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：小象
  - 新浪微博：ChinaHadoop

第8课　图像问答

Image Question Answering (QA)

主讲人：张宗健

悉尼科技大学博士

主要研究方向：　计算机视觉、视觉场景理解、图像&语言、深度学习
图像检索CbIR、Human ReID等

本章结构

- 图像问答与图像描述的关系
- 大数据集介绍（VQA，Visual Genome）
- 图像问答模型
- 模型增强：注意机制及外部知识库
- 应用案例：VQA–2LSTM Q + Norm I

图像问答与图像描述的关系

图像问答（Image QA）

- 最AI完备（AI-complete）的任务
- 回答与图片内容相关的问题
- 输入：图片 & 问题
- 输出：答案
  - 单词/词组→分类问题
  - 句子→生成问题

问答例子：
•What is the color of worker's hat? – Yellow
•How many workers are in the image? – 4
•Are they wearing gloves? – Yes
•Why are these people wearing yellow jackets on the street? – For safety

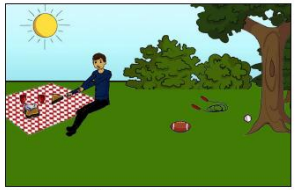
图像问答与图像描述的关系

图像问答（Image QA）

- 最AI完备（AI-complete）的任务


What color are her eyes?
What is the mustache made of?


How many slices of pizza are there?
Is this a vegetarian pizza?


Is this person expecting company?
What is just under the tree?

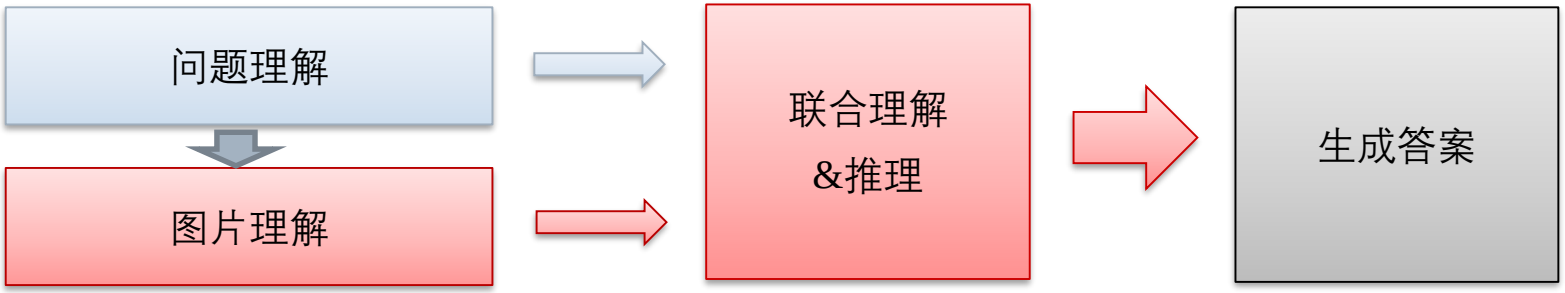

Does it appear to be rainy?
Does this person have 20/20 vision?

- 需要具备一系列AI能力
  - 细分识别（What kind of cheese is on the pizza?）
  - 物体识别（ How many bikes are there ?）
  - 动作识别（ Is this man crying?）
  - 知识库推理（ Is this a vegetarian pizza?）
  - 常识推理（ Is this person expecting company?）
  - ……

图像问答与图像描述的关系

图像问答的理解模式

- 理解问题
- 观察、理解图片
- 关注与问题相关的图片内容，并做推理
- 给出答案



图像问答与图像描述的关系

	图像问答	图像描述
任务目标	回答关于图片的问题	描述图片内容
输入	图片 & 问题	图片
输出	答案 （字，词，句子）	客观描述 （句子）
理解范围	自由&开放	显著
语言水平	读懂、生成句子	生成有语法结构的句子
知识来源	图片内容 语言 知识库/常识	图片内容 语言
对理解要求	扩展	复杂
客观的评价指标	容易	难
AI完备性	更近	近

图像问答与图像描述的关系

研究难点与挑战

- C1多模态理解与推理
  - 图片：捕捉真实世界的原始刻画
  - 自然语言：代表更高一级的抽象
- C2复合理解与推理
  - 多个元素：物体、动作、场景、事件等
  - 多步、迭代过程
  - 动态目标（自由、开放的问答）
- C3引入外部知识库
  - 数据库知识有限
  - DNN模型的容量有限

图像问答与图像描述的关系

研究方向

- 图片、语言之间的特征映射（Joint Embedding）
  - C1
- 注意机制（Attention Mechanism）
  - C2
- 动态模型
  - C2
- 外部知识库增强
  - C3

大数据集介绍

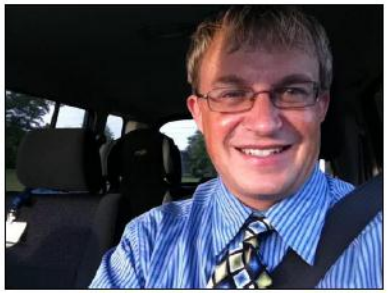
COCO-QA

- 图片集
  - 来源：MS-COCO
  - 训练：72738
  - 测试：38948
- 问答集
  - 每张图片一对QA
  - 根据Caption自动生成
  - 四类：object、number、color 和location

