GroupFormer: Group Activity Recognition with Clustered Spatial-Temporal Transformer(ICCV 2021)

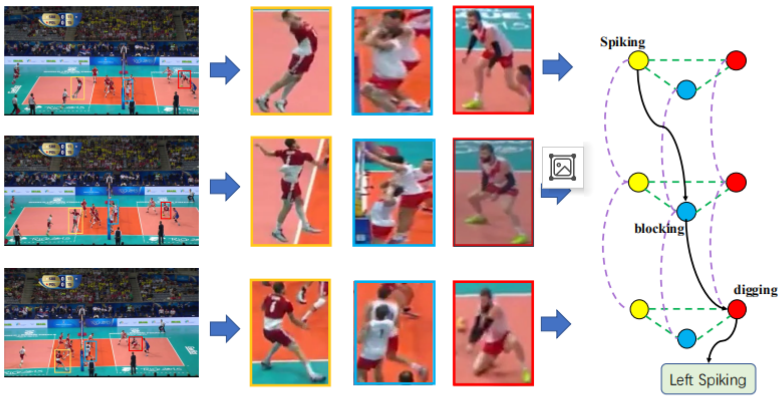


Figure 1. Examples of a clip centered at the annotated frame. The actors with 'spiking', 'blocking', 'digging' actions perform temporally, but they may perform strong spatial-temporal dependencies, which shows the importance of considering spatial and temporal interactions for reasoning about the 'Left Spiking' activity.

引言分析

论文首先说明利用个体关联来推理集体行为是很有挑战性的。有很多方法也在尝试去捕获这种关联信息。近期的很多方法都用到了注意力机制来对个体关联进行建模。但是上述方法也存在缺陷：

1. 无法对时空情景信息作为一个整体来进行建模。

   如上图所示，时间信息于空间信息是强关联的，因此需要将它们作为一个整体来进行考虑

2. 没有基于他们直接的关联关系进行分组

   不是所有的角色都对事件的类别起作用，存在一部分关键角色，他们的贡献更多。

核心亮点

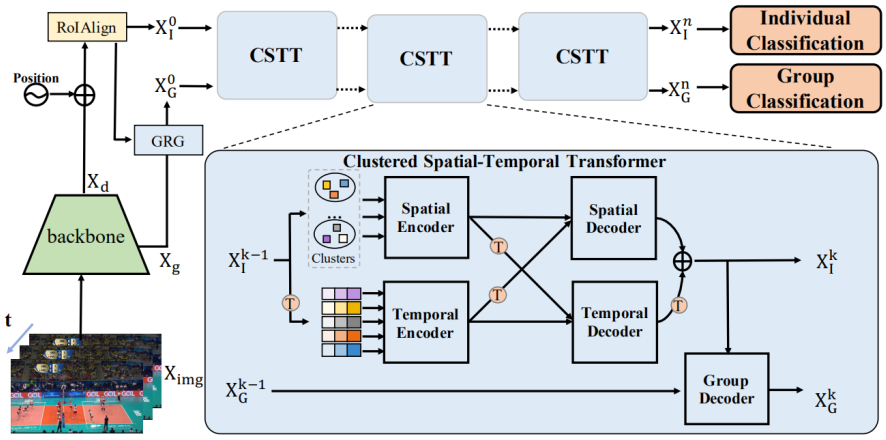


Figure 2. Illustration of our proposed GroupFomer. It contains three main components: 1) a CNN backbone that extracts feature representation of video clips. 2) a Group Representation Generator that initializes the group representation from individual and scene features. 3) a Clustered Spatial-Temporal Transformer that models the spatial-temporal relations and refines the group representation and individual representation.

