Symbiotic Attention for Egocentric Action Recognition with Object-centric Alignment

1、Background

任务：egocentric action recognition。按照文章的解释，应该是第一人称的动作识别（videos captured from a first-person viewpoint）。

egocentric action recognition要求模型去区分人正在交互的物体和其他小的物体。

数据集：EPIC-Kitchens。一个action被定义为verb和noun的结合（比如“open door”）。过往的研究中verb和noun的分类通常是分开训练的。verb branch用来对agent正在进行的行为进行分类，而noun branch则是判断出人正在交互的物体。

2、Motivation

过往的研究将verb branch和noun branch分开，只关注到了noun branch和object detection feature之间的交互，而没有关注到verb branch和noun branch之间的交互。然而，一个action是由动作本身和交互的物体共同表征的。即使是人类，只关注物体而忽略了行为同样也是很难预测action的。

因此，作者作出了如下贡献：

- 提出了object-centric feature alignment method将local-aware information集成到两个branches上
- 完成alignment后，得到了一系列候选的verb features和noun features。再通过一个symbiotic attention模块获取与action最相关的feature
- 做了丰富的实验验证模型效果

3、Approach

overview：

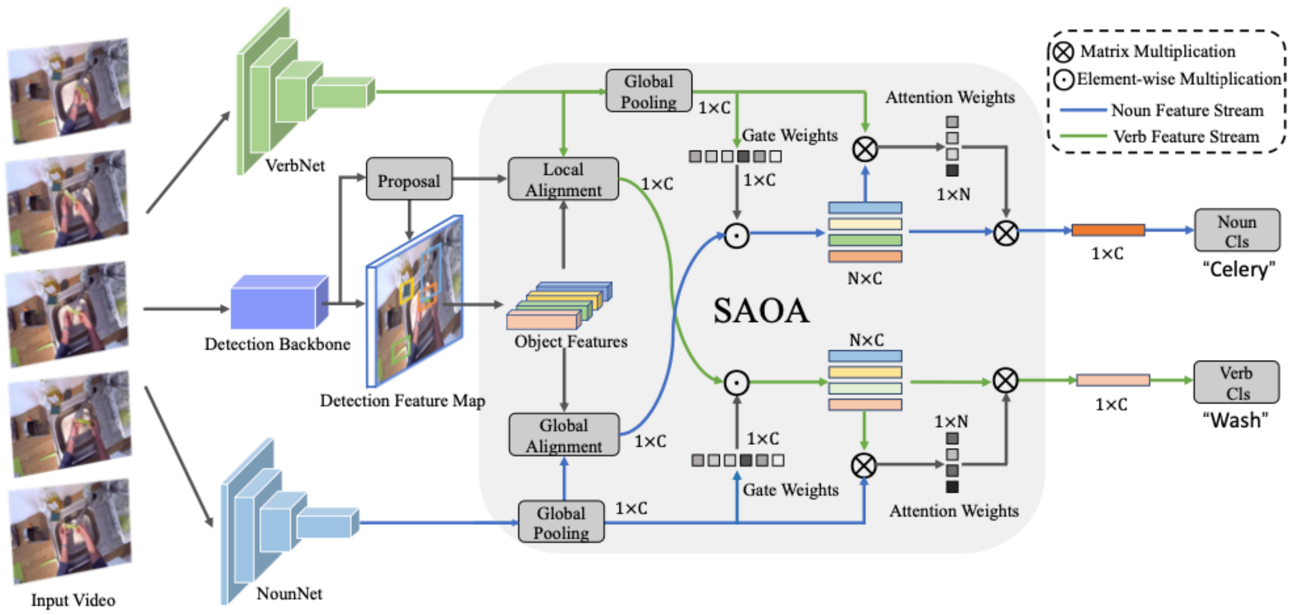

Fig. 2. The proposed SAOA framework. Our framework consists of three feature extractors and one interaction module. The detection model generates a set of local object features and location proposals. This location-aware information is injected to the two branches by an object-centric alignment method For the Verb branch, the feature map is locally aligned with the objects by combining the local motion features with corresponding object detection features. For the Noun branch, the object features are aligned with the global noun representation. Subsequently, the fused features from each branch interact with the global feature from the other branch by a symbiotic attention mechanism. The two object-centric feature matrices are first normalized by a cross-stream gating operation. After that, the matrices are attended by the other branch to select the most action-relevant information. The outputs of SAOA are used to classify the verb and noun, respectively.