大数据集介绍

Visual Question Answering（VQA）

图片集	现实图片 VQA-Real	抽象场景VQA-Abstract
COCO图片	204,721 （train/val/test）	50,000
问题	614,163	150,000
答案（10A/Q）	614,163 * 10	150,000 * 10
无图片答案	614,163 * 3	150,000 * 3



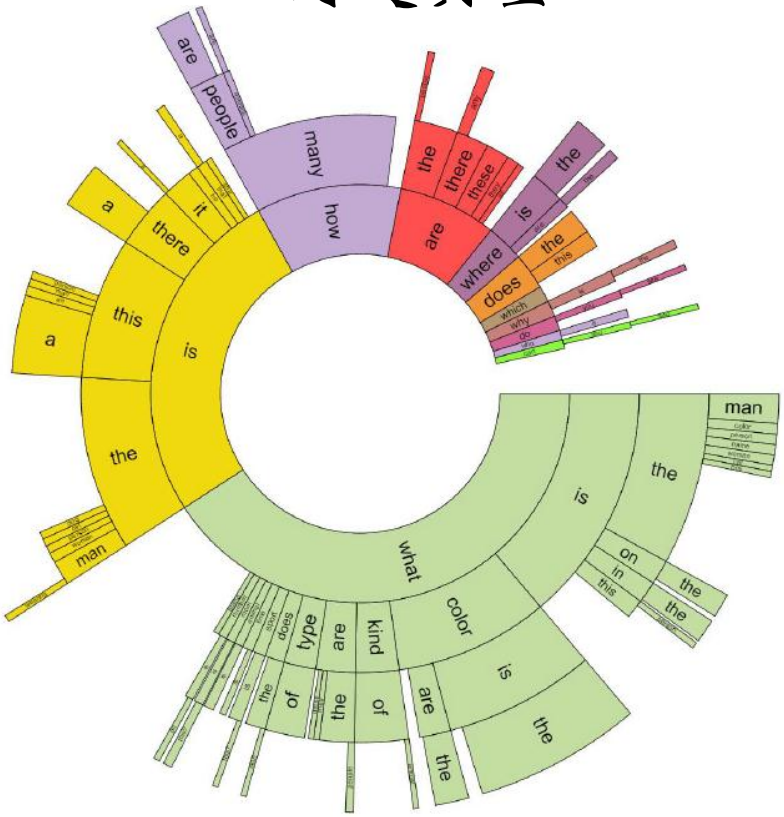
Does this man have children?	yes yes yes	yes yes yes
Is this man crying?	no no no	no yes yes

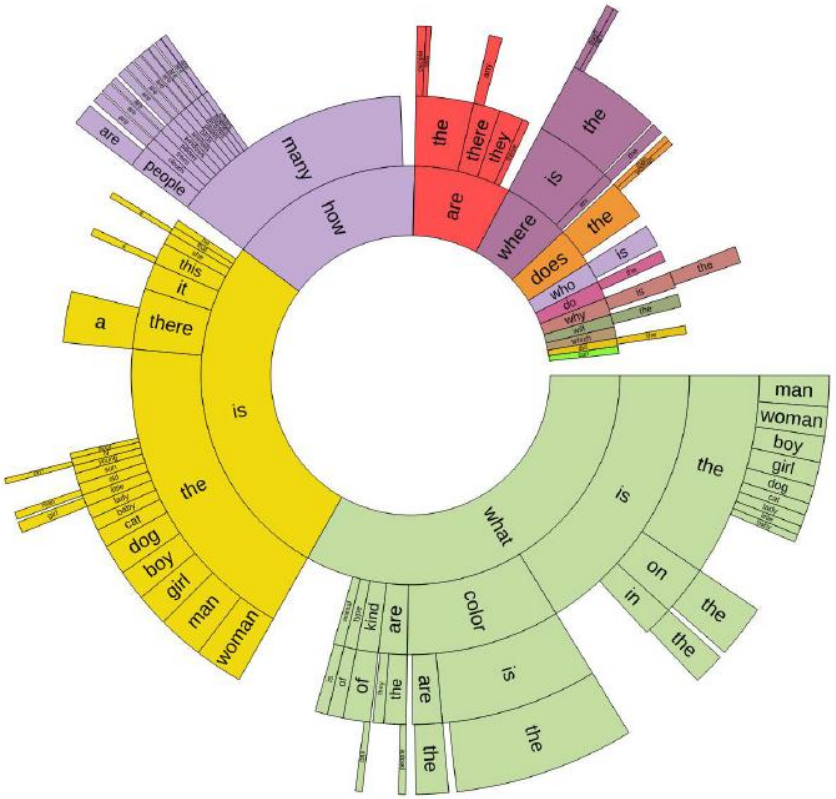
How many glasses are on the table?	3 3 3	2 2 6
What is the woman reaching for?	door handle glass wine	fruit glass remote

大数据集介绍

Visual Question Answering（VQA）

- 问题类型





大数据集介绍

Visual Question Answering（VQA）

- 2种任务
  - 开放问答（Open-ended）
  - 多选题（Multiple-choice）
    - 18个选项答案
- 答案准确性评价
  - Accuracy = min(答案在10个中出现次数/3,1)

