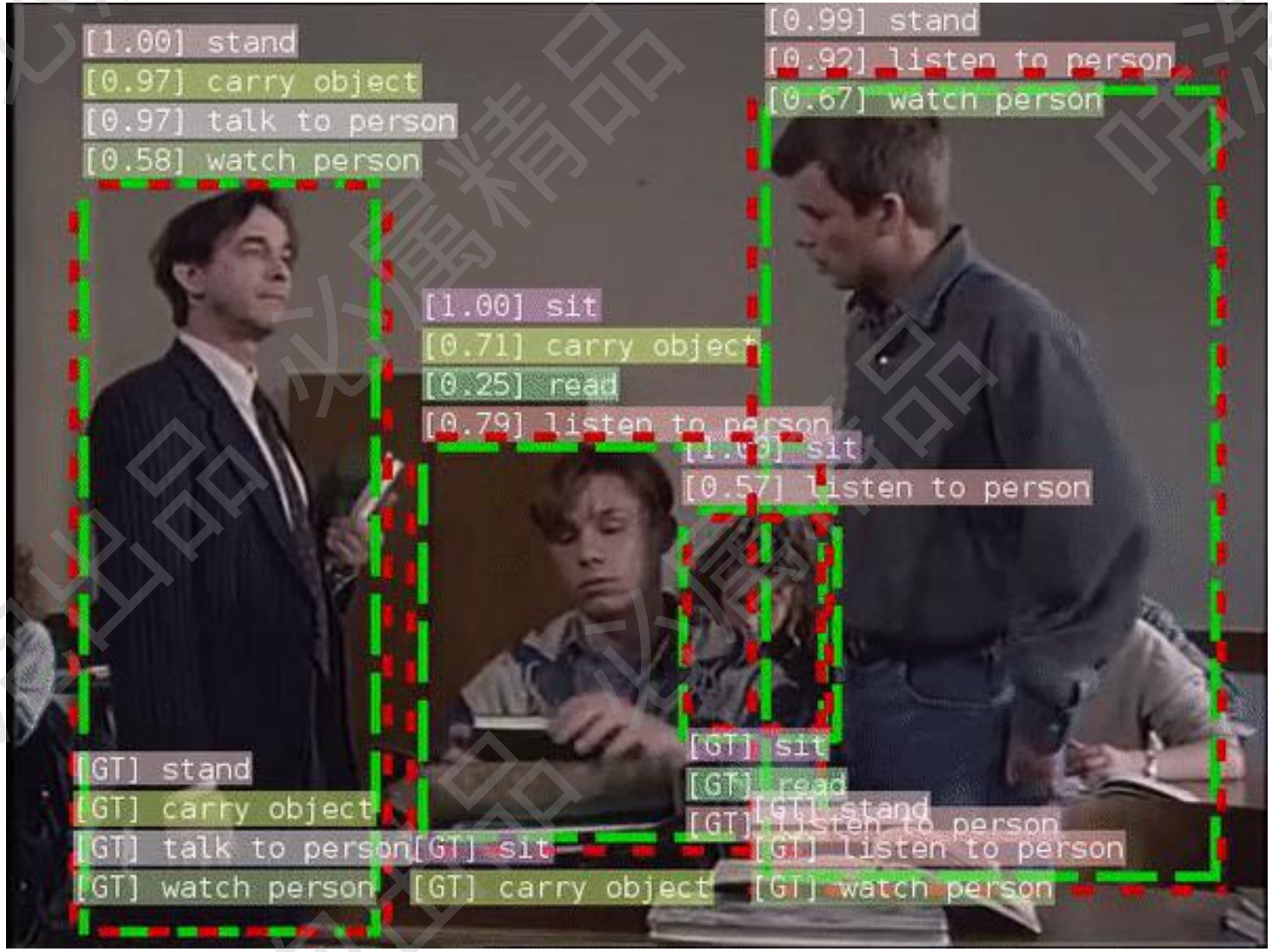
slowfast

- 基本思想
  - 动作在变，环境不变
  - 如何获取动作信息
  - 如何获取环境信息
  - 他们俩该怎么融合呢？



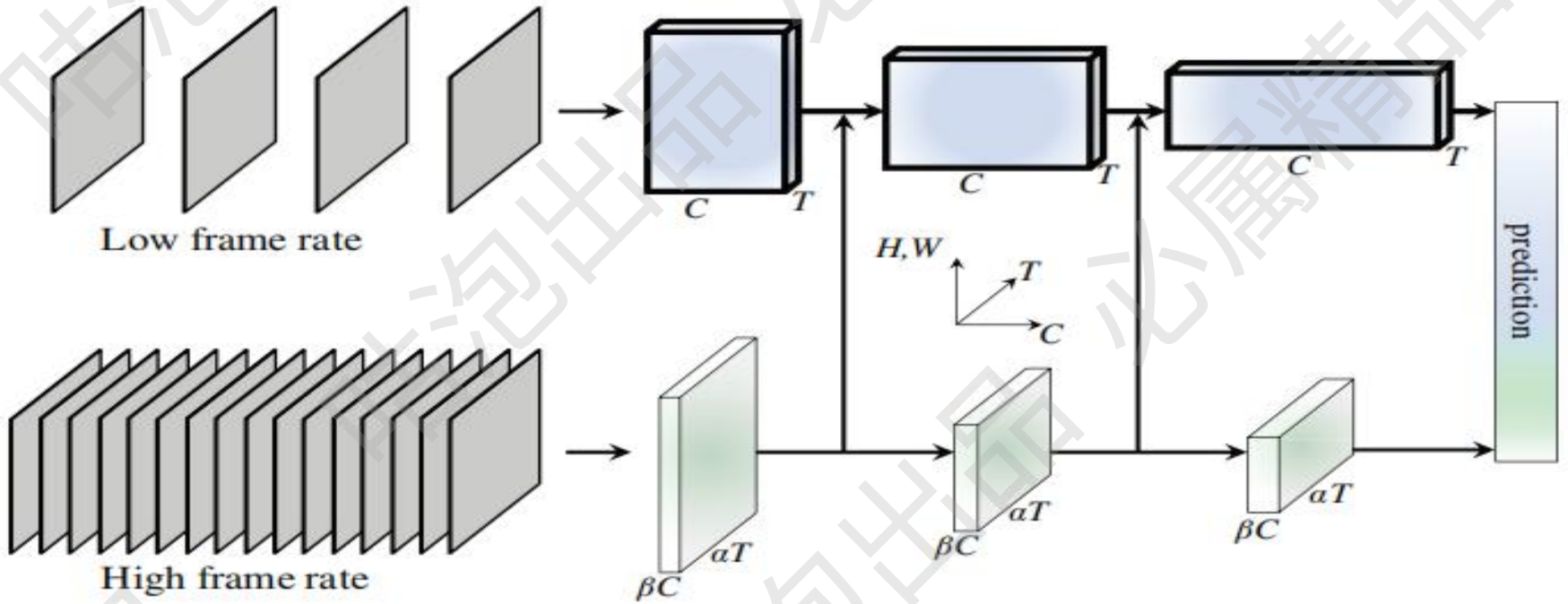
slowfast

- 主要贡献
  - 一个通用的行为识别框架（facebook），自己的项目轻松套用
  - 对比实验很丰富，网络结构设计及其预训练模均提供
  - 思想比较直接，高频与低频特征通吃，直接融合特征进行预测
  - 源码资源丰富，直接可以当作模板的项目

slowfast

- 核心网络结构

1.分别获取高频与低频图像数据；2.分别进行特征提取；3.特征融合；4.预测



slowfast

- 网络结构细节
  - datalayer：对视频进行采样
  - 不同stride得到不同帧数数据
  - 需注意stride的H和W相同
  - 输出结果：slow:4;fast:32

stage	*Slow* pathway	*Fast* pathway	output sizes $T\times S^2$
raw clip	-	-	64×224^2
data layer	stride 16, 1^2	stride **2**, 1^2	*Slow* : 4×224^2 *Fast* : 32×224^2
conv_1	1×7^2, 64 stride 1, 2^2	5×7^2, 8 stride 1, 2^2	*Slow* : 4×112^2 *Fast* : 32×112^2
pool_1	1×3^2 max stride 1, 2^2	1×3^2 max stride 1, 2^2	*Slow* : 4×56^2 *Fast* : 32×56^2
