

# 法律声明

---

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：大数据分析挖掘

■ 新浪微博：ChinaHadoop



# 卷积网络+递归网络

---

主讲人： 李伟

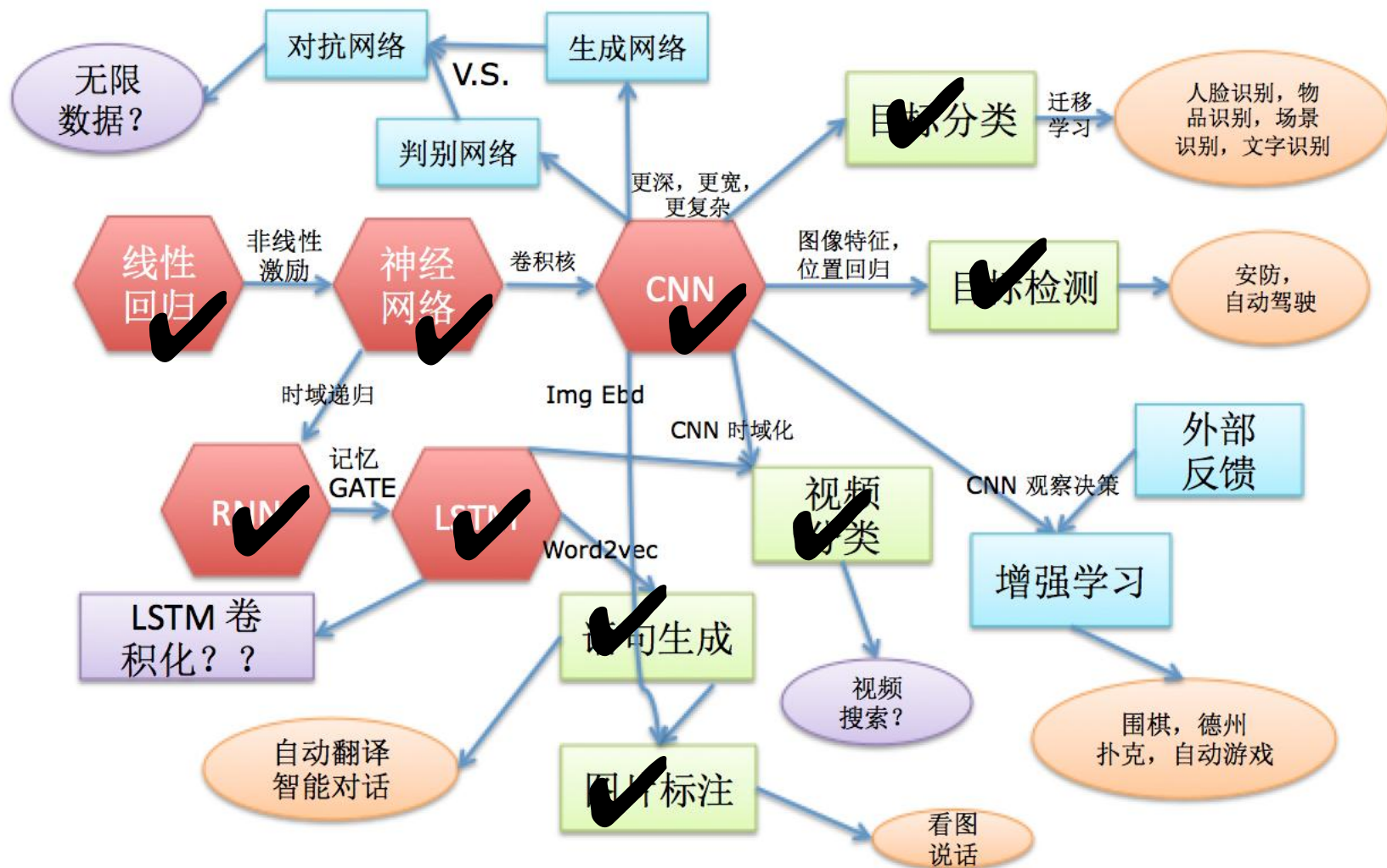
纽约城市大学博士

主要研究深度学习，计算机视觉，人脸计算  
多篇重要研究文章作者，重要会议期刊审稿人

微博ID: weightlee03 (相关资料分享)

GitHub ID: wiibrew (课程代码发布)

# 结构



# 提纲

---

- 1. CNN + RNN
- 2. 图片标注
- 3. 视频行为识别
- 4. 图片 / 视频问答
- 5. 实例学习 Image Caption 图片自动标注

# 期待目标

---

- 1. 了解传统神经网络空间时间扩展概念
- 2. CNN, RNN特征提取方面异同, 结合的特点, 在图片标注/视频分类/图片问答应用中的作用
- 3. 明白图片标注的训练流程, 能够运用现有package完成训练和测试

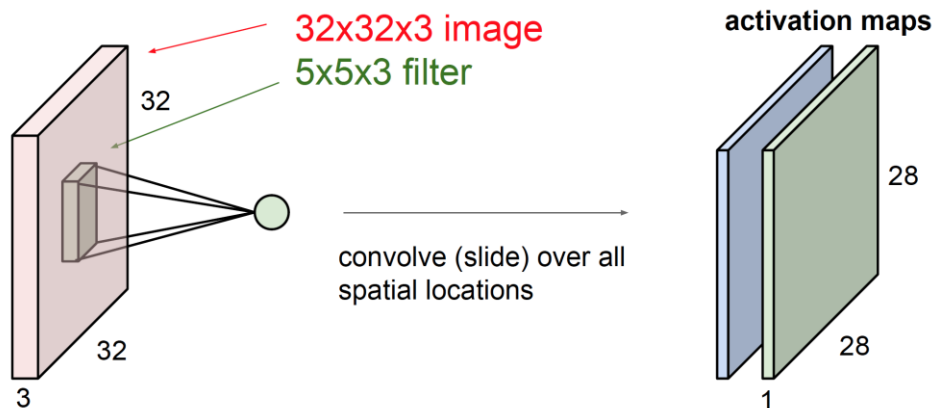
# 提纲

---

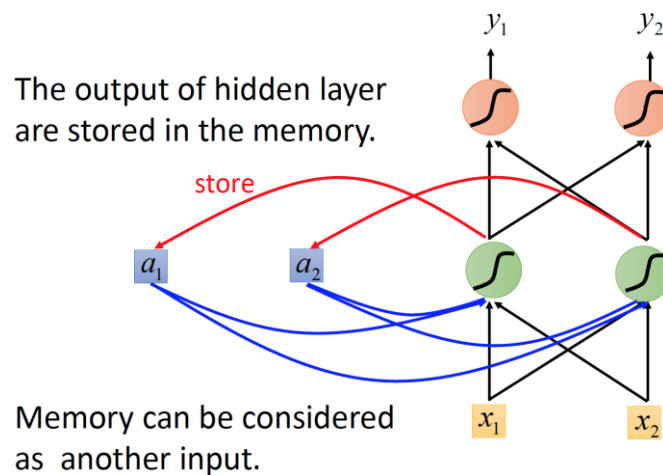
- 1. CNN + RNN
- 2. 图片标注
- 3. 视频行为识别
- 4. 图片 / 视频问答
- 5. 实例学习 Image Caption 图片自动标注

# CNN+RNN

## □ CNN卷积神经网络



## RNN递归神经网络



# CNN+RNN

---

## □ 相同点

□ 1. 传统神经网络的扩展

□ 2. 前向计算产生结果，反向计算模型更新

□ 3. 每层神经网络横向可以有多个神经元共存，纵向可以有多个神经网络连接



# CNN+RNN

---

## □ 不同点

- 1. CNN空间扩展，神经元与特征卷积；RNN时间扩展，神经元与多个时间输出计算
- 2. RNN可以用于描述时间上连续状态的输出，有记忆功能，CNN用于静态输出
- 3. CNN高级100+深度，RNN深度有限

# CNN+RNN

---

## □ 组合意义

1. 大量信息同时具有时间空间特性：视频，图文结合，真实的场景对话
2. 带有图像的对话，文本表达更具体
3. 视频相对图片描述的内容更完整

这是开门还是关门？



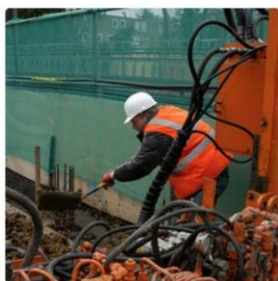
# CNN+RNN

## □ 组合方式

### 1. CNN 特征提取，用于RNN语句生成→ 图片标注



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



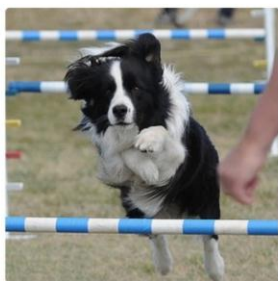
"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinaina on swina."



"man in blue wetsuit is surfing on wave."

