AI CHALLENGER 2018
全球AI挑战赛

短视频实时分类

参赛队：SSS（原火箭少女101）

周瑶　中山大学

马平川　南开大学

孟天健　匹兹堡大学

比赛难点

- 难点: 为了同时满足准确率和实时性
  - 可能不需要多种模态，如光流，音频等
    - 提取视频里面的光流信息需要额外的计算代价

比赛难点

- 难点: 为了同时满足准确率和实时性
  - 可能不需要多种模态，如光流，音频等
    - 提取视频里面的光流信息需要额外的计算代价
  - 可能不需要较大的2D或者3D模型
    - 大模型的参数量和flops都比较大

比赛难点

- 难点: 为了同时满足准确率和实时性
  - 可能不需要多种模态，如光流，音频等
    - 提取视频里面的光流信息需要额外的计算代价
  - 可能不需要较大的2D或者3D模型
    - 大模型的参数量和flops都比较大

又快又小的高性能视频理解模型

更快更高效的视频解码方案

解决方法

- 快速高效的视频在线解码方案

解决方法

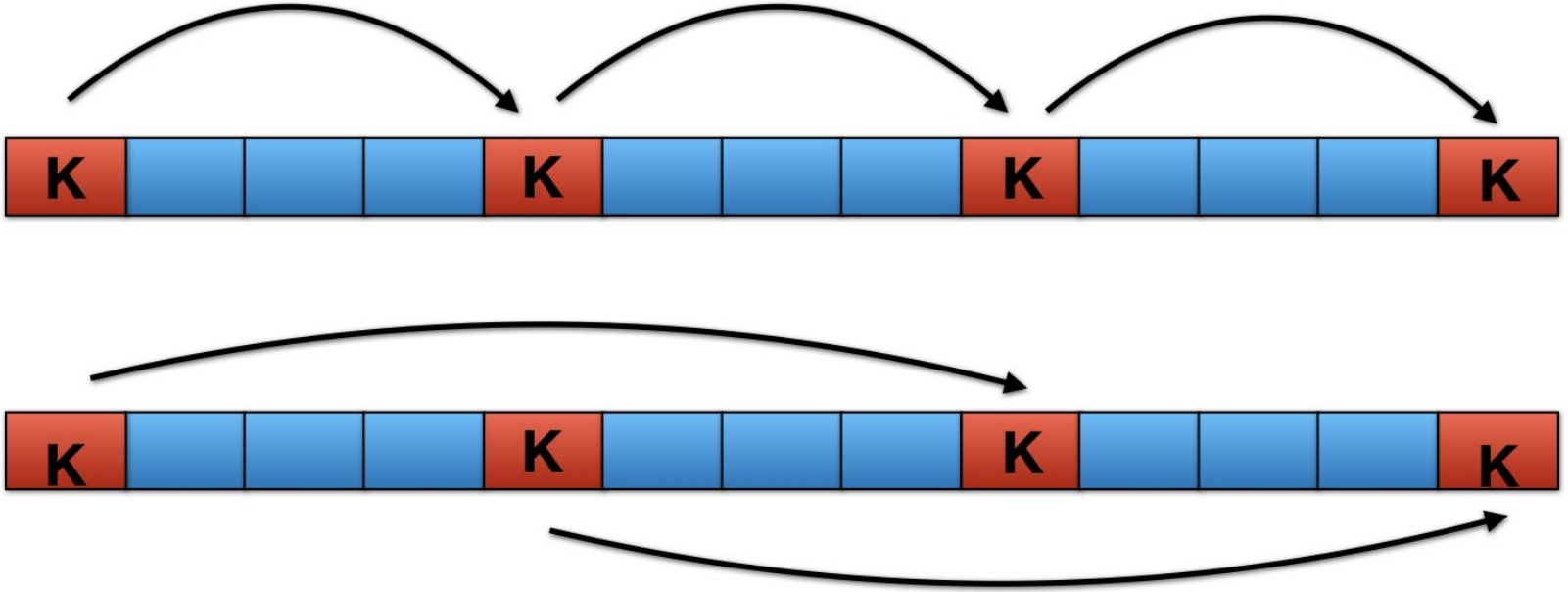
- 快速高效的视频在线解码方案
  - 从流行的FFMPEG到GPU解码的NVVL[1]

