HMDB51 dataset, we assign the memory size K to be varied in interval [16,36] with the step-size 2. And we set the memory size K in UCF101 dataset to be tuned from 8 to 72 with a step-size 8. In Fig. 6, we plot the performance curves as the increase of memory size on two datasets. They indicate that the accuracy of activity recognition improves with the increase of memory size at first; after reaching the maximum, the performance decreases gradually. This situation is reasonable because of the following two reasons: (1) a mere handful of attributes are not enough to express the characteristics of human activities in some aspects; (2) large amount of attributes are prone to overfitting the training activity samples, but may have an adverse impact on few-shot predictions over novel activity classes.



(a) HMDB51

(b) UCF101

Fig. 7. The influence of fusion dimension for three-way one-shot activity recognition.

In addition, the optimal value of memory size distinguishes over different datasets because of the diverse properties related to the datasets. When the memory size is set to 24 and 40, the performance of three-way one-shot activity recognition can be most prominent on dataset HMDB51 and UCF101 respectively.

4.7. Impact of fusion dimension

In few-shot classification module, we concatenate the textual and visual contextual embeddings together, then project it into a novel multi-modal fusion space. Apparently, the dimension d_f of fusion space might have a certain influence on the few-shot activity recognition results. In this section, with the condition memory size $K = 24$ and 40 for HMDB51 and UCF101, we study the effect of dimension d_f by assigning it to be tuned from 32 to 512 with a step-size of 32. As shown in Fig. 7, the performance of three-way one-shot activity recognition increases with the increase of d_f at first, stabilizes in interval [256,352], and then it declines gradually when d_f is greater than 352. We also find that classification performance fluctuates extensively when the space dimension is too small. The results demonstrate that a proper value of d_f can achieve the optimal few-shot activity recognition result by mining the meaningful semantics and filtering out noise information in contextual multi-modal embeddings. In the real applications, we can determine the value of d_f by cross-validation strategy if the activity videos are sufficient in base dataset.

5. Conclusion

In this paper, we propose an end-to-end deep network for few-shot activity recognition, which includes three modules: deep embedding module, cross-modal memory module, and few-shot classification module. In the training stage, the cross-modal memory is utilized to remember the visual and textual semantic information for some attributes related to human activities. After that, the learned memory provides the rich experiences for novel few-shot activity recognition tasks during testing. We conduct extensive experiments on two widely-used video datasets HMDB51 and UCF101. The experimental results indicate that our model generally obtains the comparable performance to the state-of-the-art models in terms of few-shot recognition accuracy. Even so, the Word2Vec label embeddings are insufficient to represent the textual semantics of base and novel activities, and the simple MLP network may be limited in mining the high-level textual information that related to human activities. To combat this issue, we attempt to capture some abundant articles about activity classes from website, and then replace the MLP text network with BERT network to extract the textual semantics automatically. Additionally, we also want to apply the cross-modal memory to the general few-shot classification framework in image and audio fields.

Declaration of Competing Interest

We wish to confirm that there are no conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. We confirm that the manuscript has been read and proved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

Acknowledgment

This work was supported by National Key Research and Development Program of China (2018YFB1004500), National Natural Science Foundation of China (61532004, 61532015, 61672418, 61672419 and 61877050), Innovative Research Group of the National Natural Science Foundation of China (61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Centre for Engineering Science and Technology, Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00340, Air Force Research Laboratory and DARPA under agreement number FA8750-19-2-0501, and Australian Research Council Discovery Early Career Researcher Award under grant no. DE190100626.

References

[1] J.M. Carmona, J. Climent, Human action recognition by means of subtensor projections and dense trajectories, Pattern Recognit. 81 (2018) 443–455.
[2] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, S.J. Maybank, Asymmetric 3d convolutional neural networks for action recognition, Pattern Recognit. 85 (2019) 1–12.
[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision (CVPR), 2015, pp. 4489–4497.
[4] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1933–1941.
[5] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6299–6308.
[6] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, IEEE Trans. Pattern Anal. Mach.Intell. (PAMI) 40 (6) (2018) 1510–1517.
[7] Y. Han, B. Wang, M. Idesawa, H. Shimai, Recognition of multiple configurations of objects with limited data, Pattern Recognit 43 (4) (2010) 1467–1475.
[8] N. Jackson*, Developing the concept of meta-learning, Innov. Educ. Teach. Int.(IETI) 41 (4) (2004) 391–403.

