Computer Science and Application 计算机科学与应用, 2022, 12(3), 499-507
Published Online March 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.123050

Hans汉斯

基于直推式零样本学习的动作识别方法

齐秋平

同济大学计算机科学与技术系，上海

收稿日期：2022年2月3日；录用日期：2022年2月28日；发布日期：2022年3月9日

摘　要

在物联网和智能设备飞速发展的当代社会，网络信息已从海量文本数据逐渐演变为更为直观的图像和视频数据。丰富的视频数据在为人类提供诸多便利的同时，其内容理解和分类也给人们带来诸多新的挑战。针对现有深度方法严重依赖大量标注样本，并且所学知识不可拓展的问题，零样本学习作为迁移学习的一种特殊场景，以可从可见类别拓展到未见类别的独特优势吸引了大量关注。本文提出一种基于直推式零样本学习的动作识别方法，首先将视觉信息映射到语义空间中，然后通过语义空间的最近邻搜索来完成识别任务，并且引入带有偏差的损失函数，旨在提高识别精度的同时有效缓解强偏问题。该模型在UCF101、HMDB51以及OlympicSports数据集上的识别准确率分别达到26.8%、20.3%和46.5%，充分证明了该方法的有效性。

关键词

零样本学习，直推式学习，视频动作识别

Research on Transductive Zero-Shot Learning for Action Recognition

Qiuping Qi

Department of Computer Science and Engineering, Tongji University, Shanghai

Received: Feb. 3rd, 2022; accepted: Feb. 28th, 2022; published: Mar. 9th, 2022

Abstract

With the rapid development of Internet of things and intelligent devices, Internet information has gradually evolved from massive text data to more intuitive image and video data. Rich video data not only provides conveniences, but also brings many new challenges in video understanding and classification. In view of the problem that existing deep learning methods rely on a large number of labeled data and the learned knowledge cannot be expanded, zero-shot learning as a kind of transfer learning, has attracted a lot of attention because of its unique advantage of expanding

文章引用: 齐秋平. 基于直推式零样本学习的动作识别方法[J]. 计算机科学与应用, 2022, 12(3): 499-507.
DOI: 10.12677/csa.2022.123050

from seen categories to unseen categories. In this paper, an action recognition method based on transductive zero-shot learning is proposed. Firstly, visual information is mapped into the semantic space, and then the recognition task is carried out through the nearest neighbor search in the semantic space, and the loss function with deviation is introduced to improve the recognition accuracy and effectively alleviate the problem of strong bias. Experiments on UCF101, HMDB51 and Olympic sports datasets show that the accuracy of the proposed method is 26.8%, 20.3% and 46.5% respectively, which fully proves the effectiveness of the proposed method.

Keywords

Zero-Shot Learning, Transductive Learning, Video Action Recognition

1. 引言

随着互联网技术的兴起和智能监控等传感器技术的日新月异，网络信息已从海量文本数据逐渐演变为更为直观的图像和视频数据。海量视频数据结构复杂，数据非结构化，并且富含丰富的用户画像描述和潜在特征信息。如何有效利用这些数据，实现对视频内容的合理分析与推理，从而为安防、教育、娱乐等行业提供更好的帮助，还存在大量问题亟需解决。由于深度学习技术的发展，最近关于视频中人类动作识别的研究取得了快速进展。大多数现有方法未充分利用无标注的未见类数据实例，而这类数据在现实场景中是大量真实存在并且无标注的。传统深度方法因依赖于大量标注数据，受到了可扩展性问题的困扰。因此，如何准确识别无标注数据是极具现实意义并且亟需解决的。

在学习类和创新类实例之间建立良好的连接以实现准确的知识转移是必要且具有挑战性的，零样本学习(Zero-Shot Learning, ZSL)为这个问题提出了一个解决方案，可以识别和分类从未出现在训练数据中且不属于训练类的视频。该任务旨在借助语义信息，通过利用已见动作类的数据实例获得相应知识，构建已见类和未见类之间的语义关联来识别新的类别实例。其中，语义信息包括高维向量中可见类和不可见类的标签嵌入。这种学习范式类似于人脑对新事物的识别过程。当人类识别一个新对象时，通常通过比较其描述与之前学习的概念之间的相似性来进行认知学习。

