

RACV 2016

计算机视觉研究与应用  
创新论坛

September 18-20, 2016 Shanghai China



# 爱奇艺视频标注竞赛 答辩

参赛队: DeepBrain

姓名: 司晨阳 王君波 李林

# 一、队伍介绍

---

- 队名：DeepBrain
- 姓名：司晨阳 王君波 李林
- 单位：中国科学院自动化所
- 成绩：第四名

## 二、赛题理解

---

- 问题描述：

- 通过某种算法对输入的视频进行自动的标签识别。输出该视频的所属类别标签包括：音乐、篮球、赛车等20个类别。
- 某个输入的视频可能属于多个类别，即对应于多个标签，要求算法在检测出主要的标签同时，也能同时判断出该视频所属的其他类别。

## 三、数据观察

---

- 数据

- 视频分辨率宽度最小355，高度最小288
- 视频长度 1~5分钟
- 视频格式 mp4
- 每个视频都包含预先提取好的\*.mp3文件和\*.jpg关键帧图片

- 标注

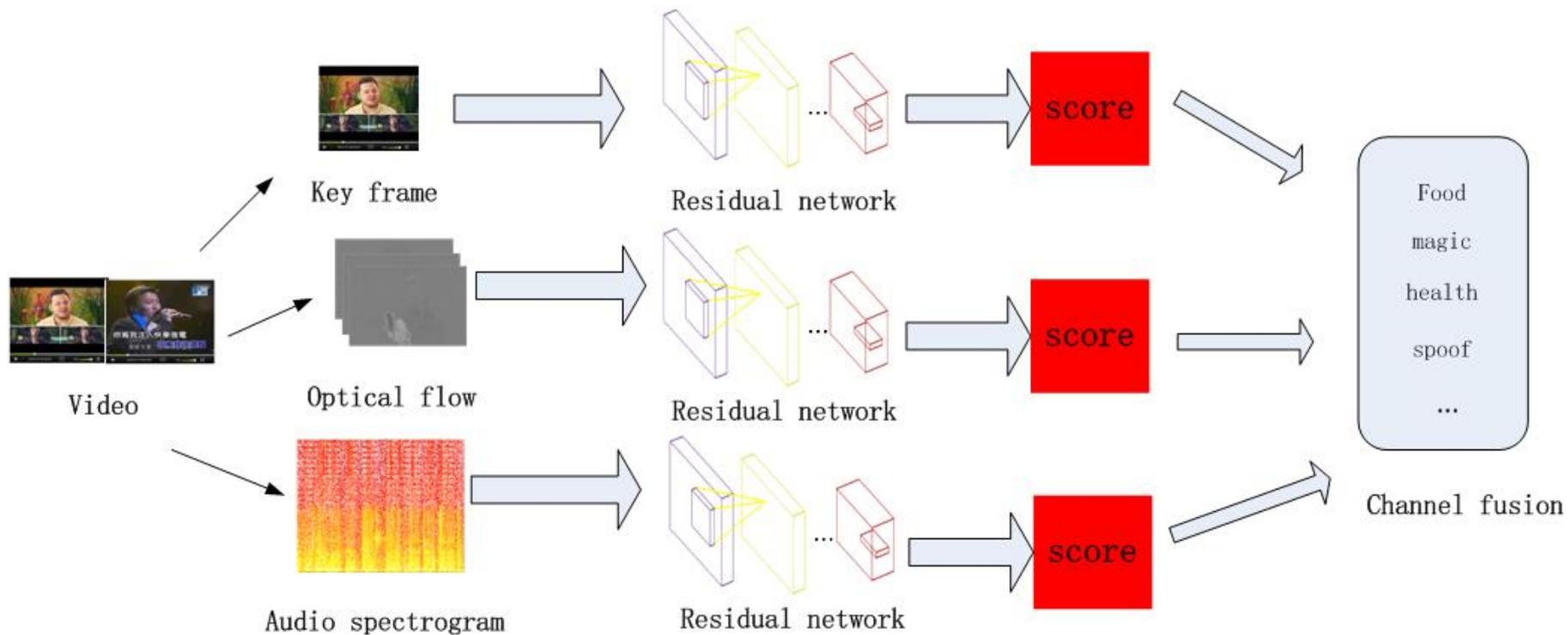
- 视频内容包括音乐、舞蹈、篮球等20个视频标签类别。
- 每个标签类型约1800个训练视频，500个验证视频；500个测试视频，共计56110个视频，约5000小时视频。

## 四、思路及切入点

---

- Appearance :一个video的每帧通常包含了丰富的信息，所以在这里我们选取key frame作为video的appearance 信息
- Motion: 除了appearance之外，更能体现一个video特征的是frame与frame之间的运动信息，我们提取video的optical flow作为motion信息
- Audio:一个video的audio同样包含了丰富的类别信息，我们提取audio的能量谱图作为特征