[9] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016, pp. 3630–3638.
[10] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2017, pp. 4077–4087.
[11] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1199–1208.
[12] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[13] M. Andrychowicz, M. Denil, S. Gomez, M.W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, N. De Freitas, Learning to learn by gradient descent by gradient descent, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016, pp. 3981–3989.
[14] T. Munkhdalai, H. Yu, Meta networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 2554–2563.
[15] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, Meta-learning with memory-augmented neural networks, in: International conference on machine learning, 2016, pp. 1842–1850.
[16] Q. Cai, Y. Pan, T. Yao, C. Yan, T. Mei, Memory matching networks for one-shot image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4080–4088.
[17] Z. Xu, L. Zhu, Y. Yang, Few-shot object recognition from machine-labeled web images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1164–1172.
[18] L. Zhu, Y. Yang, Compound memory networks for few-shot video classification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 751–766.
[19] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, J. Weston, Key-value memory networks for directly reading documents, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 1400–1409.
[20] L. Zhang, J. Liu, M. Luo, X. Chang, Q. Zheng, A.G. Hauptmann, Scheduled sampling for one-shot learning via matching network, Pattern Recognit. 96 (2019) 106962.
[21] R. Krishnan, S. Sarkar, Conditional distance based matching for one-shot gesture recognition, Pattern Recognit. 48 (4) (2015) 1302–1314.
[22] S. Zhou, C. Deng, Z. Piao, B. Zhao, Few-shot traffic sign recognition with clustering inductive bias and random neural network, Pattern Recognit (2019) 107160.
[23] L. Fe-Fei, et al., A bayesian approach to unsupervised one-shot learning of object categories, in: Proceedings of the IEEE International Conference on Computer Vision (CVPR), 2003, pp. 1134–1141.
[24] L. Fei-Fei, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Trans. Pattern Anal. Mach.Intell. (PAMI) 28 (4) (2006) 594–611.
[25] B. Du, L. Zhang, A discriminative metric learning based anomaly detection method, IEEE Trans. Geosci. Remote Sens.(TGRS) 52 (11) (2014) 6844–6857.
[26] Y. Dong, B. Du, L. Zhang, L. Zhang, Dimensionality reduction and classification of hyperspectral images using ensemble discriminative local metric learning, IEEE Trans. Geosci. Remote Sensi(TGRS) 55 (5) (2017) 2509–2524.
[27] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, ICML Deep Learning Workshop, 2, 2015.
[28] V. Garcia, J. Bruna, Few-shot learning with graph neural networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2018.
[29] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 1126–1135.
[30] J. Weston, S. Chopra, A. Bordes, Memory networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.
[31] S. Sukhbaatar, J. Weston, R. Fergus, et al., End-to-end memory networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 2440–2448.
[32] S. Chandar, S. Ahn, H. Larochelle, P. Vincent, G. Tesauro, Y. Bengio, Hierarchical memory networks (2016). arXiv: 1605.07427.
[33] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, Ask me anything: Dynamic memory networks for natural language processing, in: Proceedings of the International Conference on Machine Learning (ICML), 2016, pp. 1378–1387.
[34] G.S. Ramachandran, A. Sohmshetty, Ask me even more: Dynamic memory tensor networks (extended model) (2017). arXiv: 1703.03939.
[35] R. Socher, D. Chen, C.D. Manning, A. Ng, Reasoning with neural tensor networks for knowledge base completion, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2013, pp. 926–934.
[36] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, in: Proceedings of the International Conference on Machine Learning (ICML), 2016, pp. 2397–2406.
[37] M. Henaff, J. Weston, A. Szlam, A. Bordes, Y. LeCun, Tracking the world state with recurrent entity networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[38] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 299–307.
[39] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, Hmdb: a large video database for human motion recognition, in: Proceedings of the International Conference on Computer Vision (ICCV), 2011.
[40] K. Soomro, A.R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild (2012). arXiv: 1212.0402.

Lingling Zhang is currently an assistant professor in computer science at Xi'an Jiaotong University. She received the PhD degree in Computing Science from Xi'an Jiaotong University in 2020. She was a visiting student with the School of Computer Science, Carnegie Mellon University, working with Prof. A. Hauptmann. Her research interests include cross-media information mining, computer vision, zero-shot learning, and few-shot learning.

Xiaojun Chang was a Post-Doctoral Research Associate with the School of Computer Science, Carnegie Mellon University. He is currently a Faculty Member with the Faculty of Information Technology, Monash University, Clayton Campus, Australia. He is also affiliated with the Centre for Data Science, Monash University. He has spent most of his time working on exploring multiple signals (visual, acoustic, and textual) for automatic content analysis in unconstrained or surveillence videos. He has achieved top performance in various international competitions, such as TRECVID MED, TRECVID SIN, and TRECVID AVS. He was an ARC Discovery Early Career Researcher Award Fellow from 2019 to 2021.

Jun Liu received the B.S, M.S and Ph.D degrees in computer science in 1995, 1998, and 2004 from Xi'an Jiaotong University, China. He is currently a professor in the School of Electronic and Information Engineering at Xi'an Jiaotong University. His research interests include text mining, data mining, Intelligent Network Learning Environment and multimedia E-learning.

Minnan Luo received the Ph. D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2014. Currently, she is an Assistant Professor in the School of Electronic and Information Engineering at Xi'an Jiaotong University. She is also a Post-Doctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. Her research interests include machine learning and optimization, video analysis, cross-media retrieval and fuzzy system identification and modeling.

Mahesh Prakash received Ph. D. degree in Thermofluids Engineering from Victoria University and CSIRO, Melbourne, India, in 2000. Currently, he is a Senior Principal Research Scientist at CSIRO Data61 and leads a group of Scientists, Engineers and Software Professionals who develop and apply a diverse range of modelling and analytics techniques for urban environmental applications with a key focus on emergency and disaster management. His scientific and technical interests are in urban flood modelling, geospatial big data analytics and the confluence between environmental research and technology.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from Johns Hopkins University, Baltimore, MD, USA, the degree in computer science from the Technische Universitat Berlin, Berlin, Germany, in 1984, and the Ph.D. degree in computer science from Carnegie Mellon University(CMU), Pittsburgh, PA, USA, in 1991. He is currently with the Department of Computer Science and the Language Technologies Institute, CMU. His current research interests include natural language processing, speech understanding and synthesis, video analysis, and machine learning.