然而，大多数现有的零样本学习方法都存在强偏(Bias)问题，即训练阶段看不见(目标)类的实例在测试时往往被归类为所看到的(源)类之一。直推式零样本学习(Transductive ZSL)作为一个新兴的话题[1] [2] [3] [4]，可以在一定程度上缓解强偏问题。在该设定中，目标类的未标记数据是可用来训练的。本文提出一种新的直推式零样本视频动作识别模型，该模型首先将视觉信息映射到语义空间中，然后通过语义空间的最近邻搜索来完成识别任务。针对零样本中固有的强偏问题，在模型的训练阶段，联合使用有标签的源数据(Labeled Source Data)和无标签的目标数据(Unlabeled Target Data)进行直推式学习，以提高模型的识别准确率和泛化能力。

2. 相关工作

2.1. 动作识别研究现状

动作识别研究的是视频中目标的行为，比如判断一个人是在走路，挥手还是跳跃。虽然人体动作识别在许多场景如视频监督、视频推荐和人机交互中得到了广泛的应用，但准确、高效的人体动作识别仍

然是计算机视觉领域中一个具有挑战性的研究问题。早在深度学习兴起之前，密集轨迹特征[5] (Dense Trajectory Method, DTF)及其变体改进密集轨迹[6] (Improved Dense Trajectories, IDT)是最成功的基于手工视觉特征的方法。它们使用梯度直方图(Histogram of Gradient, HoG)，光流直方图(Histogram of Optical Flow, HoF)和运动边界直方图(Motion Boundary Histogram, MBH)描述符。随着深度学习的兴起，研究者开始将卷积神经网络(Convolutional Neural Networks, CNN)应用于视频问题。通过探索卷积运算，时间建模和多流配置来提取视觉特征[7]。Simonyan 等人[8]提出了双流网络，包括空间流和时间流。Tran 等人[9]则将 Ji 等人[10]的工作扩展到更深层的三维网络，称为 C3D 网络。Carira 等人[11]则提出 I3D 网络，综合利用了双流网络和 C3D 的优点，在 UCF101 [12]和 HMDB51 [13]数据集上均取得了最优性能。然而，在某些特定的任务中，大多数深度学习方法专注于识别在训练中已见过的实例，而在实际应用中，现有数据集所覆盖的类别数量有限，并且实例标记过程耗时费力，许多应用程序需要对以往从未见过的无标注类别进行分类。而上述方法均利用大量有标签的训练集进行训练，因此无法在无标签的数据上进行拓展。

2.2. 零样本学习研究现状

受到人类学习和认知新事物思维过程的启发，零样本学习期望能够模仿人类的推理过程，使得计算机具有识别新事物的能力。零样本学习根据训练阶段可获得的数据范围，可分为两类：归纳式零样本学习和直推式性零样本学习。在归纳式零样本学习[14]中，提供了已见类实例的标注和训练过程中未见类标签的语义嵌入；而在直推式零样本学习[15]中，除了已标注的已见类数据和其所有标注的语义嵌入外，还提供了未标注的未见类数据实例。之前关于零样本学习的研究主要集中在图像上，Rohrbach 等人[1]、Fu 等人[2]采用标签语义空间作为知识转移的中介，仅使用源图像向语义空间学习映射，仅使用目标图像生成标签；Guo 等人[3]提出了一种共享模型空间学习方法，利用属性可以有效地在类之间转移知识。近年来，一些学者开始在视频动作识别领域中研究直推式学习方法。Xu 等人[4]引入了基于分组注意力机制的图卷积网络，它将目标类动作与源类别从一个视觉连接图关联到语义空间。在本工作中，直推式的零样本学习用于动作识别，其中标记源数据用于学习视觉嵌入和语义嵌入之间的关系，而未标记的目标数据用于区分源和目标类。在本研究中，基于直推式零样本学习在提高识别准确率的同时有效缓解了零样本学习中的强偏问题。

3. 基于直推式零样本学习的动作识别方法

3.1. 模型结构

本文设计所直推式学习模型结构如图 1 所示。模型主要包括四个部分：视觉特征编码模块、语义空间的构建、视觉语义衔接模块以及视频标签预测模块。下面将会对各个部分进行详细阐述。