组特征生成器GRG

该模块是用来初始化组表示的。这个模块结合了场景特征和个体特征，其想法是通过visual tokens来总结视频中的信息，分别得到场景token和个体token，再将他们融合得到最终的组特征表示。

聚簇时空transformer CSTT

如图，STT包含了两个encoder并行化地生成时间和空间特征。之后两个decoders通过一种交叉的方式来对时间和空间特征进行解码。最后使用一个group解码器来增强组特征表示。

- 编码器：

我们使用encoders来嵌入时间和空间情景信息。公式如下：

$$\mathbf{Q}^{(t)} = \mathbf{X}_\mathbf{I}^{(\mathbf{t})} W_{tq}, \mathbf{K}^{(t)} = \mathbf{X}_\mathbf{I}^{(\mathbf{t})} W_{tk}, \mathbf{V}^{(t)} = \mathbf{X}_\mathbf{I}^{(\mathbf{t})} W_{tv} \quad (1)$$

$$\mathbf{V}'^{(t)} = \text{softmax}(\frac{\mathbf{Q}^{(t)} \mathbf{K}^{(t)^T}}{\sqrt{D}}) \mathbf{V}^{(t)} + \mathbf{V}^{(t)} \quad (2)$$

$$\mathbf{V}''^{(t)} = \text{FFN}(\mathbf{V}'^{(t)}) \quad (3)$$

最终空间编码器得到的输出大小为$V_s \in \mathbb{R}^{T \times N \times D}$，同理可得出时间编码器的输出$V_t \in \mathbb{R}^{N \times T \times D}$。

- 解码器

个体解码器是用来将空间和时间情景信息整体地进行考虑。最终两个解码器的输出融合得到了增强的个体表示X_I。

组解码器利用增强的个体表示来强化组特征。在实际中，STT模块可以进行多次堆叠，以得到最佳的建模效果。

- 聚簇注意力机制

STT使用了全连接的注意力机制，但是这样其实计算了很多冗余的联系。因此为了关注于比较关键的组联系。作者设计了聚簇注意力块，该模块可以对个体进行聚簇，并利用组内和组间联系，从而大大降低了计算量。具体来说，他们定义了C个质心向量（簇中心）。对于组间联系可以用簇中心向量$M \in \mathbb{R}^{C \times D}$来代表。

实验结果

SOTA

Method	Flow	Backbone	Group Activity	Individual Action
HDTM[25]		AlexNet	81.9	-
SBGAR[31]	✓	Inception-v3	67.6	-
CERN[37]		VGG16	83.3	69.1
stagNet[35]		VGG16	89.3	-
HRN [24]		VGG19	89.5	-
SSU [6]		Inception-v3	90.6	81.8
ARG [47]		Inception-v3	92.5	83.0
CRM [5]	✓	I3D	93.0	-
Gavrilyuk*et al.* [20]	✓	I3D	93.0	83.7
Gavrilyuk*et al.* [20]	✓	I3D+HRnet	94.4	**85.9**
Ehsanpour *et al.* [18]	✓	I3D	93.1	83.3
Pramono *et al.* [33]	✓	I3D	94.1	81.9
Pramono *et al.* [33]	✓	I3D+Pose+FPN	95.0	83.1
Ours w/o GRG		Inception-v3	93.4	83.2
Ours		Inception-v3	94.1	83.7
Ours	✓	I3D	94.9	84.0
Ours	✓	I3D+Pose	**95.7**	85.6

Table 1. Comparisons with the state-of-the-art methods on Volleyball dataset in terms of Acc.%. "Flow" denotes additional optical flow input.

单单是使用rgb输入，性能效果就比之前的大部分方法好了。

消融实验

- 时空关系建模的效果

Manner	Group Activity	Individual Action
Baseline	91.0	82.1
Spatial only	91.8	82.2
Stacked	92.6	82.8
Parallel	92.2	82.9
Ours	**94.1**	**83.7**

Table 3. Ablation study on different variants architectures.

- 聚簇方法的效果

Clusters	Group Activity	Individual Action
1(w/o cluster)	93.4	83.1
2	93.7	83.6
3	93.8	**83.8**
4	**94.1**	83.7
6	93.4	83.5

Table 4. Comparisons of different clusters choices on Volleyball dataset. Cluster set to 1 demonstrate that we adopt original Spatial-Temporal Transformer.

Intra-Attn	Inter-Attn	Group Activity	Individual Action
		93.4	83.1
✓		93.8	83.4
	✓	93.6	83.5
✓	✓	**94.1**	**83.7**

Table 5. Comparisons of different cluster attention combinations. Intra-Attn and Inter-Attn denote intra-group attention and inter-group attention respectively.