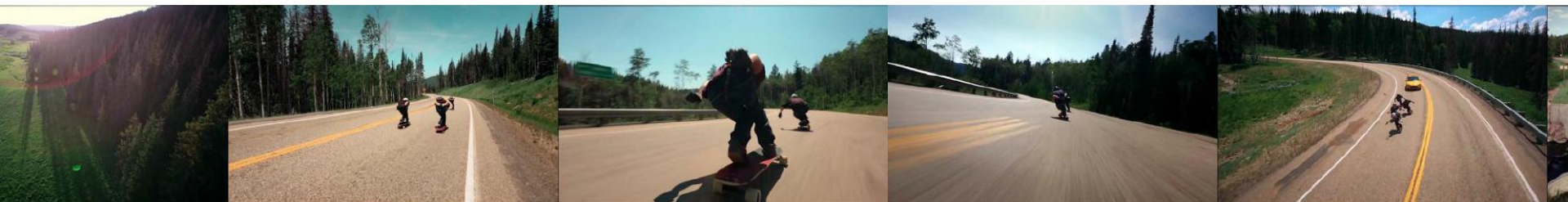
# CNN+RNN

---

## □ 组合方式

2. RNN特征提取用于CNN内容分类→ 视频分类

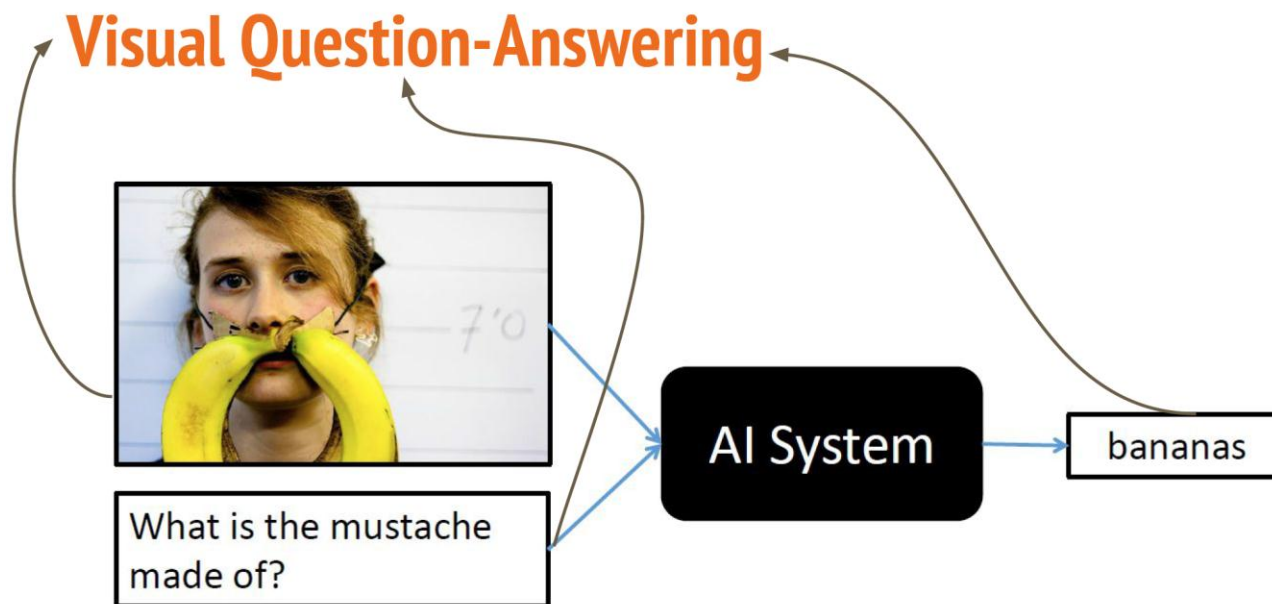
YouTube Videos



# CNN+RNN

## □ 组合方式

### 3. CNN特征提取用于对话问答→ 图片问答



# CNN+RNN

---

## □ 组合方式实现

1. 特征提取：LSTM输出，FC层输出
2. 特征合并：Concatenate 层；Attention 相乘
3. 结果输出：连续语句输出 LSTM，组合分类回归 DNN

# 提纲

---

- 1. CNN + RNN
- 2. 图片标注
- 3. 视频行为识别
- 4. 图片 / 视频问答
- 5. 实例学习 Image Caption 图片自动标注

# 图片标注

---

- 问题描述
- 拥有大量图片及标注信息，能否通过学习建立一个能够自动图片标注的模型



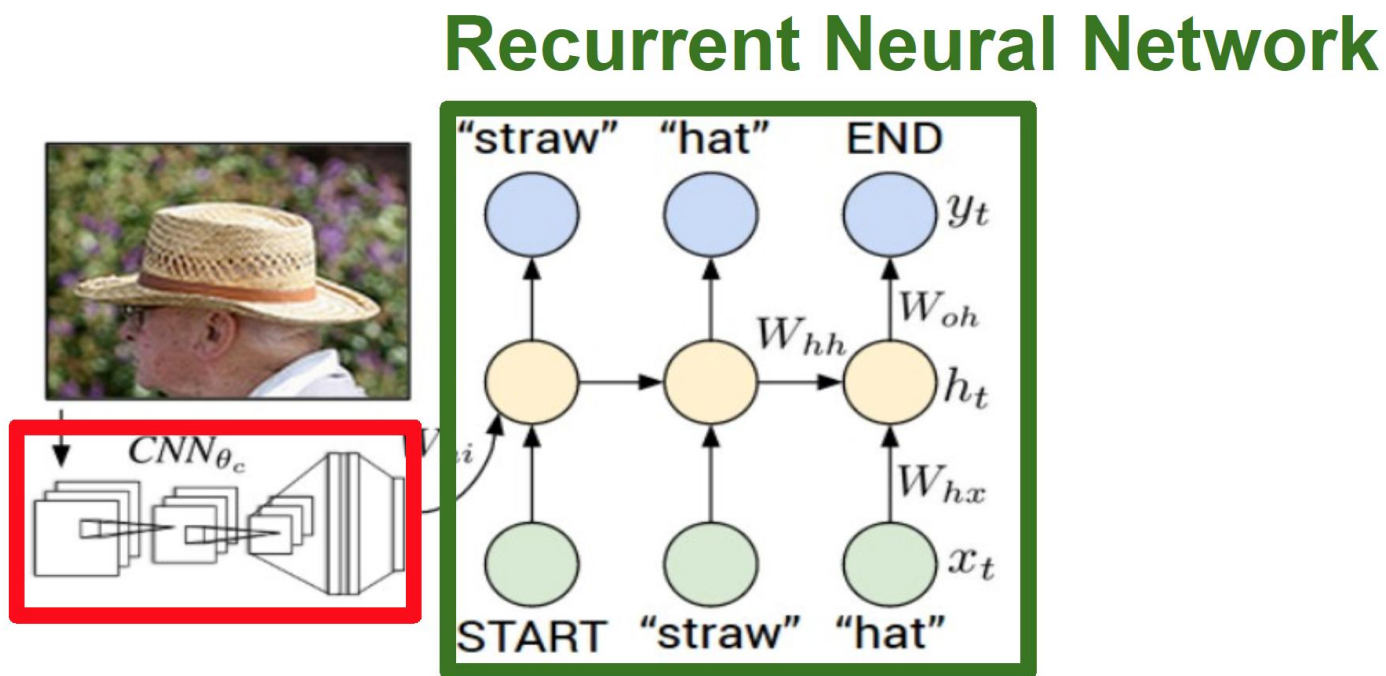
# 图片标注

---

- 基本思路
- 目标是产生标注的语句，是一个语句生成的任务，LSTM?
- 描述的对象大量图像信息，图像信息表达，CNN?
- CNN网络中全连接层特征描述图片，特征与LSTM输入结合?