	FFMPEG解码	NVVL解码
读4帧	37.3ms	11.0ms
读6帧	65.8ms	12.2ms

Intel® Core™ i7-6700K CPU @ 4.00GHz × 8

[1]https://github.com/NVIDIA/nvvl

跳跃式关键帧解码



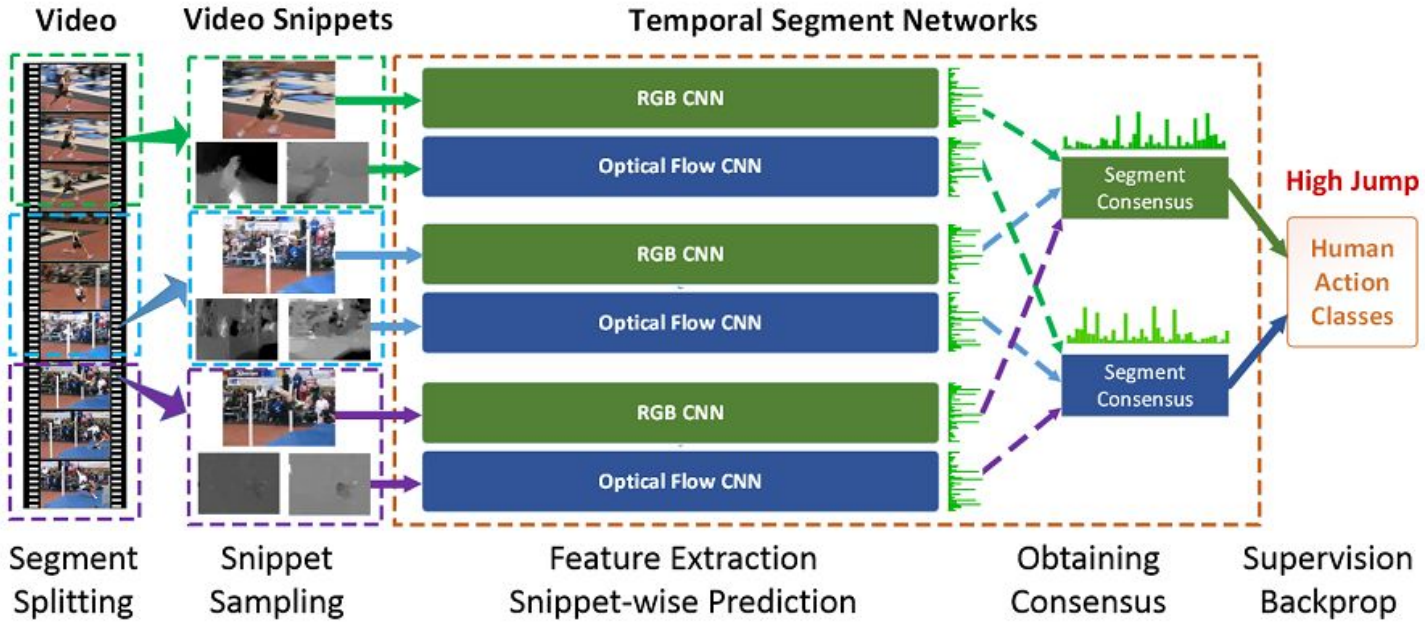
解决方法

- 快速高效的视频在线解码方案
- 数据预处理和数据增强
- 更小更快速的视频理解模型

解决方法

- 快速高效的视频在线解码方案
- 数据预处理和数据增强
- 更小更快速的视频理解模型
  - 基于性能最好的2D视频理解模型-TSN[2]

[2]Temporal Segment Networks: Towards Good Practices for Deep Action Recognition


Video
Video Snippets
Temporal Segment Networks
RGB CNN
Optical Flow CNN
RGB CNN
Optical Flow CNN
RGB CNN
Optical Flow CNN
Segment Consensus
Segment Consensus
High Jump
Human Action Classes
Segment Splitting
Snippet Sampling
Feature Extraction Snippet-wise Prediction
Obtaining Consensus
Supervision Backprop

Backbone 模型	Top-1 准确率	Inference 时间
ResNet34	90.05%	11.68ms
ResNet50	91.30%	23.48ms
ResNet101	91.37%	30.32ms
Inceptionv3	90.62%	32.46ms
DenseNet121	89.49%	20.12ms
NasNet_mobile	88.04%	25.49ms
Peleenet_exp	88.38%	13.08ms

