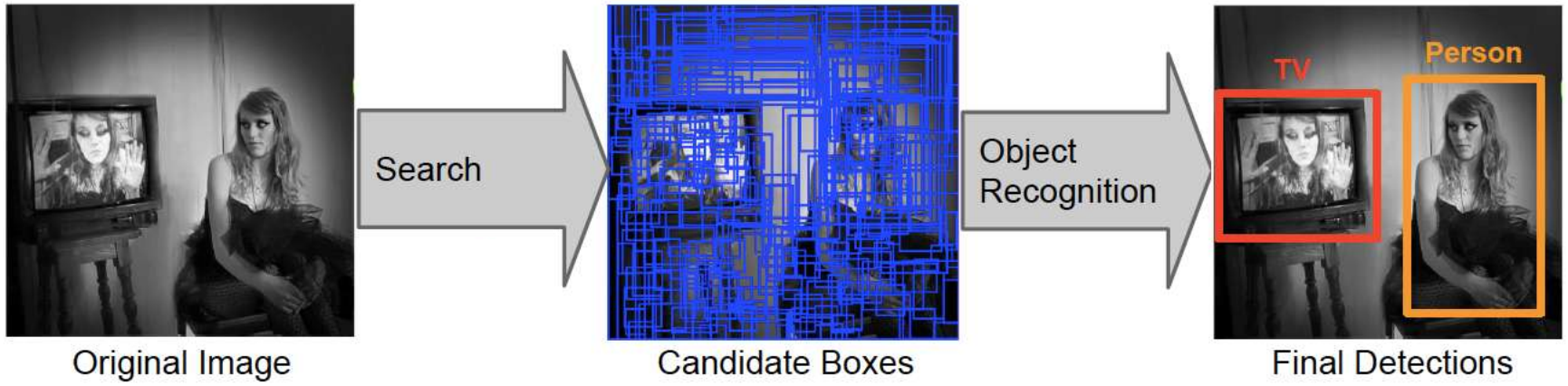
回顾——物体检测

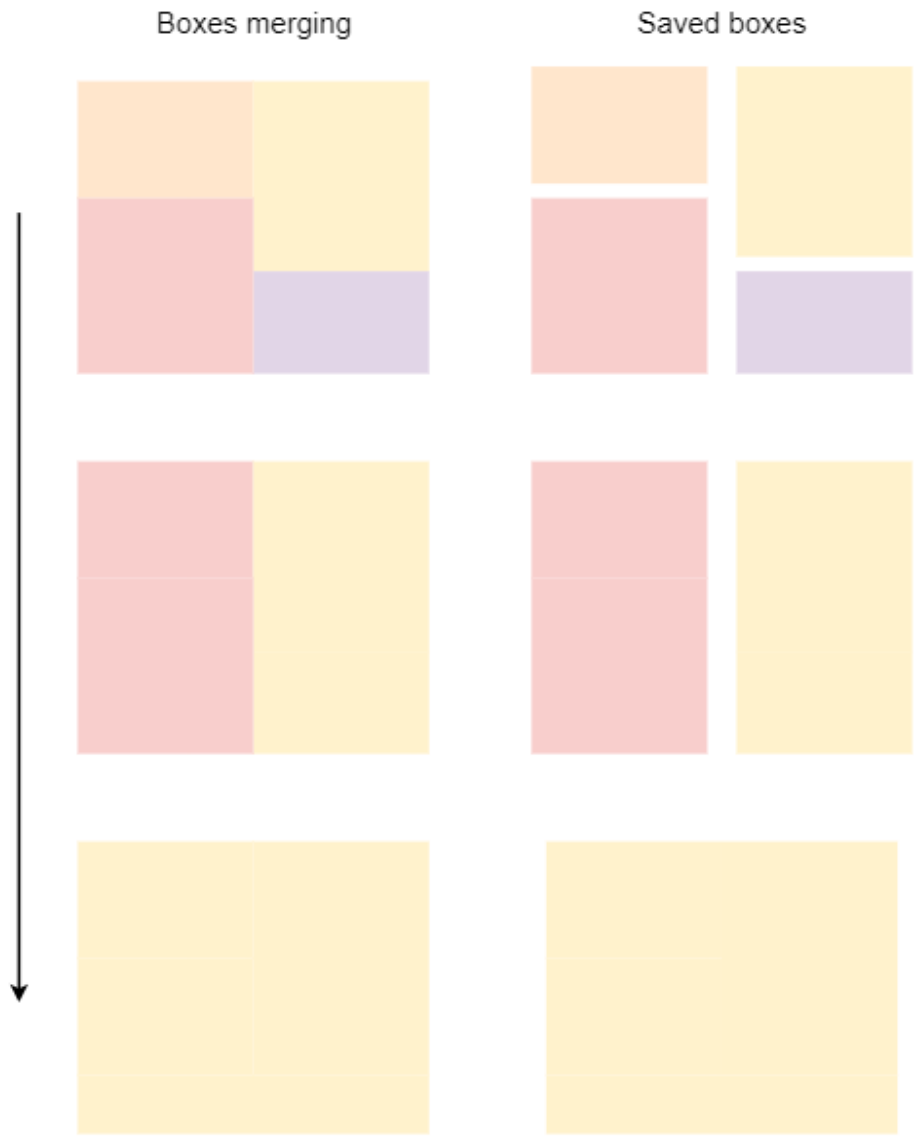
Region Proposals区域提议



- 替代滑动窗口，只选出部分候选框作为*region proposals 区域提议*
  - 谨记：特征提取和分类器速度较慢，区域提议可以减少候选区域，提升运行效率
  - 一般候选区域类别无关
  - 这个机制可以被训练（后面会讲）

Slide Credit: S. Lazebnik

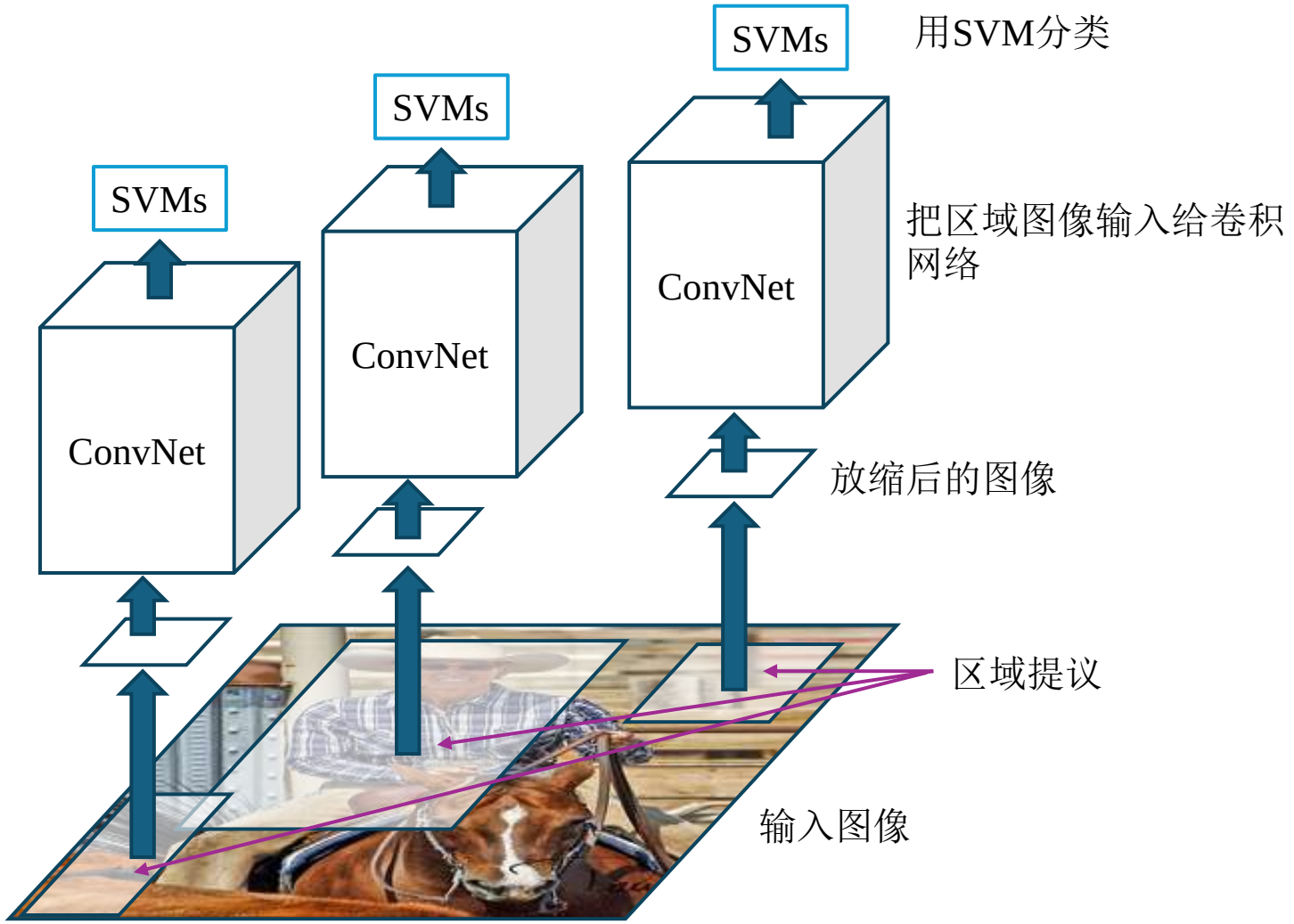
Region Proposals: Selective Search



Alexe et al, “Measuring the objectness of image windows”, TPAMI 2012
Uijlings et al, “Selective Search for Object Recognition”, IJCV 2013
Cheng et al, “BING: Binarized normed gradients for objectness estimation at 300fps”, CVPR 2014 Zitnick and Dollar, “Edge boxes: Locating object proposals from edges”, ECCV 2014

R-CNN: Region proposals + CNN features

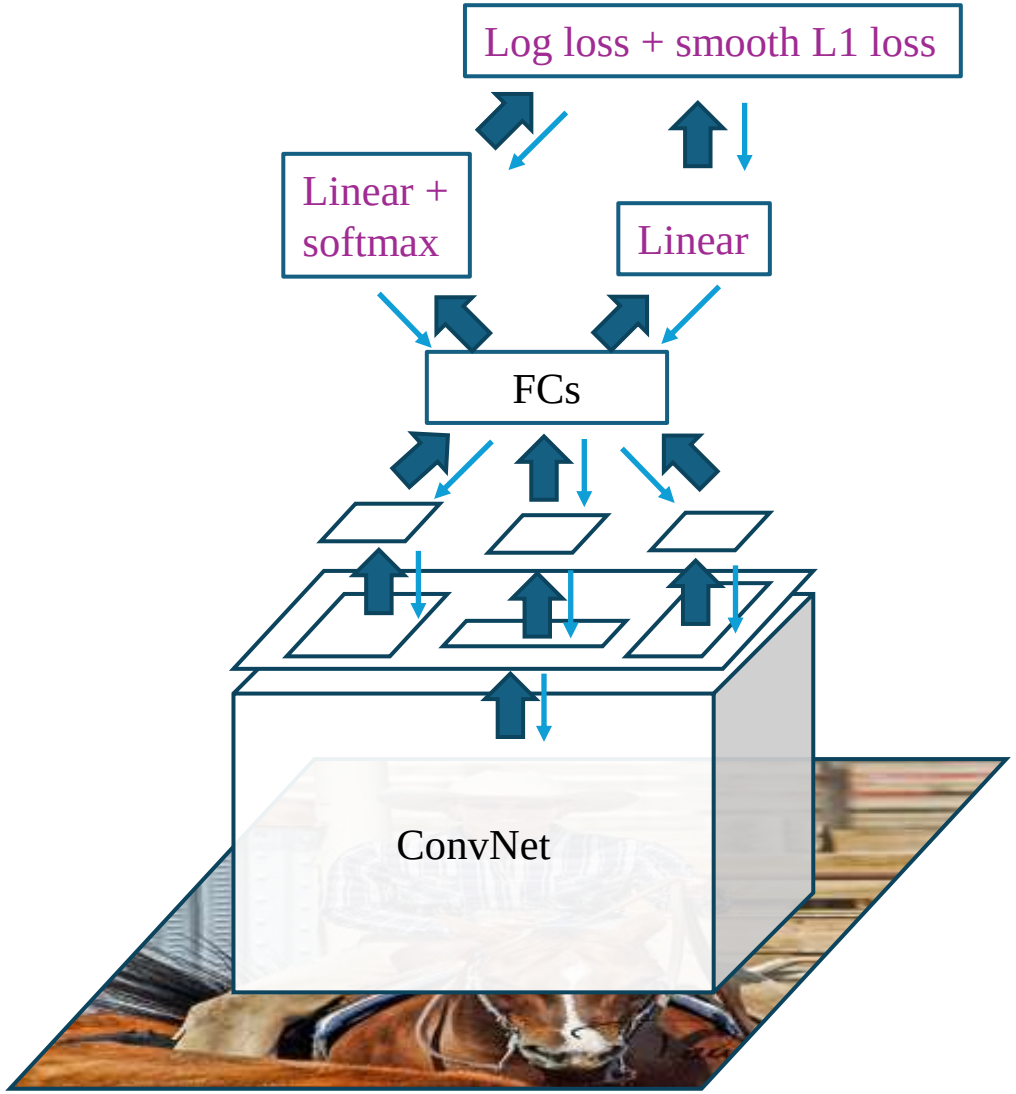
Source: R. Girshick



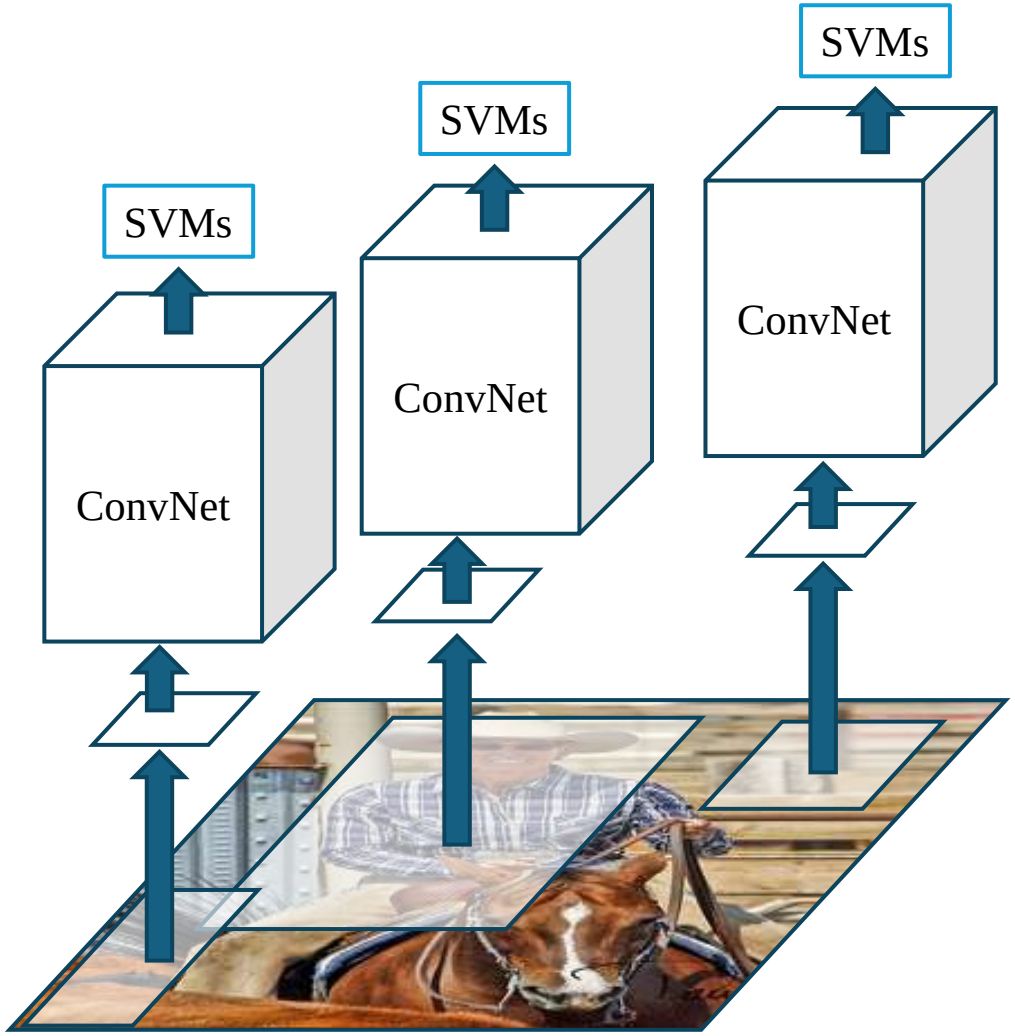
R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, CVPR 2014.

Fast R-CNN VS “Slow” R-CNN

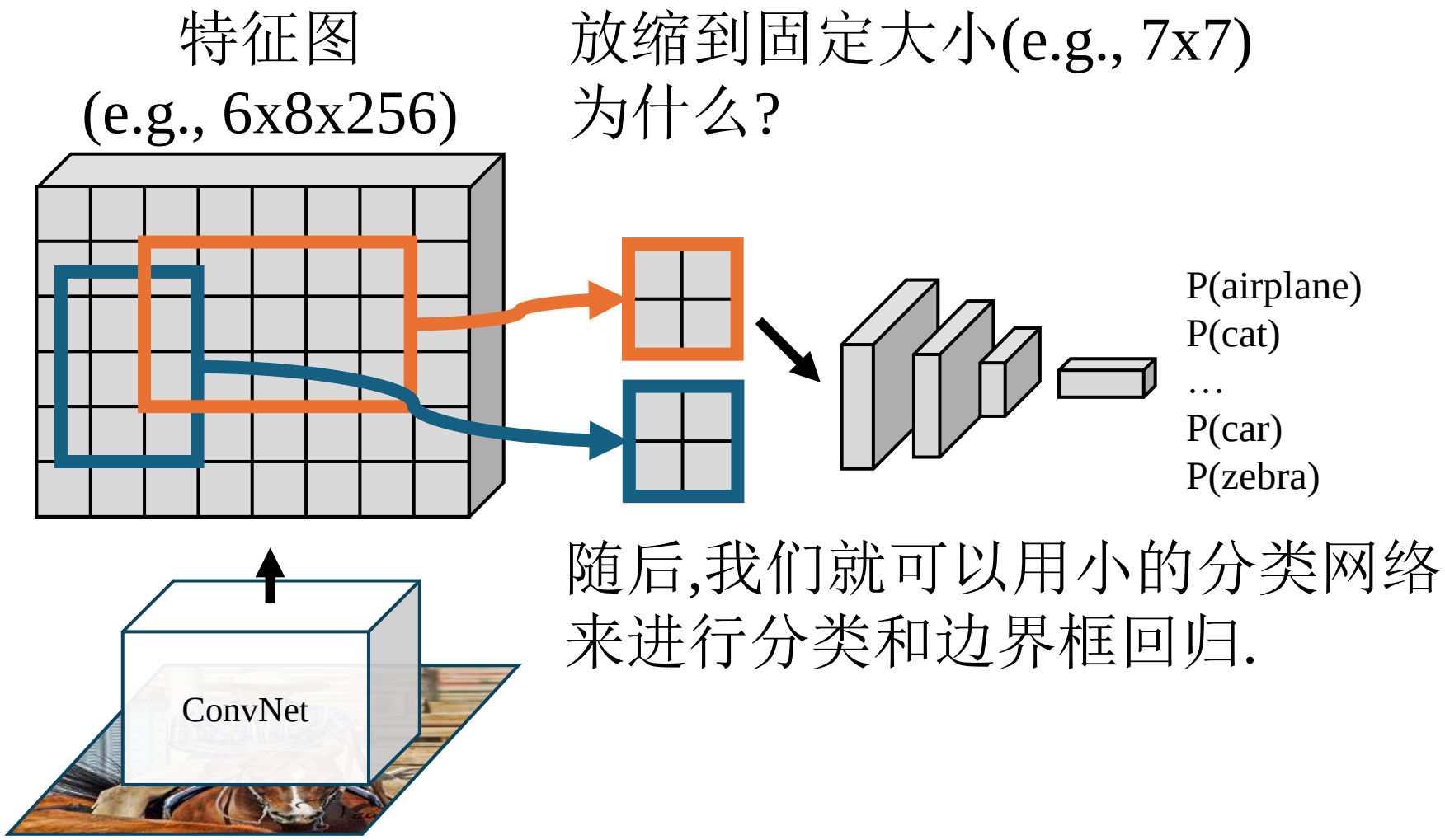
Fast R-CNN



“Slow” R-CNN



ROI Pooling/Align

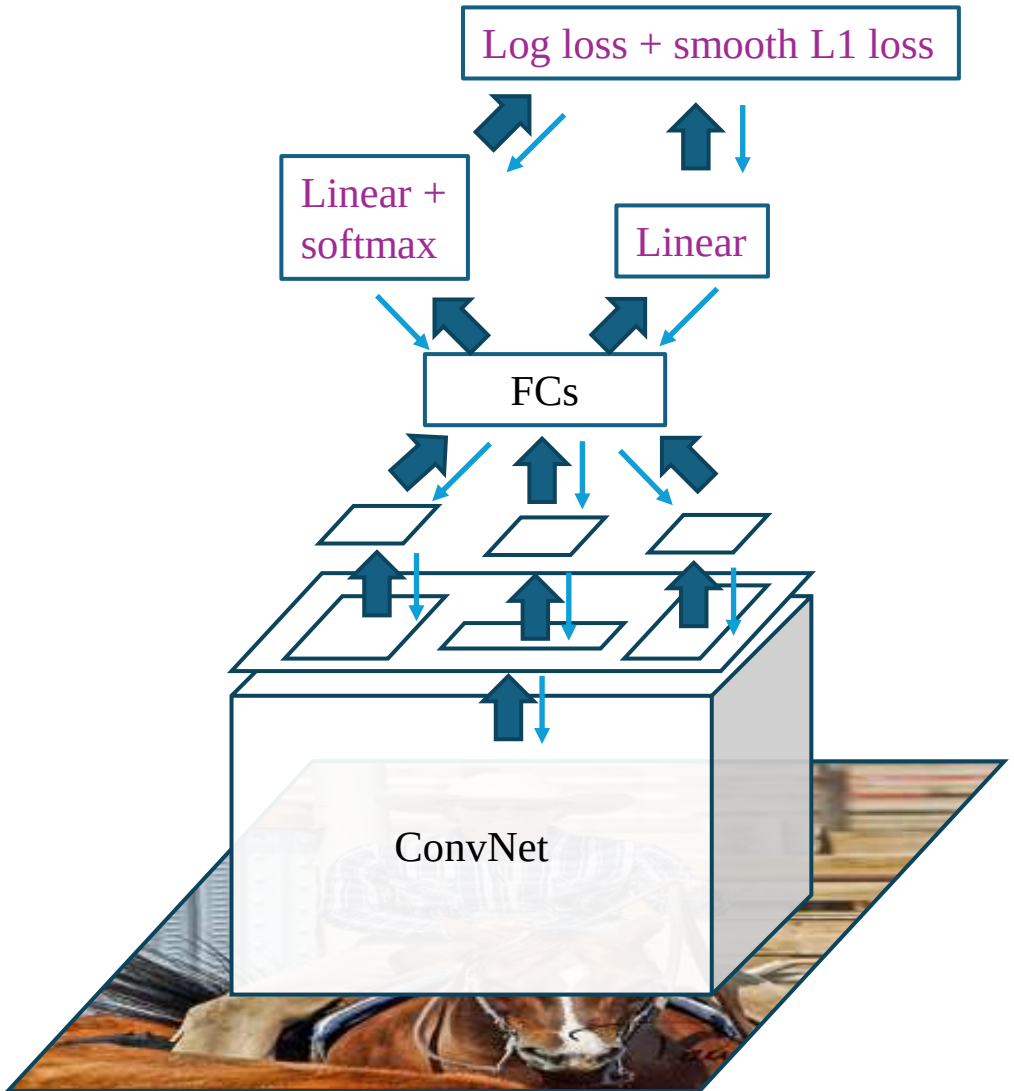


随后,我们就可以用小的分类网络来进行分类和边界框回归.