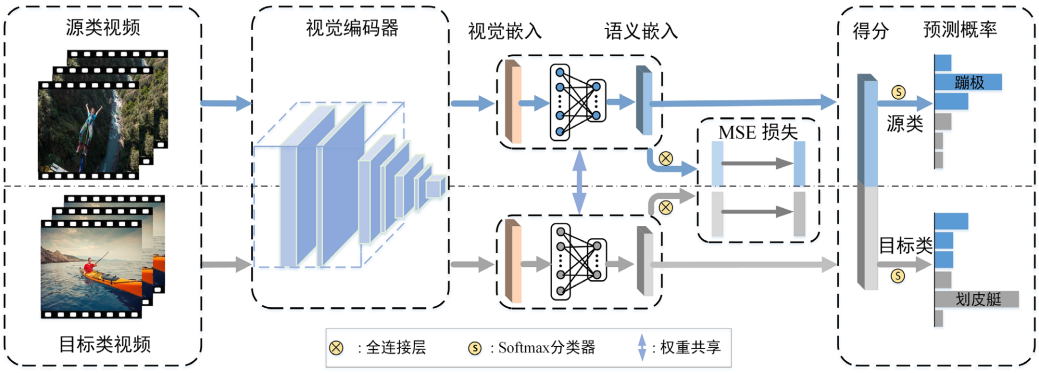

Figure 1. The model framework
图 1. 模型框架图

视觉特征编码模块，给定输入视频，将视频采样 16 帧后的片段作为视觉编码器的输入，采用可训练的三维卷积神经网络，R(2+1)D [16]在各种视频识别任务中表现优异，所以选用 R(2+1)D 作为视觉编码器。R(2+1)D 将三维卷积滤波器分解为独立的时空分量，显著提高了识别精度。不同于之前的工作使用预先训练好的三维卷积神经网络，并在优化过程中修正模型，本方法提出的视觉编码器与其他模块(即视觉语义桥接模块和标签预测模块)一起进行端到端的训练。给定任意一个视频 x，视觉编码器用 $v=F\left(x\mid w^{enc}\right)$ 提取视觉特征 v，其中 $F\left(\cdot\mid w^{enc}\right)$ 为参数为 w^{enc} 的视觉编码器。

语义空间的构建，由于在零样本中没有可用的目标示例，本文通过把源类和目标类的标签嵌入公共语义空间，实现语义连接来将可见的知识转移到不可见的知识。语义空间中的语义嵌入使用每个类名的 Word2Vec 表示形式，语义嵌入的维度是 300。对于多单词类名，将它们的 Word2Vec 嵌入做平均。

视觉语义衔接模块，零样本学习通过建立视觉特征和语义之间的联系来实现对新类别的识别，如图 2 所示。将视觉特征投射到语义空间用几个全连接层实现，每个层后面都有一个 ReLU 非线性激活层，用来连接视觉和语义嵌入。以视觉嵌入 v 为输入，视觉语义桥接模块将 v 映射到语义嵌入 z，$v=B\left(v\mid w^{vsb}\right)$，其中 $B\left(\cdot\mid w^{vsb}\right)$ 为视觉语义桥接模块 w^{vsb} 的参数。将视觉嵌入和语义嵌入连接起来后，通过语义嵌入空间的最近邻搜索来完成识别任务。

视频标签预测模块，以语义嵌入作为输入，语义嵌入和标准化语义嵌入做内积，将其实现为一个单一的全连接层，然后采用归一化指数函数分类器得到识别结果。在此模块中，以投射的语义嵌入 z 为输入，概率向量 p 由 $p=P\left(z\mid w^{pre}\right)$ 产生，其中 $P\left(\cdot\mid w^{pre}\right)$ 为参数为 w^{pre} 的标签预测模块，$p\in R^{(S+T)}$。权重 w^{pre} 是用源类和目标类的规范化语义向量初始化的。与视觉编码器和视觉语义嵌入模块不同，标签预测模块的权值是固定的，在训练阶段不会更新。

