Interactive Prototype Learning for Egocentric Action Recognition

1、Motivation

由于第一视角视频中会出现很多与action本身无关的物体，因此过往的第一人称动作识别方法通常需要使用object detection或是人体注视信息，这造成了较大的计算和人力开销。为了解决这一问题，作者引入actor的运动信息，设计了end2end的**interactive prototype learning（IPL）** 框架来学习更好的active objects表现。

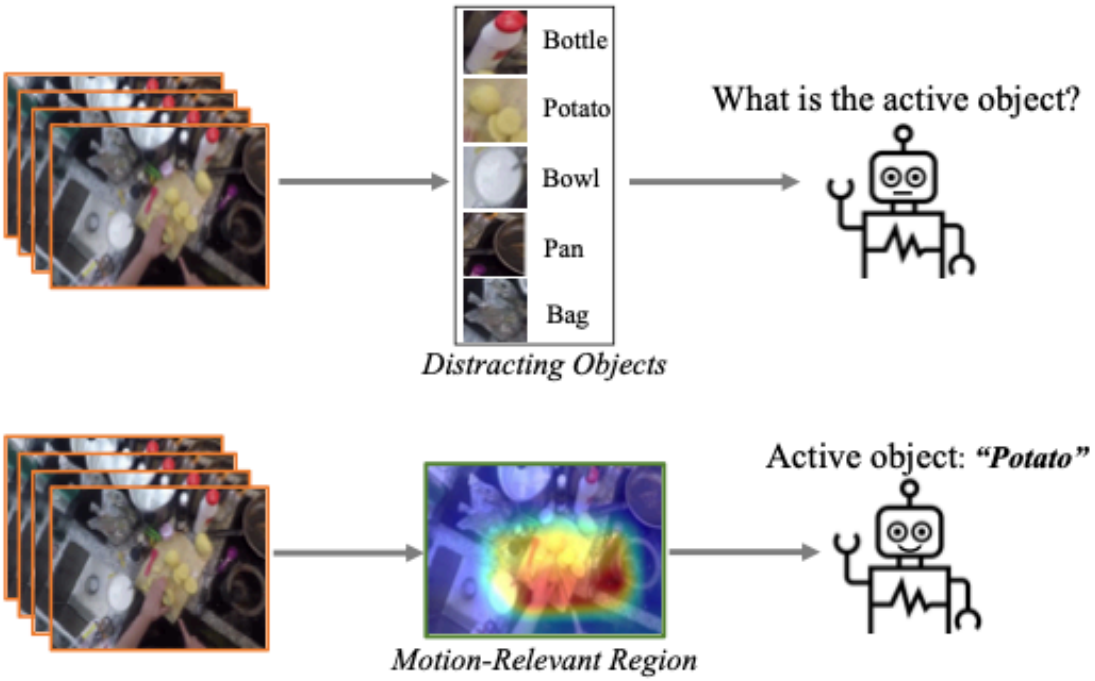
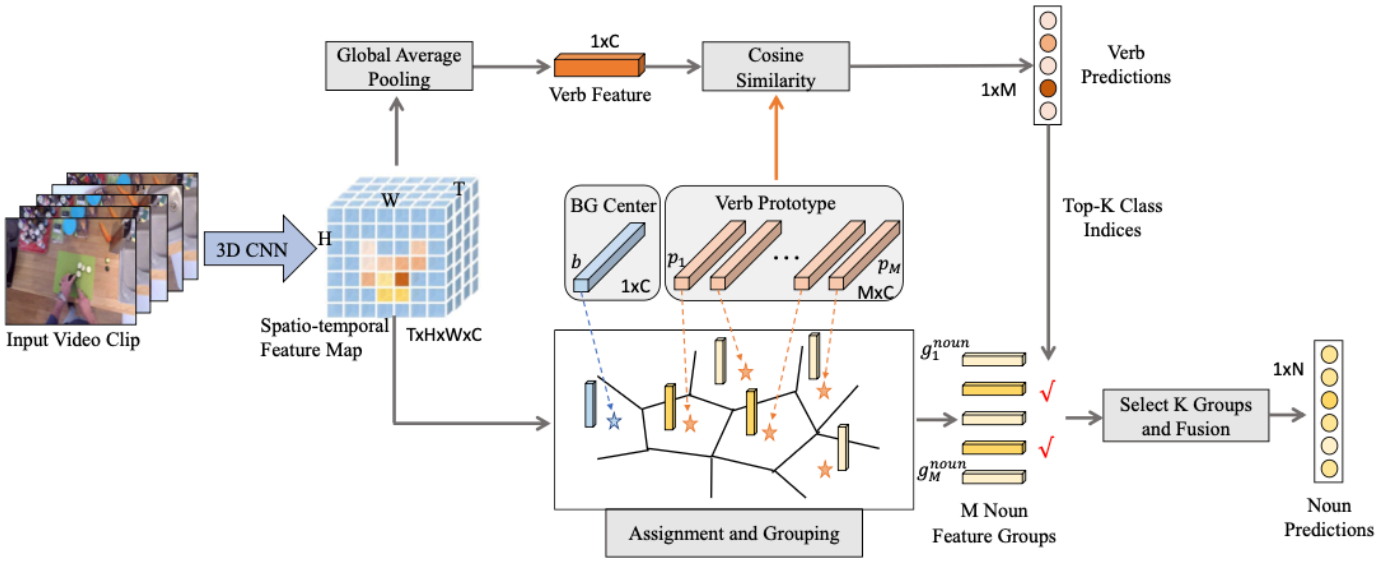