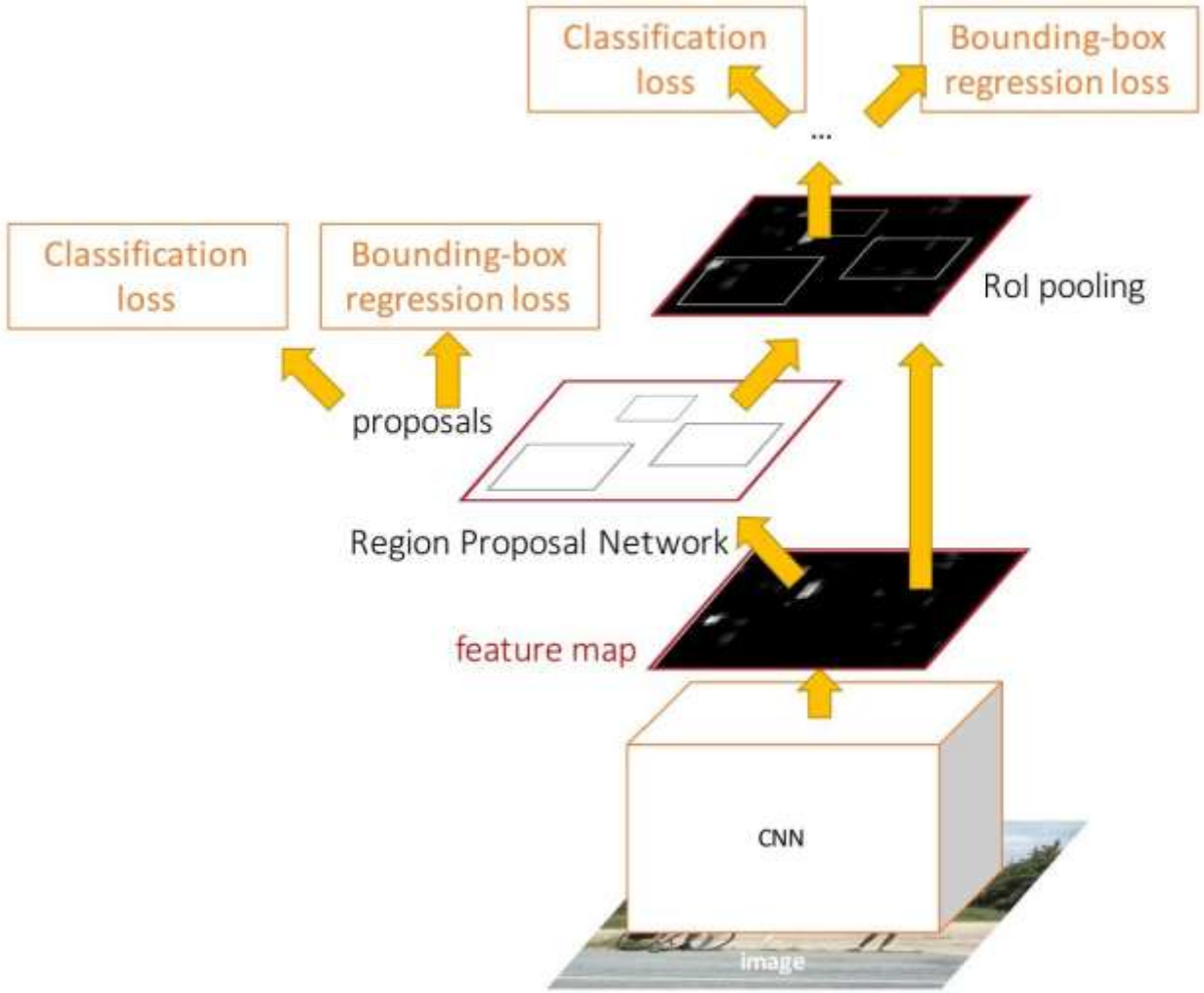
Faster R-CNN:
让CNN提Proposals！

Region Proposal Network (RPN) 预测区域proposal
其余的与Fast R-CNN一致

Fast R-CNN

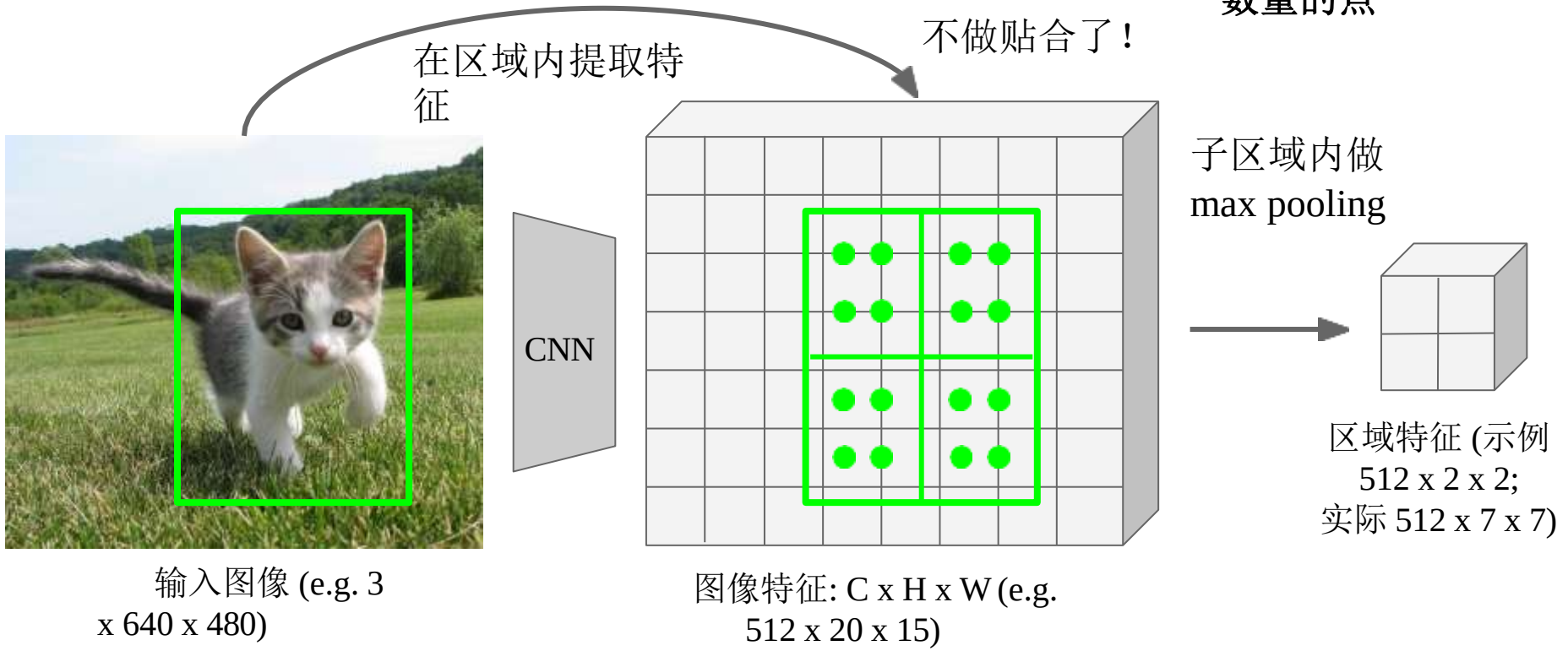


Faster R-CNN



RoI Pool升级版: RoI Align

在每个子区域采样固定数量的点



He et al, “Mask R-CNN”, ICCV 2017

MaskRCNN – 效果


traffic light.99
traffic light1.00
car.95
car.81
car.89
car.91
car.97
car.96
car.94
car.97
car.96
car.99
car.97
car.86
car1.00
parking meter


donut.86
donut.95
donut.89
donut.90
donut.89
donut.89
donut.93
donut.99
donut.96
donut.86
donut.95
donut.98
donut.89
donut1.00
donut.96
donut.98
donut.98
donut.99
donut.98
donut.89
donut.97
donut.95
donut.94
donut.98
donut1.00
donut.95
donut.99
donut.96
donut.90
donut.86
donut1.00

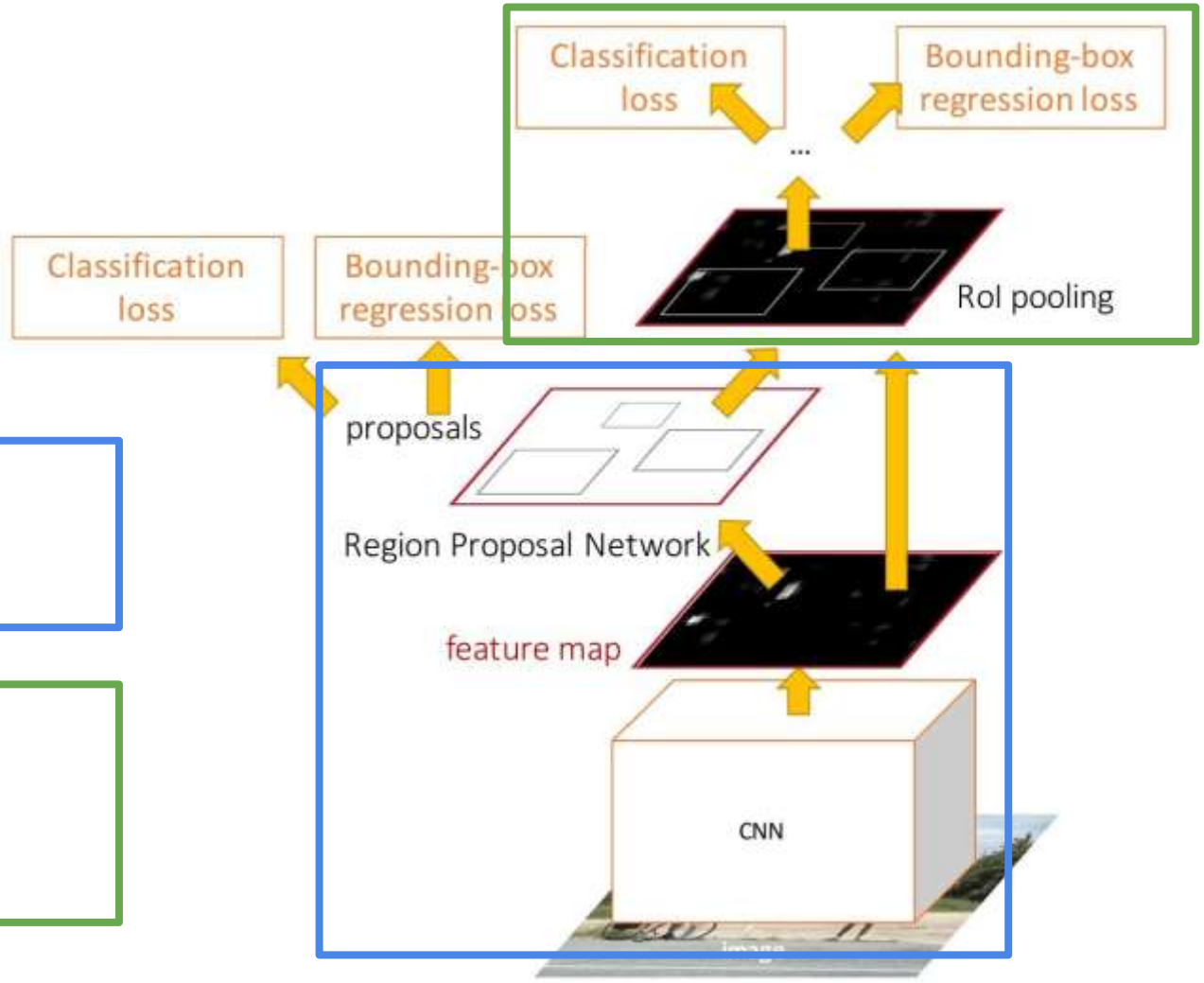
Faster R-CNN:
让CNN提Proposals！

Faster R-CNN 是一个
两阶段物体检测器

第一个阶段: 在每张图上
- Backbone network
- Region proposal network

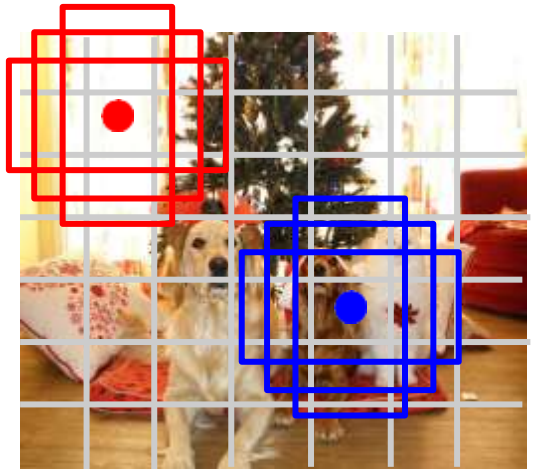
第二个阶段: 在每个区域上
- 裁剪特征: RoI pool / align
- 预测物体类别
- 预测框回归

我们真的需要第二个阶段吗？



单阶段物体检测方法: YOLO / SSD / RetinaNet

输入图像 3
x H x W



把图像分成7 x 7的格点

每张图会包含一些以格点为中心的**基础框**

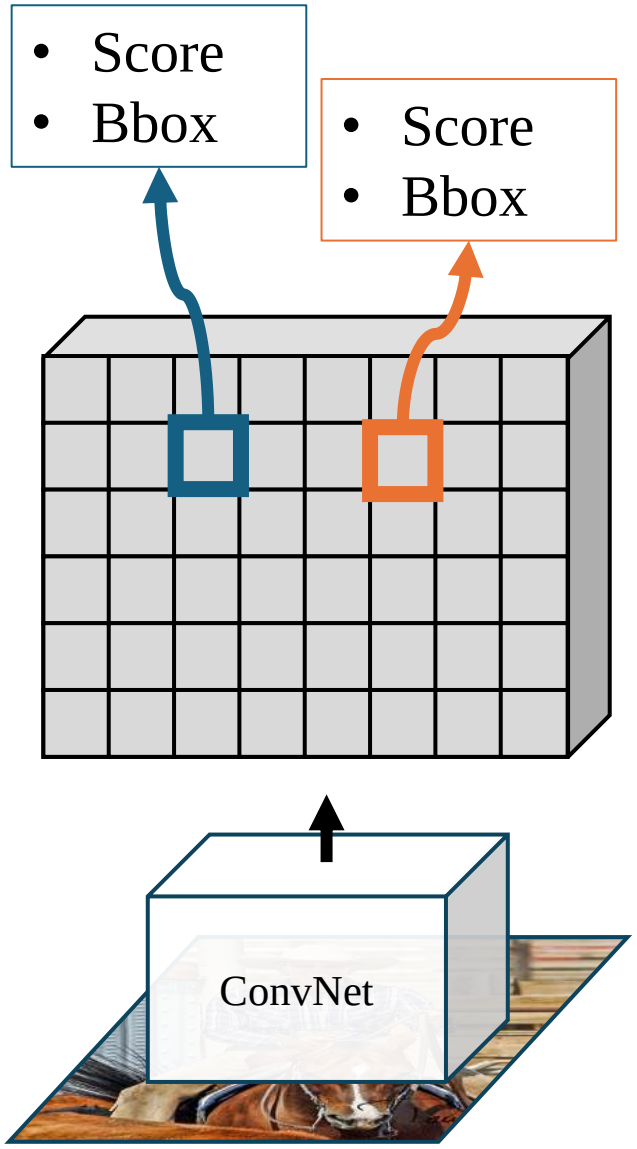
Here B = 3

对于每一个格点:

- 从B个 基础框（Anchor）来回归出最终的物体框:
  (dx, dy, dh, dw, confidence)
- 每个框对每个类都预测一个值 (包括背景)
- 类似于我们有7x7个RPN

Output:
7 x 7 x (5 * B + C)

YOLO



- 没有区域提议
- 对7x7的特征图上的每个点，预测分类评分+边界框校正
- 比 Faster-RCNN 快7倍，但精确度较低
- 由于其速度较快，在实际应用中，如机器人等，较常使用
- 众多版本

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, CVPR 2016

物体检测: 各种变种...