3.2. 训练和推理

在零样本学习中，存在两组不相交的类：源类 Y^s 和目标类 Y^t，其中 $Y^s\cap Y^t=\varnothing$，源类中类别数量 S，目标类中类别数量 T。相应的，源数据集 $D^s=\left\{\left(x_i^s y_i^s\right)\right\}_{i=1}^{N^s}$ 由 N^s 个视频组成，每个视频 x_i^s 关联一个标签 y_i^s，其中 $y_i^s\in Y^s$，目标数据集 $D^t=\left\{\left(x_i^t y_i^t\right)\right\}_{i=1}^{N^t}$ 由 N^t 个视频及其标签 $y_i^t\in Y^t$ 组成。对于直推式学习，D^s 和 D^t 的动作实例都是可用的，但是只有源类的标签 Y^s 可用于训练。源类的辅助知识 $\varphi\left(Y^s\right)$ 也可用。零样本动作识别的任务是识别目标类的标签 $y\in Y^t$。

在训练阶段，模型首先充分训练视觉特征编码模块和视觉语义衔接模块，采用均方误差作为损失函数，损失函数可计算为：

$$L_1=\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{z}_i-z_i\right\|^2 \tag{1}$$

其中 N 为样本数，$z_i=\varphi\left(y_i\right)$ 为 Word2Vec 表示的真值。

在充分训练视觉特征编码模块和视觉语义衔接模块的基础上，联合已训练模块和未训练模块，训练全部网络(包括视觉特征编码模块、视觉语义衔接模块和视频标签预测模块)，其使用的损失函数如下：

$$L_2=\frac{1}{N^s}\sum_{i=1}^{N^s}L_{ce}\left(x_i^s\right)+\frac{1}{N^t}\sum_{i=1}^{N^t}L_p\left(x_i^t\right) \tag{2}$$

其中，其中 $L_{ce}\left(\cdot\right)$ 是交叉熵损失，$L_p\left(\cdot\right)$ 是偏差损失，N^s 和 N^t 分别是源类和目标类样本的数目。就偏差损失而言，$L_p\left(x_i^t\right)=-\ln\sum_{i\in y^t}p_i$ 旨在缓解对源类的偏差，其中 p_i 是第 i 类的预测概率。给定一个未标记的目标类实例 x_i^t，如果模型错误地将其分类为源类，则 $L_p\left(\cdot\right)$ 将很大，从而阻止目标类实例映射到源类。

对于输入批处理数据 $\{(x_i)\}_{i=1}^{N^{batch}}$，$N^{batch}=N^s+N^t$ 用作批处理大小。由于预先知道数据实例 x_i 属于源类还是目标类，因此 $\{(x_i)\}_{i=1}^{N^{batch}}$ 可以分为两个子集 $\{(x_j^s)\}_{j=1}^{N^s}$ 和 $\{(x_j^t)\}_{j=1}^{N^t}$。如果 x_i 属于源类，则用 $L_{ce}(\cdot)$ 计算，否则用 $L_p(\cdot)$ 计算。最终损失将分别为 $\{(x_j^s)\}_{j=1}^{N^s}$ 和 $\{(x_j^t)\}_{j=1}^{N^t}$ 上的 $L_{ce}(\cdot)$ 和 $L_p(\cdot)$ 的和。

在推理阶段，零样本学习通过将视觉特征映射到语义中，比较语义空间中的相似性，实现了对新类别的识别。这种学习范式类似于人类通过比较对目标的描述与先前学习的概念之间的相似性来识别新动作的过程。因此，在语义空间中采用最近邻搜索进行分类，公式如下：

$$\hat{label}=\arg\min_{y\in Y}\cos\left(\hat{z},\varphi(y)\right) \tag{3}$$

其中，$\hat{label}$ 为预测类标签，$\varphi(y)$ 为类 y 的语义嵌入，用 Word2Vec 表示，$\cos(\cdot)$ 为余弦距离。给定一个视频 x，其投影语义嵌入 $\hat{z}^s$ 和 $\hat{z}^t$ 分别由 $\hat{z}^s=\vartheta^s\left(x\mid\theta^s\right)$ 和 $\hat{z}^t=\vartheta^s\left(x\mid\varphi\right)$ 生成。根据模型 $\vartheta^t\left(\cdot\mid\theta^t\right)$ 的分类结果，采用最近邻搜索法将 x 的类别进一步细分到源类和目标类中，并映射语义嵌入 $\hat{z}^s$ 和 $\hat{z}^t$。