# 图片标注

## □ 模型设计 - 整体结构



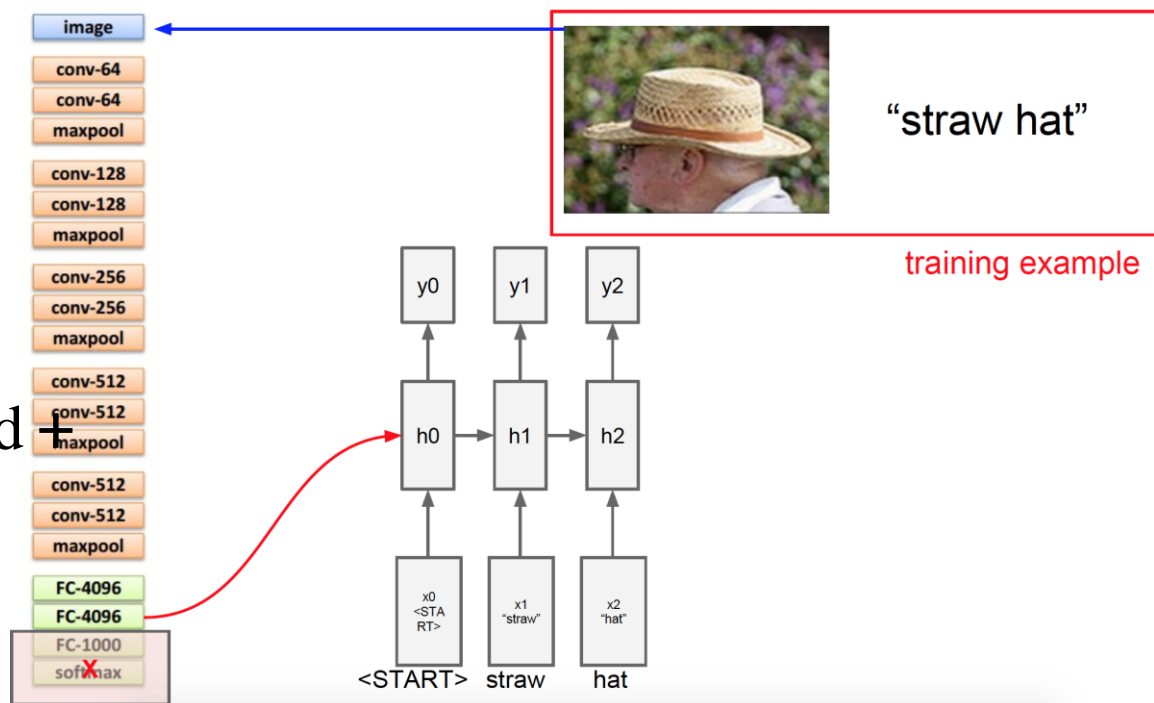
## Convolutional Neural Network

# 图片标注

## □ 模型设计 - 特征提取

□ 全连接层特征用来描述原图片

□ LSTM输入: word + 图片特征; 输出下一word



# 图片标注

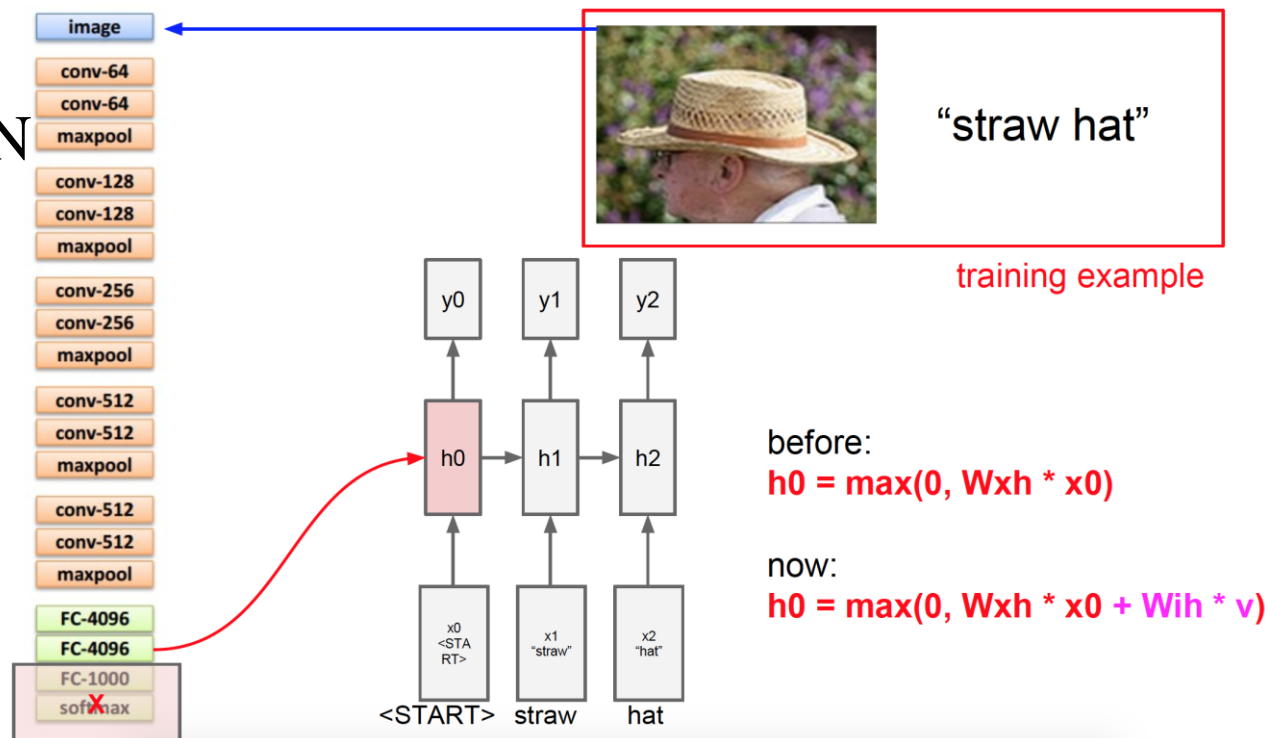
## □ 模型设计 - 特征融合

### □ 图片特征CNN

全连接提取

### □ 语言特征:

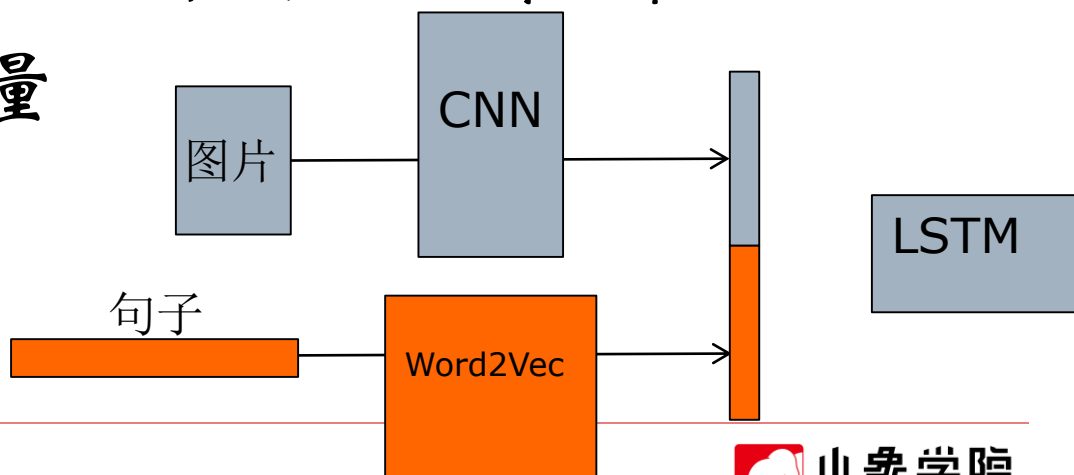
Word2Vect



# 图片标注

## □ 模型设计 - 数据准备

1. 图片CNN特征提取
2. 图片标注生成Word2Vect 向量
3. 生成训练数据：图片特征 + 第n单词向量：  
第n + 1单词向量



# 图片标注

---

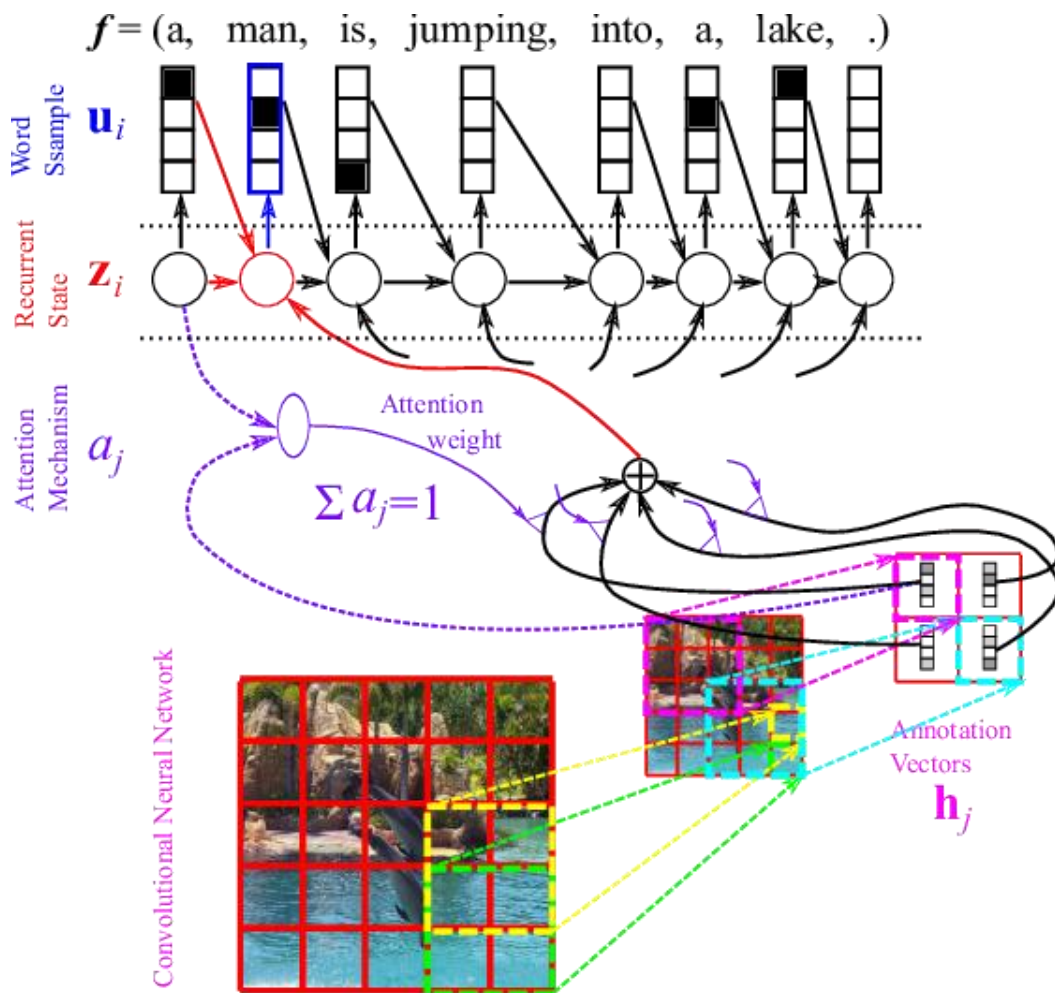
## □ 模型训练

1. 运用迁移学习，CNN特征，语句特征应用已有模型
2. 最终的输出模型是LSTM，训练过程的参数设定：梯度上限(gradient clipping)，学习率调整(adaptive learning)
3. 训练时间很长