Backbone Network
VGG16
ResNet-101
Inception V2
Inception V3
Inception
ResNet
MobileNet

“Meta-Architecture”
Two-stage: Faster R-CNN
Single-stage: YOLO / SSD
Hybrid: R-FCN

Image Size
 # Region Proposals
…

总结
Faster R-CNN 慢，但是准很多
SSD/YOLO快，但是不准
更深的网络总是表现得更好

实例分割

分类

CAT

无空间限定

语义分割

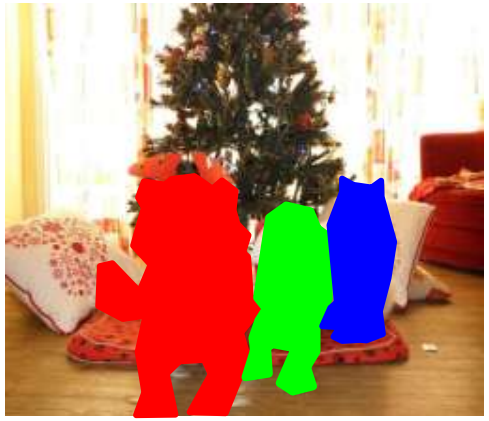
GRASS, CAT, TREE, SKY

像素级预测

物体检测

DOG, DOG, CAT

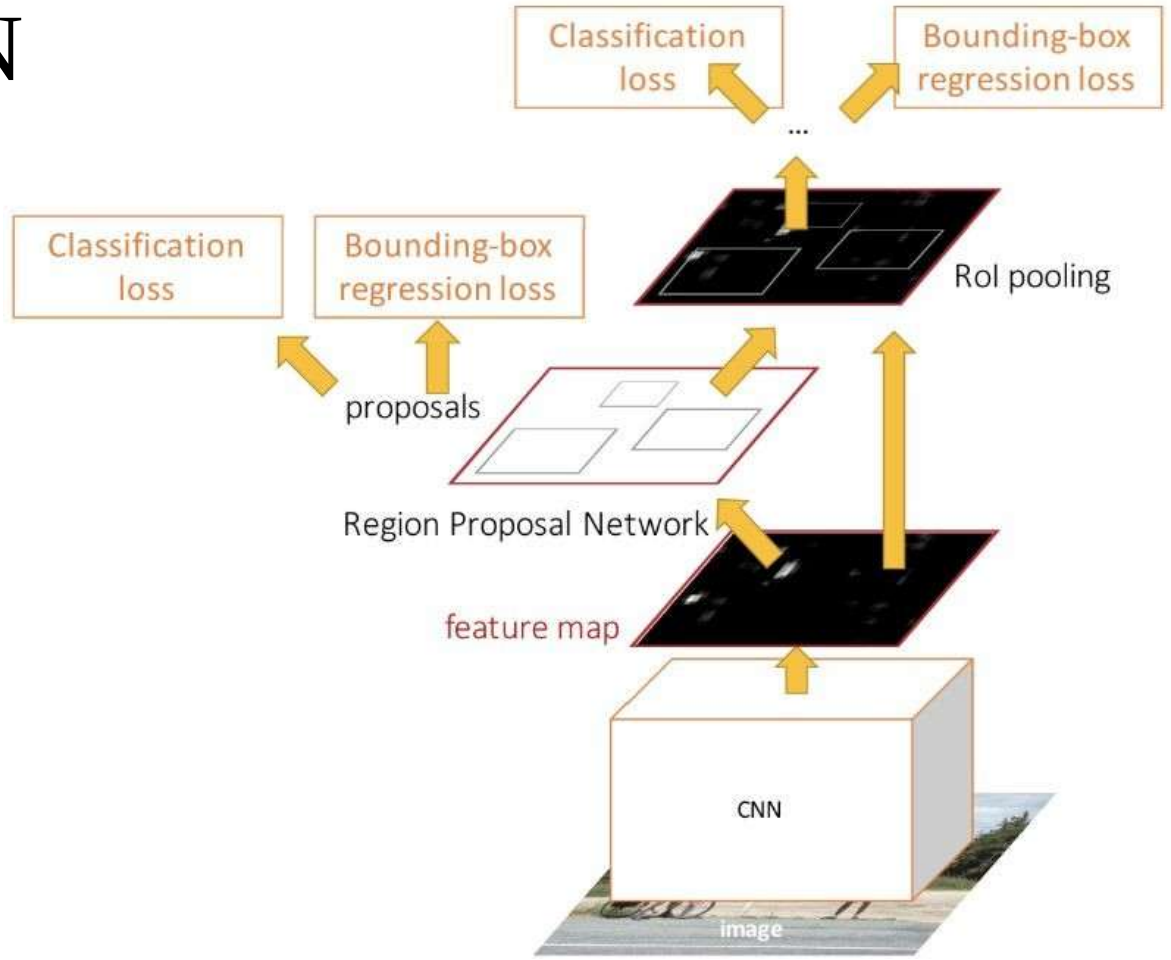
实例分割



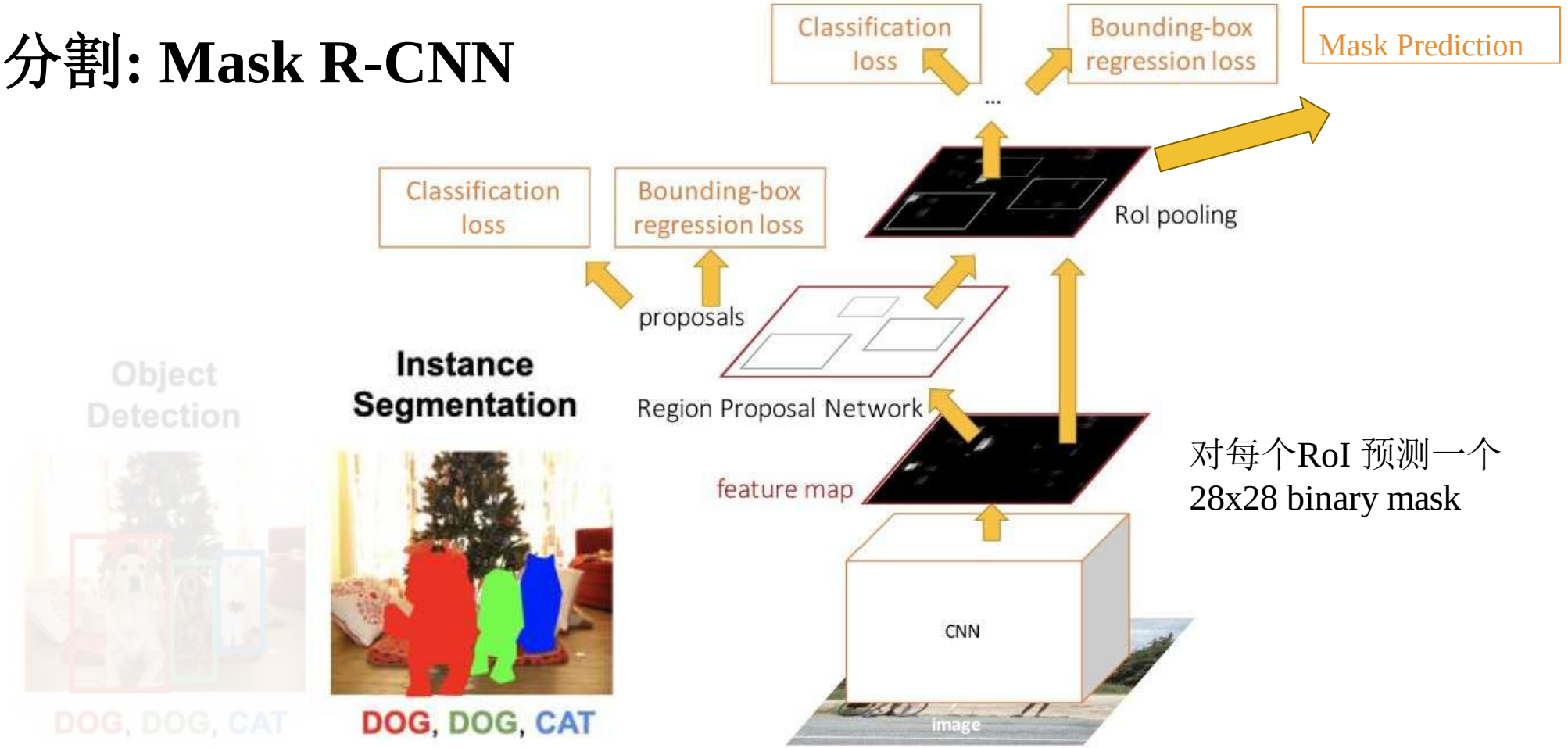
DOG, DOG, CAT

多物体

物体检测: Faster R-CNN



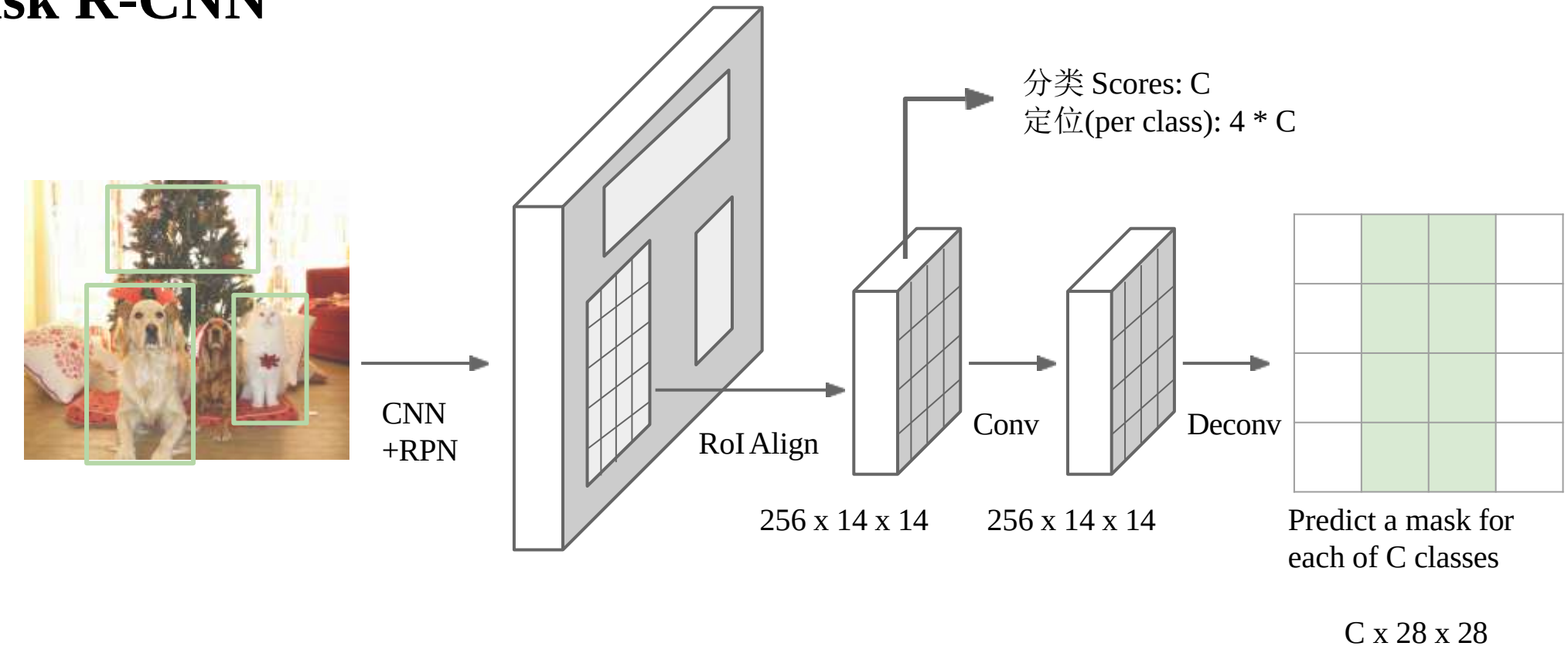
实例分割: Mask R-CNN



对每个RoI 预测一个28x28 binary mask

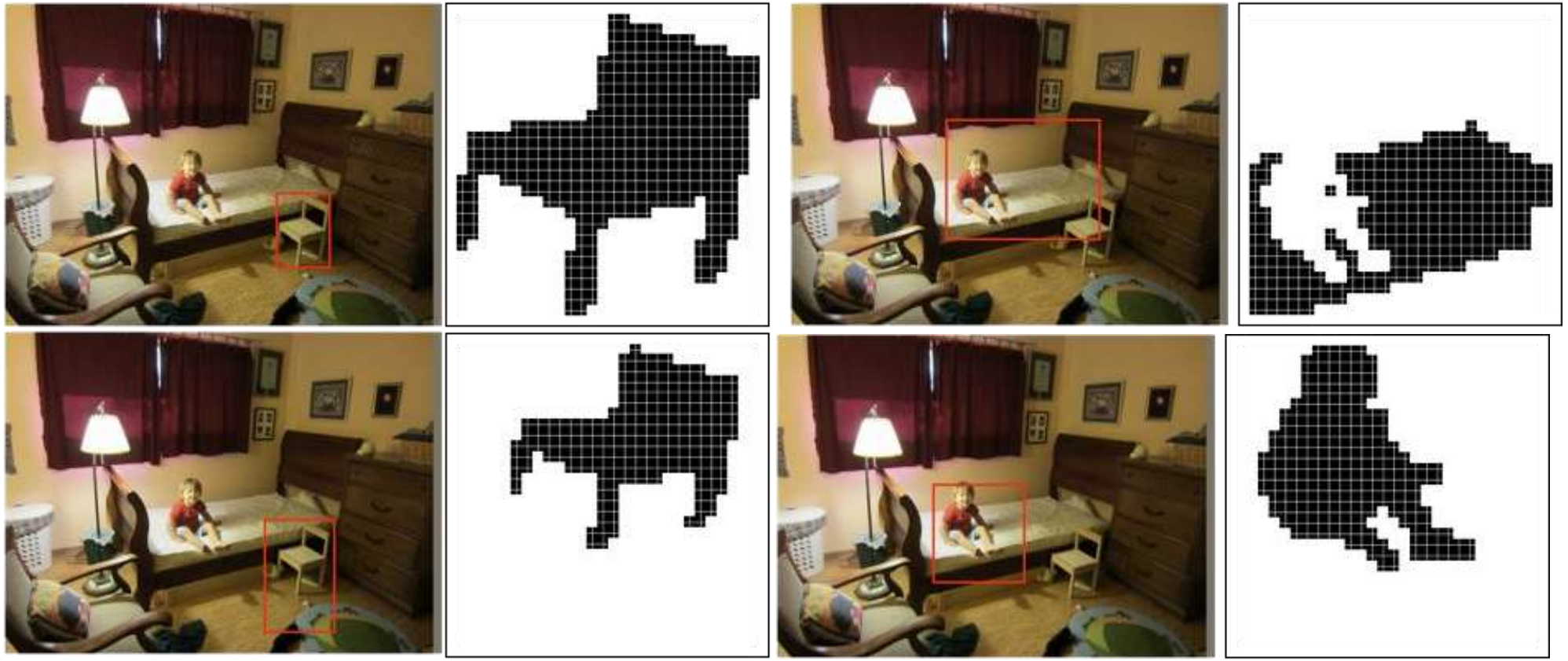
He et al, “Mask R-CNN”, ICCV 2017

Mask R-CNN

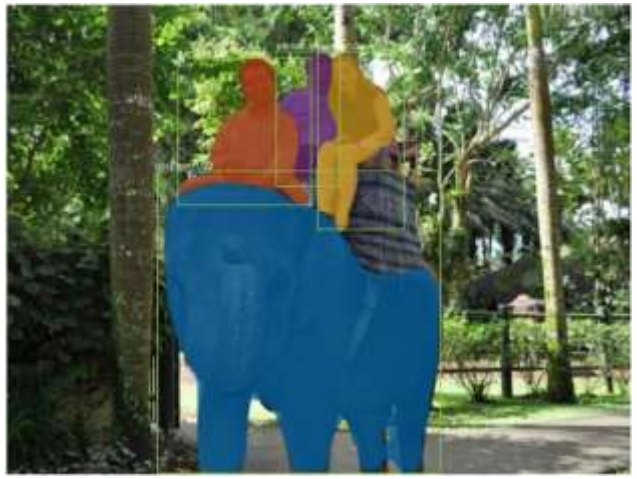
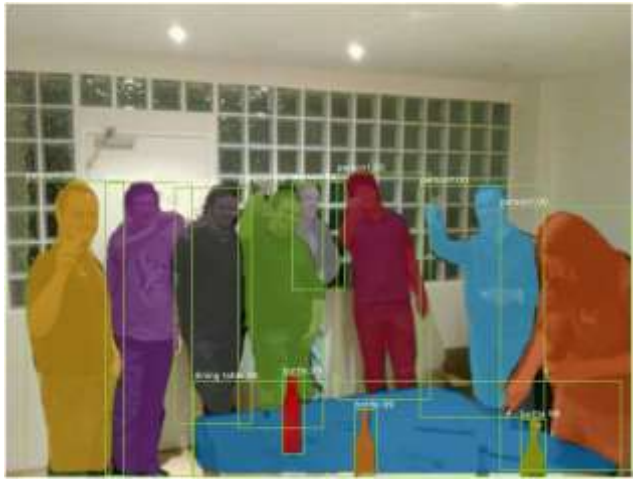


He et al, “Mask R-CNN”, arXiv 2017

Mask R-CNN: 例子



Mask R-CNN

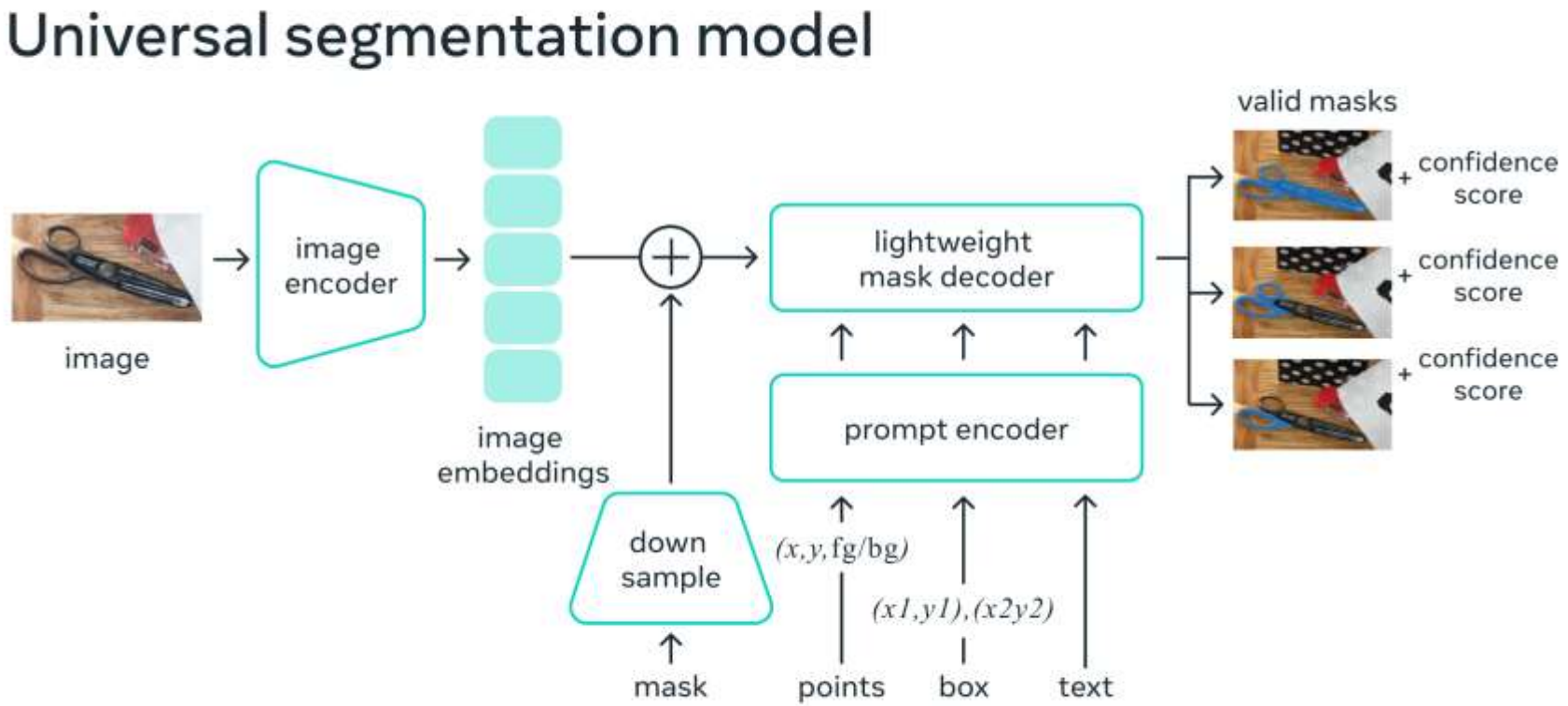
He et al, “Mask R-CNN”, ICCV 2017

Mask R-CNN 也可以做姿态估计

He et al, “Mask R-CNN”, ICCV 2017

Segment Anything

针对任何提示返回有效的分割掩码



Segment Anything

点击作为触发词

Segment Anything

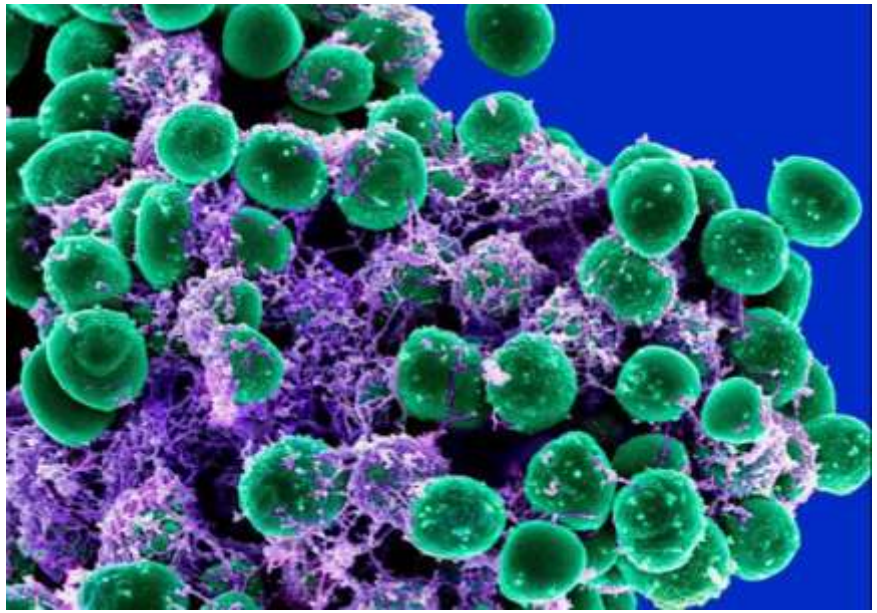
标签作为触发词

Segment Anything

“valid mask” 不同触发词对应的合理掩码

Segment Anything

零知识“zero-shot”预测



序列模型

这是啥?

Dog image credit: T. Gupta

回顾: 图像分类

图像特征向量

Cat weight vector	0.2	-0.5	0.1	2.0	1.1	56
Dog weight vector	1.5	1.3	2.1	0.0	3.2	231
Hat weight vector	0.0	0.3	0.2	-0.3	-1.2	24
						2
						1

W

x_i

-96.8	Cat score
437.9	Dog score
61.95	Hat score

Wx_i

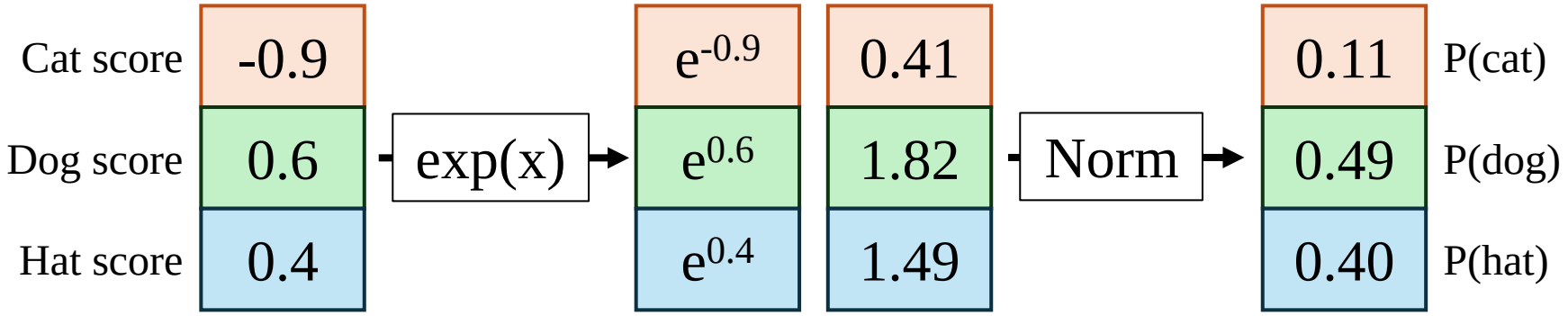
权重矩阵是每个类别的得分函数

为每一个类别预测得分

Diagram by: Karpathy, Fei-Fei

回顾: 图像分类

Softmax把得分转化为概率



$\sum$=3.72

P(class j):

$$\frac{\exp((Wx)_j)}{\sum_k \exp((Wx)_k)}$$

这样有什么问题?

这该标注为狗?还是帽子?

多分类

图像特征向量

Cat weight vector	0.2	-0.5	0.1	2.0	1.1	56
Dog weight vector	1.5	1.3	2.1	0.0	3.2	231
Hat weight vector	0.0	0.3	0.2	-0.3	-1.2	24
						2
						1

W

x_i

-96.8	Cat score
437.9	Dog score
61.95	Hat score

Wx_i

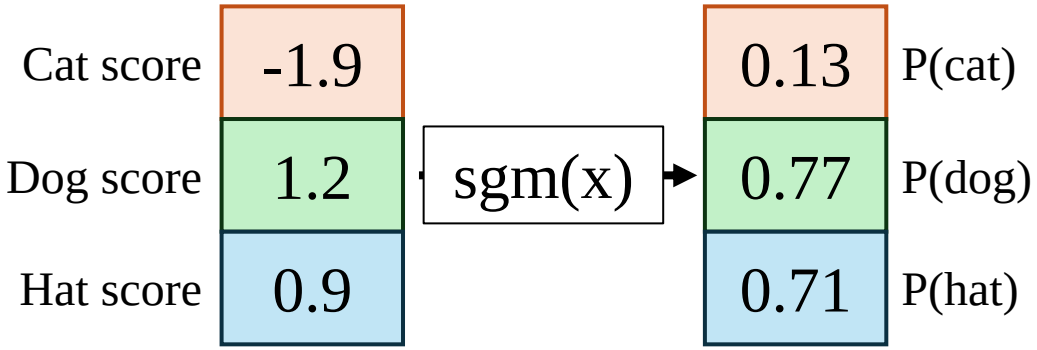
权重矩阵是每个类别的得分函数

为每一个类别预测得分

Diagram by: Karpathy, Fei-Fei

多分类

用Sigmoid把得分转化为概率



77% dog
71% hat
13% cat?

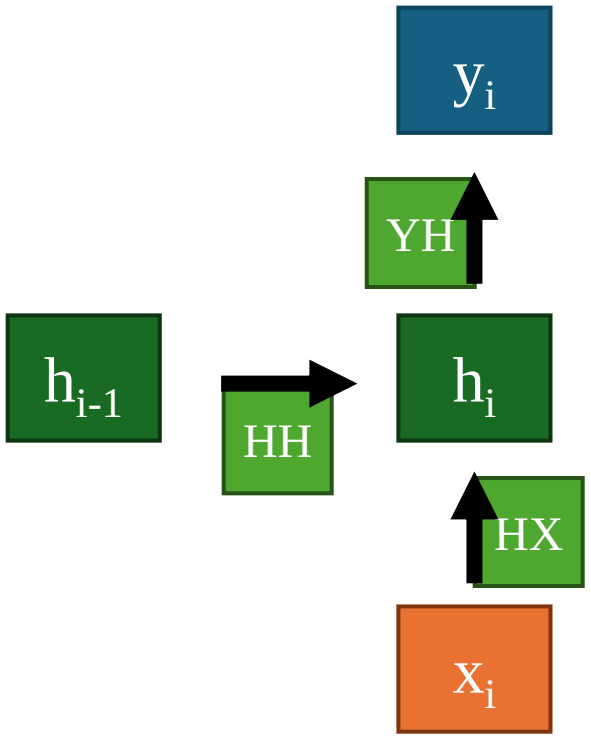
那么…

- 我们如果要做: “dog with a hat” 或者 “husky wearing a hat” 等
- 最直接的方式(给定 N 个单词,选中其中至多C 个). 有多少个方案?
- N=10k, C=5 -> 10的20次方
- 我们不可能训练一个分类器来实现这个任务

换个方式: 序列模型

- 建立N个单词的字典
- 新的目标:顺序去做C个N分类任务
- 字典应该具备:
  - 训练集中所有可能出现的单词
  - 包括一些特殊”单词”: START, END, UNK

Sequence Modeling 序列模型



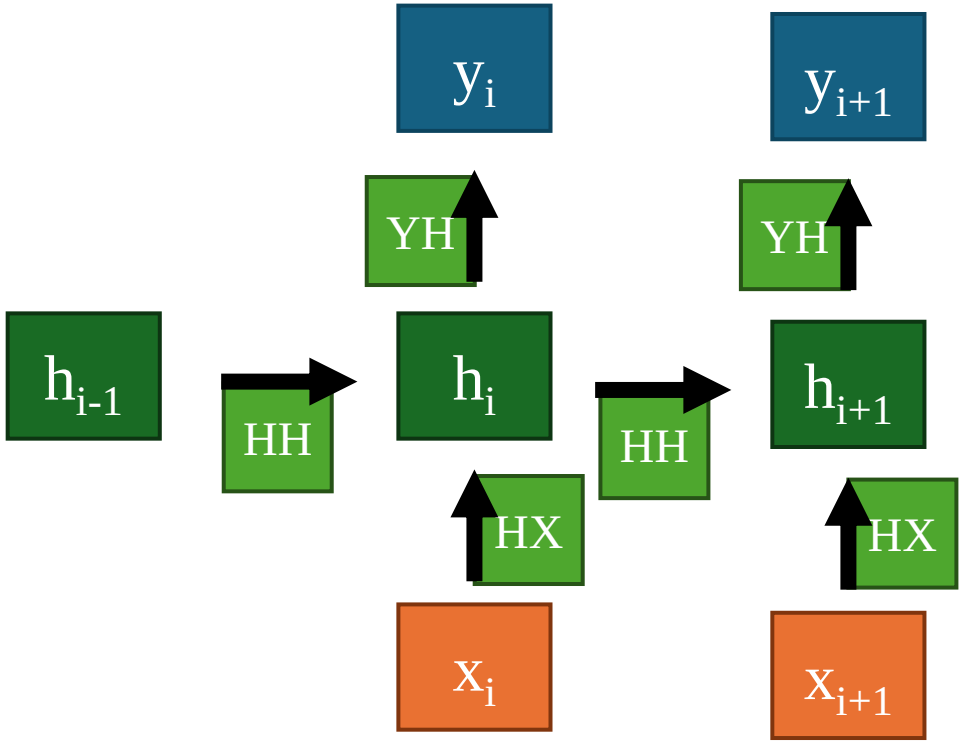
第 i 个输出$\boldsymbol{y_i}$是 隐藏状态$\boldsymbol{h_i}$ 的线性变换

$$\boldsymbol{y_i = W_{yh}\, h_i}$$

隐藏状态$\boldsymbol{h_i}$ 基于输入 $\boldsymbol{x_i}$，上一步的隐藏状态$\boldsymbol{h_{i-1}}$，加上非线性层

$$\boldsymbol{h_i = \sigma(W_{hx}x_i + W_{hh}h_{i-1})}$$

Sequence Modeling 序列模型

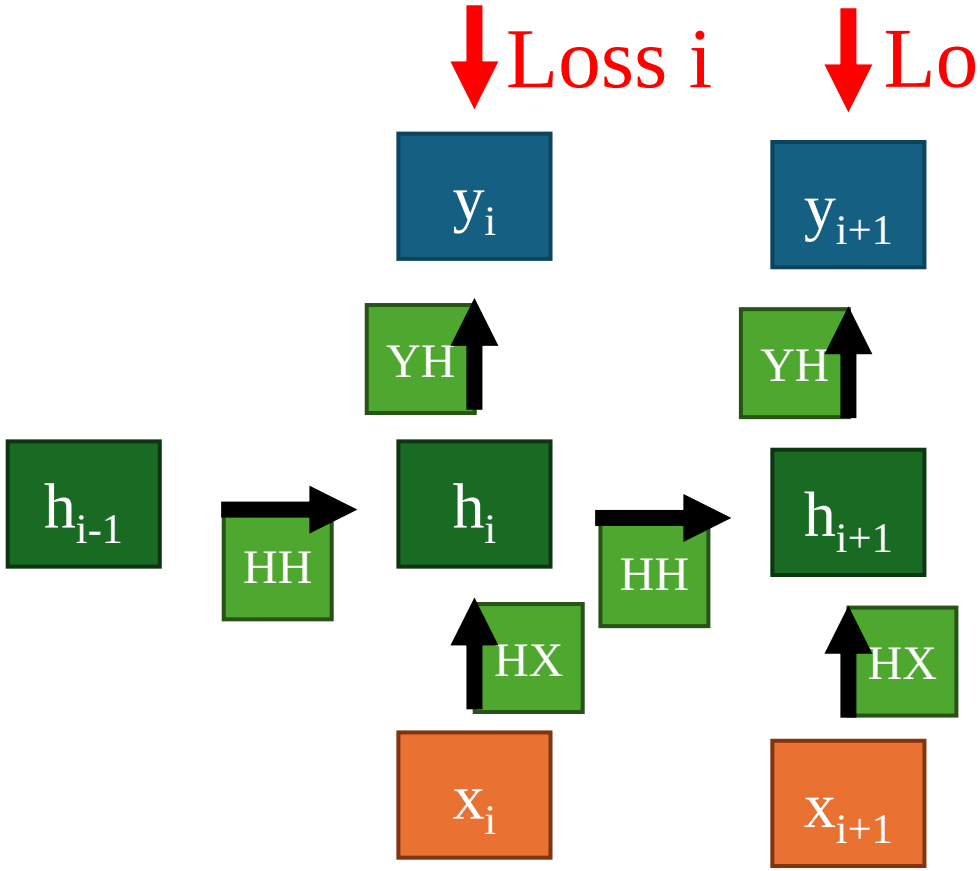


可以将多个RNN层叠加在一起，以处理更复杂的序列数据。在这种结构中，每一层的输出可以作为下一层的输入，允许网络学习更深层次的特征。模型可以用参数表达：W_{HX}, W_{HH}, W_{YH}

$$\boldsymbol{y_i = W_{yh}\, h_i}$$

$$\boldsymbol{h_i = \sigma(W_{hx} x_i + W_{hh} h_{i-1})}$$

Sequence Modeling 序列模型



可以针对每个输出定义损失，并相对于所有权重进行微分

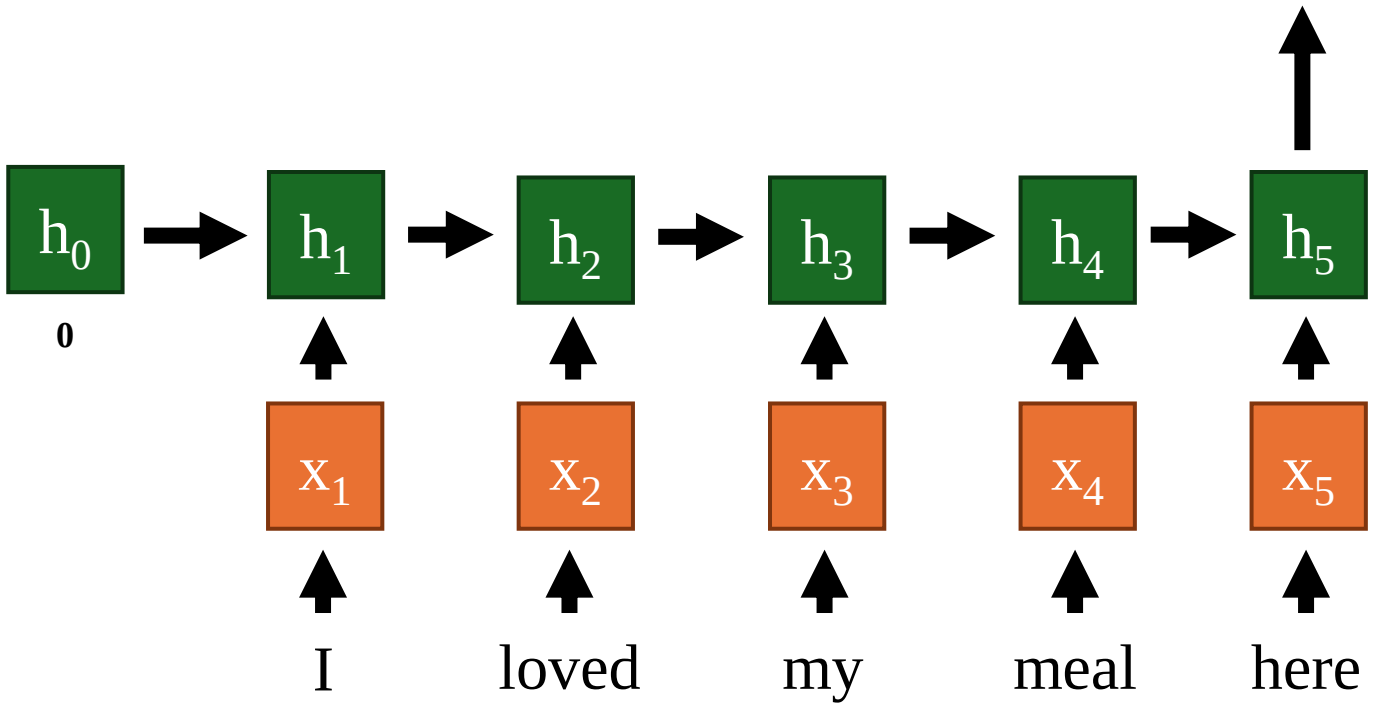
可以量化每个时间步的预测与实际值之间的误差，并通过时间上的反向传播优化网络

$$y_i = W_{yh}\, h_i$$

$$h_i = \sigma(W_{hx}x_i + W_{hh}h_{i-1})$$

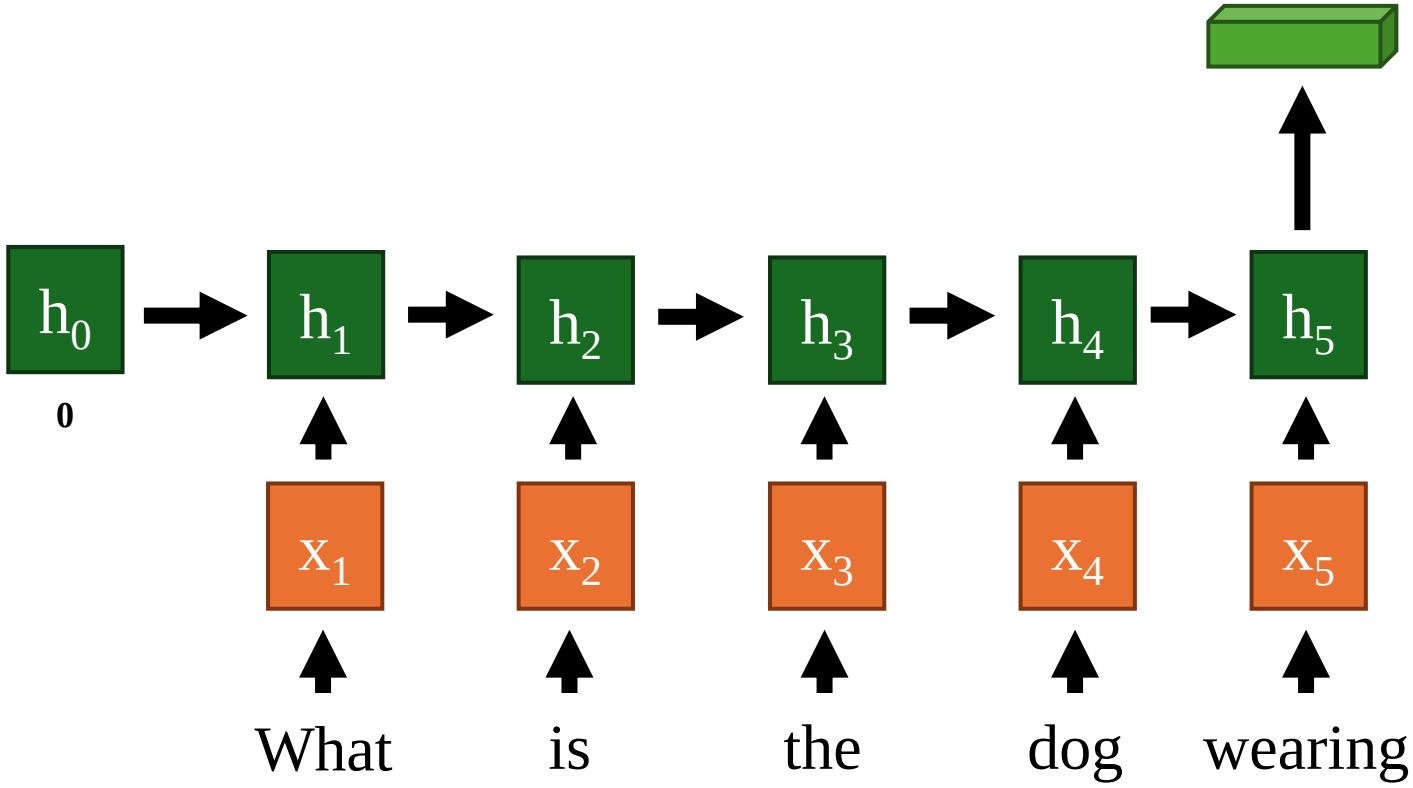
序列模型

多个输入，单个输出　　评价分数

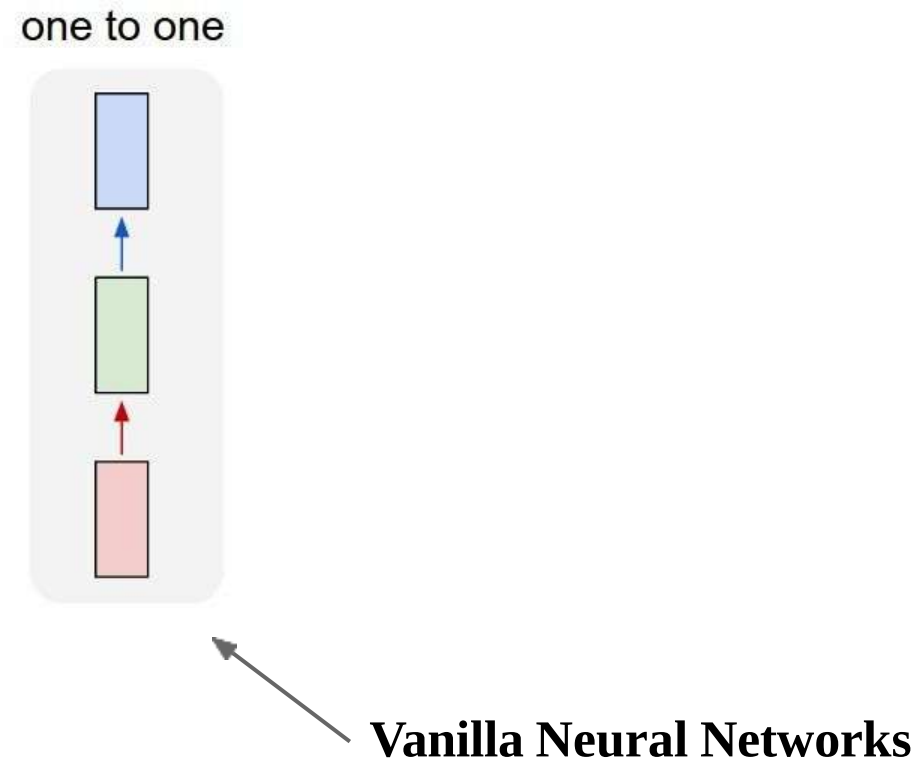


序列模型

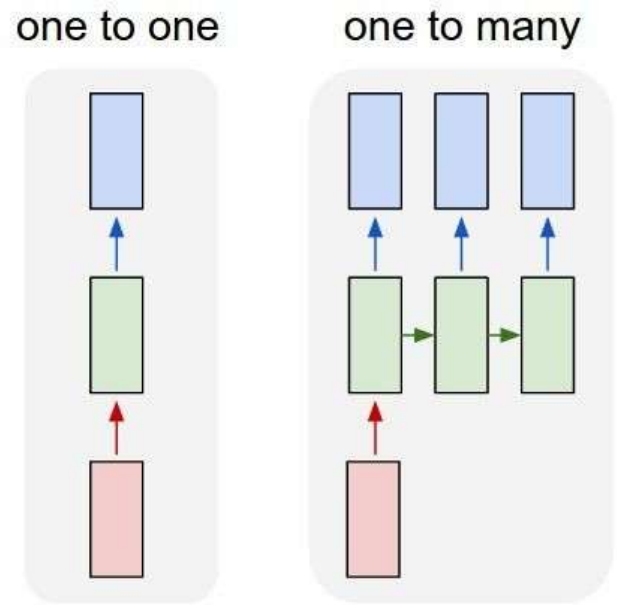
可以是一个特征



处理序列数据——不同的任务

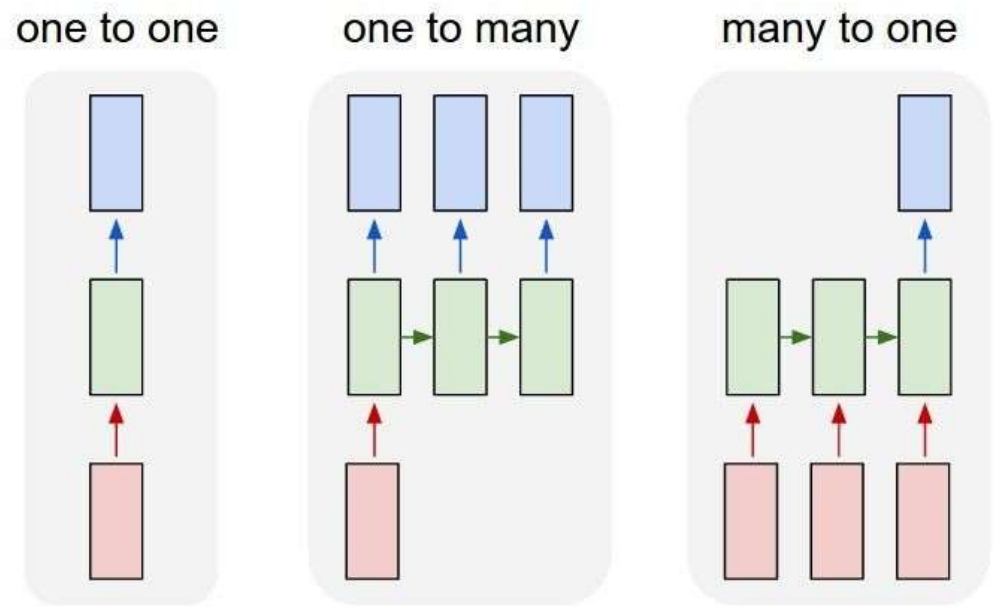


处理序列数据——不同的任务



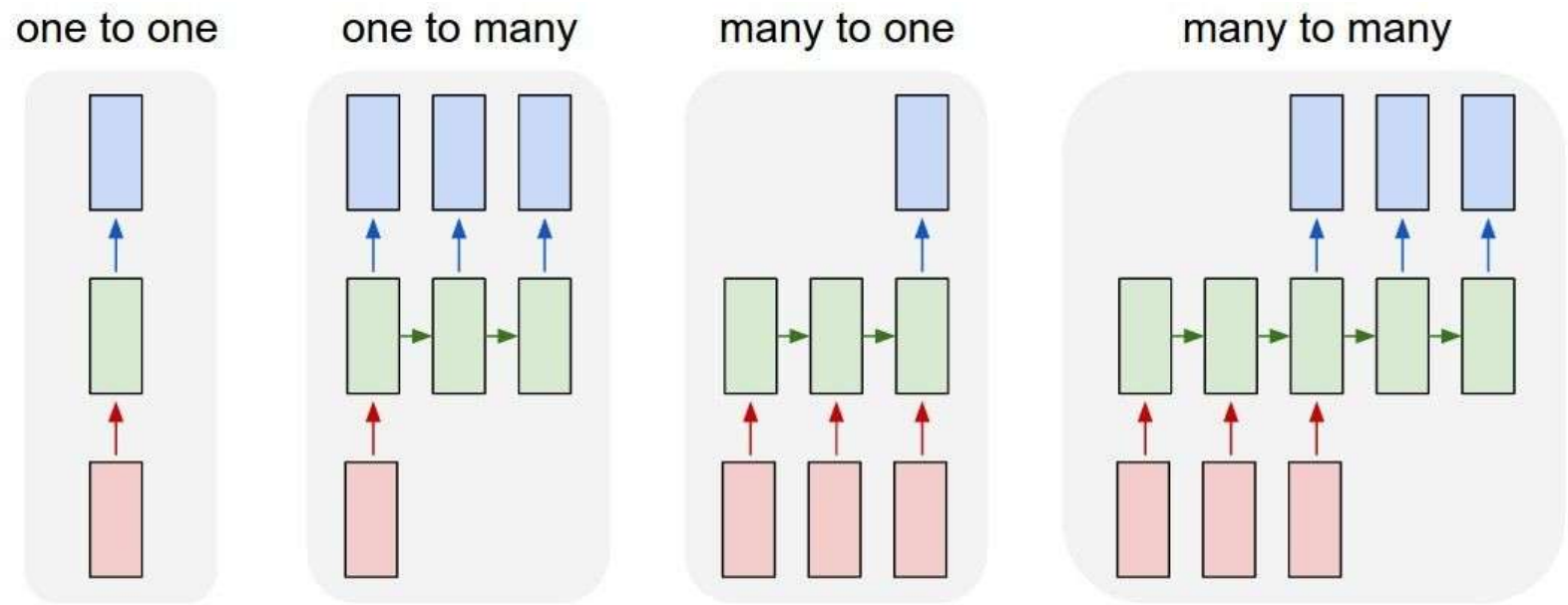
e.g. 图文描述
image -> sequence of words

处理序列数据——不同的任务



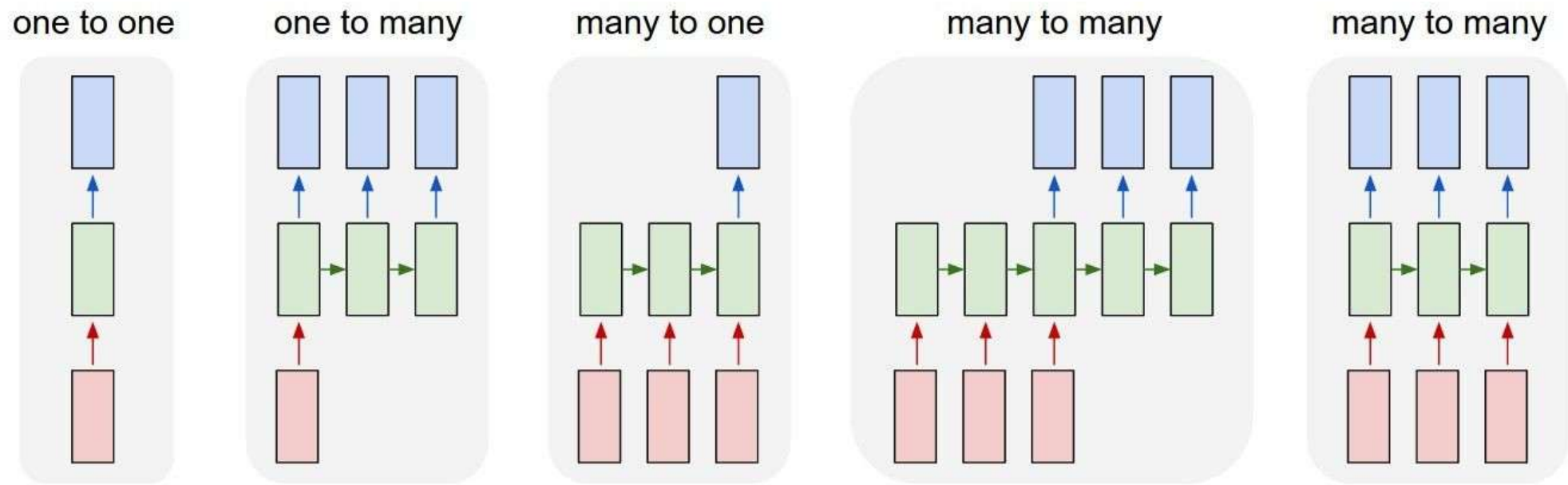
e.g. 动作/事件预测
sequence of video frames -> action class
用户评价打分
Review -> Scores

处理序列数据——不同的任务



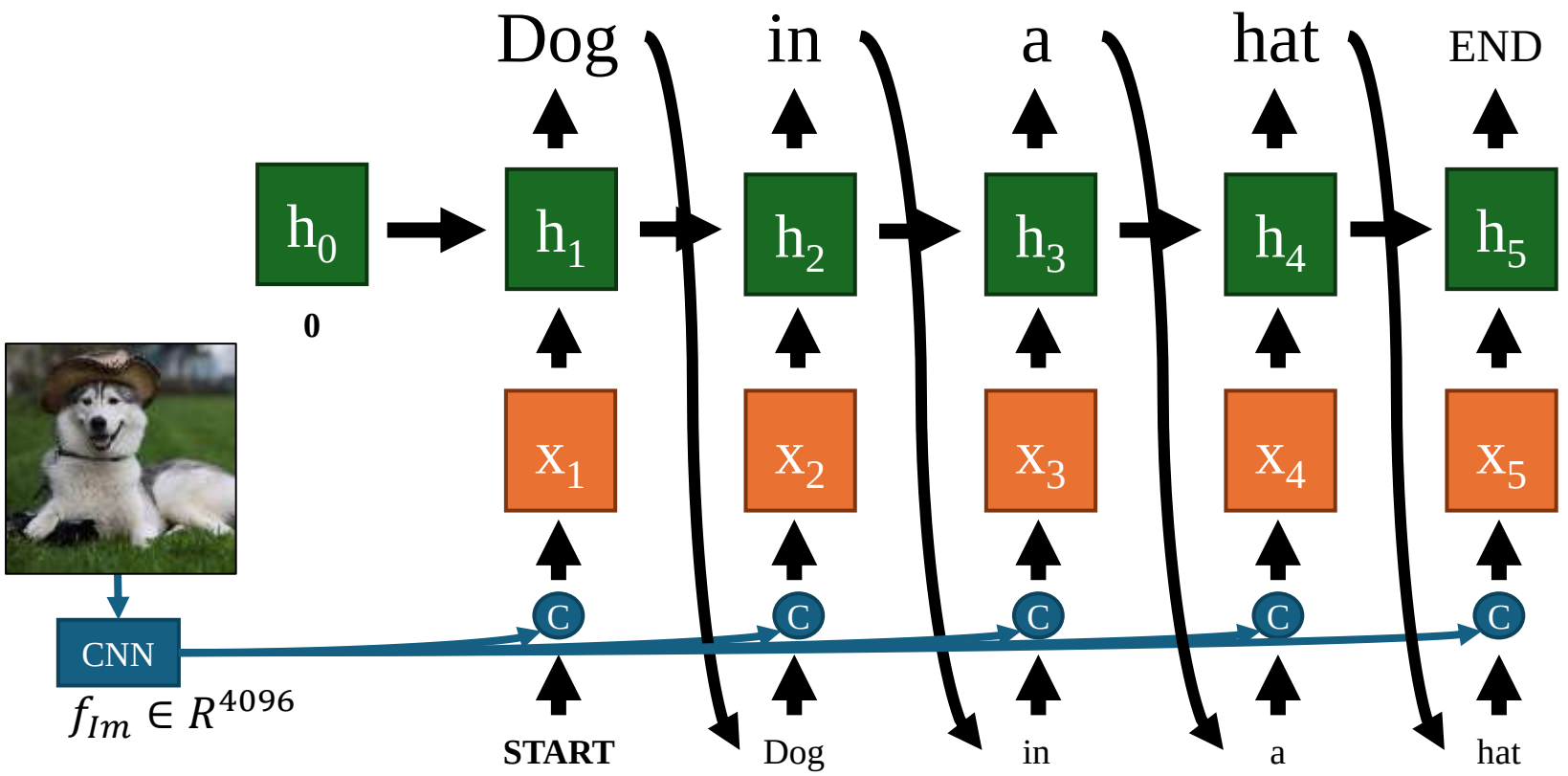
E.g. 视频配文
Sequence of video frames -> caption

处理序列数据——不同的任务



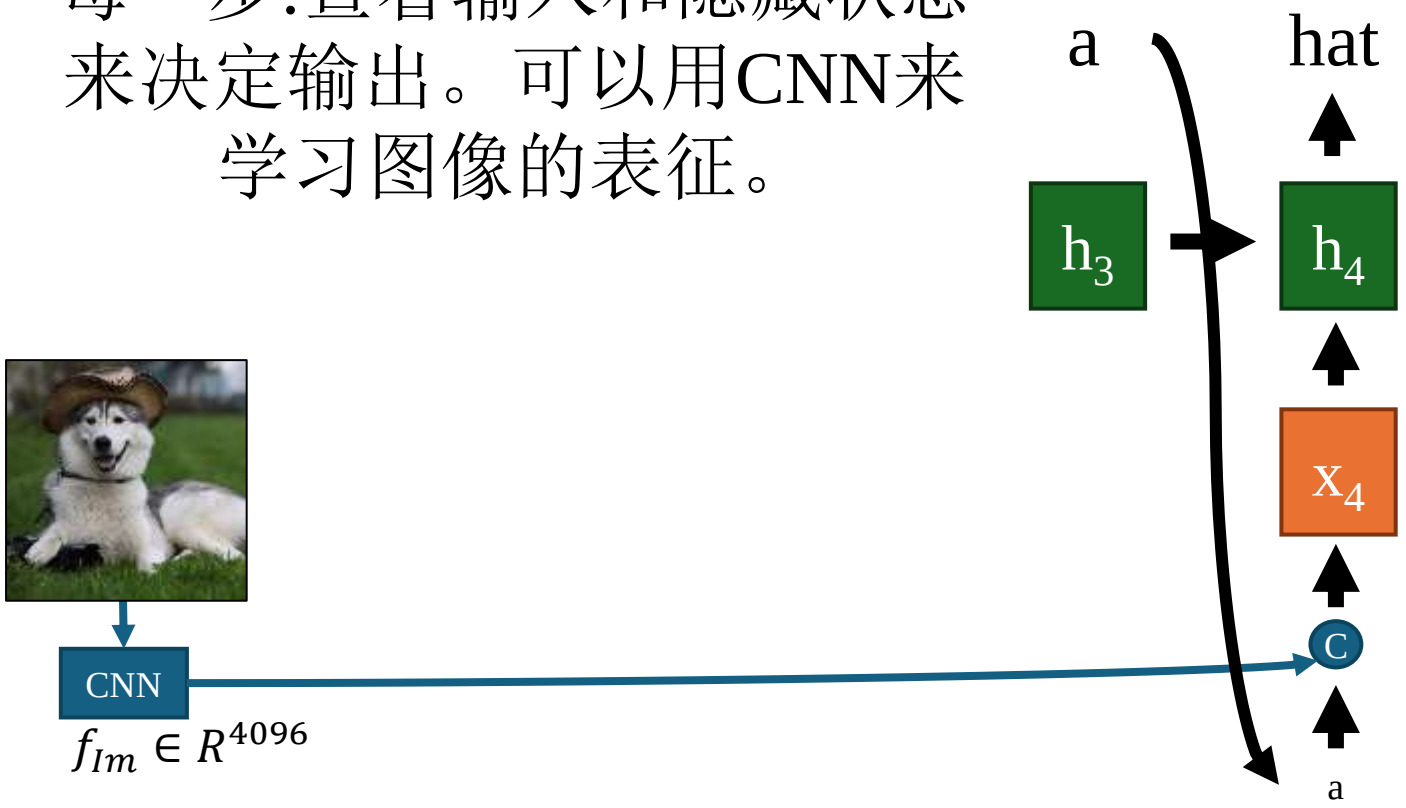
e.g. 帧级别的视频分类

Captioning：图像描述



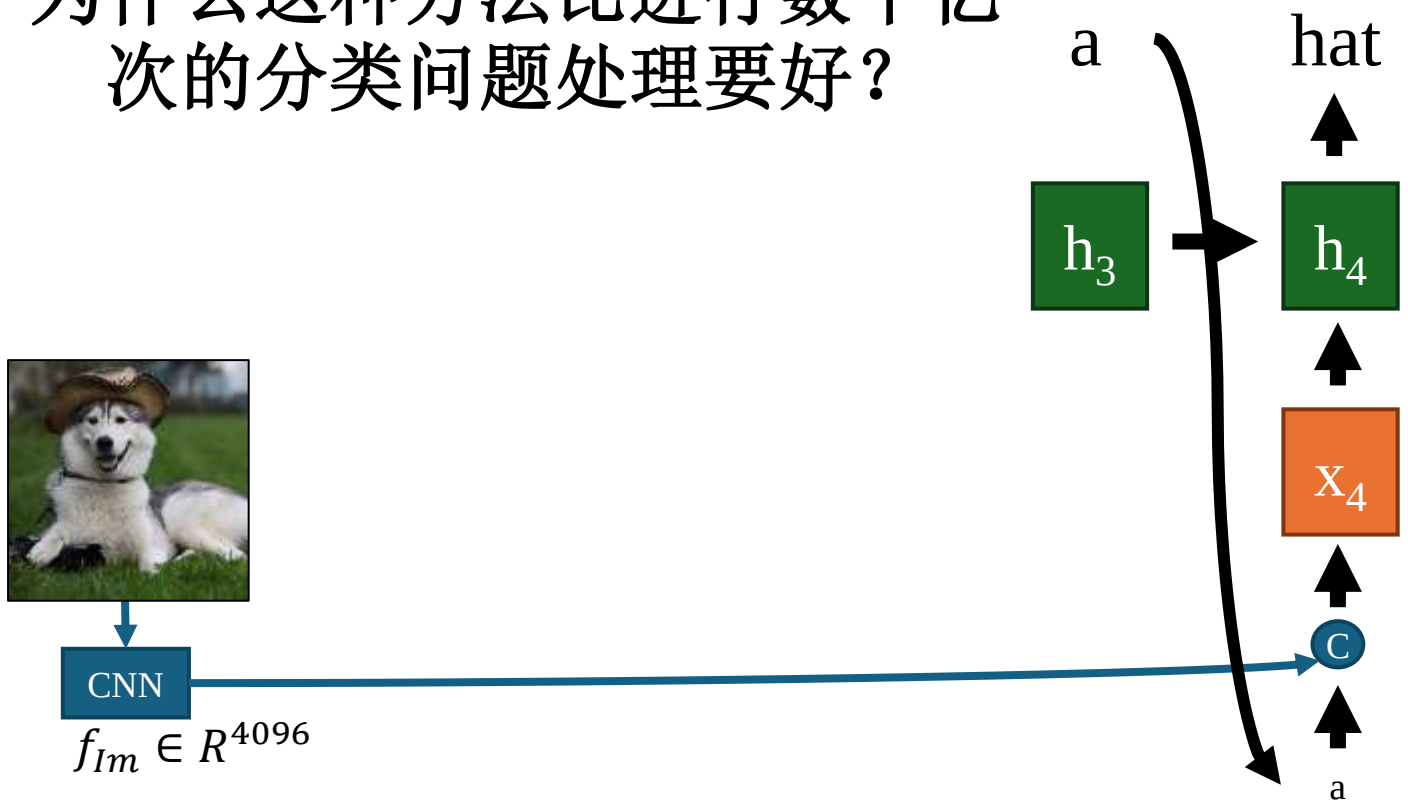
Captioning：图像描述

每一步:查看输入和隐藏状态来决定输出。可以用CNN来学习图像的表征。



图文描述

为什么这种方法比进行数十亿次的分类问题处理要好？



图像描述：效果

A female tennis player in action on the court.

A group of young men playing a game of soccer

A man riding a wave on top of a surfboard.

A baseball game in progress with the batter up to plate.

A brown bear standing on top of a lush green field.

A person holding a cell phone in their hand.

Long-term Recurrent Convolutional Networks for Visual Recognition and Description. Donahue et al. TPAMI, CVPR 2015.

图像描述：失败案例？

A close up of a person brushing his teeth.

A woman laying on a bed in a bedroom.

A black and white cat is sitting on a chair.

A large clock mounted to the side of a building.

A bunch of fruit that are sitting on a table.

A toothbrush holder sitting on top of a white sink.

Long-term Recurrent Convolutional Networks for Visual Recognition and Description. Donahue et al. TPAMI, CVPR 2015.

图像描述: 失败案例?

A woman is holding a cat in her hand

A person holding a computer mouse on a desk

A woman standing on a beach holding a surfboard

A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

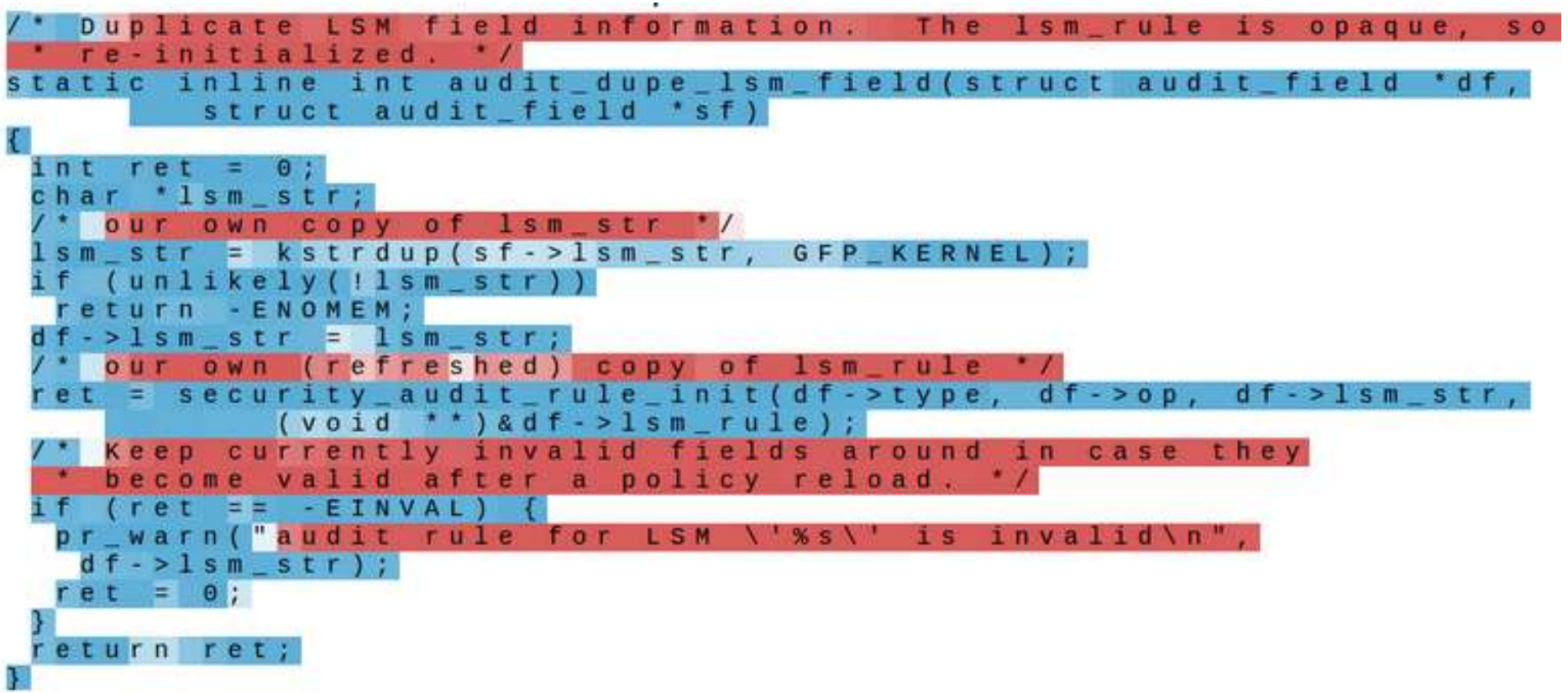
训练RNN写Linux Code

```
/*
 * If this error is set, we will need anything right after that BSD.
 */
static void action_new_function(struct s_stat_info *wb)
{
  unsigned long flags;
  int lel_idx_bit = e->edd, *sys & ~((unsigned long) *FIRST_COMPAT);
  buf[0] = 0xFFFFFFFF & (bit << 4);
  min(inc, slist->bytes);
  printk(KERN_WARNING "Memory allocated %02x/%02x, "
    "original MLL instead\n"),
    min(min(multi_run - s->len, max) * num_data_in),
    frame_pos, sz + first_seg);
  div_u64_w(val, inb_p);
  spin_unlock(&disk->queue_lock);
  mutex_unlock(&s->sock->mutex);
  mutex_unlock(&func->mutex);
  return disassemble(info->pending_bh);
}
```

Result credit: A. Karpathy

RNN

RNN的输出. 蓝色 到 红色 显示了每个代码按照时间步数生成的顺序。 这能说明什么？



```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
                                       struct audit_field *sf)
{
  int ret = 0;
  char *lsm_str;
  /* our own copy of lsm_str */
  lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
  if (unlikely(!lsm_str))
    return -ENOMEM;
  df->lsm_str = lsm_str;
  /* our own (refreshed) copy of lsm_rule */
  ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
                                 (void **)&df->lsm_rule);
  /* Keep currently invalid fields around in case they
   * become valid after a policy reload. */
  if (ret == -EINVAL) {
    pr_warn("audit rule for LSM \'%s\' is invalid\n",
            df->lsm_str);
    ret = 0;
  }
  return ret;
}
```

Result credit: A. Karpathy

RNN

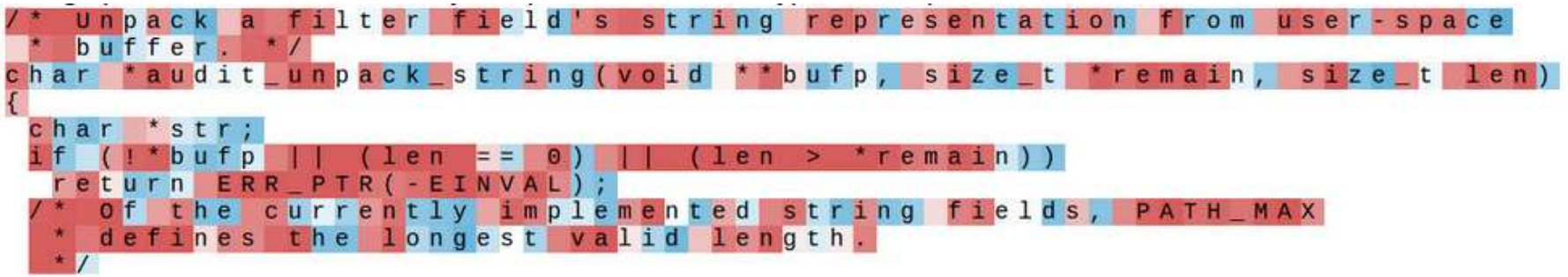
RNN的输出. 蓝色 到 红色 显示了每个代码按照时间步数生成的顺序。 这能说明什么？

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```

Result credit: A. Karpathy

RNN

RNN的输出. 蓝色 到 红色 显示了每个代码按照时间步数生成的顺序。 这能说明什么？



```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
```

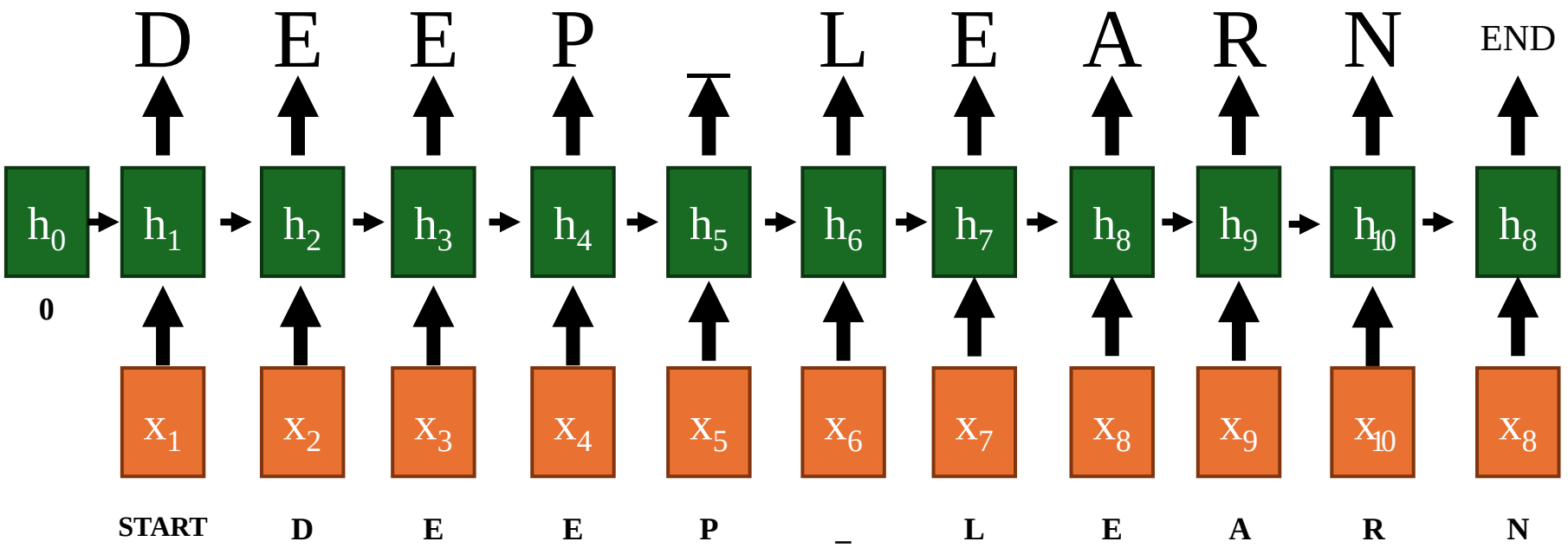
Result credit: A. Karpathy

可能出现的问题：迭代深度

如果迭代的步数过多会怎样?

考虑 $g^n \quad g \neq 1$

梯度爆炸或消失



可能出现的问题：迭代深度

- 有一些更复杂的方法 (LSTM, GRU)来避免这些问题

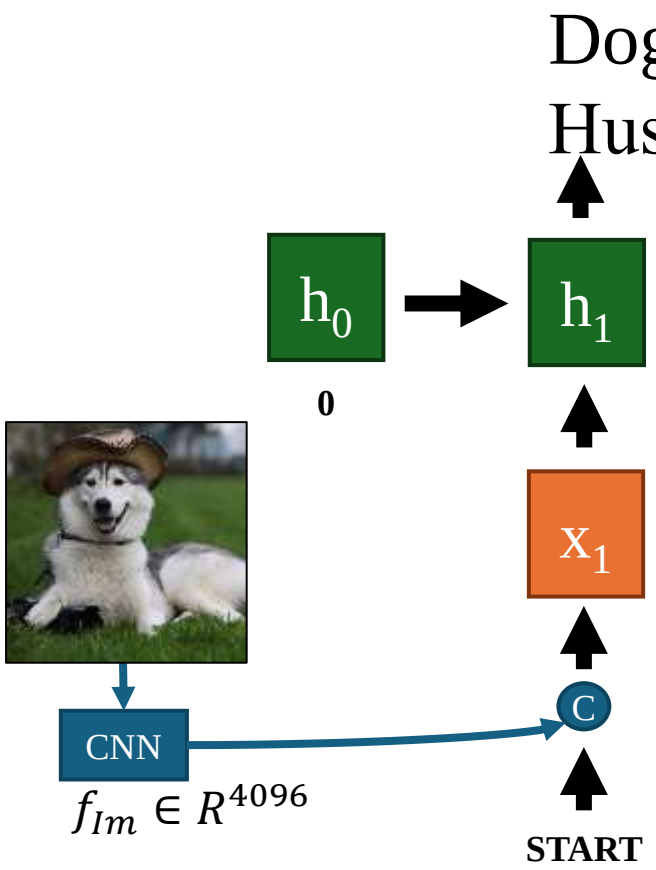
- 总体策略（残差学习）：尽量保证隐藏状态的改变比较小，可以每步只加上一小点状态更新

可能出现的问题：多个可能的结果

可能有很多描述都是对的!

- A dog in a hat
- A dog wearing a hat
- Husky wearing a hat
- Husky holding a camera, sitting in grass
- A dog that’s in a hat, sitting on a lawn with a camera

可能出现的问题：多个可能的结果

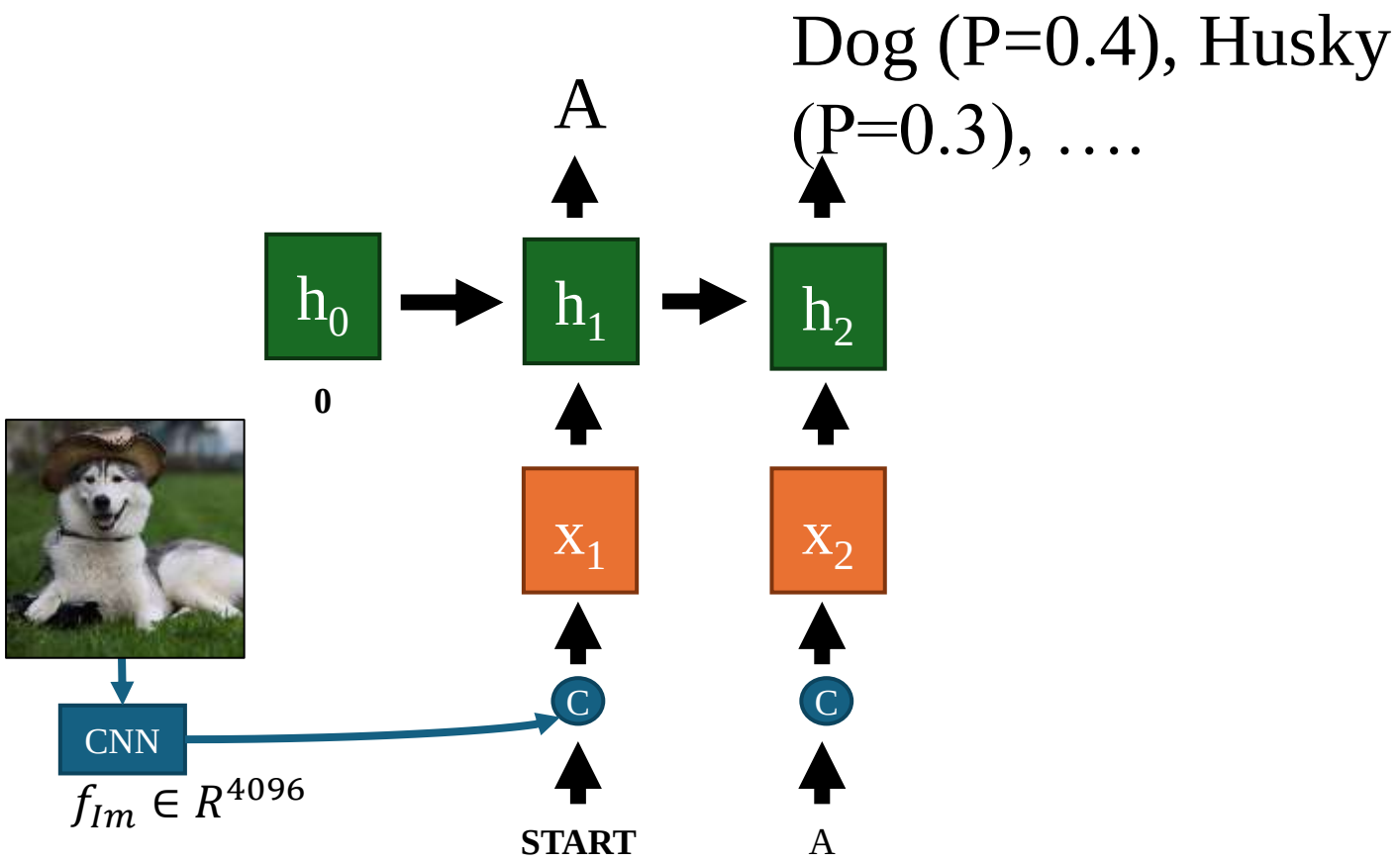


- 根据每个单词的概率来去采样
- 可以调整 “temperature”温度参数 exp(score/t) 来调节概率分布
- exp(5) / exp(1) → 54.6
- exp(5/5) / exp(1/5) → 2.2

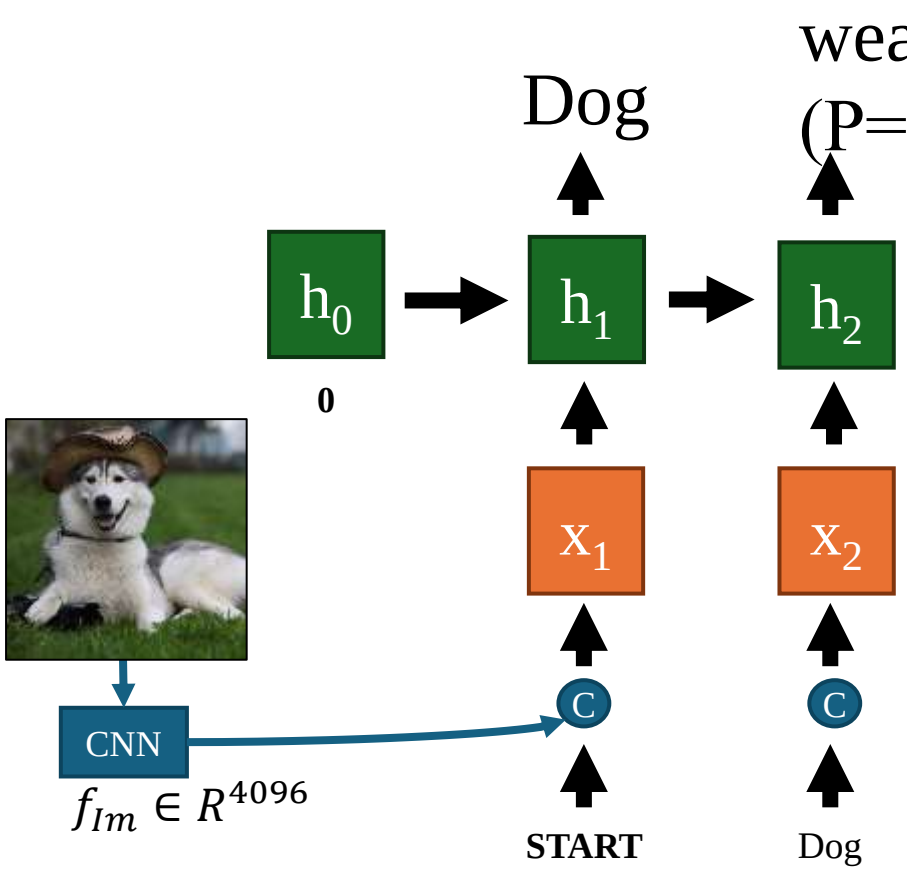
温度的效果

- 关于创业投资的段落
- **中等温度（平衡点，既有一定创造性也有可预测性）：**"创业公司正处在一个转折点，人工智能的崛起使得新的数据分析工具变得至关重要。投资者在寻找那些能够利用机器学习优化其业务流程的初创企业。在这个环境下，能够清晰展示其价值主张的公司最有可能获得资金。“
- **高温度（创造性高，随机性高）：**“在光滑的滨江道路上，有一个会说诗的机器人引发了投资者的狂热，他们纷纷拿出奇怪的艺术品作为资本。一个穿着西装的猫成为了下一个大亨，而闪电辩论赛决定了下一轮融资的赢家。“
- **低温度（保守，重复性高）：**"创业公司需要资金来发展。投资者寻找有潜力的创业公司来投资。创业公司需要资金来发展。投资者寻找有潜力的创业公司来投资。创业公司需要资金来发展，这是一个不断循环的过程。"

采样



采样



每个单词概率由条件概率 $P(w_i|w_1,\ldots,w_{i-1})$ 决定

相当于在一个字典树上选择了概率最大的分支

可能的问题：质量评估

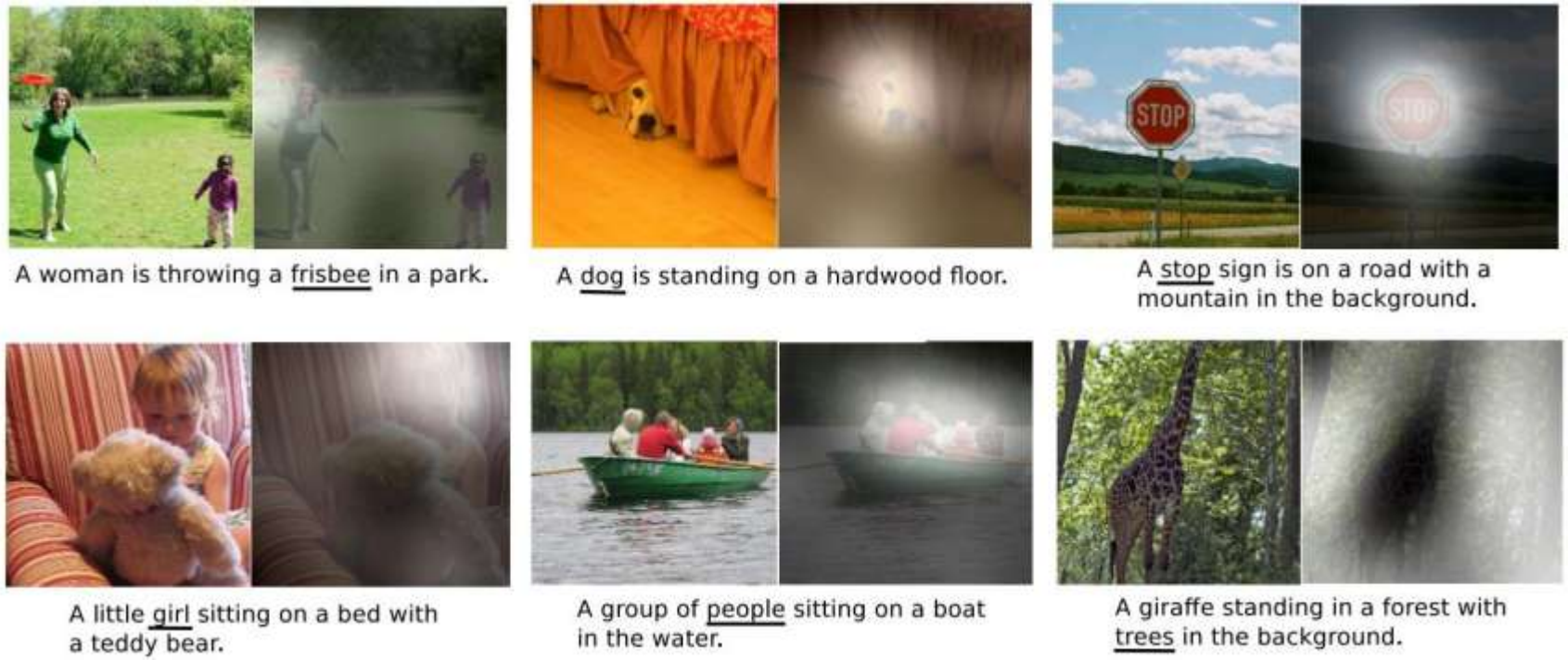
Computer: “A husky in a hat”
Human: “A dog in a hat”

哪个描述更好?

1) 人工评估
2) 避免使用简单指标导致简单输出 (e.g., “A a a a a”)
3) 实际常用：单词重叠度、词汇频率和召回率、特征匹配度。。。

图像内容： Attention

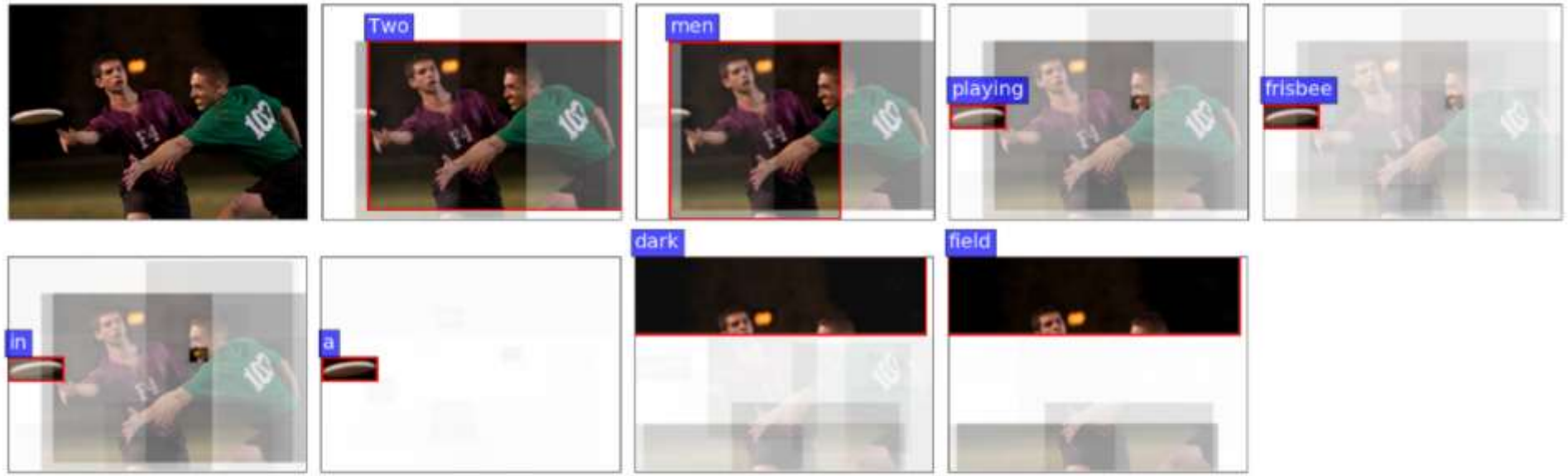
注意力机制会让我们关注不同的空间区域



Xu et al, “Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention”, ICML 2015

Attention

通过attention，关注图像当中的物体，而不是整张图，
来提升性能



Two men playing frisbee in a dark field.

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Anderson et al. 2018.

Attention

Question: What color is illuminated on the traffic light? Answer left: green. Answer right: red.

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Anderson et al. 2018.

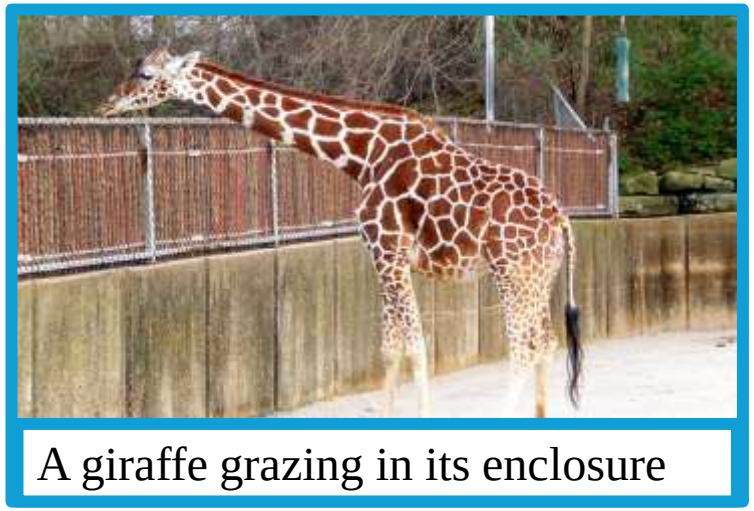
语法问题：答案空间

- 如果字典大小是10k，20个单词的句子有多少可能的答案?
- $(10k)^{20}$? **是吗?**
- 为什么不是呢？

Giraffe example credit: L. Zitnick

长颈鹿整天都在干啥

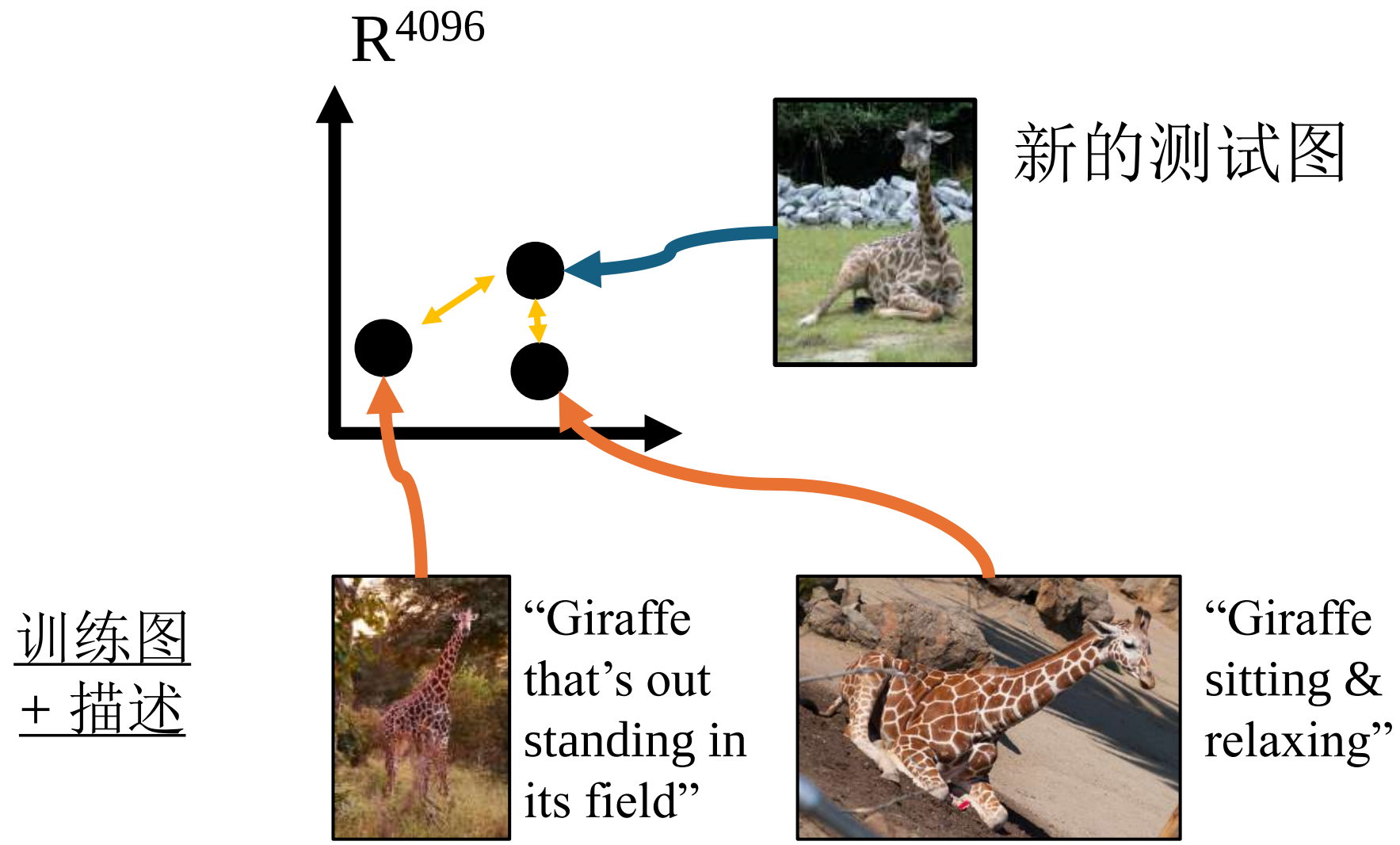
A giraffe sitting and resting



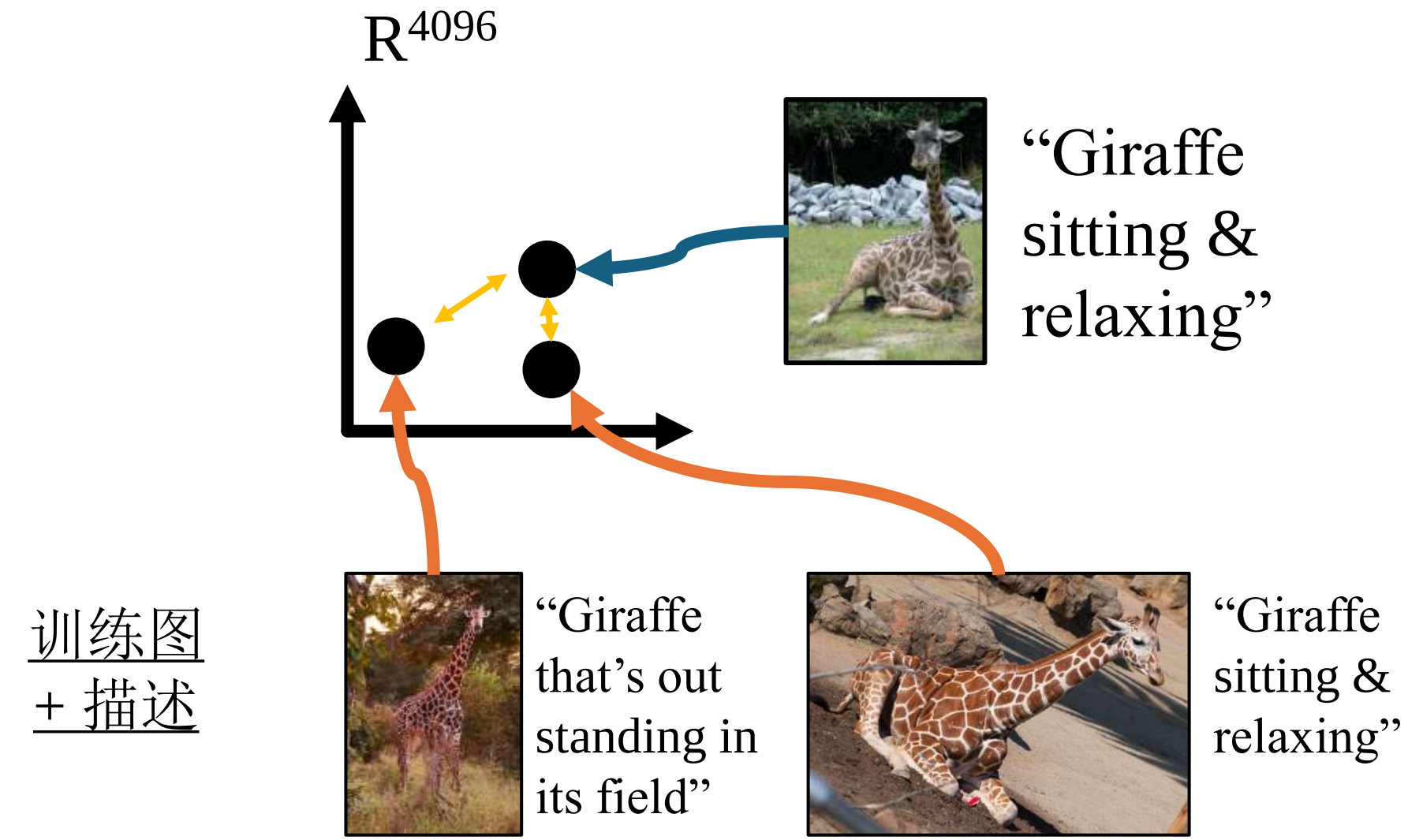
A giraffe grazing in its enclosure

A giraffe wandering around

另一种描述的方式：检索


R4096
新的测试图
训练图
+ 描述
“Giraffe that’s out standing in its field”
“Giraffe sitting & relaxing”

另一种描述的方式：检索


R4096
“Giraffe sitting & relaxing”
训练图 + 描述
“Giraffe that’s out standing in its field”
“Giraffe sitting & relaxing”

检索结果

A man riding a wave on a surfboard.

A man riding a wave on a surfboard in the ocean.

A person flying a kite in the sky.

A person flying a kite in the sky.

A cat sitting in a bathroom sink.

A black and white cat sitting in a bathroom sink.

Exploring Nearest Neighbor Approaches for Image Captioning. Devlin et al. 2015

检索结果

- 我们可能不太喜欢检索的结果
- 因为它们没有创造新的句子!

A wooden bench in front of a building.

A window display on the front of a building.

A building with a clock on the top.

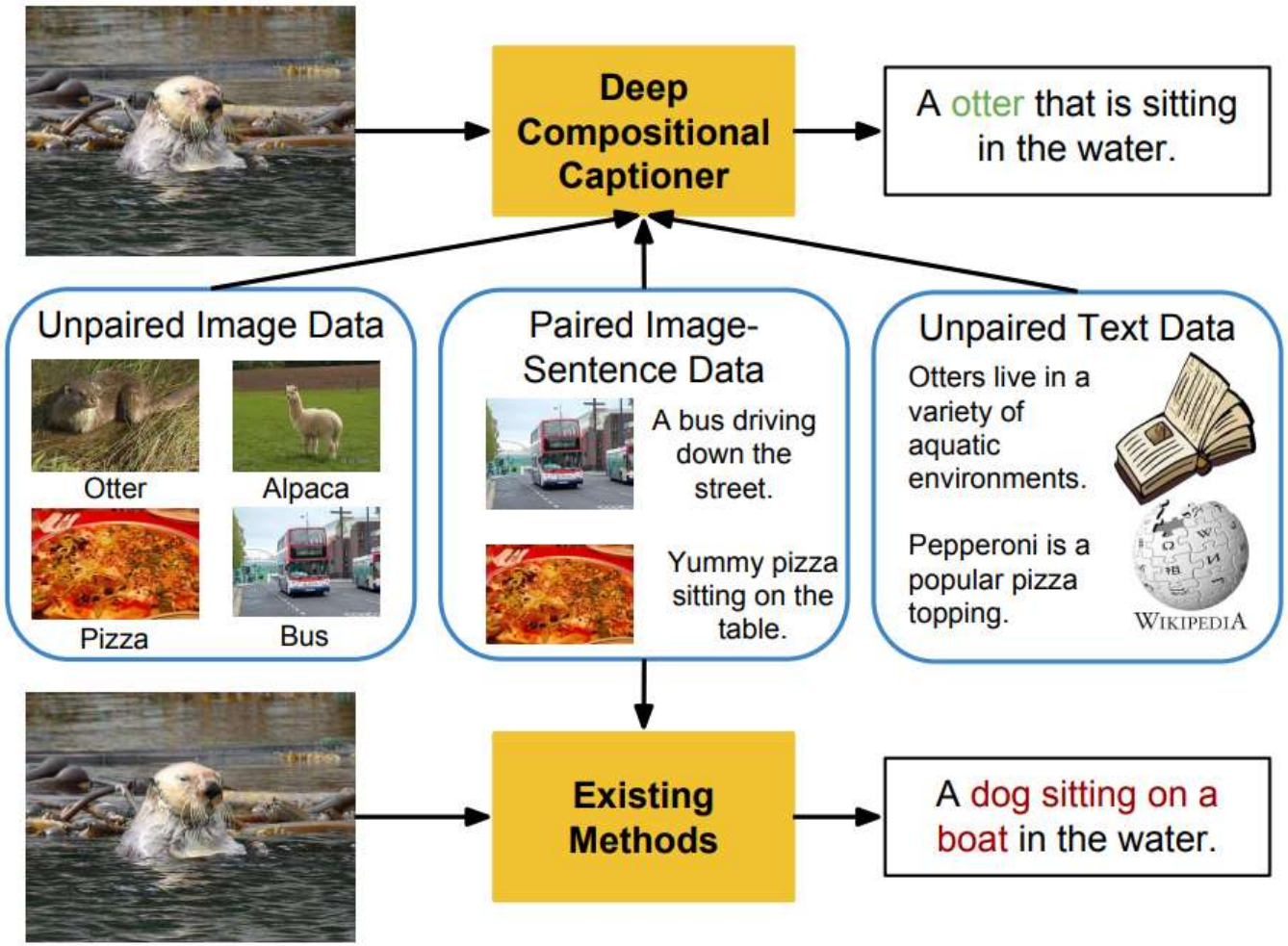
A clock tower on the top of a building.

The side of a passenger train at a train station.

A bus that is on the side of a road.

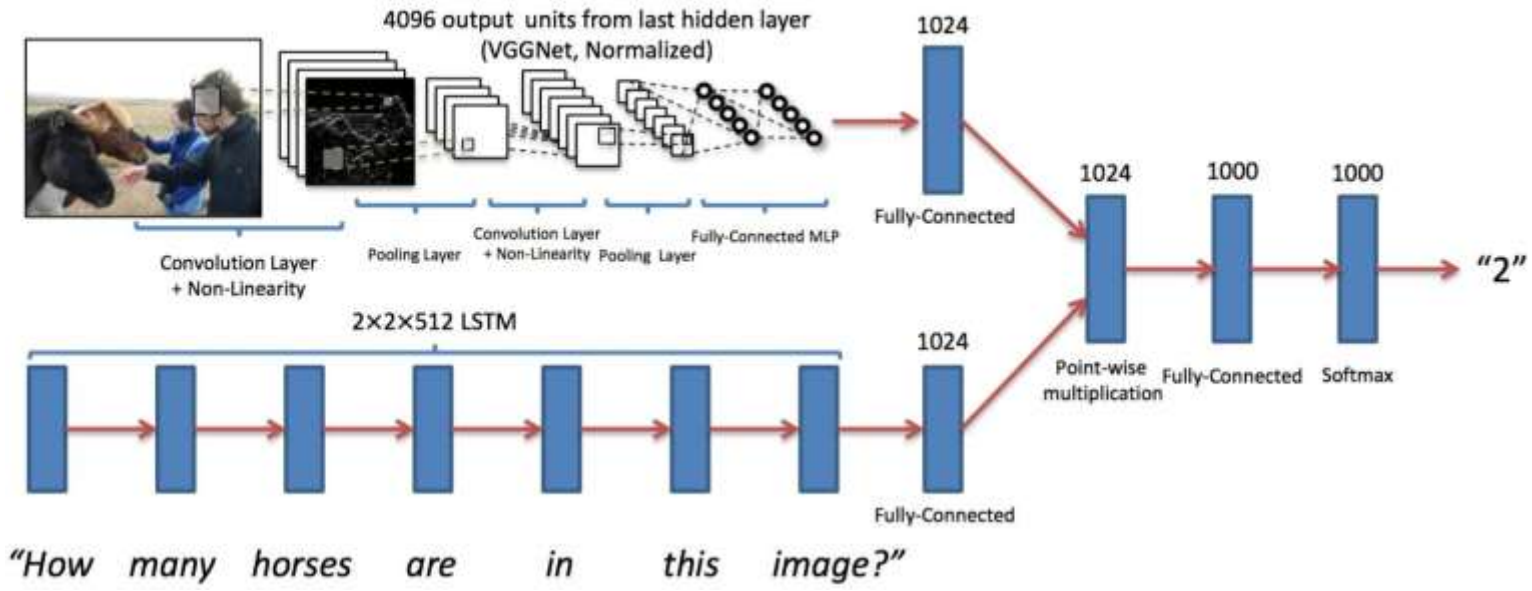
Exploring Nearest Neighbor Approaches for Image Captioning. Devlin et al. 2015

怎么获得新的描述？Unpaired Data



Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data.
L. Hendricks et al. CVPR 2016