3.1 preliminaries

使用两个3D CNN为backbones提取video clip中的verb feature和noun feature。使用Faster R-CNN提取object feature。然后使用object-centric模块将object feature和verb、noun feature分别融合。最后用symbiotic attention模块生成最后的预测结果。

3.2 Object-centric Feature Alignment

之所以要设计这样一个模块，是因为3D CNN提取到的global feature maps无法很好地表征交互信息，需要引入一些local details。考虑到verb branch和noun branch之间语义信息到不同，作者设计了两种不同的机制实现特征融合。

①global alignment for the noun branch

因为noun branch和object feature都表征的是物体的出现情况，gap较小，所以直接使用了global alignment。

首先，对于noun feature map $f^n \in \mathbb{R}^{T*H*W*C}$，进行global average pooling得到$f_g^n \in \mathbb{R}^{1*C}$。然后通过下面的式子进行global alignment：

$$f_i^{\hat{n}} = ReLU(W^n f_g^{nT} + W_o^n f_i^{oT} + b^n)\ , i \in [1 \dots N]$$

其中$W^n \in \mathbb{R}^{C*C}, W_o^n \in \mathbb{R}^{C*C_1}, b_n \in \mathbb{R}^{1*C_1}$。因此最终$f^{\hat{n}} \in \mathbb{R}^{N*C}$，每一行表示一个object-centric fearture。

②local alignment for the verb branch

对于每个f_i^o，object detection同时生成了位置信息$l_i = (x_i^0, y_i^0, x_i^1, y_i^1)$。local alignment可以表示为：

$$f_i^v = ROIAlign(f^v, l_i)$$
$$f_i^{\hat{v}} = ReLU(W^v f_i^{vT} + W_o^v f_i^{oT} + b^v)\ , i \in [1 \dots N]$$

3.3 Symbiotic Attention

这个部分将前面得到的两个branch的object-centric feature进行交互。首先使用一个门控机制将另一个branch的feature标准化，然后使用注意力机制进行融合。可以直接用下图来表示：