res_2	$\begin{bmatrix}1\times1^2, 64\\1\times3^2, 64\\1\times1^2, 256\end{bmatrix}\times3$	$\begin{bmatrix}3\times1^2, 8\\1\times3^2, 8\\1\times1^2, 32\end{bmatrix}\times3$	*Slow* : 4×56^2 *Fast* : 32×56^2
res_3	$\begin{bmatrix}1\times1^2, 128\\1\times3^2, 128\\1\times1^2, 512\end{bmatrix}\times4$	$\begin{bmatrix}3\times1^2, 16\\1\times3^2, 16\\1\times1^2, 64\end{bmatrix}\times4$	*Slow* : 4×28^2 *Fast* : 32×28^2
res_4	$\begin{bmatrix}3\times1^2, 256\\1\times3^2, 256\\1\times1^2, 1024\end{bmatrix}\times6$	$\begin{bmatrix}3\times1^2, 32\\1\times3^2, 32\\1\times1^2, 128\end{bmatrix}\times6$	*Slow* : 4×14^2 *Fast* : 32×14^2
res_5	$\begin{bmatrix}3\times1^2, 512\\1\times3^2, 512\\1\times1^2, 2048\end{bmatrix}\times3$	$\begin{bmatrix}3\times1^2, 64\\1\times3^2, 64\\1\times1^2, 256\end{bmatrix}\times3$	*Slow* : 4×7^2 *Fast* : 32×7^2
global average pool, concate, fc			# classes