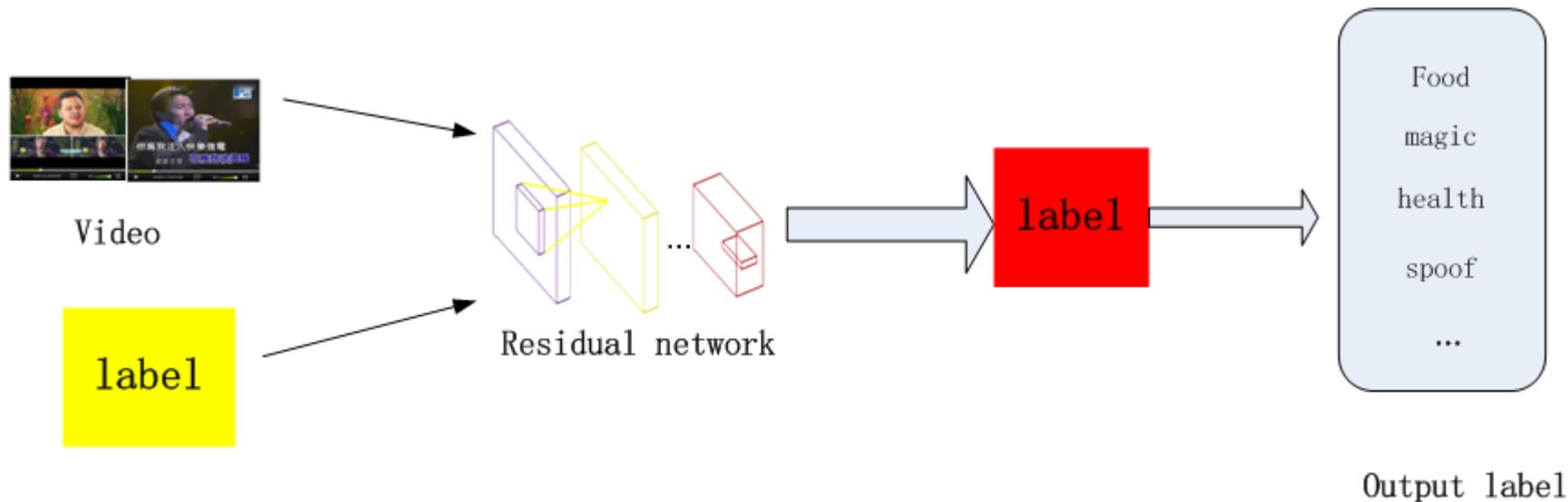
# 五、技术框架



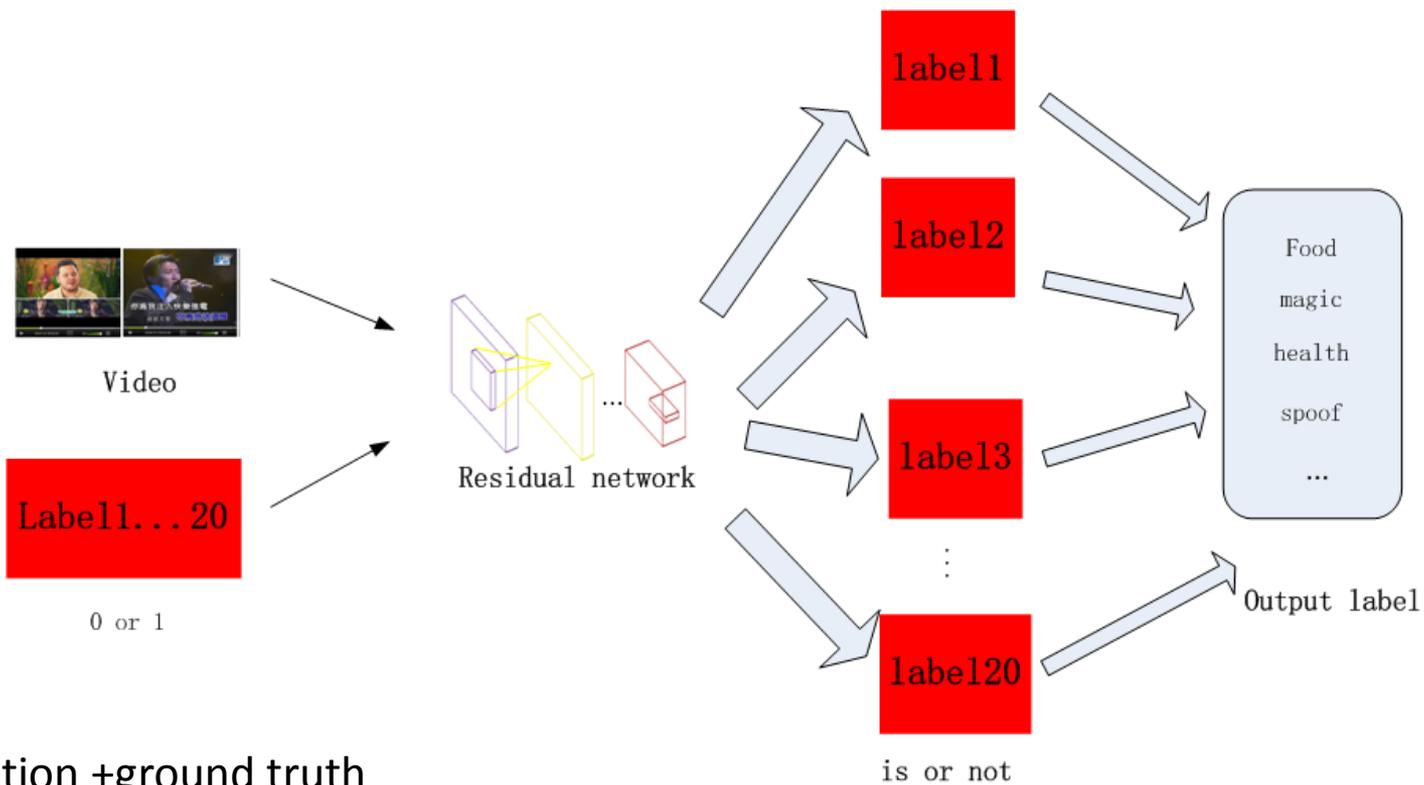
# 数据处理

- A：我们发现在训练数据中，各个类别包含的图像数量相差较多，训练的准确率一直无法提升故我们以包含图片数量较少的类别为基础，从其他类别图片中采样一部分来进行网络的训练。
- O：我们截取中间大约10s的视频提取它的optical flow，之所以提取中间的clip是因为这段clip更能描述一段video的运动信息，用的工具主要是dense\_flow
- S：我们首先截取音频中间大约30s左右的数据，进行fft变换得到它的能量谱图

# Appearance通道单label分类技术框架



# Appearance通道多label分类技术框架



validation +ground truth

## 六、实现方法与结果

方法(method) 次数(times)	单标签(single-label)	多标签(multi-label)	单标签与多标签融合(sl+ml)
第一次	74.31	68.30	74.04
第二次	74.80	68.40	74.45

### 结论：

- 单标签视频分类可以取得比较好的结果
- 多标签与单标签进行融合可以提升多标签的分类能力

## 七、创新点

---

- 网络结构：以往的视频分类任务使用的都是vggnet、alexnet等网络，这次我们使用50层的residual network，并且准备打算在flow通道和audio通道同时使用residual network，最后进行信息的融合
- 任务设置
  - 单标签(single label)
  - 多标签(multi label)

## 八、精度与速度优化

---

- 速度优化方面：
  - 我们采用比赛提供的视频的关键帧数据进行训练和测试，减少了模型的训练负担
  - 在提取光流数据和音频频谱数据时我们都是截取一段video的clip或者audio的clip，video的clip大约为中间10s的视频，audio的clip大约为中间的30s左右音频数据，这样做可以减少模型的训练负担，提高网络训练的速度。