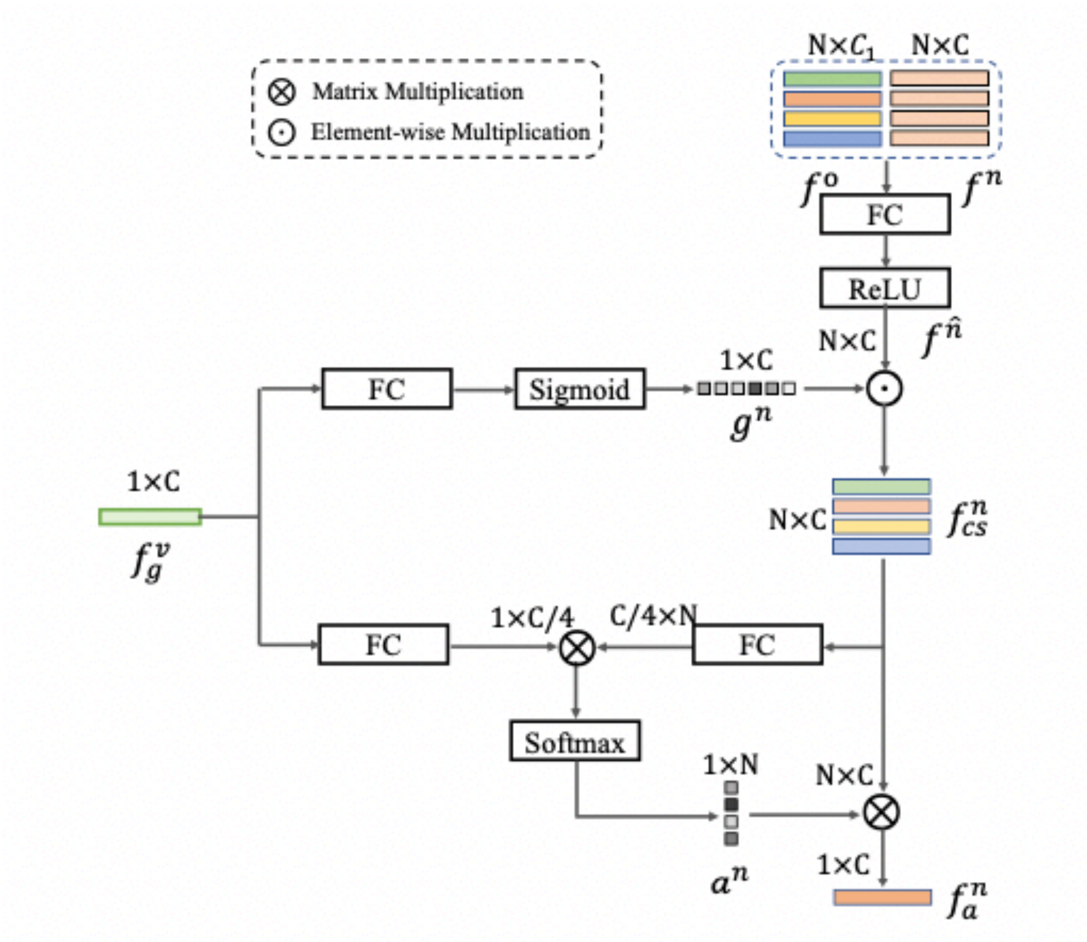

Fig. 3. The illustration of symbiotic attention on the noun branch. The object-centric noun feature matrix is first normalized by the global verb feature. After that, the feature matrix interacts with the global verb feature to generate attention weights. The final noun representation is produced by weighted-summing the normalized object-centric features.

①cross-stream gating

设计这个gating的模块是为了过滤掉action-irrelevant信息，用更正确的特征指导学习。

In noun classification：

$$f_g^v = GAP(f^v)$$
$$g^n = Sigmoid(W_g^n f_g^{vT} + b_g)$$
$$f_{cs}^n = g^n \odot f^{\hat{n}}$$

In verb classification：

$$f_g^n = GAP(f^n)$$
$$g^v = Sigmoid(W_g^v f_g^{nT} + b_g^v)$$
$$f_{cs}^v = g^v \odot f^{\hat{v}}$$

②Action-attended Relation Module

In noun classification：

$$a^n = Softmax(f_g^v W_v^a W_{cn}^a f_{cs}^{nT})$$
$$f_a^n = a^n f_{cs}^n$$

In verb classification：

$$a^v = Softmax(f_g^n W_n^a W_{cv}^a f_{cs}^{vT})$$
$$f_a^v = a^v f_{cs}^v$$

3.4 Training and Objectives

We use Faster R-CNN with the ResNeXt-101-FPN backbone as our object detector. Following the training procedure in [15], we first pre-train the detector on Visual Genome [53] and then finetune it on EPIC-Kitchens object detection set. For VerbNet and NounNet, we adopt 3D Resnet-50 [22] and I3D [5] as our backbones. The two networks are both initialized with Kinetics pre-trained weights. In the first stage, we individually train the VerbNet and NounNet with the corresponding CrossEntropy Loss, *i.e.*, $\mathcal{L}^v$ and $\mathcal{L}^n$.

$$\mathcal{L}^n = CrossEntropy(f_a^n, y^n), \quad (12)$$
$$\mathcal{L}^v = CrossEntropy(f_a^v, y^v). \quad (13)$$

After the base training stage, we freeze the weights of the backbone and cascade our SAOA module. The objective for the second stage is the same as the base training stage, and only the weights of SAOA are optimized.

4、Experiment

action calculation：引入先验，滤掉完全不可能出现的action（如：open the knife）

$$P(action = y) = \mu(y^v, y^n) P(verb = y^v) P(noun = y^n)$$

Ablation study of symbiotic attention

TABLE 1
The effectiveness of Symbiotic Attention (SA) for **verb prediction** and **noun prediction** on the EPIC-Kitchens validation set. "ARM" denotes the Action-attended Relation Module. "CSG" denotes the Cross-Stream Gating.