正确率是使用25帧，256*256的结果，inference时间是使用4帧的时间

解决方法

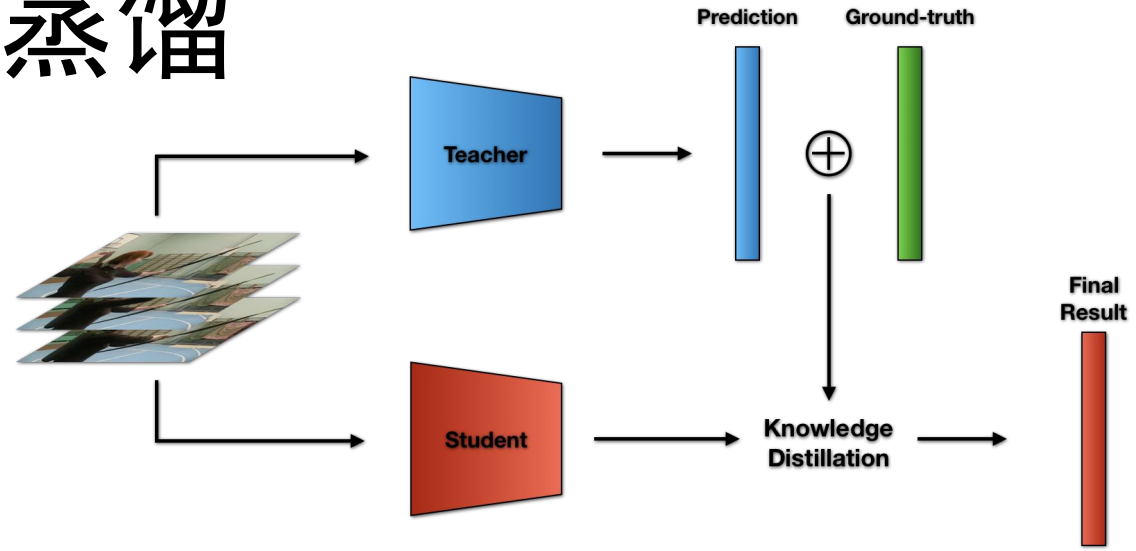
- 快速高效的视频在线解码方案
- 数据预处理和数据增强
- 更小更快速的视频理解模型
- 模型压缩和加速

解决方法

- 快速高效的视频在线解码方案
- 数据预处理和数据增强
- 更小更快速的视频理解模型
- 模型压缩和加速
  - 视频知识蒸馏[3]，使用大模型蒸馏小模型
  - TensorRT异步融合+模型校正

[3]Distilling the Knowledge in a Neural Network

知识蒸馏



$$\mathcal{L}_f = \frac{1}{N_f}((1-\alpha)\sum_{i}^{C}\sum_{j}^{N_f} H(\sigma(I_{s,ij}), y_i) + \alpha T^2 \sum_{i}^{C}\sum_{j}^{N_f} H(\sigma(I_{s,ij}/T), \sigma(I_{t,ij}/T))),$$

Student	Teacher	UCF101	Kinetics400
ResNet-18	w/o	82.55%	63.71%
	ResNet-50	85.41%	64.88%
	ResNet-101	**86.71%**	65.01%
	Inception-v3	86.66%	**65.09%**
ResNet-50	w/o	86.57%	70.94%
	ResNet-101	**88.02%**	71.43%
	Inception-v3	87.84%	**71.45%**