大数据集介绍

Balanced Visual Question Answering（VQA）

- 平衡数据集V1.9（→ V2.0）
  - 目标
    - 为了评估图片理解的在任务中的作用
    - 降低语言偏置、不均衡（Language priors）
  - 为问题补充对立图
    - 一个问题对应2个图片
      - 语义场景相似
      - 但答案不同
      - 一个图片10个答案

大数据集介绍

Balanced Visual Question Answering（VQA）

图片集	现实图片	抽象场景
COCO图片	123,287 （train/val）	31,325 （train/val）
问题	658,111	33,383
答案（10A/Q）	658,111 * 10	33,383 * 10
无图片答案	1,974,333	-

大数据集介绍

Visual7W -- Visual Genome的子集

- 图片集：47,300
- 任务类型：多选题（4个选项）
- 7W:What, Where, How, When, Who, Why, Which
  - Which: 选择跟问题相关的区域

Q: Which item is used to cut items?

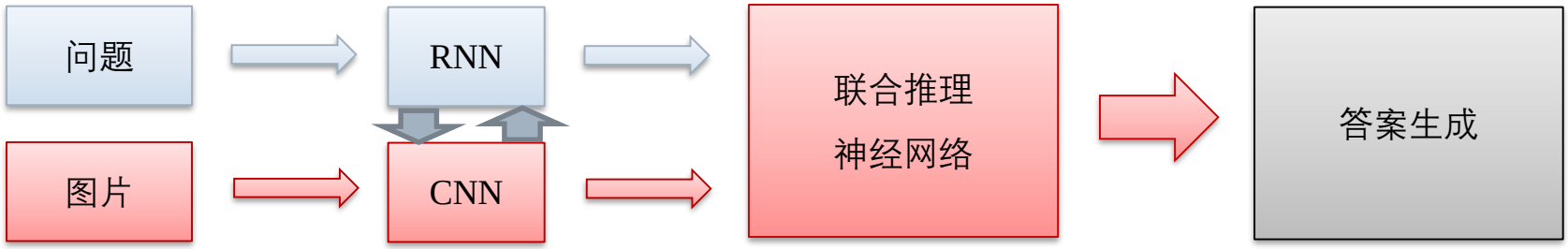
Q: Which doughnut has multicolored sprinkles?

Q: Which man is wearing the red tie?

图像问答模型

基于DNN的点对点（end-to-end）方法

- DNN模块
  - CNN→处理图片
  - RNN→处理语言
- 处理流程
  - 把图片和问题编码到同一个特征空间下
  - 结合图片和问题，理解，推理，生成答案



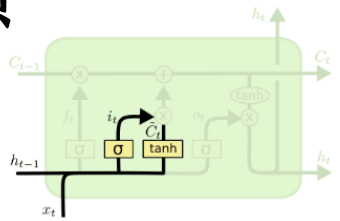
图像问答模型

Method	Joint Multimodal Embedding	Attention Mechanism	Dynamic Model	Knowledge Base Enhancement	Task Type	CNN Model	RNN Model
Neural-Image-QA	Y				Generation	GoogleNet	LSTM
2-VIS+BLSTM	Y				Classification	VGG-Net	LSTM
mQA	Y				Generation	GoogleNet	LSTM
MRN	Y				Classification	ResNet	LSTM
DualNet	Y				Classification	VGG-Net ResNet	LSTM
SANs		Y			Classification	GoogleNet	LSTM
ABC-CNN		Y			Classification	VGG-Net	LSTM
MCB	Y	Y			Classification	ResNet	LSTM
LSTM-Att		Y			Classification	VGG-Net	LSTM
Region-Sel		Y			Classification	VGG-Net	LSTM
FDA		Y			Classification	ResNet	LSTM
HieCoAtt		Y			Classification	ResNet	LSTM
DppNet	Y		Y		Classification	VGG-Net	GRU
DMN+			Y		Classification	VGG-Net	GRU
NMN			Y		Classification	VGG-Net	LSTM
Attributes-LSTM				Y	Generation	VGG-Net	LSTM
ACK				Y	Generation	VGG-Net	LSTM
Ahab				Y	Generation	VGG-Net	
Facts-VQA				Y	Generation	VGG-Net	LSTM