Methods	Verb Top-1	Noun Top-1
Baseline	54.6	23.8
SA w/o CSG	57.0	32.6
SA w/o Gating	57.2	33.6
SA w/o Cross-Stream	57.4	33.2
SA w/o ARM	56.6	32.7
SA	**57.7**	**34.8**

为什么SA w/o CSG没有SA w/o Cross-Stream好呢?

SA outperform other aggregation operations / The effectiveness of the global alignment for noun classification

TABLE 2
Comparisons between our symbiotic attention and other aggregation methods for **noun prediction** on the EPIC-Kitchens validation set. "Noun" denotes the global feature from NounNet. "Det Feat" is the location-aware object features.

Methods	Top-1 Accuracy
Det Feat+Avg Pooling	24.5
Det Feat+Max Pooling	25.6
SA (Det Feat only)	30.4
Noun + Det Feat	31.2
SA + Local Alignment	33.6
SA + Global Alignment	**34.8**

The effectiveness of the local alignment for verb classification

TABLE 3
Ablation study for **verb prediction** using **RGB** data as inputs. We evaluate the comparisons two backbones, *i.e.*, R-50 and I3D. The top-1 results are reported on the EPIC-Kitchens validation set. "Det Feat" denotes the object detection feature. "Det Box" denotes the location of the object detection proposal.

Methods	Verb	Noun	Det Feat	Det Box	R-50	I3D
Baseline (RGB)	✓	-	-	-	54.6	53.2
Verb+Noun Fusion (RGB)	✓	✓	-	-	54.7	53.7
SAP (RGB)	✓	✓	✓	-	55.9	54.3
SAOA (RGB)	✓	✓	✓	✓	**57.7**	**55.1**

Benefit of the multi-modal fusion

TABLE 4
Two-stream SAOA for both verb classification and noun classification.

Methods	Verb Top-1	Noun Top-1
Our SAOA (RGB+Obj)	55.1	34.7
Our SAOA (Flow+Obj)	56.9	35.0
Our SAOA (RGB+Flow+Obj)	**60.4**	**37.4**

Comparison with SOTA results

TABLE 5
The comparison with the baseline models and state-of-the-art methods on the EPIC-Kitchens dataset. "Obj" indicates the method leverages the information from the object detection model. ↑ indicates the improvement of our method compared to the baseline.