# 图片标注

## □ 模型运行

1. CNN特征提取
2. CNN特征+语句开头，单词逐个预测

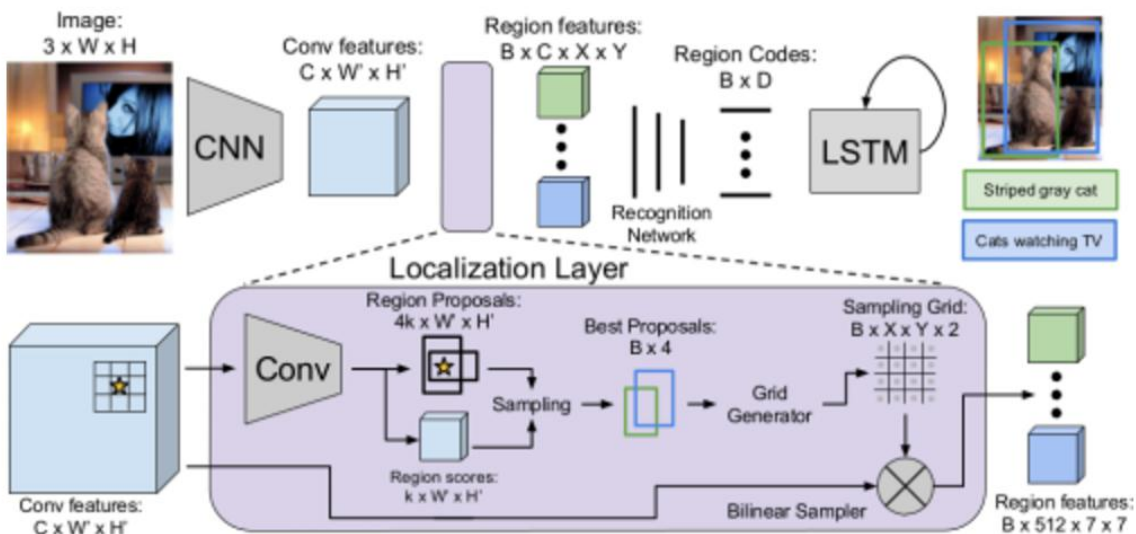


# 图片标注

## □ 图片标注升级 - 详细标注

## □ DenseCap

1. 图片一个标注?
2. 不同区域区别?
3. 标注目的
4. Dense意义



Justin Johnson, Andrej Karpathy, Li Fei-Fei CVPR 2016



# 图片标注

## □ 图片标注升级 - 详细标注

### DenseCap 训练

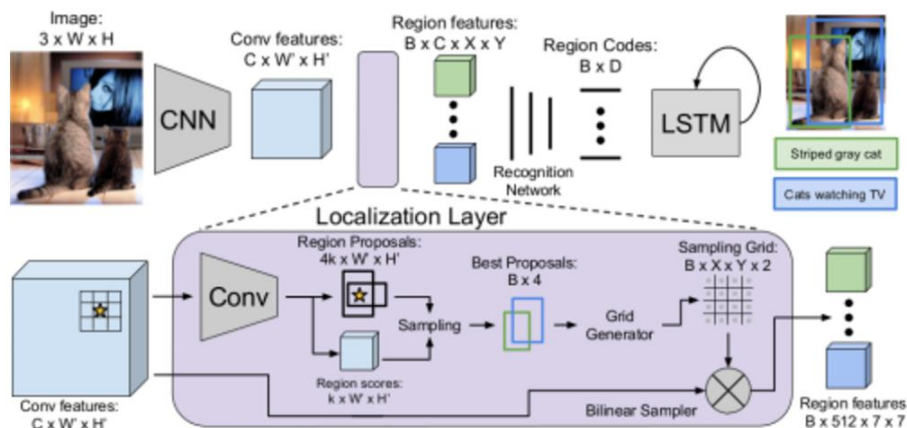
Loss: 目标探测loss

目标识别loss

区域标注loss

训练方式: End-to-end.

优化方法: SGD, Adam



# 图片标注

## □ 图片标注升级 - 详细标注

### DenseCap vs 普通标注



Our Model: plane is flying. tail of the plane. red and white plane. plane is white. engine on the plane. windows on the plane. nose of the plane.

Full Image RNN: A large jetliner flying through a blue sky.



woman wearing a black shirt. teddy bear is brown. chair is black. glass of wine. table is brown. woman with brown hair. paper on the table.

A man and a woman sitting at a table with a cake.



teddy bear is wearing a red shirt. red and white teddy bear. bear is wearing a red hat. red and white shirt. table is brown. black nose of a bear.

A teddy bear with a red bow on it.



train on the tracks. trees are green. front of the train is yellow. grass is green. green trees in the background. photo taken during the day. red train car.

A train is traveling down the tracks near a forest.

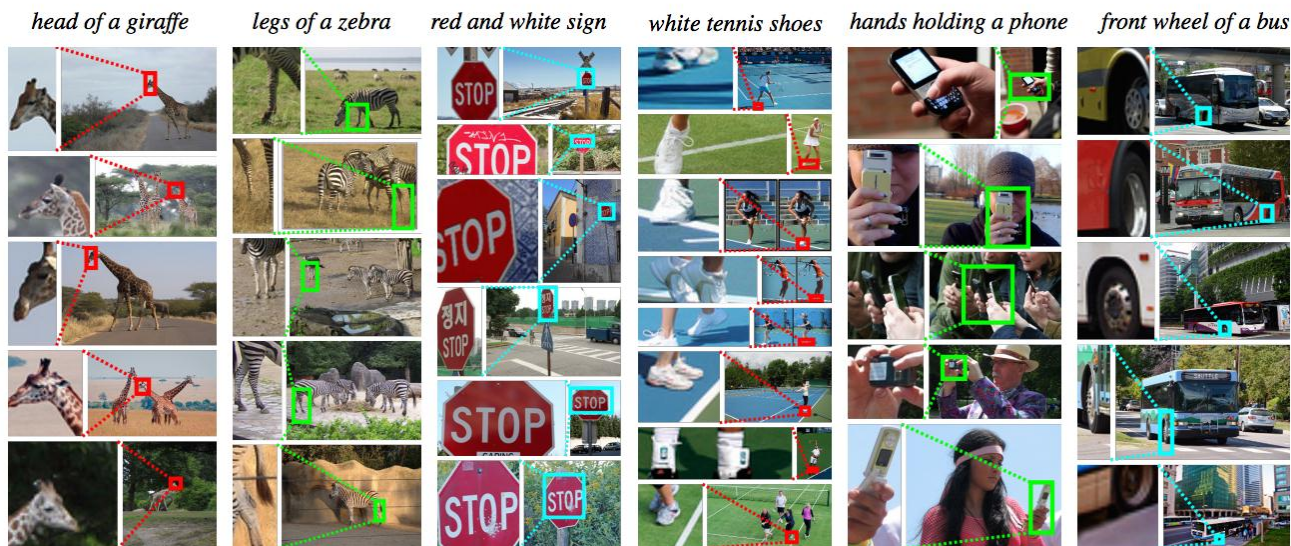
# 图片标注

## □ 图片标注升级 - 详细标注

用途：图片检索

找到一些没有定义的组合：动物的头，车的轮子，不同颜色的某物

开放性探测识别



# 提纲

---

- 1. CNN + RNN
- 2. 图片标注
- 3. 视频行为识别
- 4. 图片 / 视频问答
- 5. 实例学习 Image Caption 图片自动标注

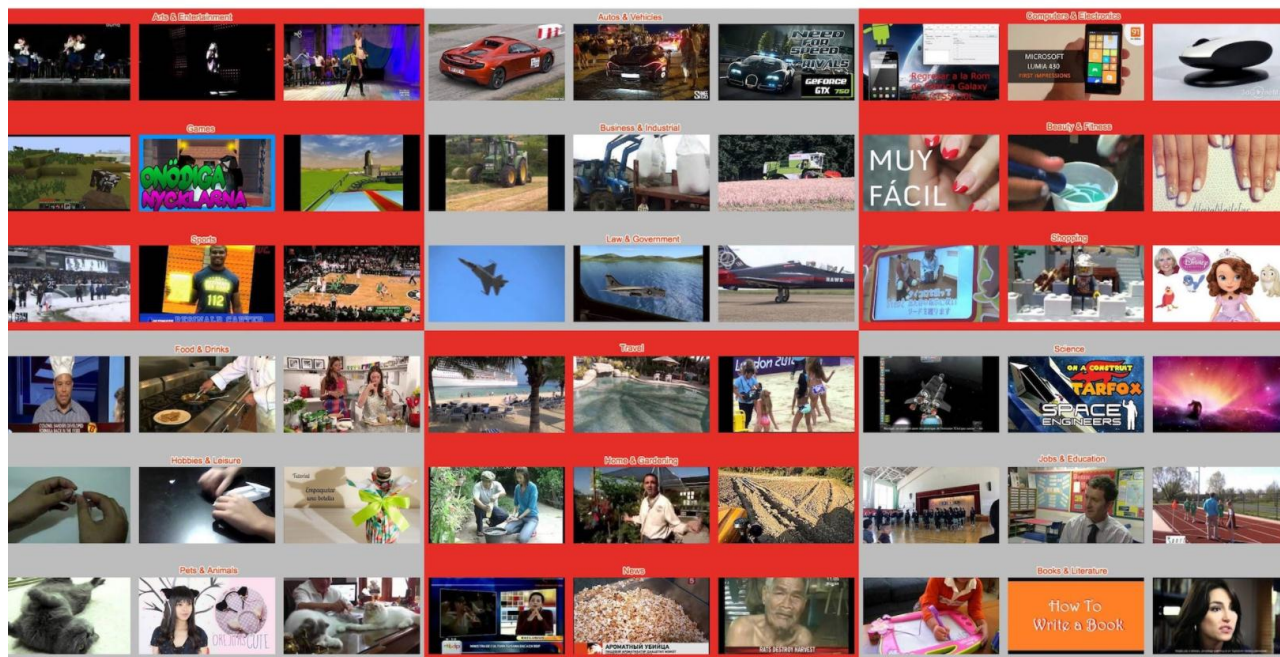


# 视频行为识别

## □ 问题定义

视频中在发  
生什么？

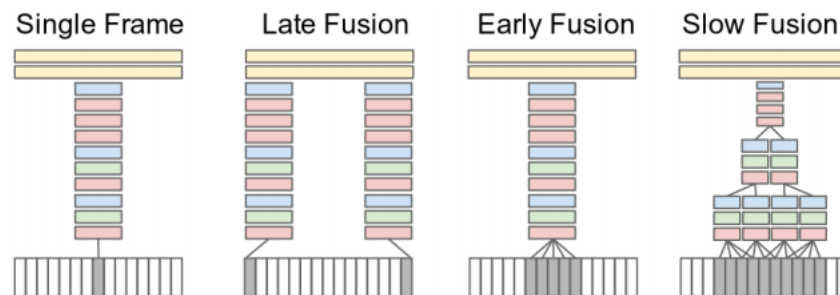
CVPR'17 Workshop on YouTube-8M Large-Scale Video Understanding



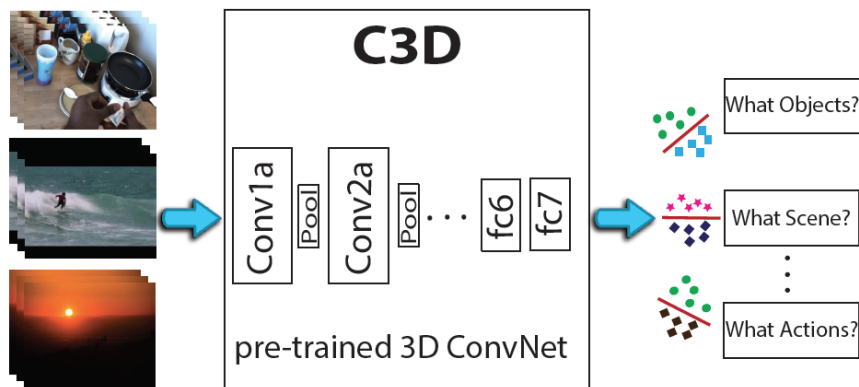
# 视频行为识别

## □ 常用方法总结

### 1. CNN 特征简单组合



### 2. 3D版本 CNN



图像特征的前后关系没有很好的学到

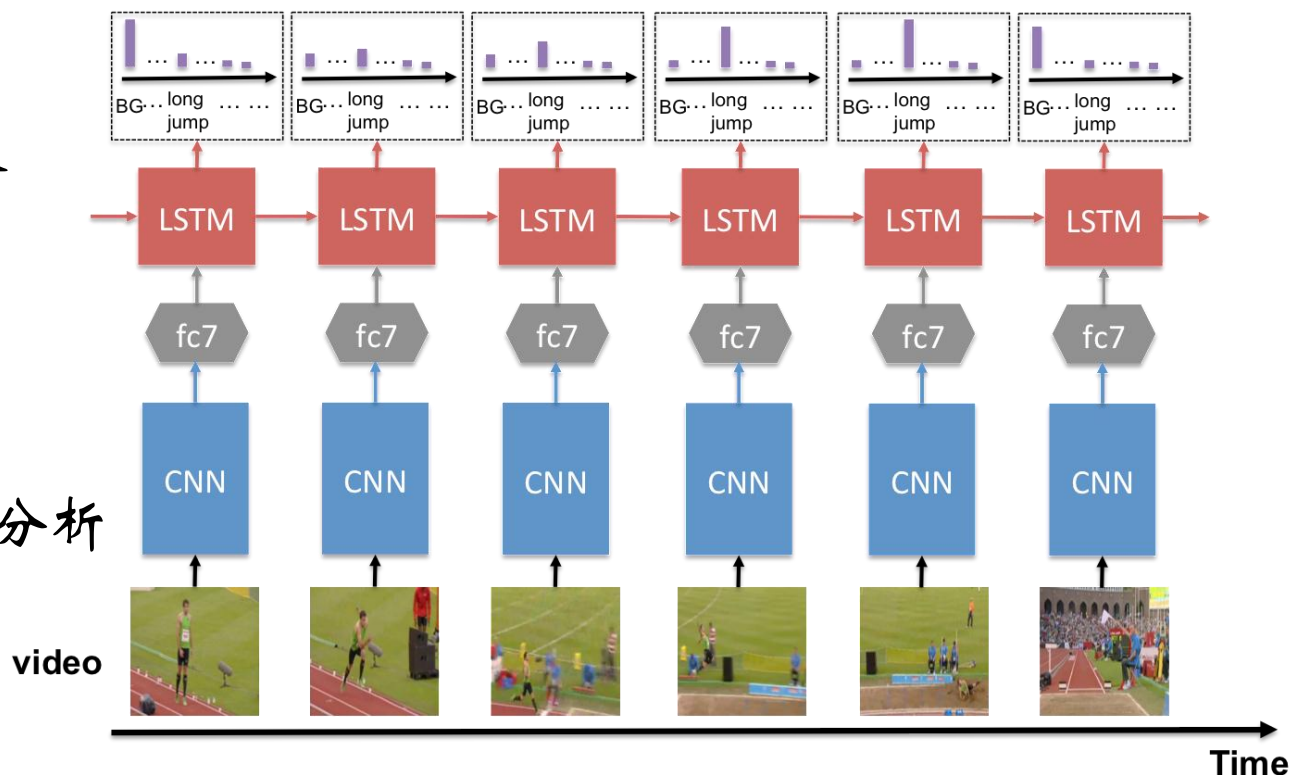
# 视频行为识别

## □ RNN用于CNN特征融合

1. CNN 特征提取

2. LSTM判断

3. 多次识别结果分析



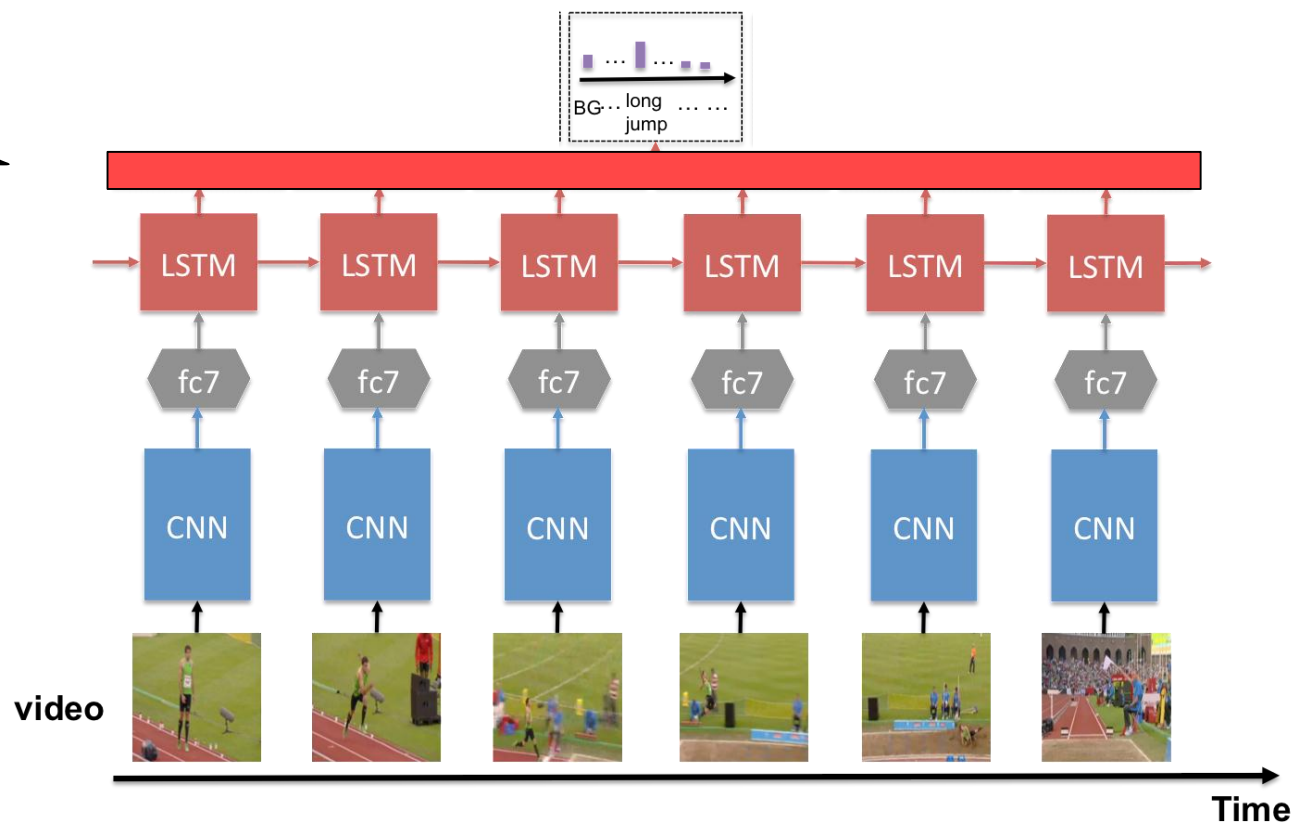
# 视频行为识别

## □ RNN用于CNN特征融合

1. CNN 特征提取

2. LSTM融合

3. Linear regr +  
Softmax 分类

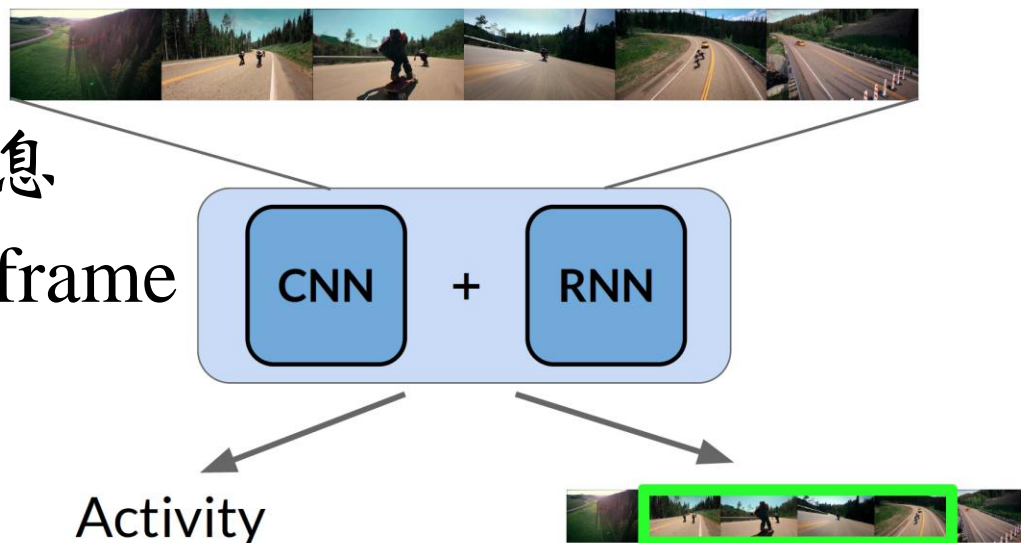




# 视频行为识别

## □ RNN用于CNN特征筛选+融合

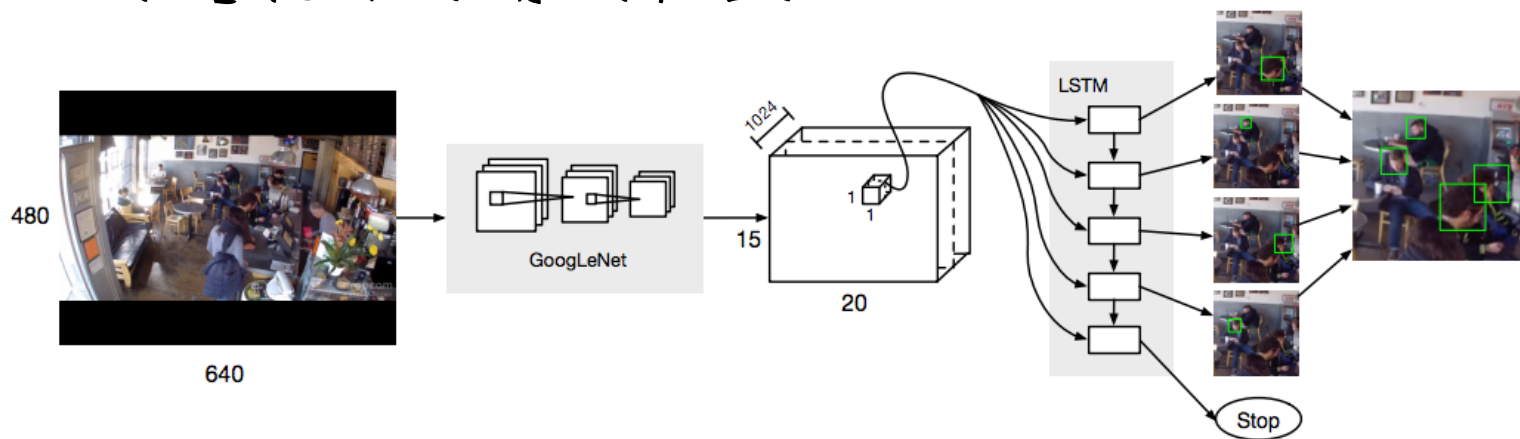
1. 并不是所有的视频图像包含确定分类信息
2. RNN用于确定哪些frame是有用的
3. 对有用的图像特征融合



# 视频行为识别

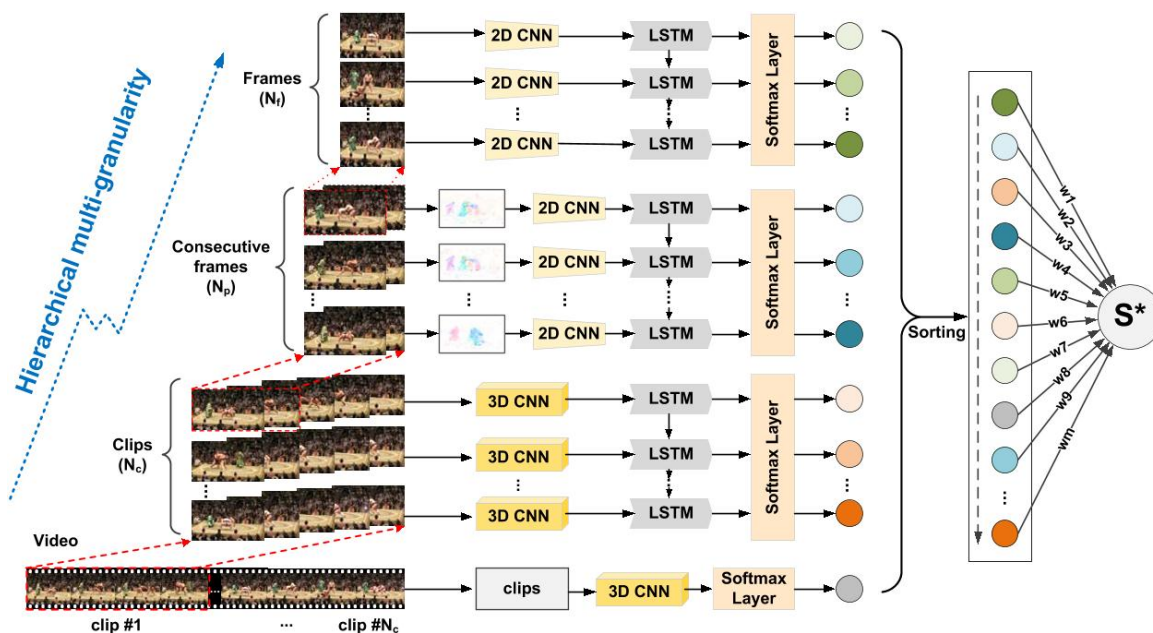
## □ RNN用于，目标检测

1. CNN直接产生目标候选区
2. LSTM对产生候选区融合（相邻时刻位置近似）
3. 确定最终的精确位置



# 视频行为识别

- 多种模型综合
- 竞赛/应用中, 为了产生最好结果, 多采用多模型ensemble形式



# 提纲

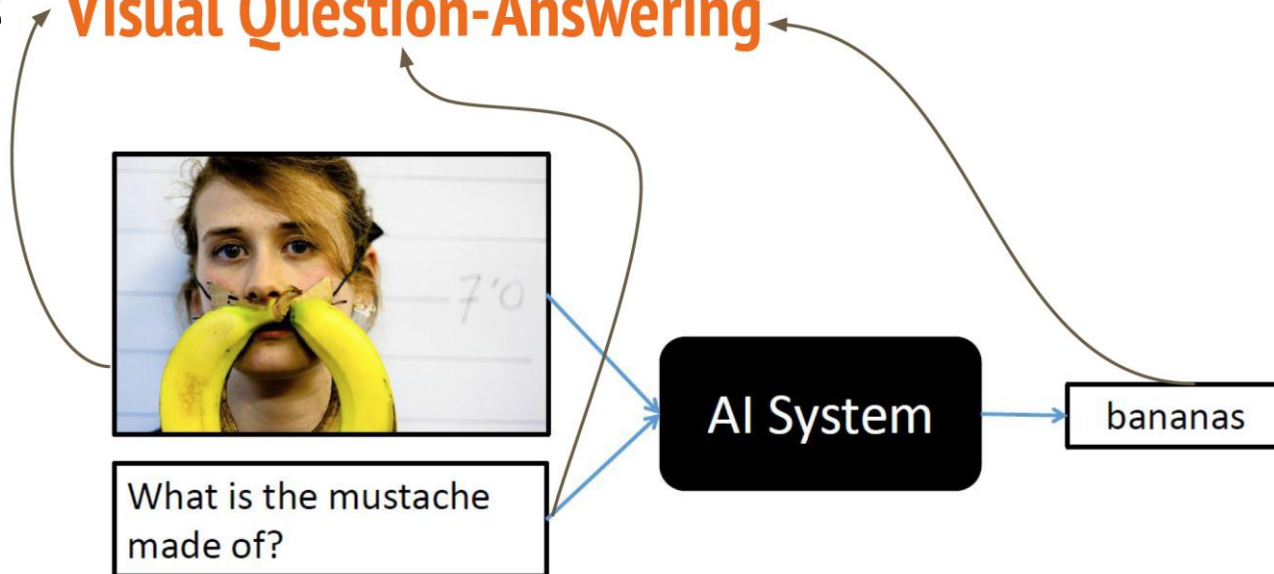
---

- 1. CNN + RNN
- 2. 图片标注
- 3. 视频行为识别
- 4. 图片 / 视频问答
- 5. 实例学习 Image Caption 图片自动标注

# 图片 / 视频问答


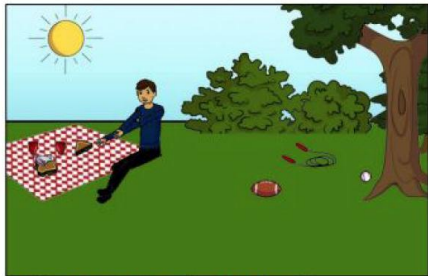


## □ 问题定义

给定一张图片，提出图片内容相关问题，问答模型给出答案 **Visual Question-Answering**



# 图片 / 视频问答

## □ 问题种类

	Real images	Abstract scenes
Open-ended	<div><p>Q: Does it appear to be rainy?</p><p>A: no</p></div>	<div><p>Q: What is just under the tree?</p><p>A: a ball</p></div>
Multi-Choice	<div><p>Q: How many slices of pizza are there?</p><p>A: 1, 2, 3, 4</p></div>	<div><p>Q: What is for desert?</p><p>A: cake, ice cream, cheesecake, pie</p></div>

# 图片 / 视频问答

---

## □ 图片问答意义

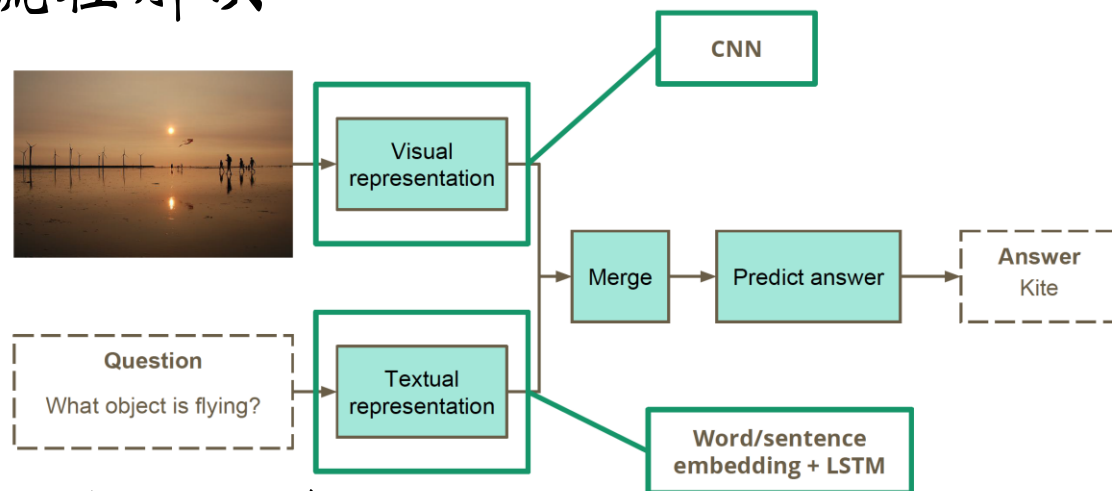
1. 是对纯文本语言问答系统的扩展
2. 图片理解和语言处理的深度融合
3. 提高人工智能应用范围—观察，思考，表达

# 图片 / 视频问答

## □ 方法流程

依旧按照语言问答流程解决

图片特征同语言  
特征融合



训练数据：问题 + 图片 - 答案

12



# 图片 / 视频问答

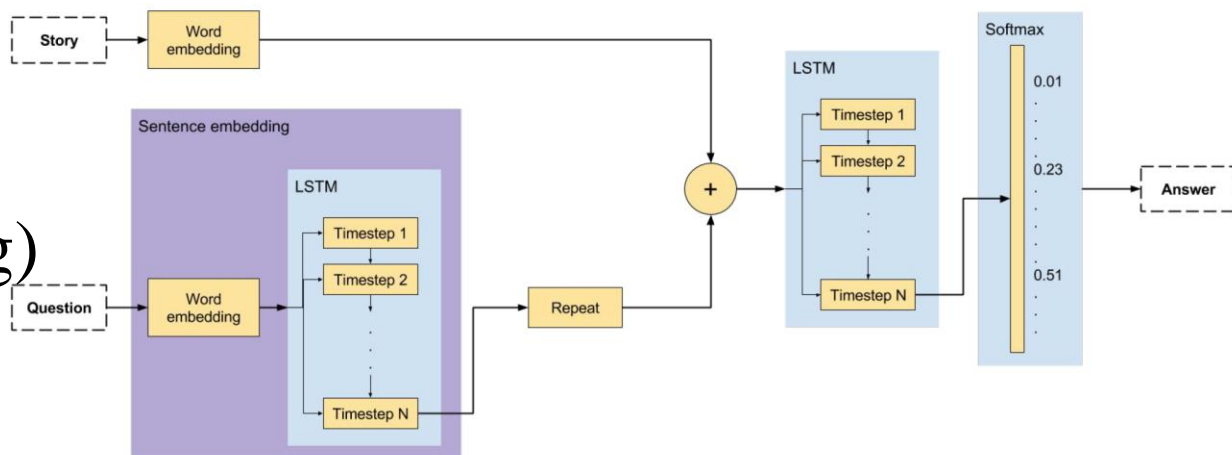
## □ 模型设计 - 纯文字问答系统

1. 背景故事

特征生成

(word embedding)

2. 问题特征生成



3. 背景，问题特征融合

4. 标准答案回归

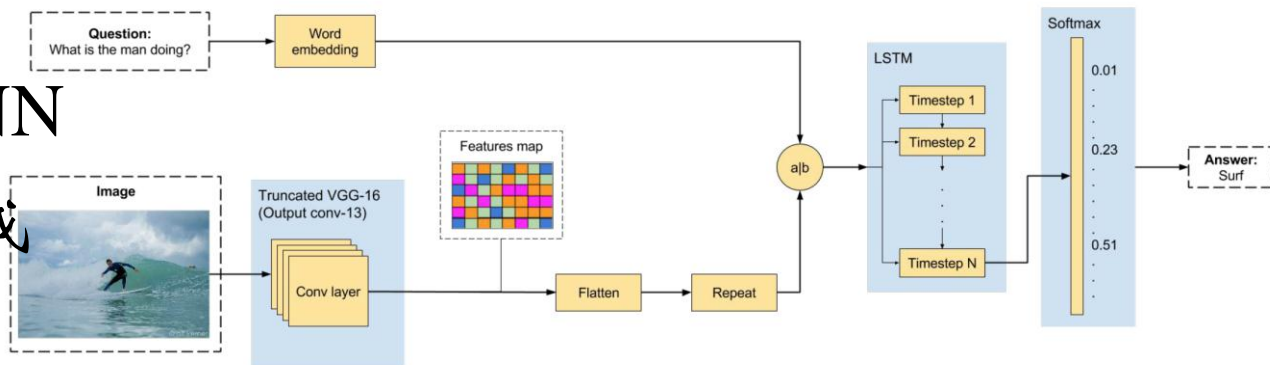
# 图片 / 视频问答

## □ 模型设计 - 图片问答系统

### 1. 背景故事

特征生成-CNN

### 2. 问题特征生成



### 3. 背景，问题特征融合

### 4. 标准答案回归

用以训练的数据：真值是什么？

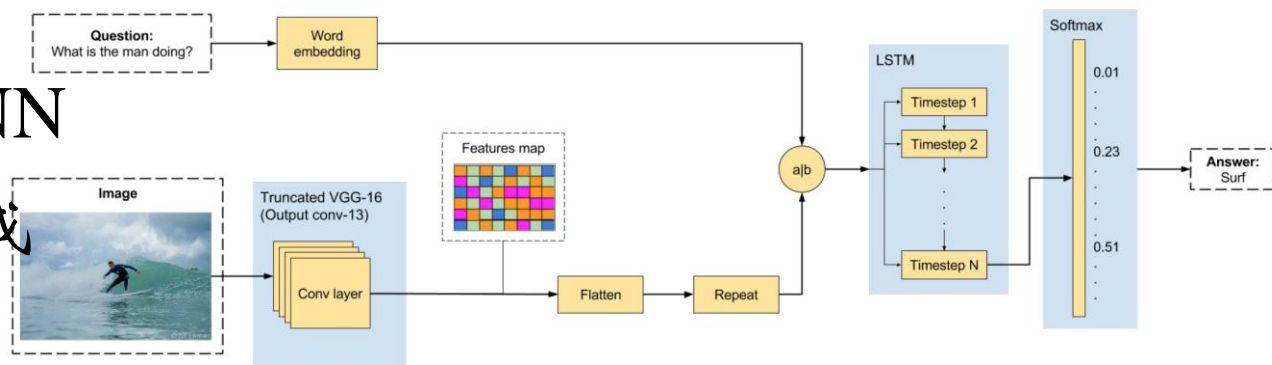
# 图片 / 视频问答

## □ 模型设计 - 图片问答系统

### 1. 背景故事

特征生成-CNN

### 2. 问题特征生成



### 3. 背景，问题特征融合

### 4. 标准答案回归

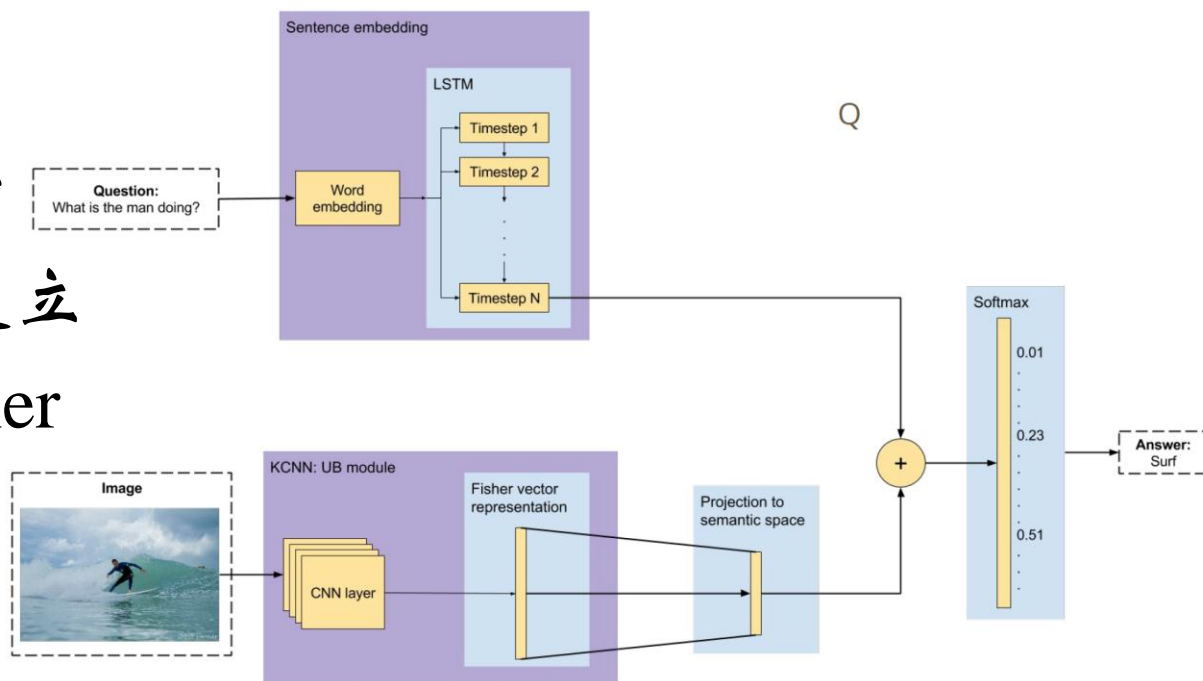
用以训练的数据：真值是什么？ 融合特征：答案

# 图片 / 视频问答

□ 模型设计 - 图片问答系统

□ 模型优化 - 1

对图片特征向量  
进一步处理，建立  
CNN 特征的fisher  
特征



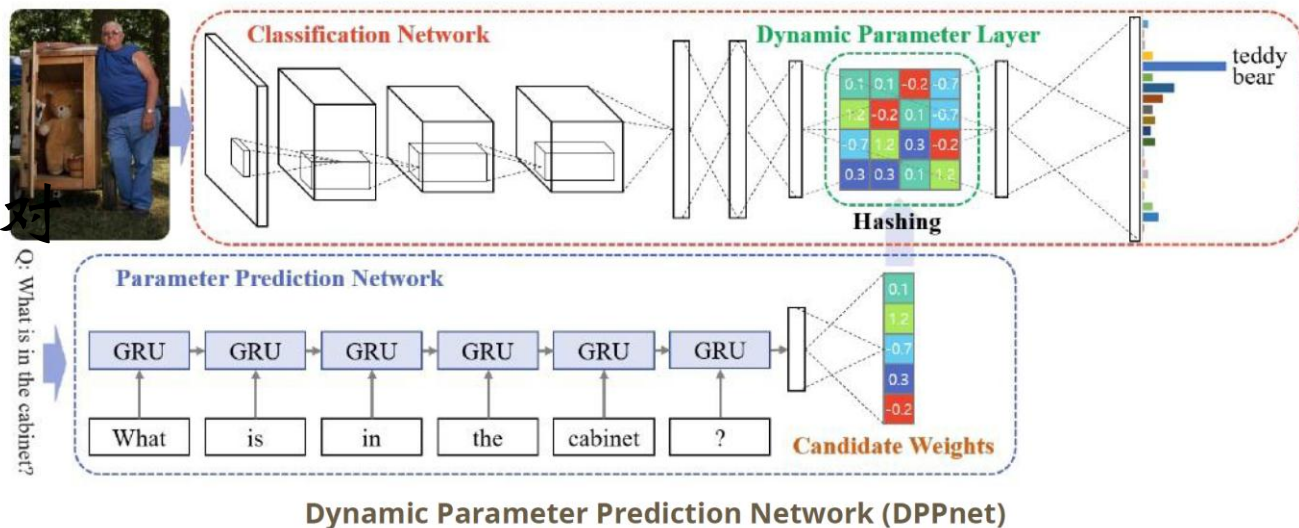
提高特征表达效率，更容易同encoding特征组合

# 图片 / 视频问答

□ 模型设计 - 图片问答系统

□ 模型优化 - 2

用问题作为  
“候选区域”对  
原始CNN特征  
图局部识别



Noh, H., Seo, P. H., & Han, B. [Image question answering using convolutional neural network with dynamic parameter prediction](#). CVPR 2016

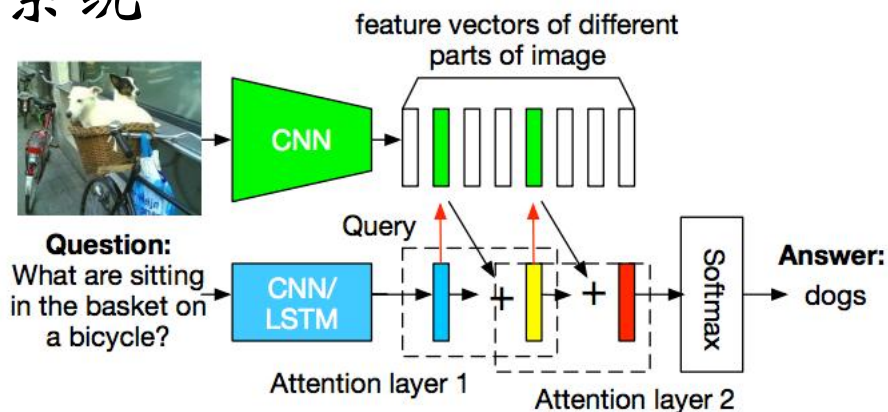
# 图片 / 视频问答

## □ 模型设计 - 图片问答系统

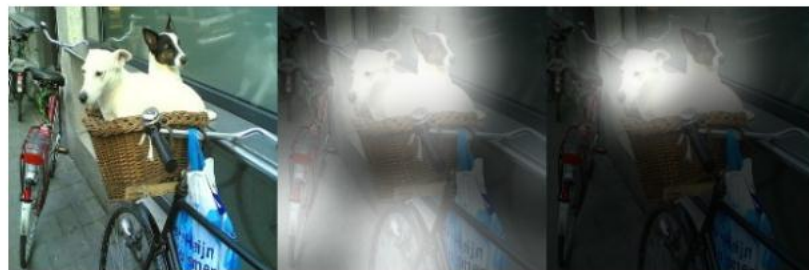
## □ 模型优化 - 3

注意力图对图片问答帮助，根据问题产生第一次注意力图，然后最终注意力图，最后进行回答  
什么在筐子里？

1 筐子范围，2 筐子里范围，3 识别



(a) Stacked Attention Network for Image QA



Original Image

First Attention Layer

Second Attention Layer

# 图片 / 视频问答

---

- 模型设计 - 图片问答系统
- 模型优化小结
- 不同的优化结构方便不同类型的问题回答，数字 / 种类 / 抽象 / 二值判断
- 仍然是很新的研究问题，上述例子来源于CVPR2016，学术价值应用价值都很大
- 人机交互中图片问答在 盲人辅助 / 教育 / 智能助手等方面大有可为

# 提纲

---

- 1. CNN + RNN
- 2. 图片标注
- 3. 视频行为识别
- 4. 图片 / 视频问答
- 5. 实例学习 Image Caption 图片自动标注



# 总结

---

- 1. 了解传统神经网络空间时间扩展概念
- 2. CNN, RNN特征提取方面特征, 结合的特点, 在图片标注/视频分类/图片问答应用中的作用
- 3. 明白图片标注的训练流程, 能够运用现有package完成训练和测试

# 总结

---

□ 有问题请到课后交流区

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他回答人回答问题

□ 课堂QQ群438285995，微信群

□ 讲师微博：weightlee03，每周不定期分享DL资料 weixin: buaacunywei

□ GitHub ID: wiibrew（课程代码发布）

<https://github.com/wiibrew/DeepLearningCourseCodes>