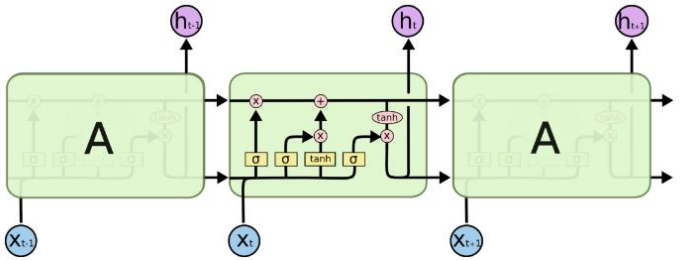
图像问答模型

LSTM回顾

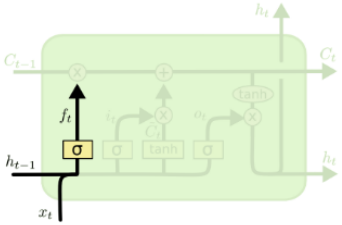
- 输入门→



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

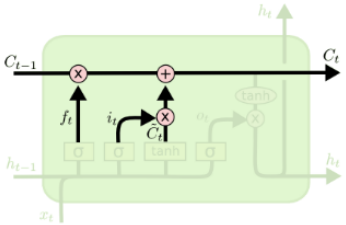


- 忘记门→

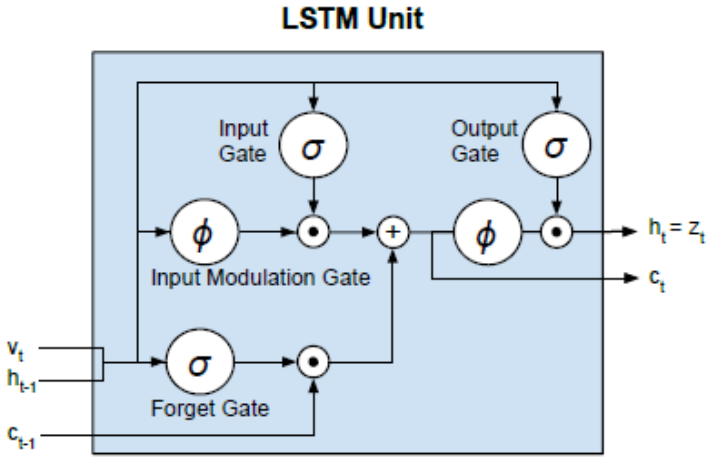


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

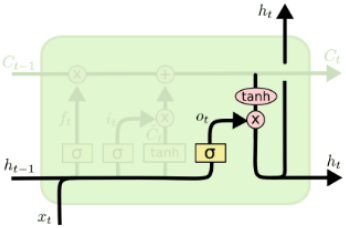
- 记忆更新→



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



- 输出门→

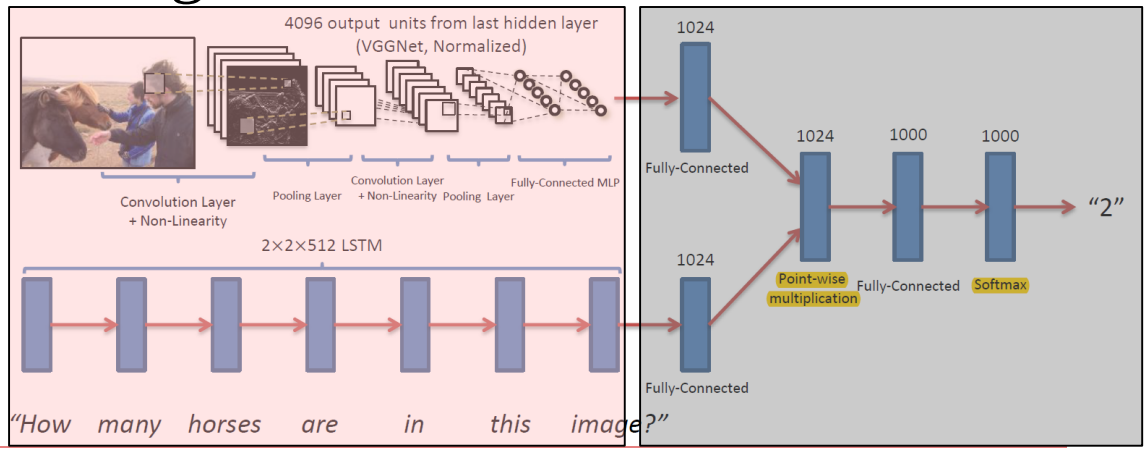


$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

图像问答模型

基本模型结构

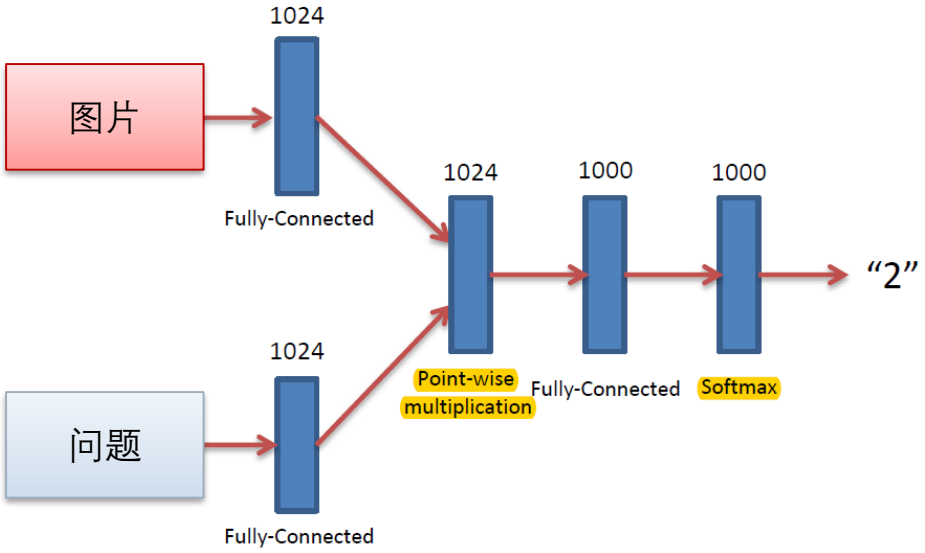
- CNN提取图片特征
  - VGG倒数第二个全连接层（4096）输出
- RNN提取问题特征
  - LSTM最后一个时刻的隐含层
- 文本特征Embedding
  - One-hot
  - 全链接层