在学术数据集UCF101和Kinetics上验证有效性

TensorRT

INT8量化

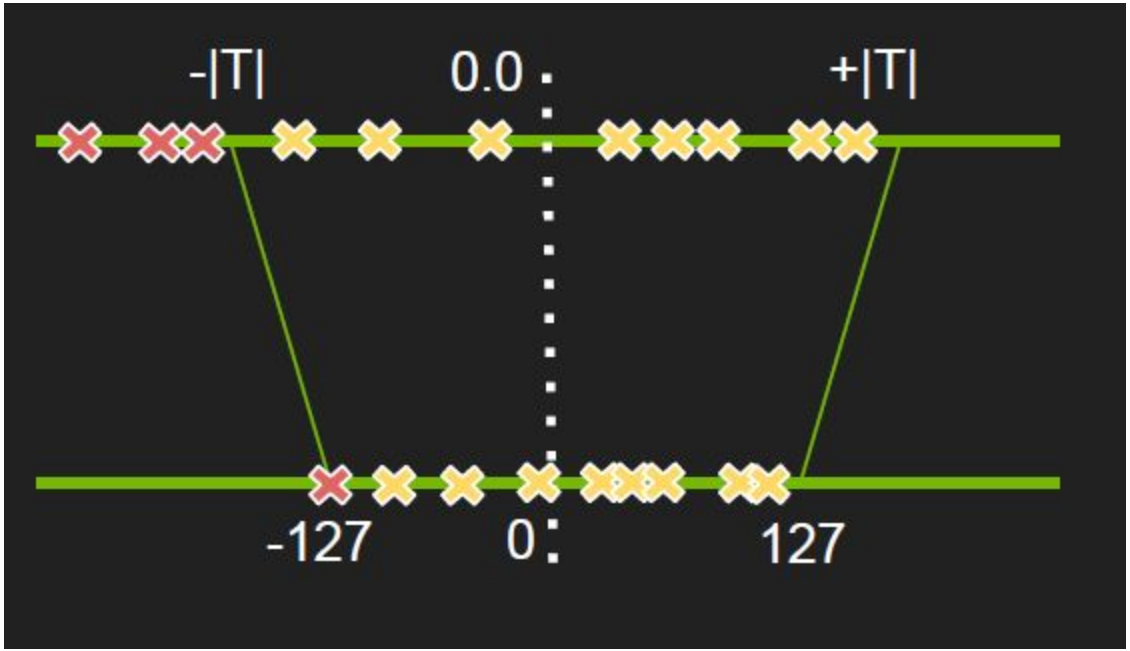
张量合并

自适应平台

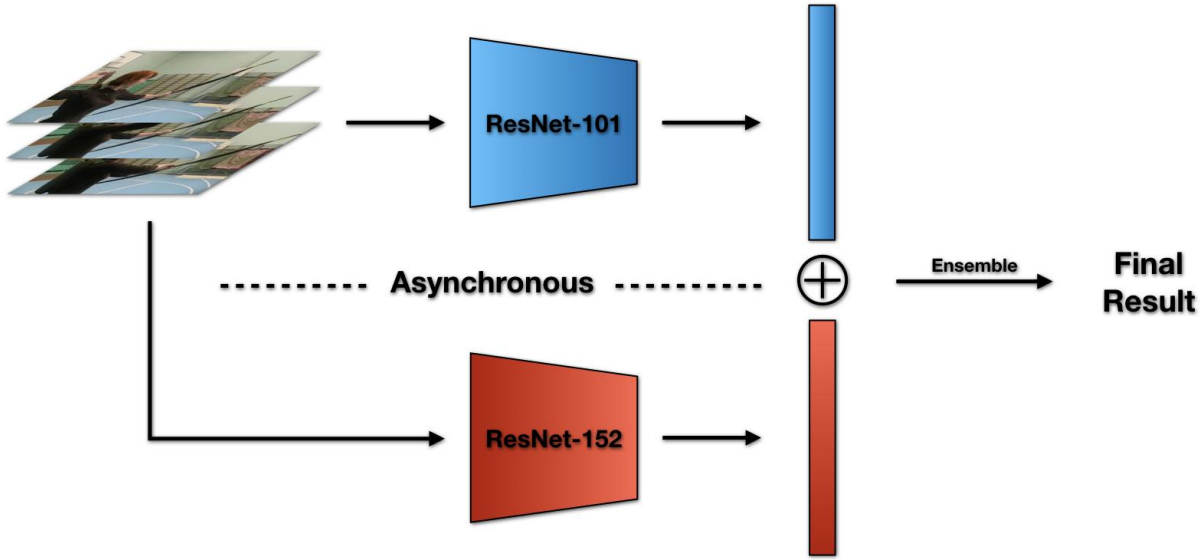
动态内存

异步多流

Float32 to INT8



异步模型融合



解决方法

- 快速高效的视频在线解码方案
- 数据预处理和数据增强
- 更小更快速的视频理解模型
- 模型压缩和加速
- 多标签解决方案

多标签解决方案

- 基于效率考虑, 采取了最简单的阈值方案
  - 引入每个短视频最多三个标签的先验, 对于模型输出的logits先取top-3, 再过sigmoid层并使用[0, 0.5, 0.5]的阈值, 使输出的标签在1到3个之间

比赛结果

- 最好提交模型
  - 蒸馏过的ResNet50模型和ResNet152模型异步融合，NVVL读4帧输入，使用TensorRT进行INT8量化。
- 未来工作
  - 训练更好模型
  - 使用时序融合方案

谢谢大家！