- 精度方面：
  - 我们打算采用single-label和multi-label结合的方法来提高精度
  - 我们采用深度神经网络，即50层的residual network来提高精度

## 九、困难与收获

---

困难：

- 处理数据花费了大量时间，准备工作做得不是很充分
- 我们在训练多标签任务时发现只有验证集提供了多标签的ground truth，这也是导致我们多标签分类结果不是很准确的原因

收获：

- 通过比赛提高了自己的动手能力，对于视频分类问题有了一定的理解
- 提高了团队合作能力
- 提高了并行工作的能力

# 十、总结

---

- 视频分类总结：

我们主要从视频的appearance通道提取信息进行深度网络的训练，从实验结果我们可以看出只利用appearance通道本身就可以达到74%的准确率，说明在进行视频标注时，video frame的appearance本身就包含了丰富的类别信息。

- 比赛总结：

- 前期准备工作：

在比赛数据公布之前，需要做好充分的准备工作，收集现有的文章与方法在已有的数据集上进行训练，才能在比赛时进展的更顺利

- 团队合作能力：

团队的合作能力对于比赛的成功与否也有着很大的影响，一个团队中有人去处理数据，有人进行网络的训练与调试参数

## 十一、对大赛的建议

---

- 希望能提供更为完整的数据，在比赛过程中我们发现现在做多标签分类任务时比赛数据不是很完整，只有验证集提供了多标签的ground truth
- 希望可以公布比赛的数据，我们可以继续完善我们的模型

Questions ?

The bottom of the slide features a decorative graphic consisting of several overlapping, semi-transparent green shapes. These shapes form a jagged, mountain-like silhouette against the white background, with varying shades of green from light to dark.