与现有技术相比，本模型能够更好地弥合视觉语义鸿沟和零样本学习中的固有偏差问题，并且具有识别准确度高、泛化能力强和训练速度快等优点。

4. 实验

本章将详细介绍本文的实验细节。首先对零样本视频动作识别数据集做相关介绍；然后对于本文所涉及的实验环境和相关参数设置，以及模型的评价指标进行介绍；最后，将本模型的实验结果与一些先进方法进行比较，并以模型在 UCF101 数据集上对不同类别的分类准确率为例进行结果展示。

4.1. 数据集

UCF101 数据集[12]于 2012 年引入，数据大都是从 YouTube 收集的 13320 个真实动作视频，包含 101 种人类动作，并把这 101 个动作类别的视频划分为 25 组，每组包含 4 到 7 个动作视频。

HMDB 数据集[13]是当前识别动作研究领域最为重要的几个数据集之一，包含 51 个动作类别共 6849 个片段，每个类别至少包含 101 个片段。

Olympic Sports 数据集[17]中包含了练习不同运动的运动员的视频，采集自 YouTube 网站，包含严重的遮挡、摄像机运动，压缩伪影等。当前版本中包含 16 类体育项目的 783 个视频，每类大约 50 个视频。

4.2. 实验设置

在本文中，采用 PyTorch 中实现的 R(2+1)D [16]作为视觉编码器。和[8] [18]中保持一致，采样视频的 16 帧片段作为视觉编码器的输入。每个类名的语义嵌入使用 Word2Vec，使用 Python 中的 gesim [19]实现，语义嵌入维数为 300。对于含有多个单词的类名，把它们的 Word2Vec 嵌入向量做平均。模型首先在 Kinetics 400 数据集上做预训练。在训练过程中，使用 Adam 进行优化，Batch 大小设置为 16，使用 4 个 NVIDIA TITANX 的 GPU。在预训练过程中，该模型以 1e−3 的学习速率开始训练，在 10 和 20 轮后分别衰减为 1e−3 和 1e−4。在进行直推式学习阶段，学习速率设置为 1e−5，并加入附加的偏置损失[20]以减轻[21]中提到的对已见类别的偏置。

4.3. 评价指标

为对本文模型的效果作出评价，采用准确率(Top-1 Accuracy)，其定义为：

$$acc_y = \frac{1}{|y|}\sum_{i=1}^{y}\frac{\#\text{正确样本数}}{\#\text{所有样本数}} \tag{4}$$

4.4. 对比实验

该方法现有的零样本视频动作识别方法进行了比较，结果如表 1 所示。

Table 1. Comparison with other methods under the transductive setting of ZSL
表 1. 在零样本学习的直推式设定下与其他方法比较

设置	方法	视觉	语义	Olympic Sports	HMDB51	UCF101
归纳式	SJE [22]	FV	A	47.5 ± 14.8	-	12.0 ± 1.2
	SJE [22]	FV	W	28.6 ± 4.9	13.3 ± 2.4	9.9 ± 1.4
	SVE [23]	BoW	W	-	18.0 ± 3.0	12.7 ± 1.6
	PDA [24]	FV	W	44.3 ± 8.1	19.7 ± 1.6	15.8 ± 1.3
	TOM [25]	2Stream	W	-	-	26.8 ± 4.4
	MR [26]	FV	W	35.7 ± 8.8	14.5 ± 2.7	11.7 ± 1.7
	ZSECOC [27]	FV	ECOC	59.8 ± 5.6	22.6 ± 1.2	15.1 ± 1.7
	GA [28]	C3D	A	50.4 ± 11.2	-	22.7 ± 1.2
	GA [28]	C3D	W	34.1 ± 10.1	19.3 ± 2.1	17.3 ± 1.1
	UR [29]	FV	W	-	24.4 ± 1.6	17.5 ± 1.6
	VDS [30]	FV	GloVe	43.9 ± 7.9	25.3 ± 4.5	25.4 ± 3.1
直推式	UDA [31]	FV	A	-	-	13.2 ± 1.9
	UDA [31]	FV	A+W	-	-	14.0 ± 1.8
	SVE [23]	BoW	W	-	21.2 ± 3.0	18.6 ± 2.2
	PDA [24]	FV	W	56.6 ± 7.7	24.8 ± 2.2	22.9 ± 3.3
	MR [26]	FV	W	43.2 ± 8.3	24.1 ± 3.8	22.1 ± 2.5
	BiDiLEL [32]	C3D+IDT	W	-	22.3 ± 1.1	23.0 ± 0.9
	ZSECOC [27]	FV	ECOC	**59.8 ± 5.6**	-	22.6 ± 1.2
	GA [28]	C3D	A	57.9 ± 14.1	-	24.5 ± 2.9
	GA [28]	C3D	W	41.3 ± 11.4	20.7 ± 3.1	20.3 ± 1.9
	UR [29]	FV	W	-	**28.9 ± 1.2**	20.1 ± 1.4
	本方法	C3D	W	46.5 ± 5.3	20.3 ± 1.4	**26.8 ± 1.7**

本文提出的模型与其他先进的零样本视频动作识别方法在归纳式设定(ID)和直推式设定(TD)下进行比较，在识别准确度和和标准差方面，结果如表 1 所示。从结果可以看出，SJE 和 GA 方法在将属性作为语义嵌入的 Olympic Sports 数据集上表现得更好，这是因为人工设计的属性在类别较小的数据集上更有效。VDS 在 Olympic Sports 和 HMDB51 上表现较好，但在 UCF101 上表现较差。ZSECOC 在归纳设置

上在 Olympic Sports 数据集上效果领先，平均准确率为 59.8%，但在 UCF101 上表现较差，说明纠错输出码无法处理复杂的数据分布。

值得注意的是，UCF101 数据集具有 101 个动作类别，远超 HMDB51 数据集的 51 个类别和 Olympic Sports 数据集的 16 个类别，因此现有方法在 UCF101 上大多不能取得较为理想的结果。本文提出的模型显著提高了 UCF101 上传统零样本动作识别方法的基线，在直推式设定下相比当前最好方法改进 2.3%平均准确率。还可以观察到，从 Olympic Sports，HMDB51 到 UCF101，随着数据集规模的增大，所提出模型的优势变得更加明显。该模型利用先验知识，有效地弥补了视觉特征和语义之间的差距，在 UCF101 数据集上取得了更好的性能。

4.5. 结果展示

该方法在 UCF101 数据集上，在不同类别之间进行比较，分类准确率如下图 2 所示。从图中可以看出，在“BlowDryHair”、“Punch”、“Surfing”等动作类别上识别准确率较高(大于 70%)，但是在“FrisbeeCatch”“TableTennisShot”等动作类别上识别效果还有待提高。可以看出，所提出方法在各种类别上均有一定效果，与现有技术相比够能够更好的弥合视觉语义鸿沟和零样本学习中的固有偏差问题。

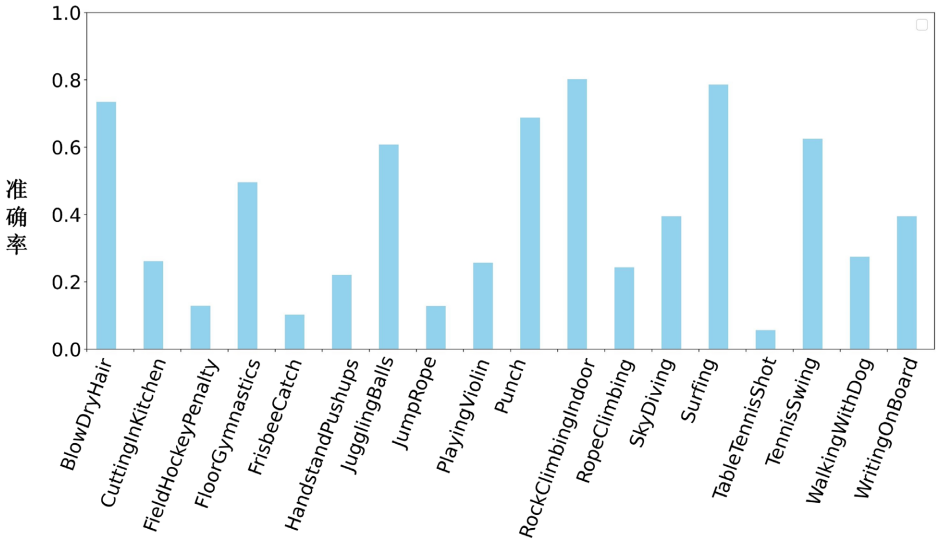

Figure 2. Classification accuracy of different categories of the model on UCF101 dataset
图 2. 模型在 UCF101 数据集上不同类别分类准确率结果展示