Figure 1. The motivation of **I**nteractive **P**rototype **L**earning (IPL) framework. The *noun* classification is difficult as the active object can be surrounded by a considerable number of distracting objects. Our framework aims to collaboratively learn judicious motion-relevant spatio-temporal features for more accurate *noun* (active object) classification.

2、Iteractive Prototype Learning



2.1 overview

对于输入的视频clip，首先使用3DCNN提取T ×H ×W ×C的feature map。

引入了M个verb prototype表示M个verb的pattern，表示为$P = \{p_1, p_2, \ldots, p_M\} \in \mathbb{R}^{M \times C}$

对verb branch，对feature map进行global avg pooling得到c维度对verb feature。然后将其与M个prototype计算余弦相似度，取相似度最高的动作作为分类结果。

对noun branch，作者引入了两个交互操作让模型学习到更丰富的位置感知特征。在**noun-to-verb配准**操作中，将feature map转换为THW个C维向量，并分配至M个verb prototype外加一个background prototype中（加入background是为了滤掉无用信息）。这将THW个特征分成M+1个组。在**verb-to-noun选择**操作中，选出上面做分类时得到的top-K verb classes对应的组，将组的feature融合起来后进行最后的noun prediction。

2.2 verb classification

verb branch可以用下面的式子表示：

$$v = GAP(\phi_\theta(X))$$

$$q_i = \frac{exp(\bar{v}\bar{p_i}^T/\tau)}{\sum_{j=1}^{M} exp(\bar{v}\bar{p_j}^T/\tau)}$$

其中，$\bar{v} = \frac{v}{||v||}, \bar{p}_i = \frac{p_i}{||p_i||}$

2.3 noun classification

2.3.1 feature assignment and grouping

这一部分的难点在于如何将特征分配到不同的prototypes上。

这里为觉得比较妙的地方是不仅用了M个行为prototypes，还引入了与动作无关的Background prototype。我认为这可能是模型能够将非交互物体与交互物体分开的关键因素。

作者定义了一个soft assignment matrix $A' \in \mathbb{R}^{B \times (M+1)}$，并将feature map拉平成B个C维2D tensor构成的矩阵Z，其中B=T×H×W。A'的计算方法如下：

$$a_{i,j} = \frac{exp(z_i {c_j}^T)}{\sum^{M} +1_{k=1} exp(z_i {c_k}^T)}$$

完成计算后将BG prototype对应的列去掉，得到矩阵$A \in \mathbb{R}^{B \times M}$。

（这个地方不太理解为什么要去掉，因为不去掉好想也没有特别大影响？）

完成后对features进行分组：

$$G = A^T Z$$

我的理解是，A^T的i，j个元素表示对第i个verb prototype，第j个特征对应的强度。因此计算出来的G的i行表示对于第i个verb prototype，与之交互的noun所表现出的C维特征。