图像问答模型

基本模型结构

- 特征映射（Feature embedding）
  - 2个1024维的全连接层
- 特征融合
  - 点乘：2*1024→1024
- 特征推理
  - 1个全连接层
  - 1个Softmax层



图像问答模型

模型实验

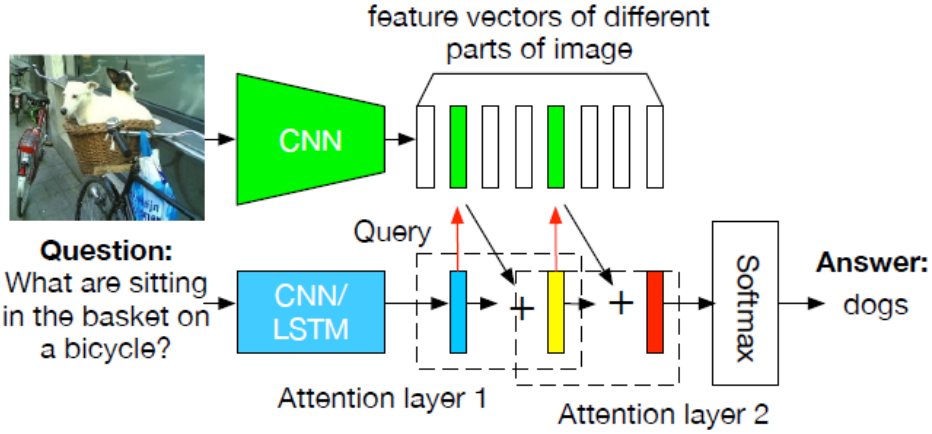
- 训练要点
  - VGG使用在ImageNet预训练的模型
  - 只训练非VGG部分网络，不fine-tuneVGG
- VQA-Real训练集上的性能

	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
I (Image)	28.13	64.01	0.42	3.77	30.53	69.87	0.45	3.76
LSTM Q (Question)	48.76	78.2	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q + I	**53.74**	**78.94**	**35.24**	**36.42**	**57.17**	**78.95**	**35.8**	**43.41**
2LSTM Q + I	50.39	78.41	34.68	30.03	55.88	78.45	35.91	41.13
2LSTM Q + Norm I	**57.75**	**80.5**	**36.77**	**43.08**	**62.7**	**80.52**	**38.22**	**53.01**

模型增强：注意机制(Attention Mechanism)

基本模型结构

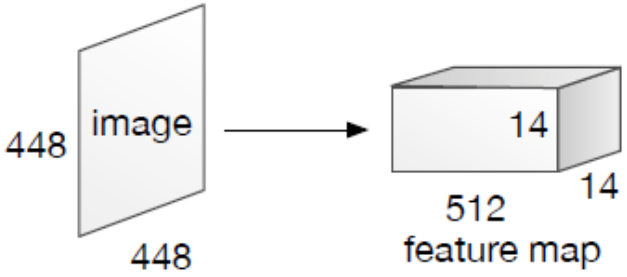
- 图片特征→ VGG
- 问题特征→LSTM
- 堆栈注意网络（Stacked Attention Network）
  - 查询语义相关的区域，虑除噪声区域
  - 多步推理，精华查询



模型增强：注意机制(Attention Mechanism)

基本模型结构

- 图片模型结构 → VGG
  - 输入图片尺寸488*488
  - 输出最后一个池化层
    - 带有空间信息14*14区域
    - 单个区域尺寸32*32
    - 特征维数512
- 图片特征Embedding模块
  - 单层神经网络
  - 将图片特征映射到共享特征空间



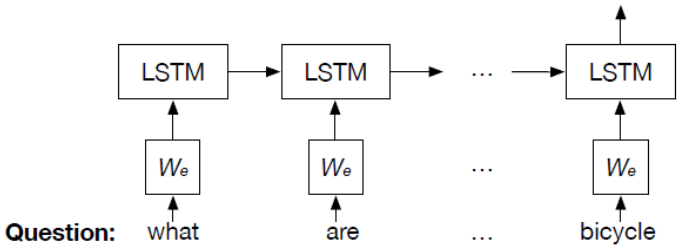
$$f_I = \text{CNN}_{vgg}(I)$$

$$v_I = \tanh(W_I f_I + b_I)$$

模型增强：注意机制(Attention Mechanism)

基本模型结构

- 问题模型结构 → LSTM
  - 问题序列中的不同单词对用不用时刻
  - 输出最后一时刻的隐含层
- 文本特征Embedding模块
  - 文本One-hot编码
  - 映射矩阵
  - 将One-hot编码映射到共享特征空间



$$x_t = W_e q_t, t \in \{1, 2, ...T\},$$
$$h_t = \text{LSTM}(x_t), t \in \{1, 2, ...T\}.$$

模型增强：注意机制(Attention Mechanism)