Method	Input Type	Pre-training	Actions		Verbs		Nouns	
			top-1	top-5	top-1	top-5	top-1	top-5
Validation								
ORN [19]	RGB+Obj	ImageNet	-	-	40.9	-	-	-
R(2+1)D-34 [55]	RGB	IG-Kinetics	22.5	39.2	56.6	**83.5**	32.7	55.5
LFB Max [15]	RGB+Obj	Kinetics+ImageNet	22.8	41.1	52.6	81.2	31.8	56.8
SAP (R-50) [21]	RGB+Obj	Kinetics	25.0	44.7	55.9	81.9	35.0	60.4
Baseline (R-50)	RGB	Kinetics	19.5	36.0	54.6	80.9	23.8	45.1
SAOA (R-50)	RGB+Obj	Kinetics	25.7 (6.2↑)	45.9	57.7	82.3	34.8	59.7
Baseline (R-50)	Flow	Kinetics	16.6	32.8	53.2	79.6	19.7	40.7
SAOA (R-50)	Flow+Obj	Kinetics	24.7 (8.1↑)	43.0	56.1	81.3	33.6	58.7
Baseline (R-50)	RGB+Flow	Kinetics	22.0	40.2	59.3	83.3	27.7	50.9
Our SAOA (R-50)	RGB+Flow+Obj	Kinetics	27.9 (5.9↑)	47.5	**61.0**	**83.8**	36.1	61.6
Baseline (I3D)	RGB	Kinetics+ImageNet	20.5	39.2	53.2	80.4	26.2	51.3
Our SAOA (I3D)	RGB+Obj	Kinetics+ImageNet	24.3 (3.8↑)	44.3	55.1	80.1	34.7	61.4
Baseline (I3D)	Flow	Kinetics+ImageNet	17.9	35.6	54.5	79.9	22.7	45.6
Our SAOA (I3D)	FLow+Obj	Kinetics+ImageNet	25.2 (7.3↑)	43.1	56.9	79.7	35.0	59.7
Baseline (I3D)	RGB+Flow	Kinetics+ImageNet	23.3	43.1	59.7	83.2	29.9	56.0
Our SAOA (I3D)	RGB+Flow+Obj	Kinetics+ImageNet	**28.8** (5.5↑)	**48.4**	60.4	82.8	**37.4**	**63.8**
Test seen								
TSN RGB [56]	RGB	ImageNet	22.4	44.8	48.0	87.0	38.9	65.5
TSN Flow [56]	Flow	ImageNet	16.8	33.8	51.7	84.6	26.8	50.6
TSN Fusion [56]	RGB+Flow	ImageNet	25.4	45.7	54.7	87.2	40.1	65.8
R(2+1)D-34 [55]	RGB	IG-Kinetics	34.4	54.2	63.3	87.5	46.3	69.6
LSTA [33]	RGB+Flow	ImageNet	30.2	-	-	-	-	-
LFB Max [15]	RGB+Obj	Kinetics+ImageNet	32.7	55.3	60.0	88.4	45.0	71.8
TBN [20]	RGB+Flow	Kinetics+ImageNet	30.3	51.8	60.9	89.7	42.9	68.6
TBN [20]	RGB+Flow+Audio	Kinetics+ImageNet	34.8	56.7	64.8	**90.7**	46.0	71.3
SAP R-50 [21]	RGB+Obj	Kinetics	34.8	55.9	63.2	86.1	48.3	71.5
Our SAOA (R-50)	RGB+Obj	Kinetics	37.0	58.3	64.0	88.0	**49.6**	**73.2**
Our SAOA (I3D)	RGB+Obj	Kinetics+ImageNet	33.8	55.3	63.6	87.4	46.1	70.0
Our SAOA (I3D)	Flow+Obj	Kinetics+ImageNet	33.4	54.7	63.8	86.8	45.7	69.2
Our SAOA (I3D)	RGB+Flow+Obj	Kinetics+ImageNet	**37.7**	**59.2**	**67.6**	89.2	47.8	71.8
Test Unseen								
TSN RGB [56]	RGB	ImageNet	11.3	26.3	36.5	74.4	22.6	46.9
TSN Flow [56]	Flow	ImageNet	13.5	27.5	47.4	77.0	21.2	42.5
TSN Fusion [56]	RGB+Flow	ImageNet	14.8	29.8	46.1	76.7	24.3	49.3
R(2+1)D-34 [55]	RGB	IG-Kinetics	23.7	39.1	55.5	80.9	33.6	56.7
LSTA [33]	RGB+Flow	ImageNet	15.9	-	-	-	-	-
LFB Max [15]	RGB+Obj	Kinetics+ImageNet	21.2	39.4	50.9	77.6	31.5	57.8
TBN [20]	RGB+Flow	Kinetics+ImageNet	16.8	32.6	49.6	78.4	25.7	50.9
TBN [20]	RGB+Flow+Audio	Kinetics+ImageNet	19.1	36.5	52.7	79.9	27.9	53.8
SAP R-50 [21]	RGB+Obj	Kinetics	23.9	40.5	53.2	78.2	33.0	58.0
Our SAOA (R-50)	RGB+Obj	Kinetics	23.3	41.2	55.1	79.9	32.3	57.1
Our SAOA (I3D)	RGB+Obj	Kinetics+ImageNet	21.9	42.1	52.9	79.9	31.7	58.5
Our SAOA (I3D)	Flow+Obj	Kinetics+ImageNet	23.2	42.4	55.5	80.1	32.6	58.1
Our SAOA (I3D)	RGB+Flow+Obj	Kinetics+ImageNet	**25.8**	**45.1**	**58.1**	82.6	**34.4**	**60.4**