slowfast

- 网络结构细节
  - resnet层：特征提取
  - slow与fast提取特征目的不同
  - 均使用3D卷积计算
  - fast计算要更轻量级

stage	*Slow* pathway	*Fast* pathway	output sizes $T\times S^2$
raw clip	-	-	64×224^2
data layer	stride 16, 1^2	stride **2**, 1^2	*Slow* : 4×224^2 *Fast* : 32×224^2
conv_1	1×7^2, 64 stride 1, 2^2	5×7^2, 8 stride 1, 2^2	*Slow* : 4×112^2 *Fast* : 32×112^2
pool_1	1×3^2 max stride 1, 2^2	1×3^2 max stride 1, 2^2	*Slow* : 4×56^2 *Fast* : 32×56^2
res_2	$\begin{bmatrix}1\times1^2, 64\\1\times3^2, 64\\1\times1^2, 256\end{bmatrix}\times3$	$\begin{bmatrix}3\times1^2, 8\\1\times3^2, 8\\1\times1^2, 32\end{bmatrix}\times3$	*Slow* : 4×56^2 *Fast* : 32×56^2
res_3	$\begin{bmatrix}1\times1^2, 128\\1\times3^2, 128\\1\times1^2, 512\end{bmatrix}\times4$	$\begin{bmatrix}3\times1^2, 16\\1\times3^2, 16\\1\times1^2, 64\end{bmatrix}\times4$	*Slow* : 4×28^2 *Fast* : 32×28^2
res_4	$\begin{bmatrix}3\times1^2, 256\\1\times3^2, 256\\1\times1^2, 1024\end{bmatrix}\times6$	$\begin{bmatrix}3\times1^2, 32\\1\times3^2, 32\\1\times1^2, 128\end{bmatrix}\times6$	*Slow* : 4×14^2 *Fast* : 32×14^2
res_5	$\begin{bmatrix}3\times1^2, 512\\1\times3^2, 512\\1\times1^2, 2048\end{bmatrix}\times3$	$\begin{bmatrix}3\times1^2, 64\\1\times3^2, 64\\1\times1^2, 256\end{bmatrix}\times3$	*Slow* : 4×7^2 *Fast* : 32×7^2
global average pool, concate, fc			# classes

slowfast

- 特征融合

- slow与fast的特征图如何融合呢？ slow: $\{T, S^2, C\}$
  fast: $\{\alpha T, S^2, \beta C\}$

- 文中给出3种方案，然后选择了最直接了当的

(i) *Time-to-channel*: We reshape and transpose $\{\alpha T, S^2, \beta C\}$ into $\{T, S^2, \alpha\beta C\}$, meaning that we pack all α frames into the channels of one frame.

(ii) *Time-strided sampling*: We simply sample one out of every α frames, so $\{\alpha T, S^2, \beta C\}$ becomes $\{T, S^2, \beta C\}$.

(iii) *Time-strided convolution*: We perform a 3D convolution of a 5×1^2 kernel with $2\beta C$ output channels and stride $= \alpha$.

slowfast

效果分析

效果还是非常不错的，原论文也做了多项对比实验

model	flow	pretrain	top-1	top-5	GFLOPs×views
I3D [5]		ImageNet	72.1	90.3	108 × N/A
Two-Stream I3D [5]	✓	ImageNet	75.7	92.0	216 × N/A
S3D-G [61]	✓	ImageNet	77.2	93.0	143 × N/A
Nonlocal R50 [56]		ImageNet	76.5	92.6	282 × 30
Nonlocal R101 [56]		ImageNet	77.7	93.3	359 × 30
R(2+1)D Flow [50]	✓	-	67.5	87.2	152 × 115
STC [9]		-	68.7	88.5	N/A × N/A
ARTNet [54]		-	69.2	88.3	23.5 × 250
S3D [61]		-	69.4	89.1	66.4 × N/A
ECO [63]		-	70.0	89.4	N/A × N/A
I3D [5]	✓	-	71.6	90.0	216 × N/A
R(2+1)D [50]		-	72.0	90.0	152 × 115
R(2+1)D [50]	✓	-	73.9	90.9	304 × 115
SlowFast 4×16, R50		-	75.6	92.1	36.1 × 30
SlowFast 8×8, R50		-	77.0	92.6	65.7 × 30
SlowFast 8×8, R101		-	77.9	93.2	106 × 30
SlowFast 16×8, R101		-	78.9	93.5	213 × 30
SlowFast 16×8, R101+NL		-	**79.8**	**93.9**	234 × 30