基本模型结构

- 堆栈注意网络
  - 不断的增强跟答案相关区域的特征权重，跟问题特征融合为增强特征，用于生成答案
  - 部分1→为14*14区域生成注意权重
    - 单层神经网络 $h_A^k = \tanh(W_{I,A}^k v_I \oplus (W_{Q,A}^k u^{k-1} + b_A^k))$
    - Softmax层 $p_I^k = \text{softmax}(W_P^k h_A^k + b_P^k)$
  - 部分2→生成增强特征
    - 图片区域特征权重相加 $\tilde{v}_I = \sum_i p_i v_i,$
    - 合并图片、问题特征 $u = \tilde{v}_I + v_Q$
  - 部分3→预测答案（K步推理之后）
    - 单层神经网络 $p_{\text{ans}} = \text{softmax}(W_u u^K + b_u)$

模型增强：注意机制(Attention Mechanism)

实验

- SGD
  - Momentum 0.9
  - Batch size 100
  - 梯度剪切&Dropout
- 2步性能最好

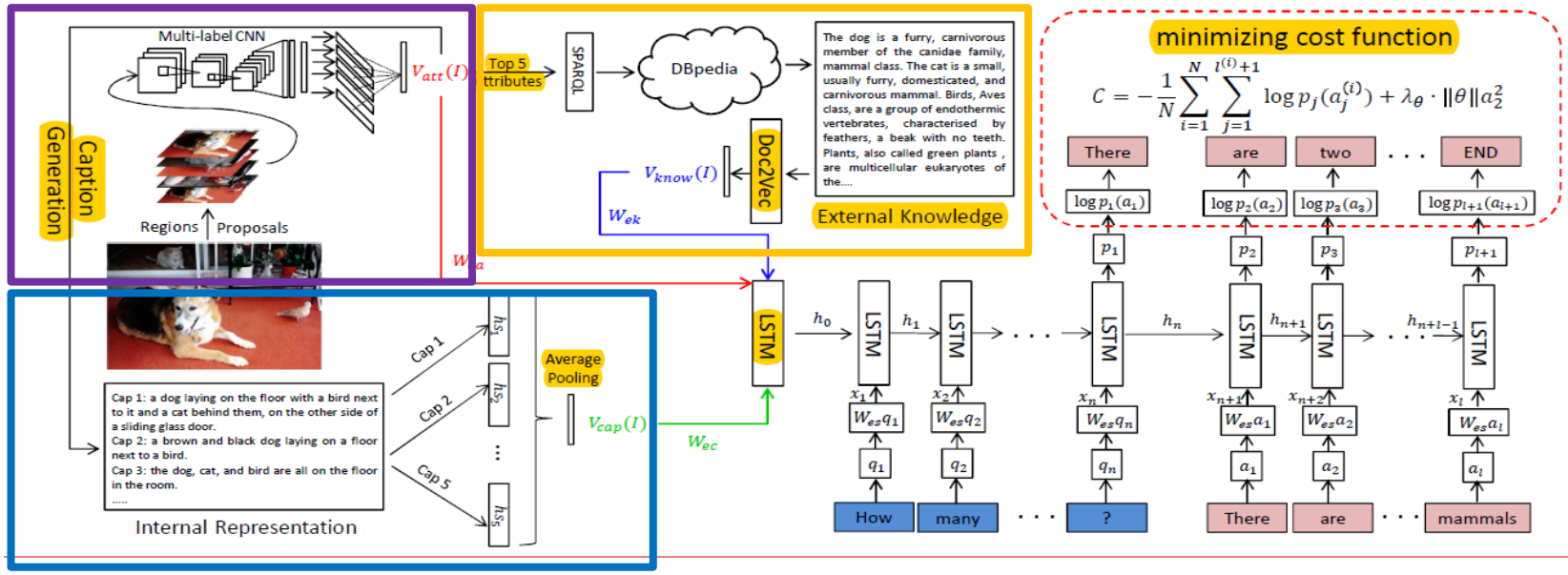
	test-dev				test-std
Methods	All	Yes/No	Number	Other	All
VQA: [1]					
Question	48.1	75.7	36.7	27.1	-
Image	28.1	64.0	0.4	3.8	-
Q+I	52.6	75.6	33.7	37.4	-
LSTM Q	48.8	78.2	35.7	26.6	-
LSTM Q+I	53.7	78.9	35.2	36.4	54.1
SAN(2, CNN)	**58.7**	**79.3**	**36.6**	**46.1**	**58.9**

Table 5: VQA results on the official server, in percentage

模型增强：外部知识库(Knowledge Base)