5. 结论

在直推式零样本学习模型中，有标记的源数据和无标记的目标数据均可用于训练。从实验结果还可以观察到，随着数据集规模的增大，所提模型的优越性也越来越明显。所提出的基于直推式零样本学习的动作识别方法，有效地缓解了零样本学习中源类和目标类之间固有的强偏问题，最终在 Olympic Sports、HMDB51 以及 UCF101 数据集上的实验结果证明了该模型的有效性。

参考文献

[1] Rohrbach, M., Ebert, S. and Schiele, B. (2013) Transfer Learning in a Transductive Setting. 2013 *NIPS Workshops*, Lake Tahoe, 5-8 December 2013, 46-54.

[2] Fu, Y., Hospedales, T.M., Xiang, T., *et al*. (2015) Transductive Multi-View Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 2332-2345. https://doi.org/10.1109/TPAMI.2015.2408354

[3] Guo, Y., Ding, G., Jin, X., *et al*. (2016) Transductive Zero-Shot Recognition via Shared Model Space Learning. *AAAI*-16: *Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, 12-17 February 2016, 3434-3500.

[4] Xu, Y., Han, C., Qin, J., *et al*. (2021) Transductive Zero-Shot Action Recognition via Visually Connected Graph Convolutional Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 3761-3769. https://doi.org/10.1109/TNNLS.2020.3015848

[5] Wang, Q. and Chen, K. (2020) Multi-Label Zero-Shot Human Action Recognition via Joint Latent Ranking Embedding. *Neural Networks*, **122**, 1-23. https://doi.org/10.1016/j.neunet.2019.09.029

[6] Wang, H., Oneata, D., Verbeek, J.J., *et al*. (2016) A Robust and Efficient Video Representation for Action Recognition. *International Journal of Computer Vision*, **119**, 219-238. https://doi.org/10.1007/s11263-015-0846-5

[7] Kong, Y. and Fu, Y. (2018) Human Action Recognition and Prediction: A Survey. CoRR, abs/1806.11230.

[8] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. 2014 *NIPS Workshops*, Nevada, December 2014, 568-576.

[9] Tran, D., Bourdev, L.D., Fergus, R., *et al*. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[10] Ji, S., Xu, W., Yang, M., *et al*. (2013) 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 221-231. https://doi.org/10.1109/TPAMI.2012.59

[11] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Proceedings CVPR*'17, Honolulu, 21-26 July 2017, 4724-4733. https://doi.org/10.1109/CVPR.2017.502

[12] Soomro, K., Zamir, A.R. and Shah, M. (2012) UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild.

[13] Kuehne, H., Jhuang, H., Garrote, E., *et al*. (2011) HMDB: A Large Video Database for Human Motion Recognition. *ICCV* 2011, Barcelona, 6-13 November 2011, 2556-2563. https://doi.org/10.1109/ICCV.2011.6126543

[14] Xian, Y., Lorenz, T., Schiele, B., *et al*. (2018) Feature Generating Networks for Zero-Shot Learning. *Proceedings CVPR*'18, Salt Lake City, 18-22 June 2018, 5542-5551. https://doi.org/10.1109/CVPR.2018.00581

[15] Verma, V.K. and Rai, P. (2017) A Simple Exponential Family Framework for Zero-Shot Learning. *ECML*/*PKDD*, Skopje, 18-22 September 2017, Vol. 10535, 792-808. https://doi.org/10.1007/978-3-319-71246-8_48

[16] Tran, D., Wang, H., Torresani, L., *et al*. (2018) A Closer Look at Spatiotemporal Convolutions for Action Recognition. *Proceedings CVPR*'18, Salt Lake City, 18-22 June 2018, 6450-6459. https://doi.org/10.1109/CVPR.2018.00675

[17] Niebles, J.C., Chen, C. and Li, F. (2010) Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. *ECCV* 2010: 11*th European Conference on Computer Vision*, Heraklion, 5-11 September 2010, 392-405. https://doi.org/10.1007/978-3-642-15552-9_29

[18] Wang, L., Qiao, Y. and Tang, X. (2015) Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 4305-4314. https://doi.org/10.1109/CVPR.2015.7299059

[19] Tsochantaridis, I., Joachims, T., Hofmann, T., *et al*. (2005) Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*, **6**, 1453-1484.