但是，作者认为G中既包含了action motion的信息，也包含了active object的信息。因此，进一步地，作者计算了G与P之间的加权残差：

$$g_i^{noun} = g_i - \sum_{k=1}^{B} a_{k,i} p_i$$

其中，a_{k,i}是A中的元素。

2.3.3 group selection and noun classification

完成上面的步骤后，最终得到$G^{noun} = \{g_1^{noun}, \ldots, g_M^{noun}\}$

后面的分类过程比较好理解，不做详细阐述。不过值得注意的是，这里的fusion用了1D卷积+BN+ReLU的组合。**关于最后的分类方法，作者说也用了余弦相似度，但是问题是要跟什么计算余弦相似度呢？这里作者没有做很细致的阐述**

3、Experiment

3.1 数据集

1.EPIC-KITCHENS-55

2.EPIC-KITCHENS-100

3.EGTEA

3.2 Comparison with State of the Arts

1. EPIC-KITCHENS-100

Method	Overall Top-1 Accuracy			Unseen Participants Top-1 Accuracy			Tail Classes Top-1 Accuracy		
	Verb	Noun	Act.	Verb	Noun	Act.	Verb	Noun	Act.
Chance [7]	10.68	1.79	0.55	9.37	1.90	0.59	0.97	0.39	0.12
TSN [39]	59.03	46.78	33.57	53.11	42.02	27.37	26.23	14.73	11.43
TRN [46]	63.28	46.16	35.28	57.54	41.36	29.68	28.17	13.98	12.18
TBN [19]	62.72	47.59	35.48	56.69	43.65	29.27	30.97	19.52	14.10
SlowFast [10]	63.79	48.55	36.81	57.66	42.55	29.27	29.65	17.11	13.45
TSM [24]	65.32	47.80	37.39	59.68	42.51	30.61	30.03	16.96	13.45
IPL I3D	65.66	49.74	38.43	59.12	45.26	32.17	32.17	**20.34**	**15.51**
IPL R(2+1)D-34	65.74	**50.45**	**39.17**	**61.22**	**46.01**	**33.70**	**33.02**	18.97	15.22

Table 1. The comparison with the state-of-the-art methods on the **EPIC-KITCHENS-100** Test set.

Method	Act@1	Verb@1	Noun@1
Chance [7]	0.51	10.42	1.70
TSN [39]	33.19	60.18	46.03
TRN [46]	35.34	65.88	45.43
TBN [19]	36.72	66.00	47.23
SlowFast [10]	38.54	65.56	50.02
TSM [24]	38.27	67.86	49.01
I3D† [3]	37.58	66.84	48.48
IPL I3D	39.87	67.82	50.87 (+**2.39**)
R(2+1)D-34† [12]	37.62	67.28	47.55
IPL R(2+1)D-34	40.98	68.61	51.24 (+**3.69**)

Table 2. The comparison with the baselines and state-of-the-arts on the **EPIC-KITCHENS-100** validation set. "†" indicates our implementation with two separate classifiers for noun and verb.

对I3D，作者采用了RGB+optical flow。而对R(2+1)D-34则只采用RGB。

这里的baseline就是用backbone + 两个FC来分别对noun和verb进行分类。

2. EPIC-KITCHENS-55

Method	Act@1	Verb@1	Noun@1
R50-NL [42]	19.0	49.8	26.1
R(2+1)D-34† [12]	22.5	56.6	32.7
SlowFast [43]	21.9	55.8	27.4
I3D [3]	23.5	59.6	31.3
IPL I3D	24.5	59.8	33.2 (+**1.9**)
R(2+1)D-34 [12]	23.6	60.5	31.1
IPL R(2+1)D-34	25.4	60.7	35.5 (+**4.4**)

Table 3. Comparison of 3D CNN backbones on the **EPIC-KITCHENS-55** validation set. "†" indicates [12] uses two R(2+1)D-34 backbones, one for verb classification and the other for noun. Our "IPL R(2+1)D-34" and "R(2+1)D-34" use a shared backbone for both tasks.