- 堆叠的CSTT块数量的效果

Blocks	Group Activity	Individual Action
0	91.0	82.1
1	93.6	83.4
2	93.8	83.7
3	**94.1**	**83.7**
4	93.9	83.6

Table 6. Comparisons of different setting choices for the number of CSTT blocks.

可视化

- t-SNE

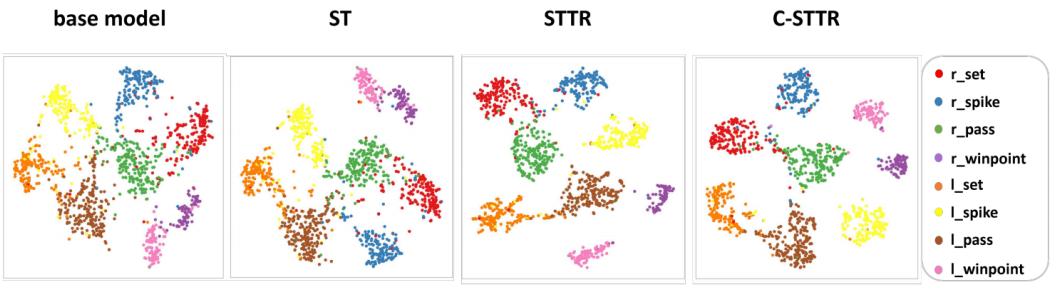


Figure 3. Feature embedding visualizations of the validation set of Volleyball dataset using t-SNE [42] by different model variants. Each clip is visualized as a point and clips belonging to the same group activity have the same color. Best viewed in color.

代码实验阅读

获取光流

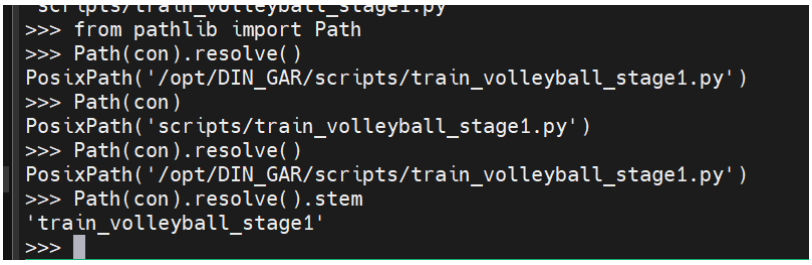
- 配光流网络的环境，获取光流数据

在官方提供的MPI-Sintel数据集上跑了下网络的推理：

```
[0.014s] batch_size: 8
[0.014s] crop_size: [256, 256]
[0.014s] fp16: False
[0.014s] fp16_scale: 1024.0
[0.014s] gradient_clip: None
[0.014s] inference: True
[0.014s] inference_batch_size: 1
[0.014s] inference_dataset: MpiSintelClean
[0.014s] inference_dataset_replicates: 1
[0.014s] inference_dataset_root: ../data/MPI-Sintel/training
[0.014s] inference_n_batches: -1
[0.014s] inference_size: [-1, -1]
[0.014s] inference_visualize: True
[0.014s] log_frequency: 1
[0.014s] loss: L1Loss
[0.014s] model: FlowNet2
[0.014s] model_batchNorm: False
[0.014s] model_div_flow: 20.0
[0.014s] name: run
[0.014s] no_cuda: False
[0.014s] number_gpus: 2
[0.014s] number_workers: 8
[0.014s] optimizer: Adam
[0.014s] optimizer_amsgrad: False
[0.014s] optimizer_betas: (0.9, 0.999)
[0.014s] optimizer_eps: 1e-08
[0.014s] optimizer_lr: 0.001
[0.014s] optimizer_weight_decay: 0
[0.015s] render_validation: False
[0.015s] resume: ./FlowNet2_checkpoint.pth.tar
[0.015s] rgb_max: 255.0
[0.015s] save: ./work
[0.015s] save_flow: True
```

下图是对生成光流图的一个可视化：

获取expriment dir




```
# basedir: experiment dir
config['basedir'] = os.getcwd() + '/experiments/' + Path(args.config).resolve().stem
```