[20] Song, J., Shen, C., Yang, Y., *et al*. (2018) Transductive Unbiased Embedding for Zero-Shot Learning. *Proceedings CVPR*'18, Salt Lake City, 18-22 June 2018, 1024-1033. https://doi.org/10.1109/CVPR.2018.00113

[21] Gao, J., Zhang, T. and Xu, C. (2019) I Know the Relationships: Zero-Shot Action Recognition via Two-Stream Graph Convolutional Networks and Knowledge Graphs. *The Thirty-Third AAAI Conference on Artificial Intelligence* (*AAAI*-19), Honolulu, 27 January-1 February 2019, 8303-8311. https://doi.org/10.1609/aaai.v33i01.33018303

[22] Akata, Z., Reed, S.E., Walter, D., *et al*. (2015) Evaluation of Output Embeddings for Fine-Grained Image Classification. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 2927-2936. https://doi.org/10.1109/CVPR.2015.7298911

[23] Xu, X., Hospedales, T.M. and Gong, S. (2015) Semantic Embedding Space for Zero-Shot Action Recognition. 2015 *IEEE International Conference on Image Processing*, *ICIP* 2015, Quebec City, 27-30 September 2015, 63-67. https://doi.org/10.1109/ICIP.2015.7350760

[24] Xun, X., Hospedales, T.M. and Gong, S.G. (2016) Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation. *ECCV* 2016, Amsterdam, 8-16 October 2016, 343-359. https://doi.org/10.1007/978-3-319-46475-6_22

[25] Li, Y., Hu, S. and Li, B. (2016) Recognizing Unseen Actions in a Domain-Adapted Embedding Space. 2016 *IEEE International Conference on Image Processing*, Phoenix, 25-28 September 2016, 4195-4199. https://doi.org/10.1109/ICIP.2016.7533150

[26] Xu, X., Hospedales, T.M. and Gong, S. (2017) Transductive Zero-Shot Action Recognition by Word-Vector Embedding. *International Journal of Computer Vision*, **123**, 309-333. https://doi.org/10.1007/s11263-016-0983-5

[27] Qin, J., Liu, L., Shao, L., *et al*. (2017) Zero-Shot Action Recognition with Error-Correcting Output Codes. *Proceedings CVPR*'17, Honolulu, 21-26 July 2017, 1042-1051. https://doi.org/10.1109/CVPR.2017.117

[28] Mishra, A., Verma, V.K., Reddy, M.S.K., *et al*. (2018) A Generative Approach to Zero-Shot and Few-Shot Action Recognition. 2018 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Lake Tahoe, 12-15 March 2018, 372-380. https://doi.org/10.1109/WACV.2018.00047

[29] Zhu, Y., Long, Y., Guan, Y., *et al*. (2018) Towards Universal Representation for Unseen Action Recognition. *Proceedings CVPR*'18, Salt Lake City, 18-22 June 2018, 9436-9445. https://doi.org/10.1109/CVPR.2018.00983

[30] Zhang, C. and Peng, Y. (2018) Visual Data Synthesis via GAN for Zero-Shot Video Classification. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, 13-19 July 2018, 1128-1134. https://doi.org/10.24963/ijcai.2018/157

[31] Kodirov, E., Xiang, T., Fu, Z., *et al*. (2015) Unsupervised Domain Adaptation for Zero-Shot Learning. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 2452-2460. https://doi.org/10.1109/ICCV.2015.282

[32] Wang, Q. and Chen, K. (2017) Zero-Shot Visual Recognition via Bidirectional Latent Embedding. *International Journal of Computer Vision*, **124**, 356-383. https://doi.org/10.1007/s11263-017-1027-5