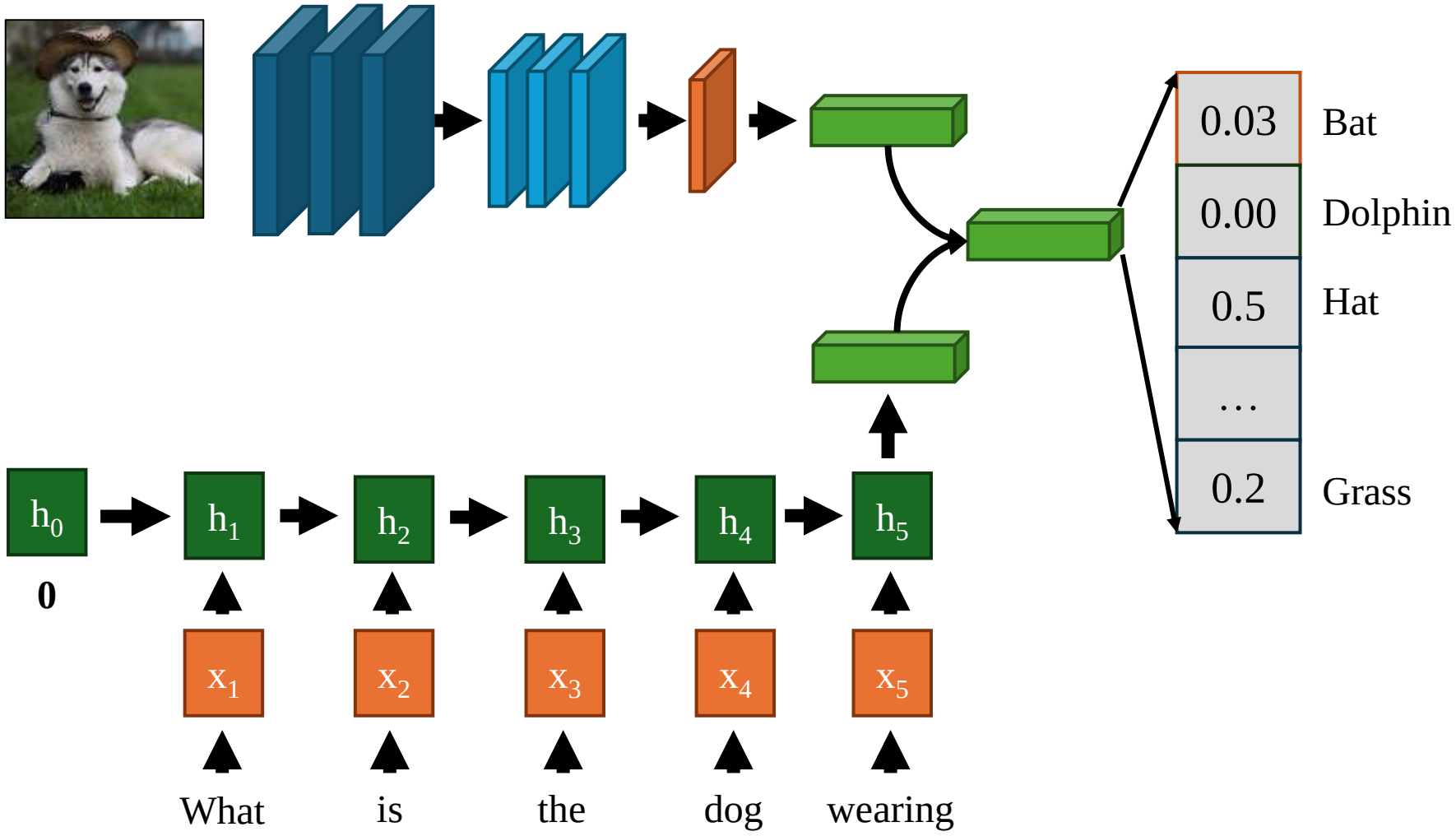
Visual Question Answering (VQA)



需要匹配文本和图像特征

Agrawal et al, "Visual 7W: Grounded Question Answering in Images", CVPR 2015

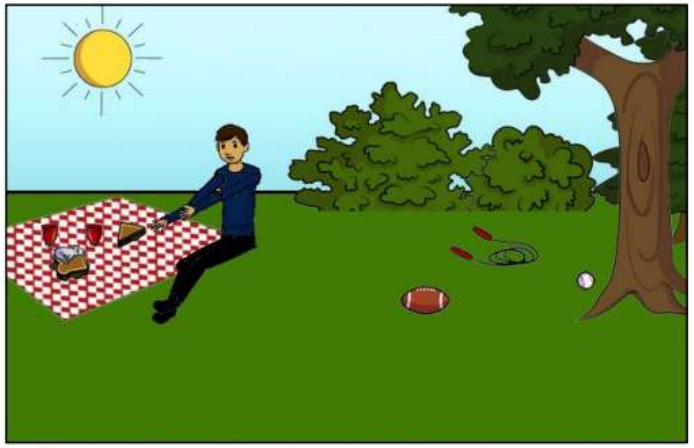
VQA模型


0.03
Bat
0.00
Dolphin
0.5
Hat
…
0.2
Grass
h0
h1
h2
h3
h4
h5
0
x1
x2
x3
x4
x5
What
is
the
dog
wearing

VQA

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?

VQA: Visual Question Answering. S. Antol, A. Agrawal et al. ICCV 2015

另一种VQA

Q: What endangered animal is featured on the truck?

A: A bald eagle.
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?

A: Onto 24 ¾ Rd.
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?

A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service

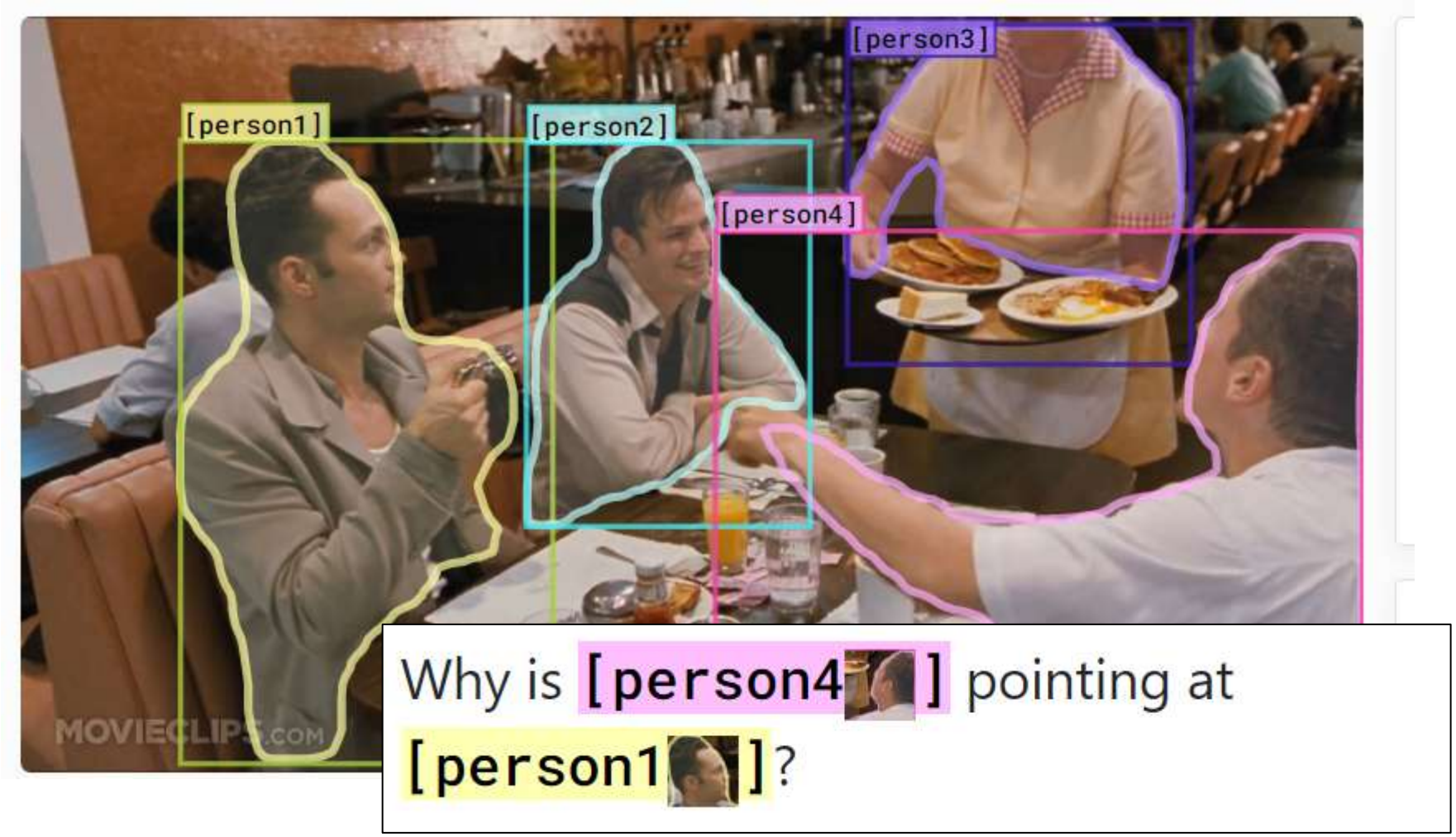
Q: Who is under the umbrella?

A: Two women.
A: A child.
A: An old man.
A: A husband and a wife.

Agrawal et al, "VQA: Visual Question Answering", ICCV 2015
Zhu et al, "Visual 7W: Grounded Question Answering in Images", CVPR 2016

根据图片和文字提问回答问题

Visual Commonsense Reasoning (VCR)

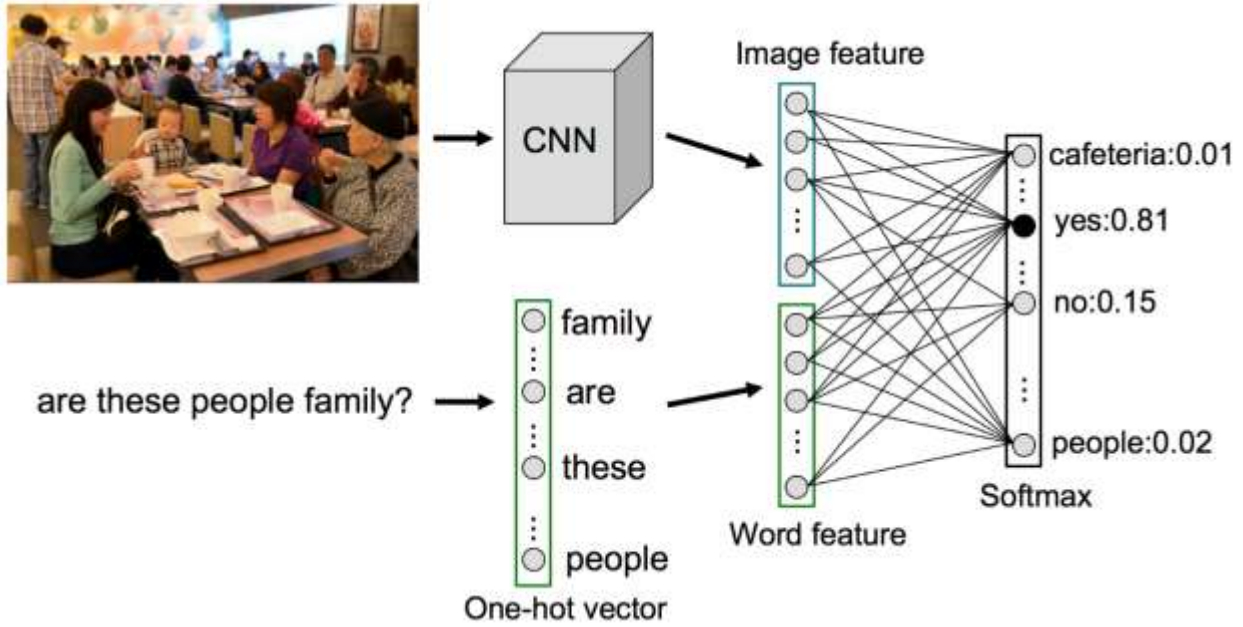


From Recognition to Cognition: Visual Commonsense Reasoning. Zellers et al. CVPR 2019

VCR

From Recognition to Cognition: Visual Commonsense Reasoning. Zellers et al. CVPR 2019

一个简单的VQA实现方式



- 建立字典，包括5000个最常用的回答所包含的词汇
- 用CNN提取图像的特征I
- 用BoW表示问题Q
- 根据概率 $P(A|Q,I)$得到回答

Zhou, Bolei, et al. "Simple baseline for visual question answering." arXiv preprint arXiv:1512.02167 (2015).

Slide credit: T. Gupta

效果

Question: what are they doing
Predictions:
playing baseball (score: 10.67 = 2.01 [image] + 8.66 [word])
baseball (score: 9.65 = 4.84 [image] + 4.82 [word])
grazing (score: 9.34 = 0.53 [image] + 8.81 [word])

Based on image only: umpire (4.85), baseball (4.84), batter (4.46)
Based on word only: playing wii (10.62), eating (9.97),
playing frisbee (9.24)

Question: how many people inside
Predictions:
3 (score: 13.39 = 2.75 [image] + 10.65 [word])
2 (score: 12.76 = 2.49 [image] + 10.27 [word])
5 (score: 12.72 = 1.83 [image] + 10.89 [word])

Based on image only: umpire (4.85), baseball (4.84), batter (4.46)
Based on word only: 8 (11.24), 7 (10.95), 5 (10.89)

Zhou, Bolei, et al. "Simple baseline for visual question answering." arXiv preprint arXiv:1512.02167 (2015).

Slide credit: T. Gupta

效果

Question: which brand is the laptop
Predictions:
apple (score: 10.87 = 1.10 [image] + 9.77 [word])
dell (score: 9.83 = 0.71 [image] + 9.12 [word])
toshiba (score: 9.76 = 1.18 [image] + 8.58 [word])

Based on image only: books (3.15), yes (3.14), no (2.95)
Based on word only: apple (9.77), hp (9.18), dell (9.12)

- 语言先验可以极大程度减少回答的空间

Zhou, Bolei, et al. "Simple baseline for visual question answering." arXiv preprint arXiv:1512.02167 (2015).

Slide credit: T. Gupta

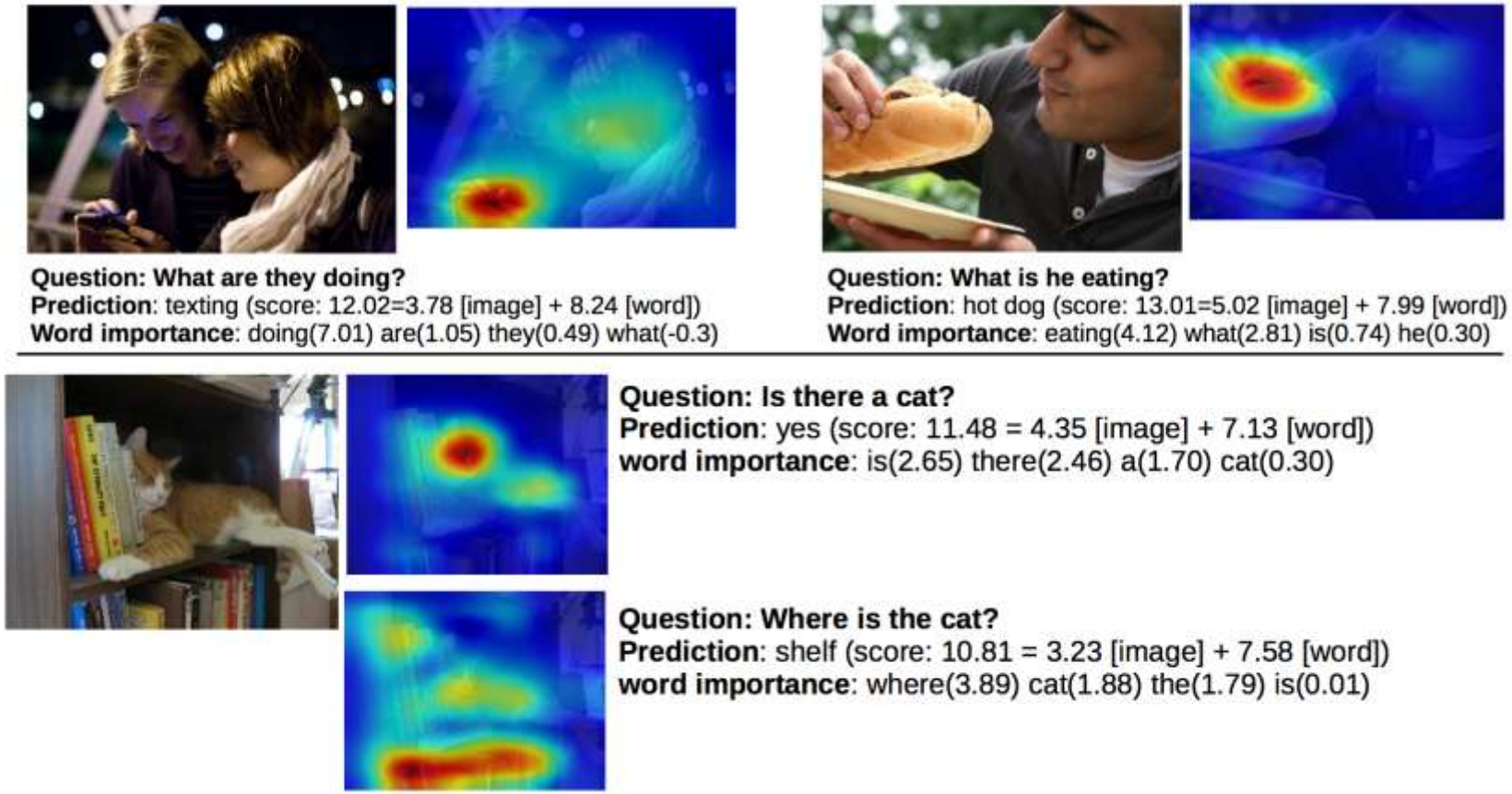
定量评价

	Open-Ended				Multiple-Choice			
	Overall	yes/no	number	others	**Overall**	yes/no	number	others
IMG [2]	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
BOW [2]	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
BOWIMG [2]	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTMIMG [2]	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
CompMem [6]	52.62	78.33	35.93	34.46	-	-	-	-
NMN+LSTM [1]	54.80	77.70	37.20	39.30	-	-	-	-
WR Sel. [13]	-	-	-	-	60.96	-	-	-
ACK [16]	55.72	79.23	36.13	40.08	-	-	-	-
DPPnet [11]	**57.22**	80.71	37.24	41.69	**62.48**	80.79	38.94	52.16
iBOWIMG	55.72	76.55	35.03	42.62	61.68	76.68	37.05	54.44

Slide credit: T. Gupta

模型可以良好地定位吗?

Class Activation Mapping （CAM）



Slide credit: T. Gupta

增强视觉分析

能更好地分析复杂的视觉场景，但是数据更难收集

Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. Goyal et al. 2017

RNN的问题

任务：机器翻译

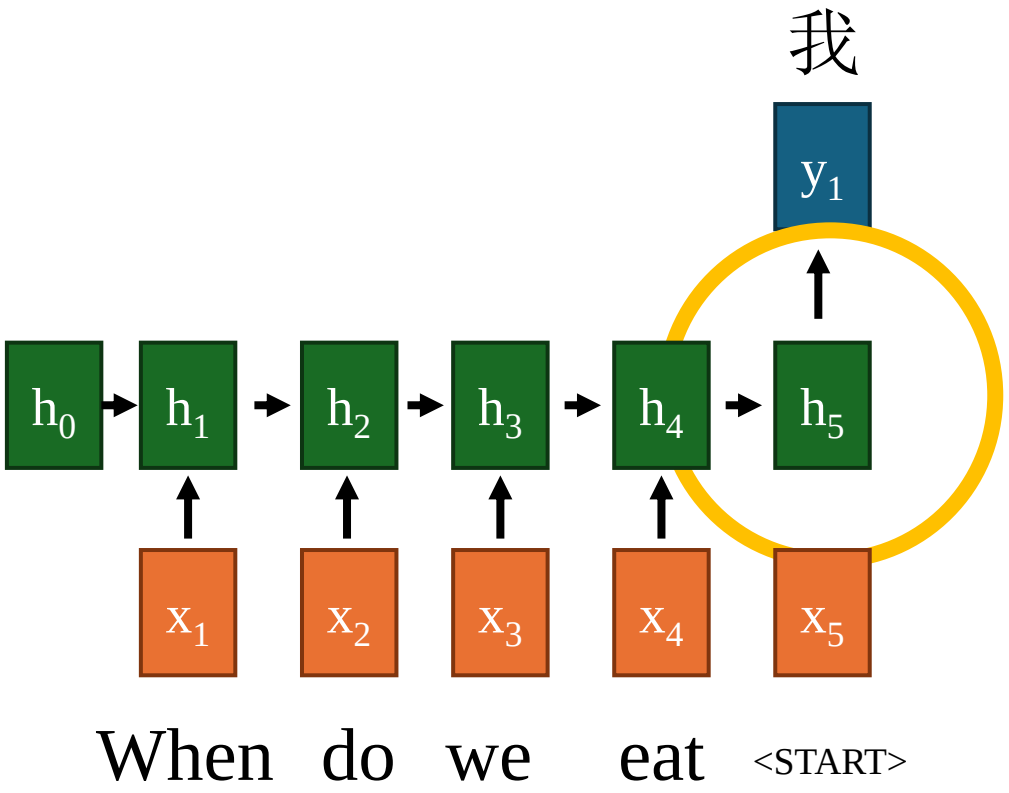


Diagram is remake of one from J. Johnson

我们需要记住每一个英文单词。

随着序列长度的增加，模型需要记住更多的信息，这使得保持长期依赖变得非常困难，这种现象通常称为**“长期依赖问题”**。