基本模型结构

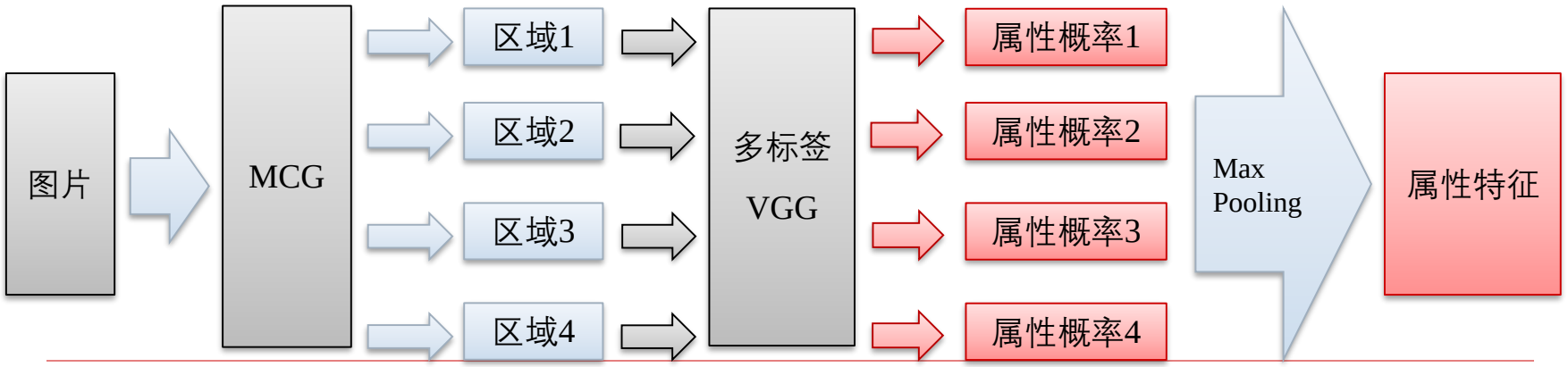
- Multi-Label CNN→属性特征$V_{att}(I)$
- Caption LSTM→内部特征$V_{cap}(I)$
- DBpedia→外部特征$V_{know}(I)$
- QA LSTM→基于3个特征，解析问题，生成答案



模型增强：外部知识库(Knowledge Base)

属性预测模型

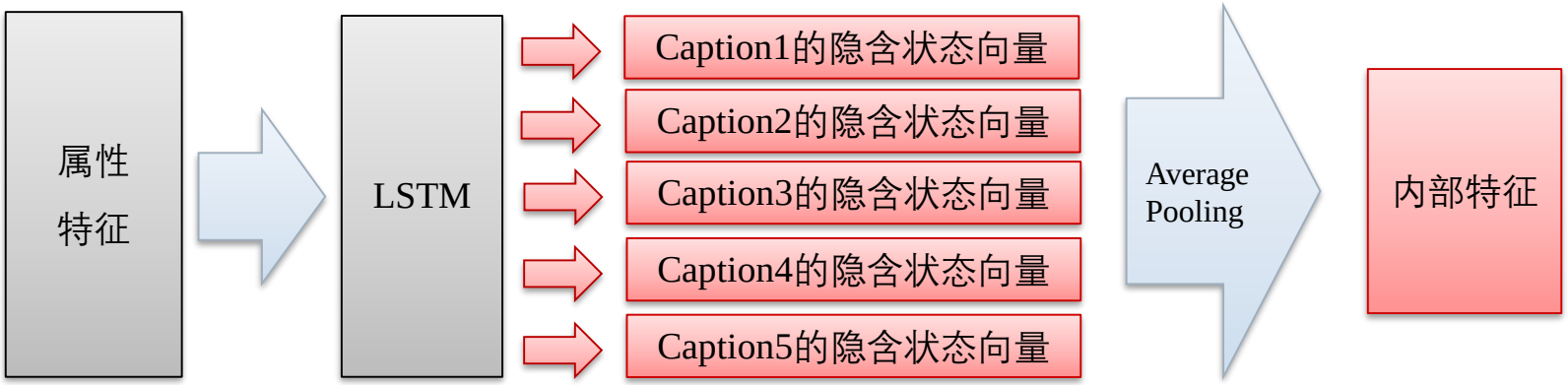
- 基于区域的多标签分类
  - 模型结构
    - 图片→候选区域→CNN分类器→属性特征
  - 属性字典
    - MS-COCO的Caption数据集中出现频度Top256的词
    - 词性：名词、动词、形容词



模型增强：外部知识库(Knowledge Base)

图说模型

- CNN属性特征+LSTM
  - 模型结构
    - 图片→属性特征→LSTM→5个Captions
  - 基于Caption的内部特征
    - 生成完最后一个词的隐含状态向量
    - 使用Average pooling将5个特征合成1个

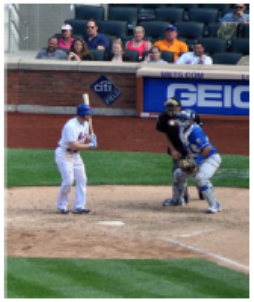


模型增强：外部知识库(Knowledge Base)

知识库模型

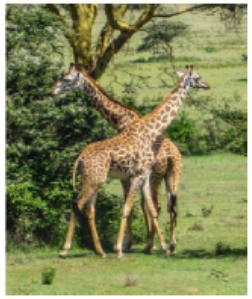
- Dbpedia知识库
  - 基于Wikipedia结构化的信息数据库
  - 可使用类SQL语言SPARQL查询
  - 使用Top5属性词分别查询出5段comments
- Doc2Vec
  - 提取不定长段落的定长语义特征

模型增强：外部知识库(Knowledge Base)



Top 5 Attributes:
players, catch, bat, baseball, swing

Generated Captions:
A baseball player swing a bat at a ball.
A baseball player holding a bat on a field.
A baseball player swinging a bat on a field.
A baseball player is swinging a bat at a ball.
A batter catcher and umpire during a baseball game.



Top 5 Attributes:
field, two, tree, grass, giraffe

Generated Captions :
Two giraffes are standing in a grassy field.
A couple of giraffe standing next to each other.
Two giraffes standing next to each other in a field.
A couple of giraffe standing next to each other on a lush green field.