Method	Obj	Act@1	Verb@1	Noun@1	GFLOPs
LFB Max [42]	✓	22.8	52.6	31.8	6664
SAP [40]	✓	25.0	55.9	35.0	2871
IPL R(2+1)D-34	✗	**25.4**	**60.7**	**35.5**	**153**

Table 4. Compare with the state-of-the-art methods using object detection annotations on the **EPIC-KITCHENS-55** validation set.

这里作者还与用object detection的方法进行了对比。

3. EGTEA

该数据集提供了人眼注视的标注

Methods	Mean Class Accuracy			
	Split1	Split2	Split3	Avg
EgoIDT+Gaze [23]	42.55	37.30	37.60	39.13
I3D (joint) [3]	55.76	53.14	53.55	54.15
I3D+Gaze [21]	53.74	50.30	49.63	51.22
I3D+EgoConv [34]	54.19	51.45	49.41	51.68
Ego-RNN-2S [36]	52.40	50.09	49.11	50.53
LSTA-2S [35]	53.00	-	-	-
Mutual Context-2S [15]	55.70	-	-	-
Prob-ATT [22]	56.50	53.52	53.58	54.53
Prob-ATT+Gaze [22]	57.20	53.75	54.13	55.03
I3D†	56.78	54.92	53.94	55.21
IPL I3D	**60.15**	**59.03**	**57.98**	**59.05**

Table 5. The comparison with the state-of-the-art methods on the **EGTEA** dataset. "†" indicates our implementation with two separate classifiers.

3.3 Ablation study

Method	Act@1	Verb@1	Noun@1
R(2+1)D Baseline	37.62	67.28	47.55
R(2+1)D + NetVLAD	38.80	67.39	49.38
IPL R(2+1)D w/o Selection	38.50	66.82	49.68
IPL R(2+1)D w/o BG Center	40.02	68.20	50.30
IPL R(2+1)D	**40.98**	**68.61**	**51.24**

Table 6. Ablation studies on the EPIC-KITCHENS-100 Val. set.

3.4 Qualitative Results

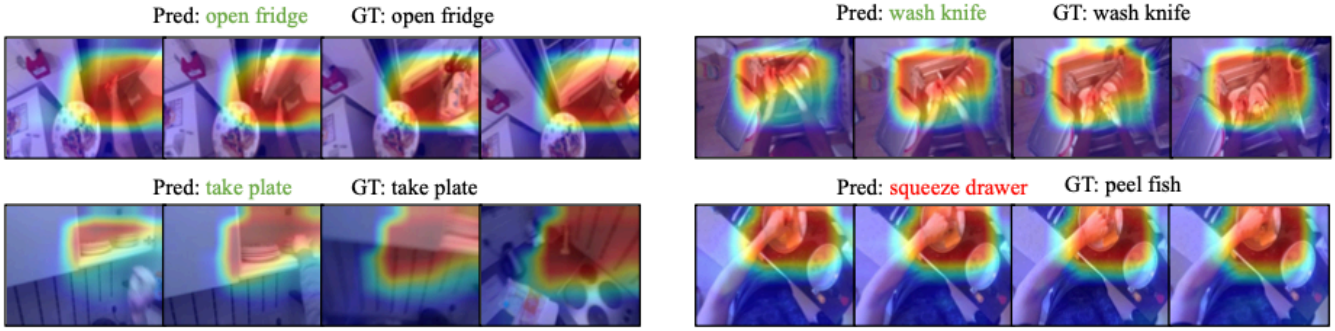


Figure 3. Assignments visualization of the IPL R(2+1)D model. We illustrate the sum of assignments on the top-K verb prototypes for each feature vector on the spatio-temporal feature map. For each input clip, we uniformly sample four frames. Higher assignment values shows in red. We also print the predictions and the ground-truth above the frames (Green for correct predictions and Red for failure cases).

GT	wash spoon	turn-on tap	pour-up oil	skin carrot	put-down knife	scoop coffee	pick-up utensil
Baseline	wash saucepan	turn-on gas	pour-up rice	put-down pasta	pick-up spoon	open lid	pick-up bin
IPL	wash spoon	turn-on tap	pour-up oil	cut carrot	pick-up knife	open coffee	pick-up lid

Figure 4. Qualitative results of the IPL R(2+1)D model and the baseline model.