Top 5 Attributes:
pizza, bottle, sitting, table, beer

Generated Captions :
A large pizza sitting on top of a table.
A pizza sitting on top of a white plate.
A pizza sitting on top of a table next to a beer.
A pizza sitting on top of a table next to a bottle of beer.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
sparql SELECT DISTINCT ?comment WHERE {
  ?entry rdfs: label "Dog"@en.
  ?entry rdfs: comment ?comment.
}
```

The domestic dog is a furry, carnivorous member of the canidae family, mammal class. Domestic dogs are commonly known as "man's best friend". The dog was the first domesticated animal and has been widely kept as a working, hunting, and pet companion. It is estimated there are between 700 million and one billion domestic dogs, making them the most abundant member of order Carnivora.

模型增强：外部知识库(Knowledge Base)

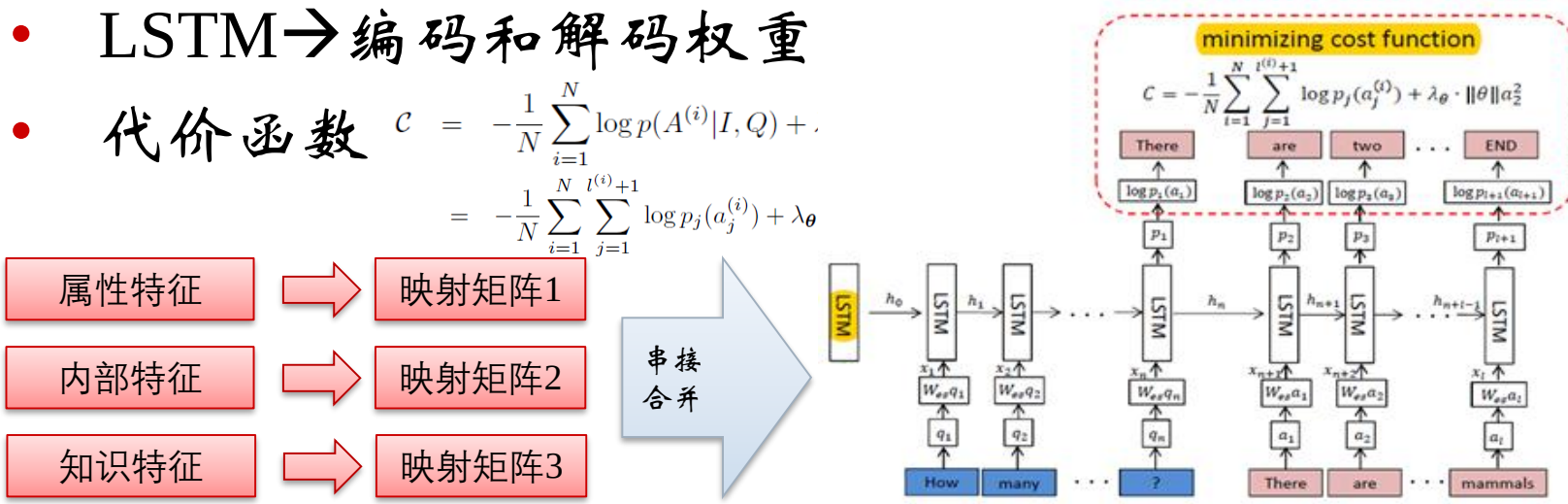
QA-LSTM模型 $\log p(A|I,Q) = \sum_{t=1}^{l} \log p(a_t|a_{1:t-1}, I, Q)$

- 输入
  - 初始时刻→三个特征映射到共享空间后做串接合并

$$x_{initial} = [W_{ea}V_{att}(I), W_{ec}V_{cap}(I), W_{ek}V_{know}(I)]$$

  - 剩余时刻→问题+答案序列 $\{q_1, ..., q_n, a_1, ..., a_l, a_{l+1}\}$
- LSTM→编码和解码权重
- 代价函数 $C = -\frac{1}{N}\sum_{i=1}^{N} \log p(A^{(i)}|I,Q) + \cdot$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{l^{(i)}+1} \log p_j(a_j^{(i)}) + \lambda_\theta$$



模型增强：外部知识库(Knowledge Base)

结果演示




	What color is the tablecloth?	How many people in the photo?	What is the red fruit?	What are these people doing?
Ours:	white	2	apple	eating
Vgg+LSTM:	red	1	banana	playing
Ground Truth:	white	2	apple	eating




	Why are his hands outstretched?	Why are the zebras in water?	Is the dog standing or laying down?	Which sport is this?
Ours:	balance	drinking	laying down	baseball
Vgg+LSTM:	play	water	sitting	tennis
Ground Truth:	balance	drinking	laying down	baseball

演示环节

- Github
  - https://github.com/349zzjau
- 百度网盘
  - http://pan.baidu.com/s/1gfpCCwj
- 代码演示
  - VQA

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

Q & A

小象账号：349zzjau

课程名：基于深度学习的计算机视觉

课后调查问http://cn.mikecrm.com/ZysMVWx

Reference List

[1] Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D. and Parikh, D., 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5014-5022).

[2] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C. and Parikh, D., 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2425-2433).

[3] Yang, Z., He, X., Gao, J., Deng, L. and Smola, A., 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 21-29).

[4] Wu, Q., Wang, P., Shen, C., Dick, A. and van den Hengel, A., 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4622-4630).

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop