图像Caption-看图说话

指导教师：张华平　　汇报人：刘云皓，李桐，邱小尧，刘天锐，王中琦

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

德以明理　学以精工

Contents

结构大纲

目录

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

1

Image Caption 简介

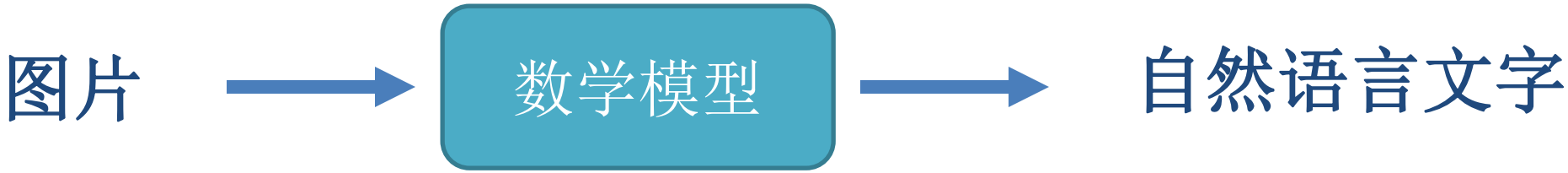
汇报人：刘云皓

什么是Image caption?

- 篮球场上有三个小人，一个人在灌篮，两个人在对抗

看图说话！！

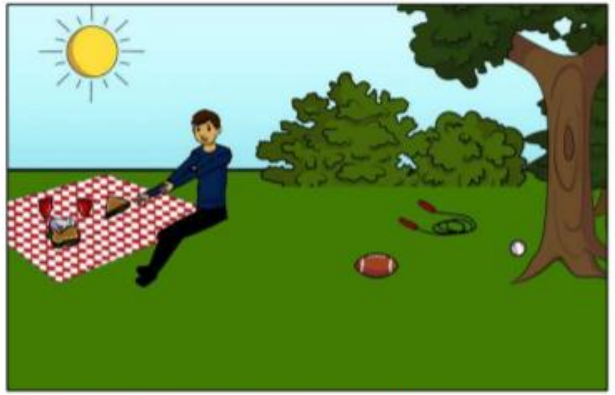
Image caption形式化定义



CV+NLP

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

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

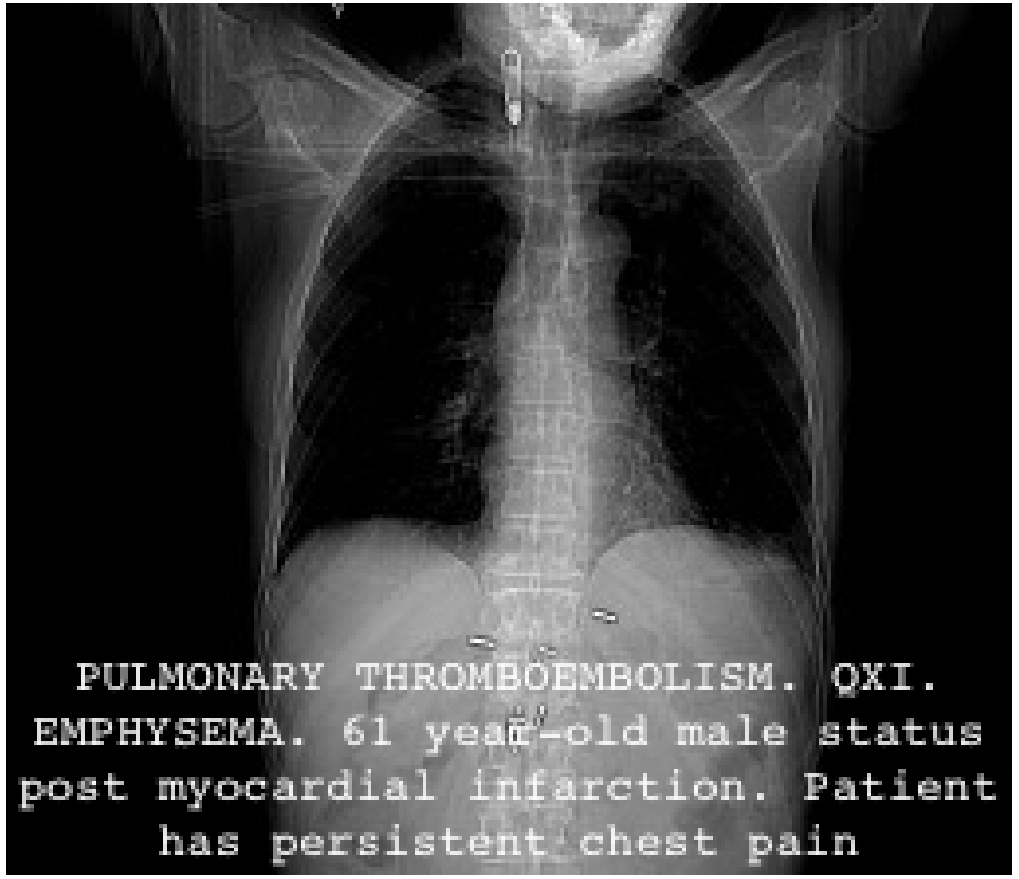


Is this person expecting company?
What is just under the tree?

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

视觉问答（VQA）

图片源自VQA: Visual Question Answering



医疗图像描述

图片源自A survey on biomedical image captioning

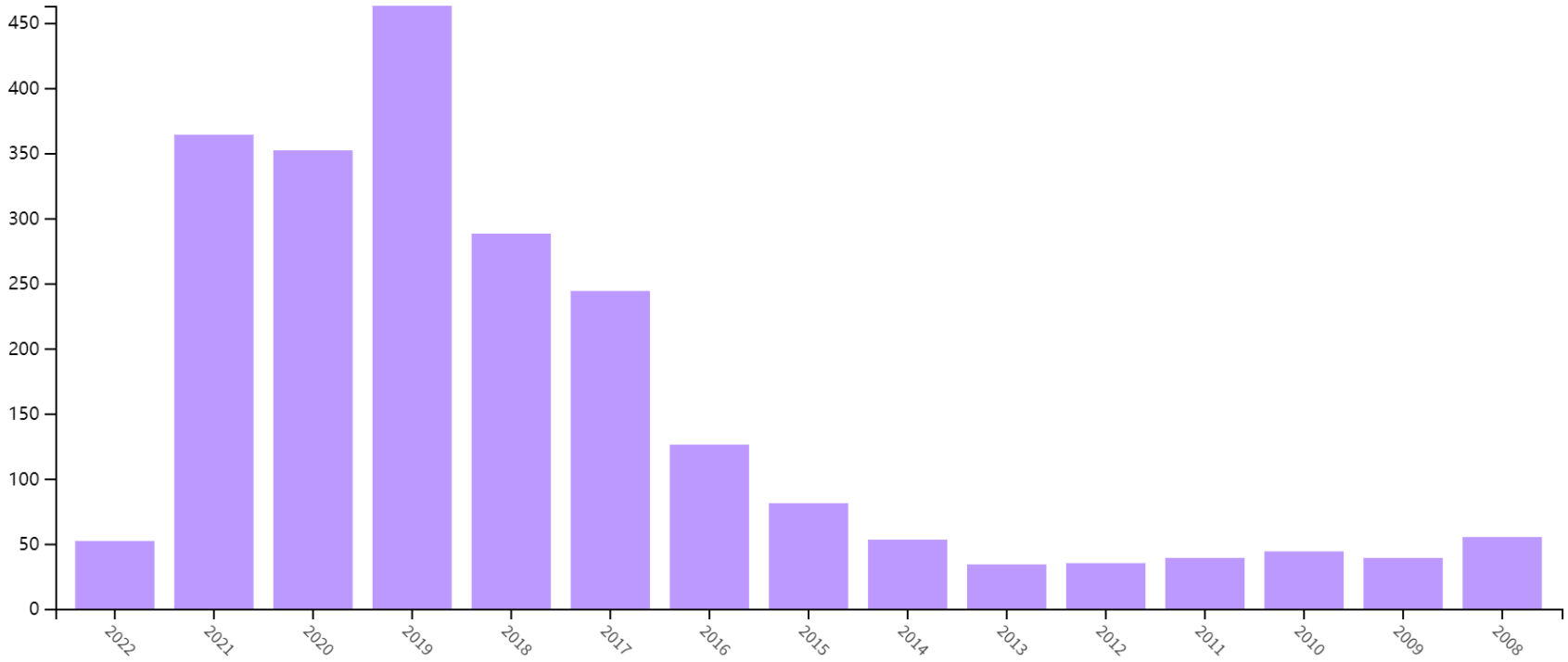
A digital thermometer resting on a wooden table, showing 38.5 degrees Celsius.

图片源自数据集VizWiz Captions

针对视障人群的辅助技术

面对视障人群的应用：SeeingAI

论文发表趋势



数据来源：web of science

面临的问题和挑战

Viewpoint variation

Intra-class variation

图片源自付莹的“计算机视觉”课程PPT

Background clutter

Occlusion

图片源自付莹的“计算机视觉”课程PPT

早期：　　不符合语法、语义的问题

深度学习时代：

1. 曝光偏差
2. 损失评估不匹配
3. 梯度消失
4. 梯度爆炸
5. 幻觉

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

2

数据集和评价指标

汇报人：刘云皓

- Standard captioning datasets：标准数据集
- e.g. MS COCO, Flickr8k, Flickr30k

- Pre-training datasets：大型数据集。用做预训练
- e.g. SBU-Captions, CC3M, CC12M, 悟空

- Domain-specific datasets：专用数据集。用于某一特定任务
- e.g. VizWiz Captions, CUB200, Oxford-102, Fashion Captioning, BreakNews, GoodNews, Loc. Narratives

MS COCO Captions

- 328000 images
- 91 objects
- 5 captions/image

应用最广泛！！

https://cocodataset.org/#home

Table 1: An overview of datasets for VLP model pre-training.

Dataset	Language	Availability	Image-text pairs
Flickr30k (Young et al., 2014)	English	✓	31,783
CxC (Parekh et al., 2020)	English	✓	247,315
SBU Captions (Ordonez et al., 2011b)	English	✓	1,000,000
Product1M (Zhan et al., 2021)	Chinese	✓	1,000,000
CC12M (Changpinyo et al., 2021)	English	✓	12,000,000
YFCC100M (Thomee et al., 2016)	English	✓	99,200,000
WIT (Srinivasan et al., 2021)	multilingual	✓	11,500,000
LAION-400M (Schuhmann et al., 2021)	English	✓	400,000,000
JFT-300M (Sun et al., 2017)	English	✗	300,000,000
JFT-3B (Zhai et al., 2021a)	English	✗	3,000,000,000
IG-3.5B-17k (Mahajan et al., 2018)	English	✗	3,500,000,000
M6-Corpus (Lin et al., 2021)	Chinese	✗	60,500,000
Wukong (Ours)	Chinese	✓	101,483,885

开源、亿级、中文

VizWiz Captions

A computer screen with a Windows message about Microsoft license terms.

A can of green beans is sitting on a counter in a kitchen.

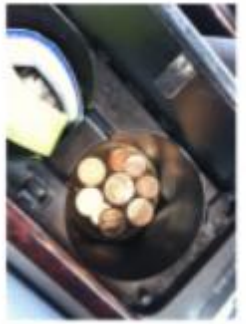
A photo taken from a residential street in front of some homes with a stormy sky above.

A blue sky with fluffy clouds, taken from a car while driving on the highway.

A hand holds up a can of Coors Light in front of an outdoor scene with a dog on a porch.

A digital thermometer resting on a wooden table, showing 38.5 degrees Celsius.

A Winnie The Pooh character high chair with a can of Yoohoo sitting on it in front of a white wall.



A cup holder in a car holding loose change from Canada.

2 数据集下载

Flickr8k: https://academictorrents.com/details/9dea07ba660a722ae1008c4c8afdd303b6f6e53b

Flickr30k: https://pan.baidu.com/s/1rORVUwctJsIOiNuVXHQ6kA code: hrf3

COCO: https://pan.baidu.com/s/1QT-sOiwVYlClMThVySzuTQ code: 2pyr

SBU Captions: http://www.cs.virginia.edu/~vicente/sbucaption/

CC3M: https://ai.google.com/research/ConceptualCaptions/download

CC12M: https://github.com/google-research-datasets/conceptual-12m

VizWiz: https://vizwiz.org/tasks-and-datasets/image-captioning/

CUB 200: https://pan.baidu.com/s/10L3s7XmzoaYmbBocYJjYyQ code: 7jeb

Oxford-102: https://s3.amazonaws.com/fast-ai-imageclas/oxford-102-flowers.tgz

Fashion Cap.: https://github.com/xuewyang/Fashion_Captioning

GoodNews: https://github.com/furkanbiten/GoodNews

Text Caps: https://textvqa.org/textcaps/

Localized Narratives: https://google.github.io/localized-narratives/

悟空: https://wukong-dataset.github.io/wukong-dataset/download.html

- 人工评价：

- 表达的流畅度
- 与图片的相关度
- 表达的多样化程度

- 基于规则的自动化评价方法：

- BLEU（机器翻译）
- METEOR（机器翻译）
- ROUGE（文本自动摘要）
- CIDEr（图像描述）
- SPICE（图像描述）

BLEU：从n-gram说起

1-gram=5/6

3-gram=2/4

C: the the the the

S: The cat is standing on the ground

1-gram=1???

C: the the the the
S: The cat is standing on the ground

1-gram=1???

$$Count_{clip} = \min(Count, Max\_Ref\_Count)$$

$$P_N = \frac{\sum_i \sum_k \min(h_k(c_i), \max_{j \in m} h_k(s_{ij}))}{\sum_i \sum_k \min(h_k(c_i))}$$

$$BLEU = BP * \exp(\sum_{n=1}^{N} W_n \log P_N)$$

$$BP = \begin{cases} 1 & if \quad l_c > ls \\ e^{1-\frac{l_s}{l_c}} & if \quad l_c <= l_s \end{cases}$$

优点：使用方便、快捷，评价结果接近人类评价
缺点：不考虑语法，也没有涉及同义词或相似表达，长语句表现不佳

- BLEU的改进版，更关注召回率而非精确率。

$$R_{LCS} = \frac{LCS(C,S)}{len(S)}$$

- ROUGE-N 将BLEU的精确率优化召回率（常用）
- ROUGE-L 将BLEU的n-gram优化为最长公共子序列（常用）
- ROUGE-W 将ROUGE-L的连续匹配给予更高的奖励
- ROUGE-S 允许n-gram出现跳词(skip)

$$P_{LCS} = \frac{LCS(C,S)}{len(C)}$$

$$F_{LCS} = \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2 P_{LCS}}$$

$$ROUGE-N = \frac{\sum_{S\in\{ReferenceSummaries\}} \sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}} \sum_{gram_n\in S} Count(gram_n)}$$

$$METEOR = (1 - pen) \times F_{means}$$

$$F_{means} = \frac{PR}{\alpha P + (1-\alpha)R}$$

$$P = \frac{m}{c}$$

$$R = \frac{m}{r}$$

$$Pen = \frac{\#chunks}{m}$$

C: the president spoke to the audience.
S: the president then spoke to the audience.

#chunks=2

← 相邻匹配的个数

优点：召回率和精确率的调和平均；增加了基于WordNet的同义词库，解决同义词匹配
缺点：参数过多，计算复杂；需要外部知识源；非语义对应

$$CIDEr(c_i, S_i) = \frac{1}{N} \sum_{n=1} CIDEr_n(c_i, S_i)$$

$$CIDEr_N(c_i, S_i) = \frac{1}{m} \sum_{j} \frac{g_n(c_i) * g_n(s_{ij})}{\| g_n(c_i) \| * \| g_n(s_{ij}) \|}$$

$$g_n(\cdot) = \frac{h_n(\cdot)}{\sum_{\omega_l \in \Omega} h_l(\cdot)} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_n(\cdot))}\right)$$

TF-IDF+余弦相似度

优点：对出现频率不同的单词赋予不同的权重，区分了重要性
缺点：非语义对应

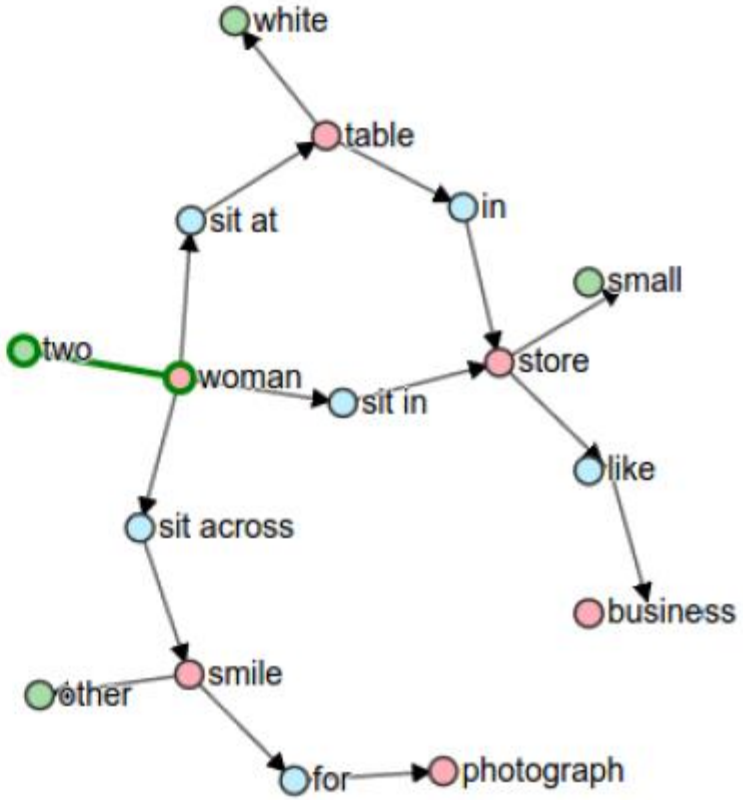
"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"



$$SPICE(c,S) = F_1(c,S) = \frac{2 \cdot P(c,S) \cdot R(c,S)}{P(c,S) + R(c,S)}$$

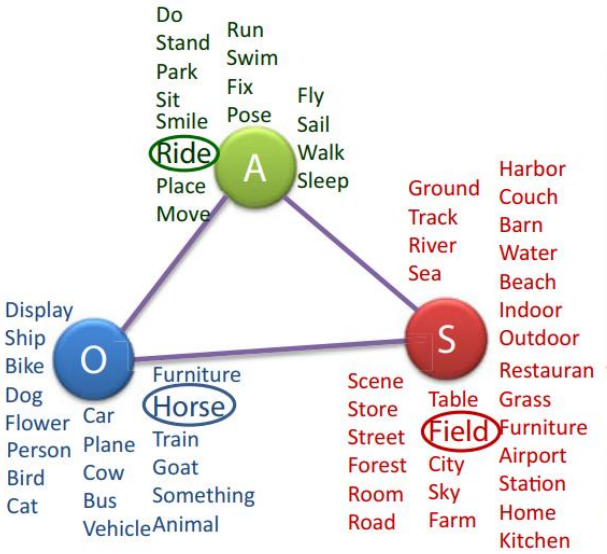
$$P(c,S) = \frac{|T(G(c) \otimes T(G(S))|}{|T(G(c))|}$$

$$R(c,S) = \frac{|T(G(c) \otimes T(G(S))|}{|T(G(S))|}$$

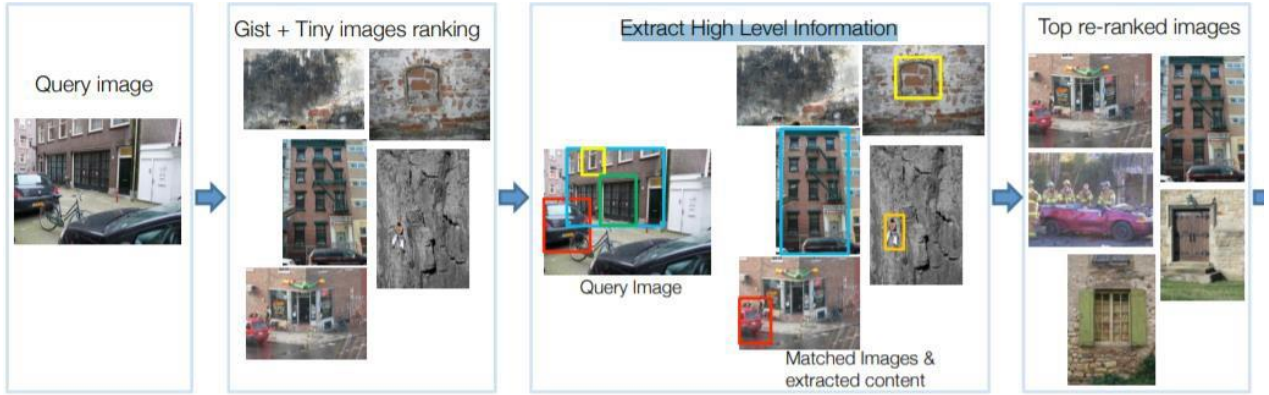
优点：基于图的语义表示
缺点：不能准确判断语法结构错误

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

基于**模板**的图像描述方法



基于**检索**的图像描述方法



存在种种问题，表现并不好……

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

3

经典综述

汇报人：李桐

bottom-up模型

bench
WOODEN
sitting
Tree

A wooden bench sitting net to a tree.

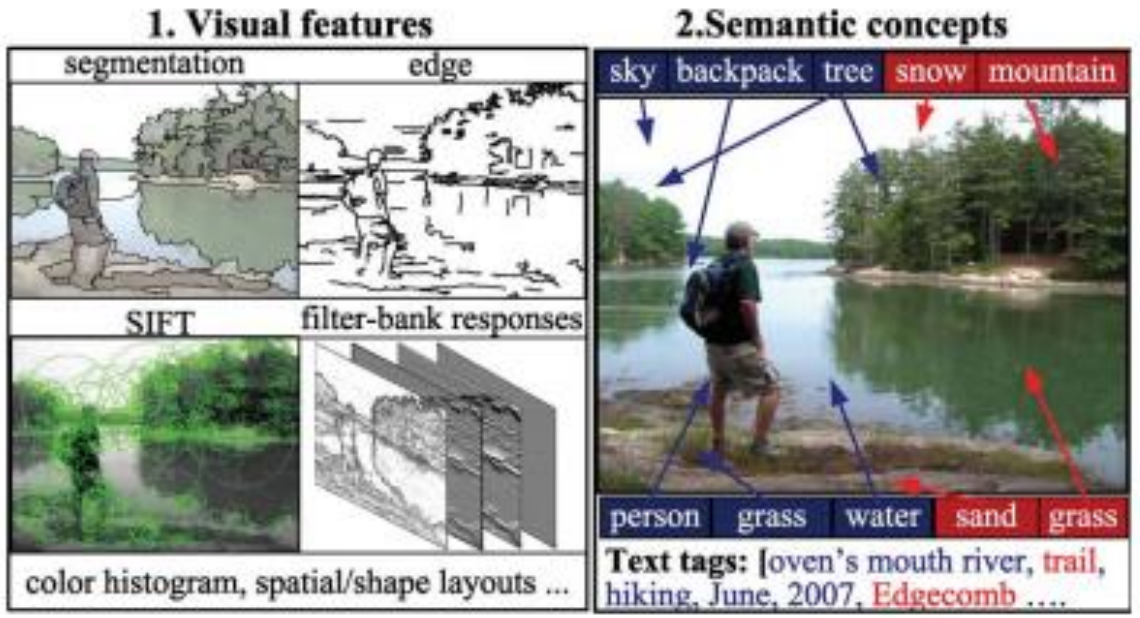
top-down模型

Deep learning

A wooden bench sitting net to a tree.

bottom-up模型

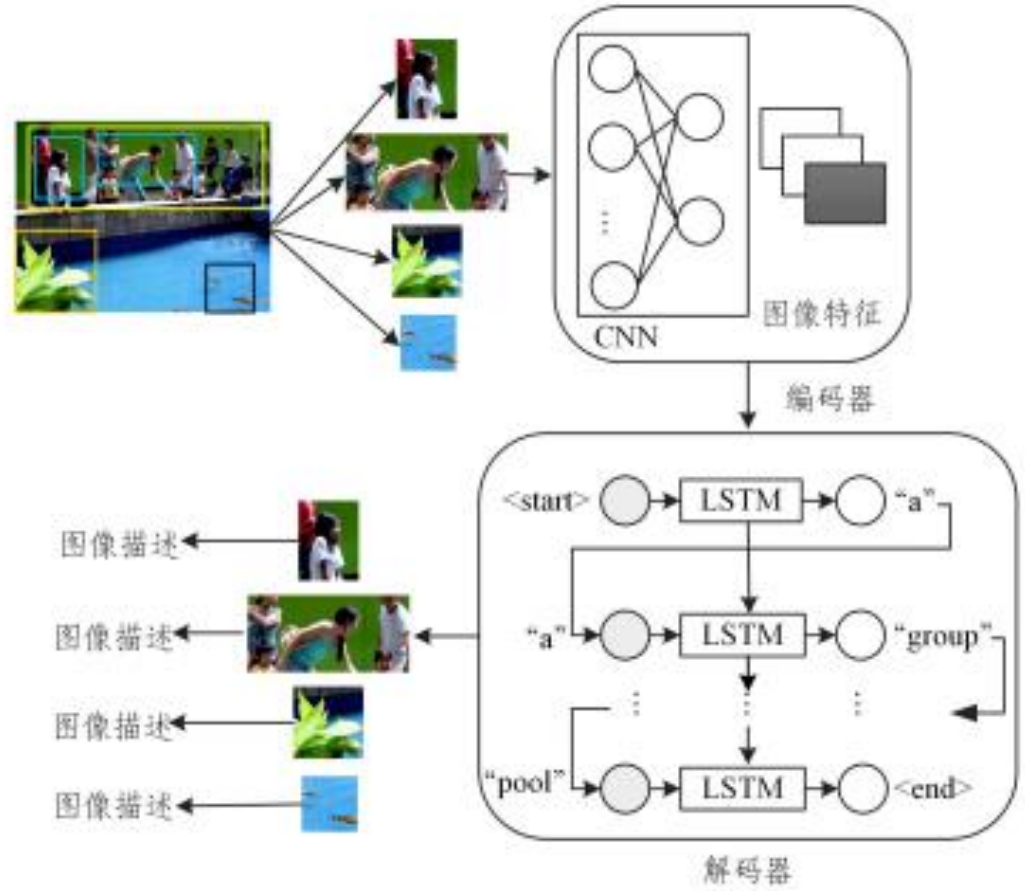
- 传统手工设计
- 早期深度学习



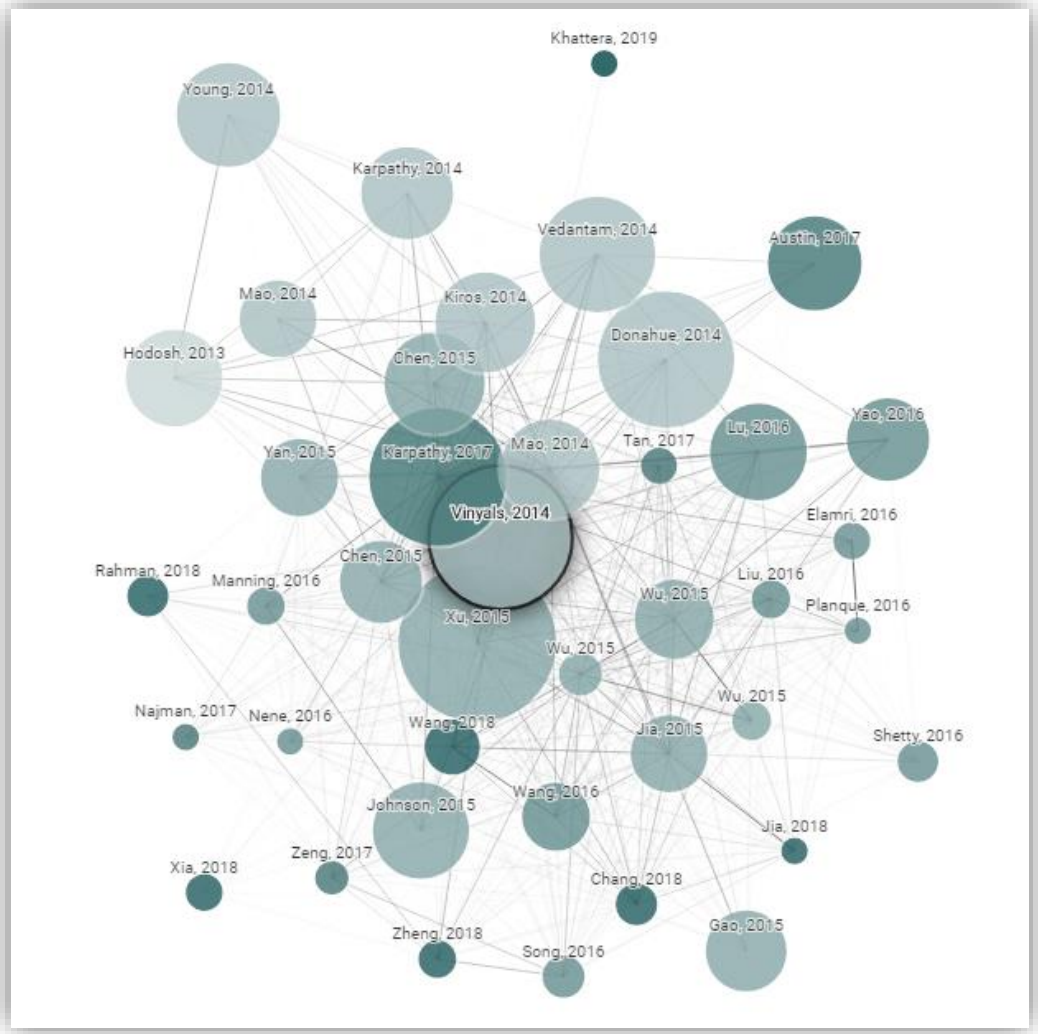
top-down模型

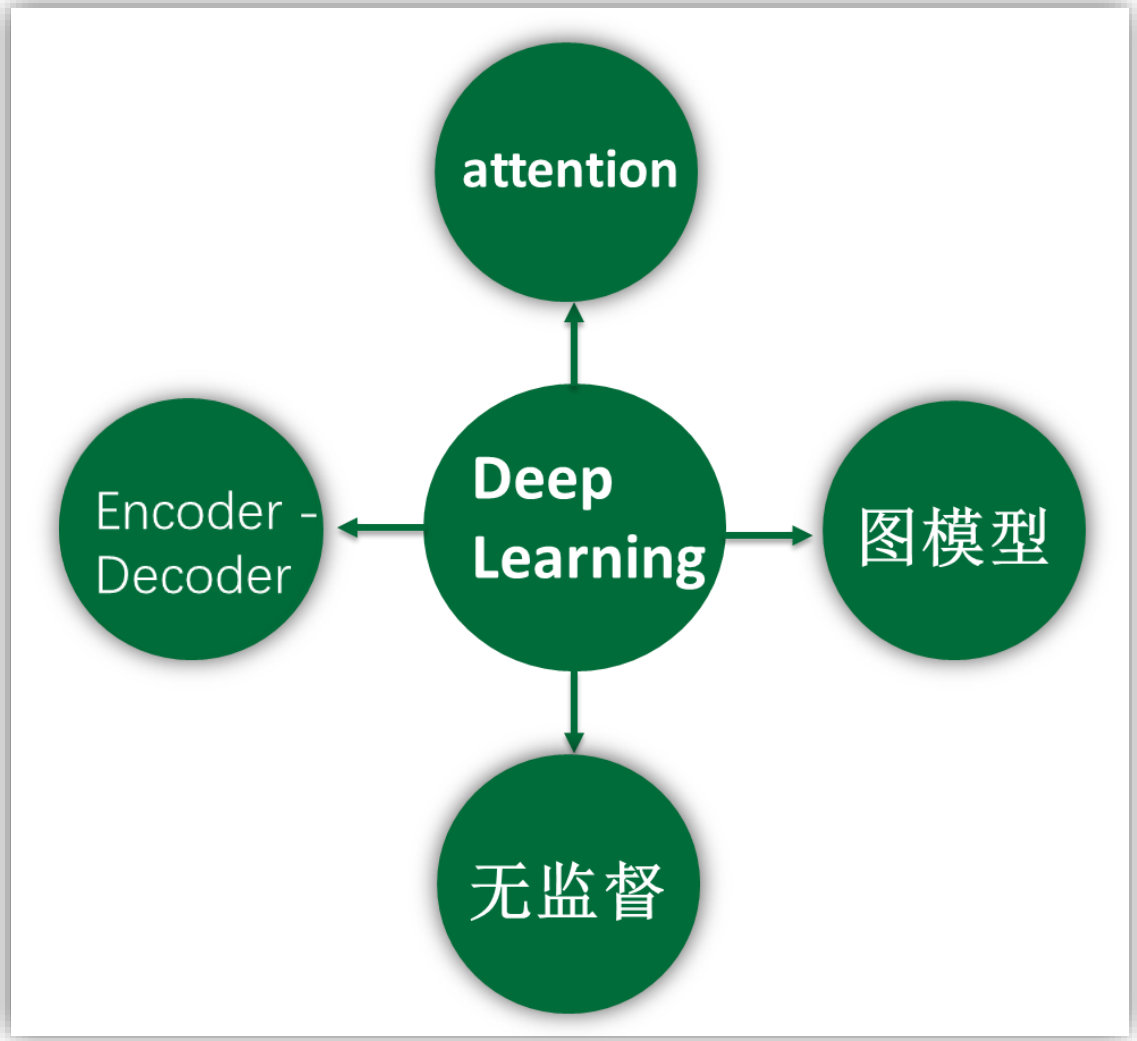
端到端模型

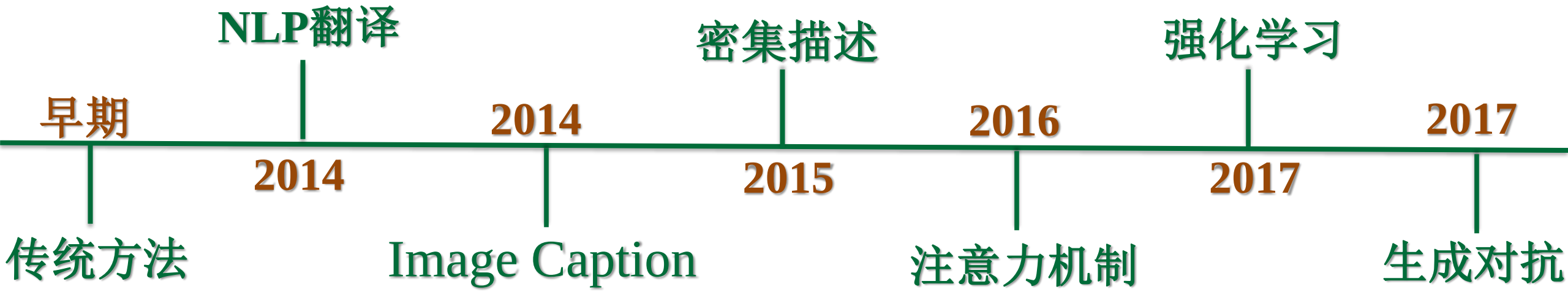
密集描述方法



分类——网络结构


Khattera, 2019
Young, 2014
Karpathy, 2014
Vedantam, 2014
Austin, 2017
Mao, 2014
Kiros, 2014
Donahue, 2014
Hodosh, 2013
Chen, 2015
Yao, 2016
Lu, 2016
Tan, 2017
Mao, 2014
Yan, 2015
Karpathy, 2017
Vinyals, 2014
Elamri, 2016
Rahman, 2018
Manning, 2016
Chen, 2015
Liu, 2016
Wu, 2015
Planque, 2016
Xu, 2015
Wu, 2015
Wu, 2015
Najman, 2017
Nene, 2016
Wang, 2018
Jia, 2015
Shetty, 2016
Wang, 2016
Johnson, 2015
Jia, 2018
Zeng, 2017
Xia, 2018
Chang, 2018
Zheng, 2018
Gao, 2015
Song, 2016


attention
Encoder - Decoder
Deep Learning
图模型
无监督


早期
传统方法
2014
NLP翻译
2014
Image Caption
2015
密集描述
2016
注意力机制
2017
强化学习
2017
生成对抗

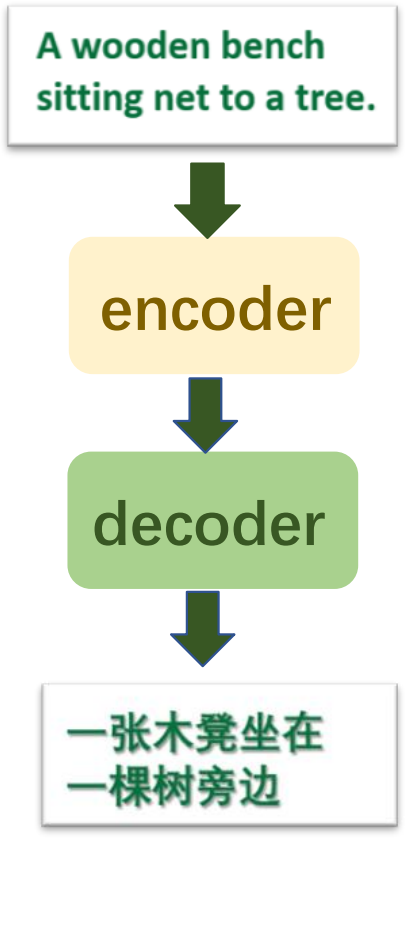
发展——深度学习的引入

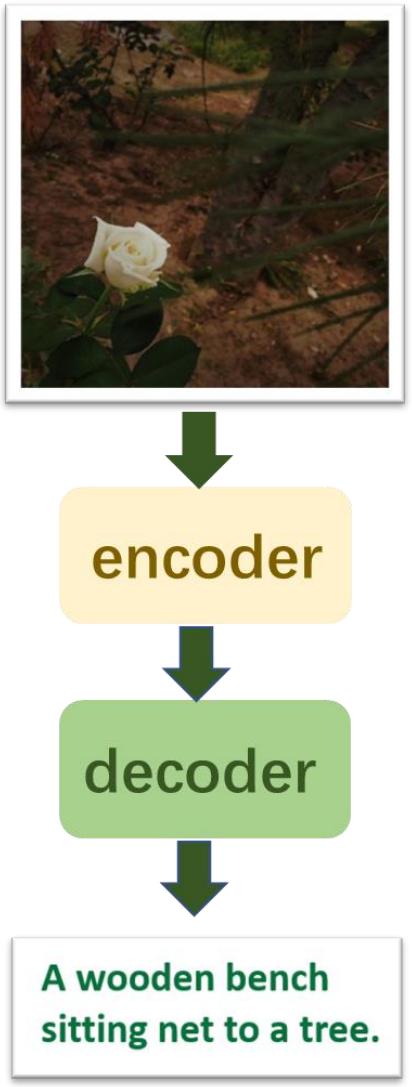
2012 —— Alexet

2014 —— VGG

2014 —— Translation

2014 —— Image Caption





2014 —— Image Caption

2014 —— Show and Tell

2014 —— Neural Talk

CNN → RNN → A wooden bench sitting net to a tree.

Show and Tell

VGG → RNN → A wooden bench sitting net to a tree.

Neural Talk

发展——2014-2015 解码端改善

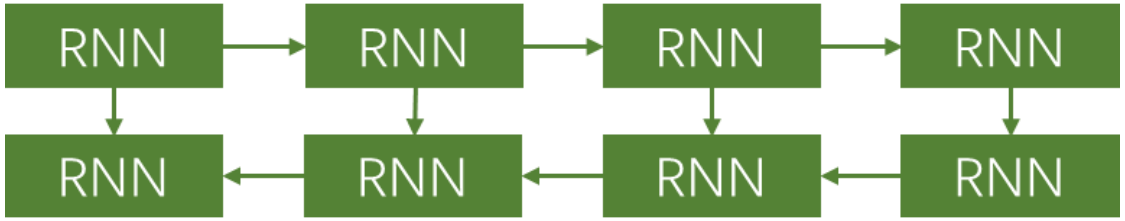
2014

Encoder

Decoder

2015

改善



双向RNN[1]

双向双层LSTM[2]

phi-LSTM[3]

……

[1]:Mind's Eye: A Recurrent Visual Representation for Image Caption Generation
[2]:Image Captioning with Deep Bidirectional LSTMs
[3]:A Phrase-based Hierarchical LSTM Model for Image Captioning

发展——2014-2015 解码端改善

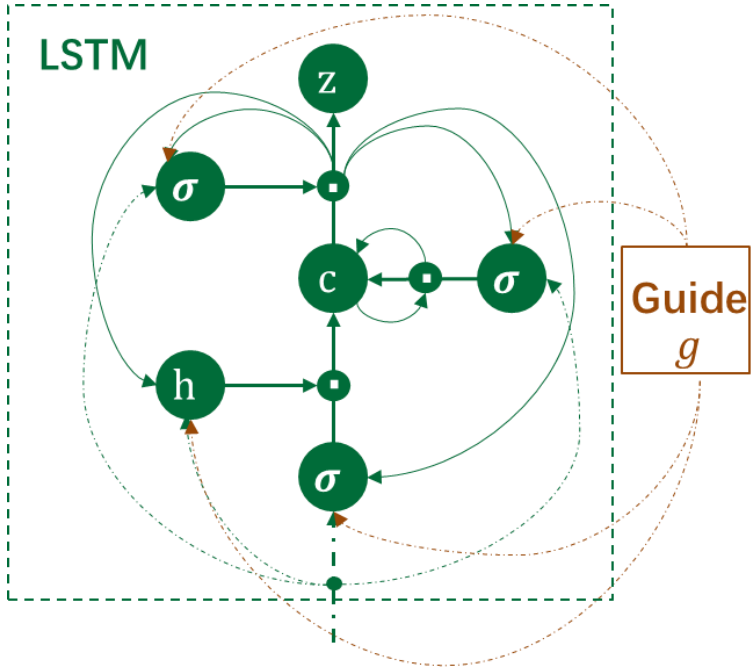
2014

2015

Encoder

Decoder

改善



g-LSTM[1]

CNN解码模型[2][3]

[1]:Guiding the Long-Short Term Memory Model for Image Caption Generation
[2]:Convolutional image captioning
[3]:CNN+ CNN: Convolutional Decoders for Image Captioning

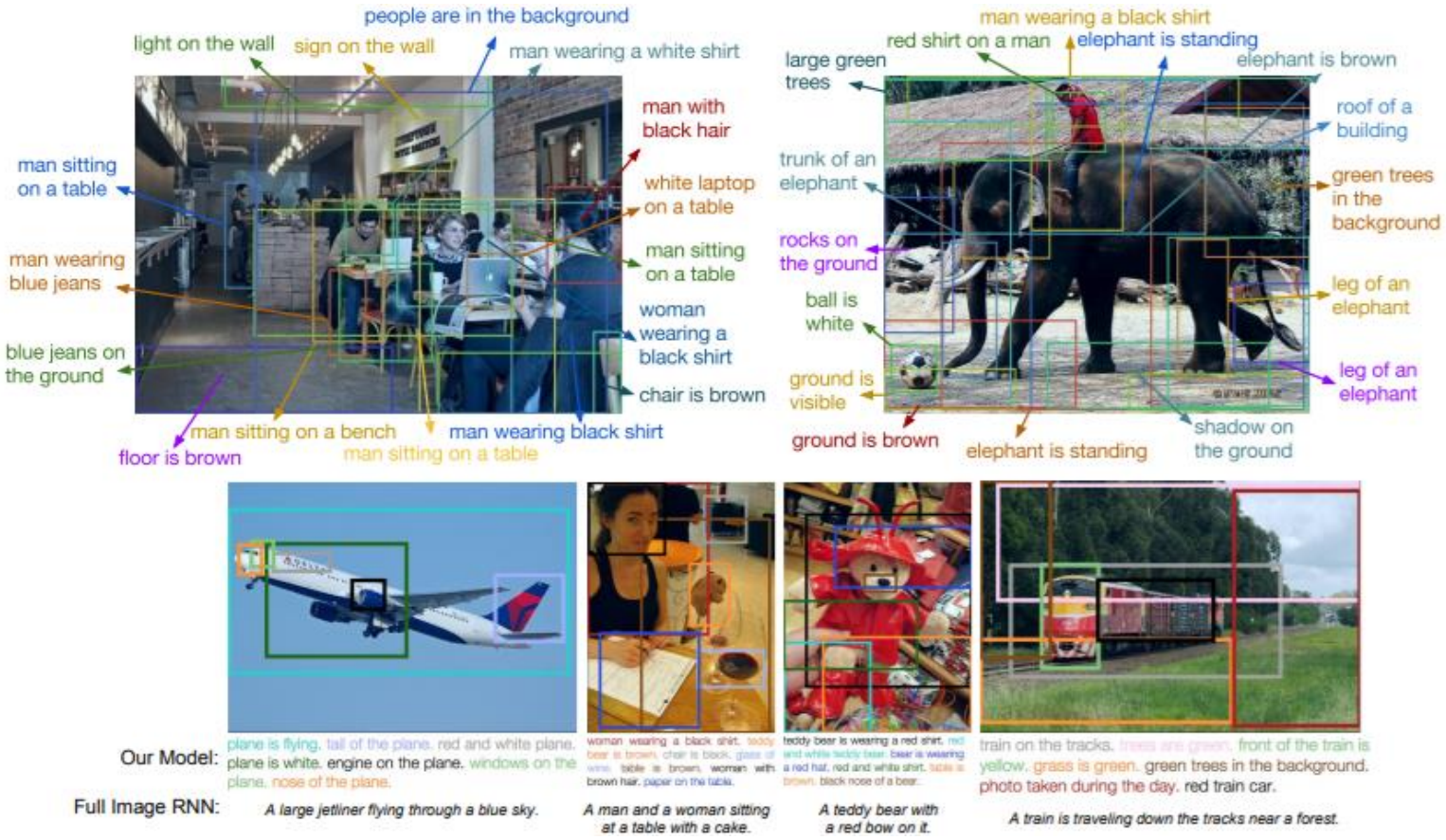
发展——密集描述引入

2015 —— DenseCap

2017 —— joint inference

2019 —— Dense relational captioning

DenseCap



DenseCap: Fully Convolutional Localization Networks for Dense Captioning CVPR2015

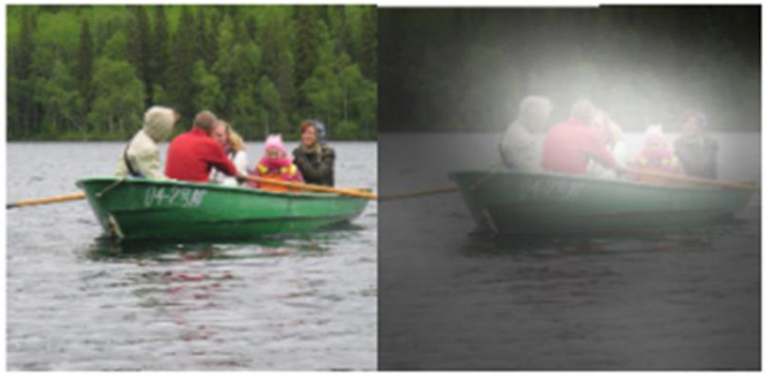
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

2015 — **Attention + Translation**

2016 — **Attention + Image Caption**

Show, Attend and Tell

A dog is standing on a hardwood floor.



A group of people sitting on a boat in the water.

发展——Show, Attend and Tell

2015 — **Attention + Translation**

2016 — **Attention + Image Caption**

Show and Tell

CNN

RNN

A wooden bench sitting net to a tree.

2014 → 2016

Show, Attend and Tell

CNN

RNN + Attention

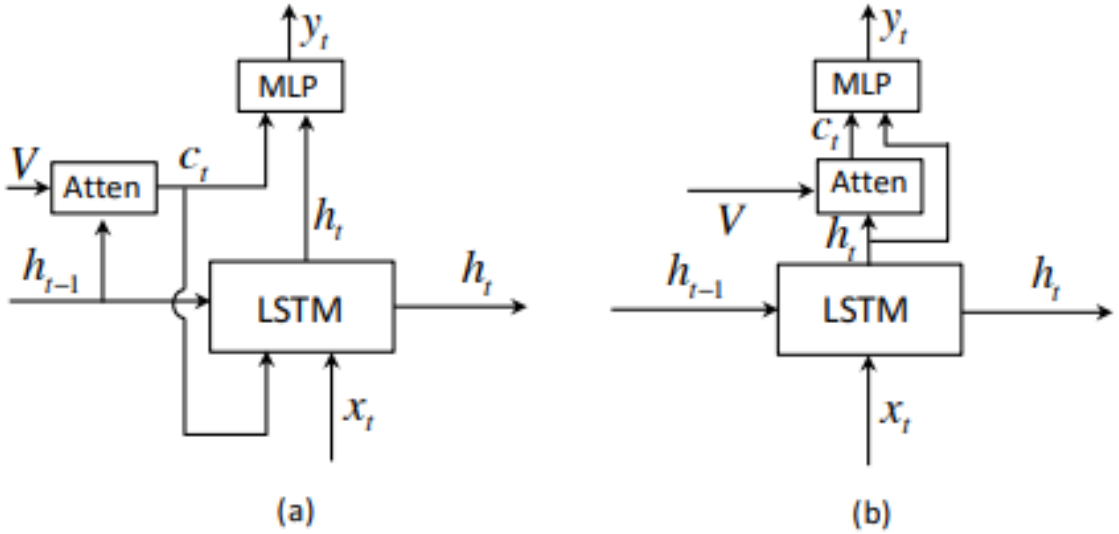
A wooden bench sitting net to a tree.

2016

2017

Attention 改善

Knowing When to Look(CVPR 2017)



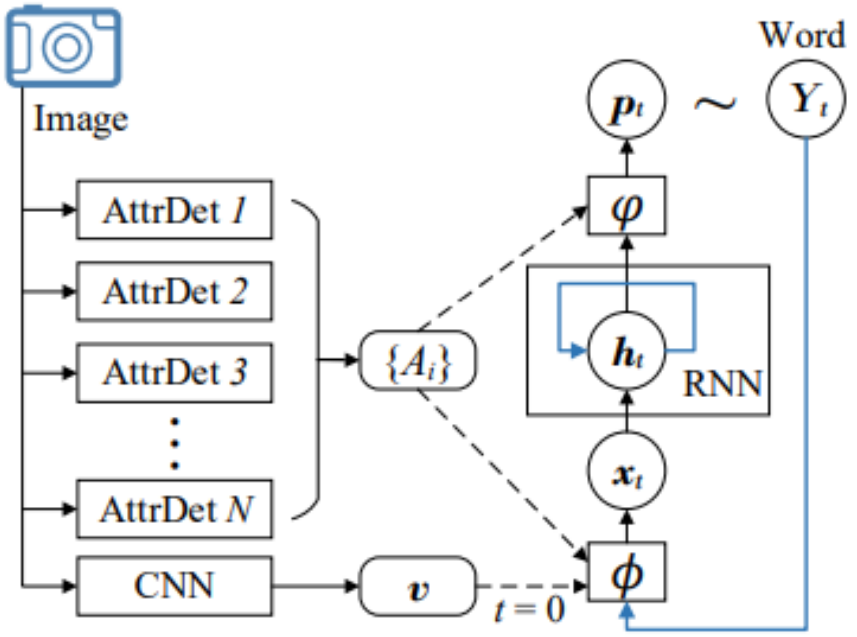
2016

2017

Attention 改善

Image Captioning with Semantic Attention

(CVPR 2016)

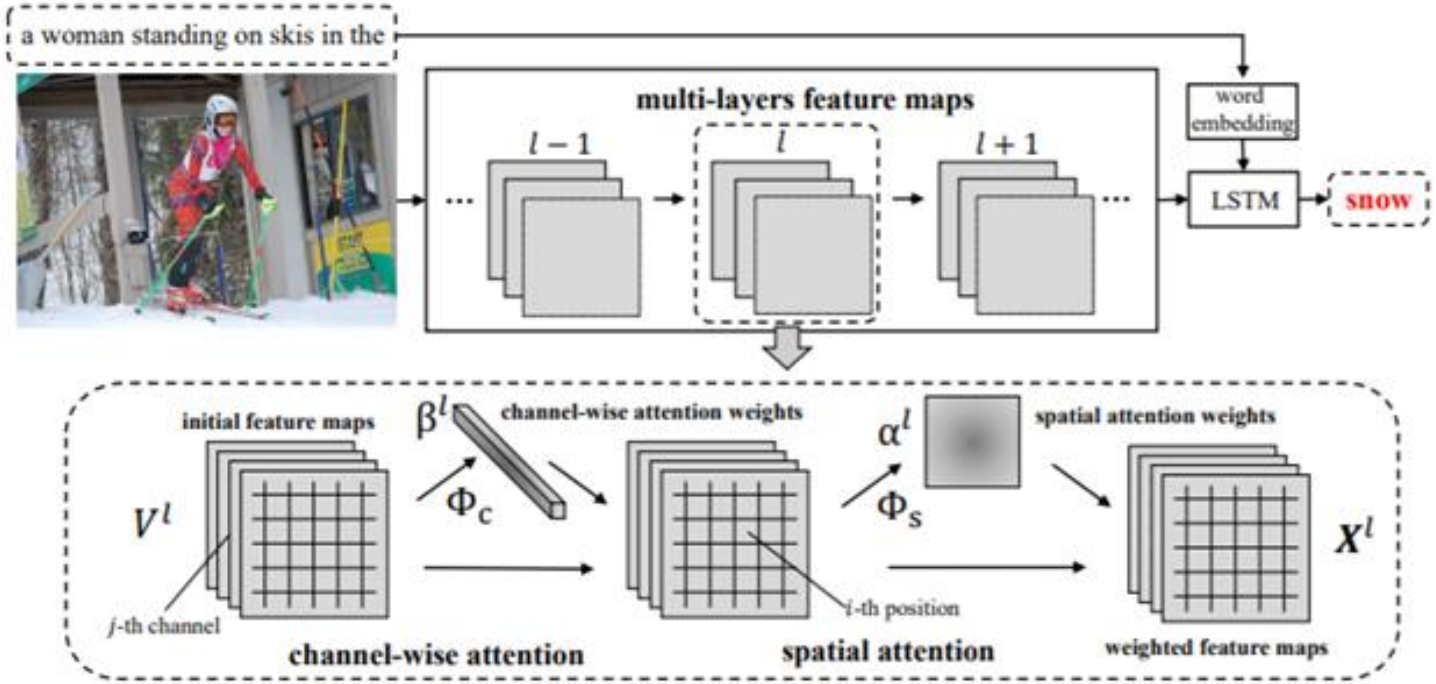


2016

2017

Attention 改善

SCA-CNN(CVPR2017)



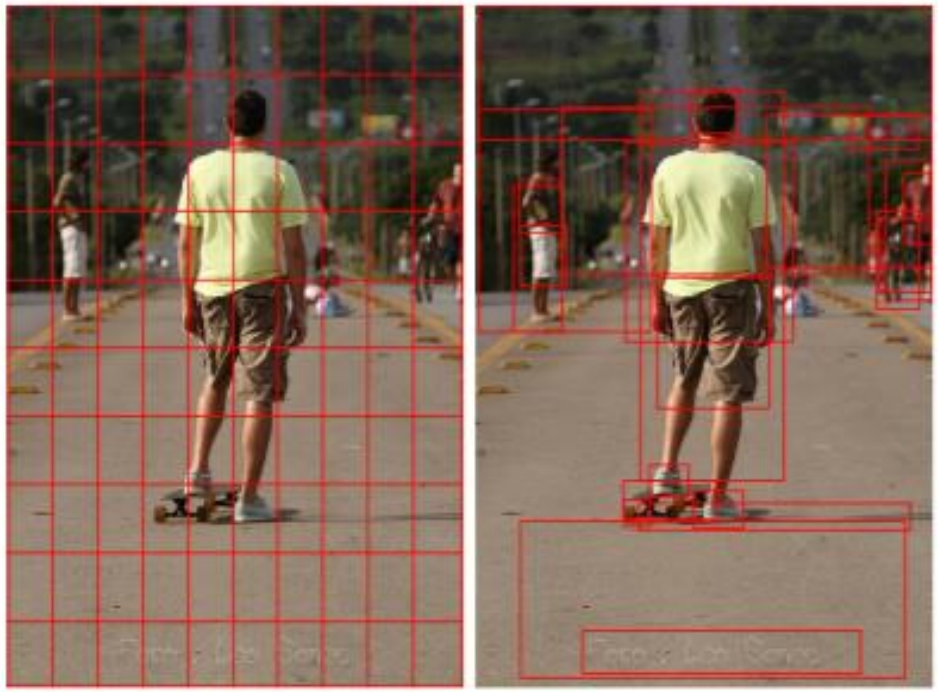
2016

2017

Attention 改善

Bottom-Up and Top-Down Attention

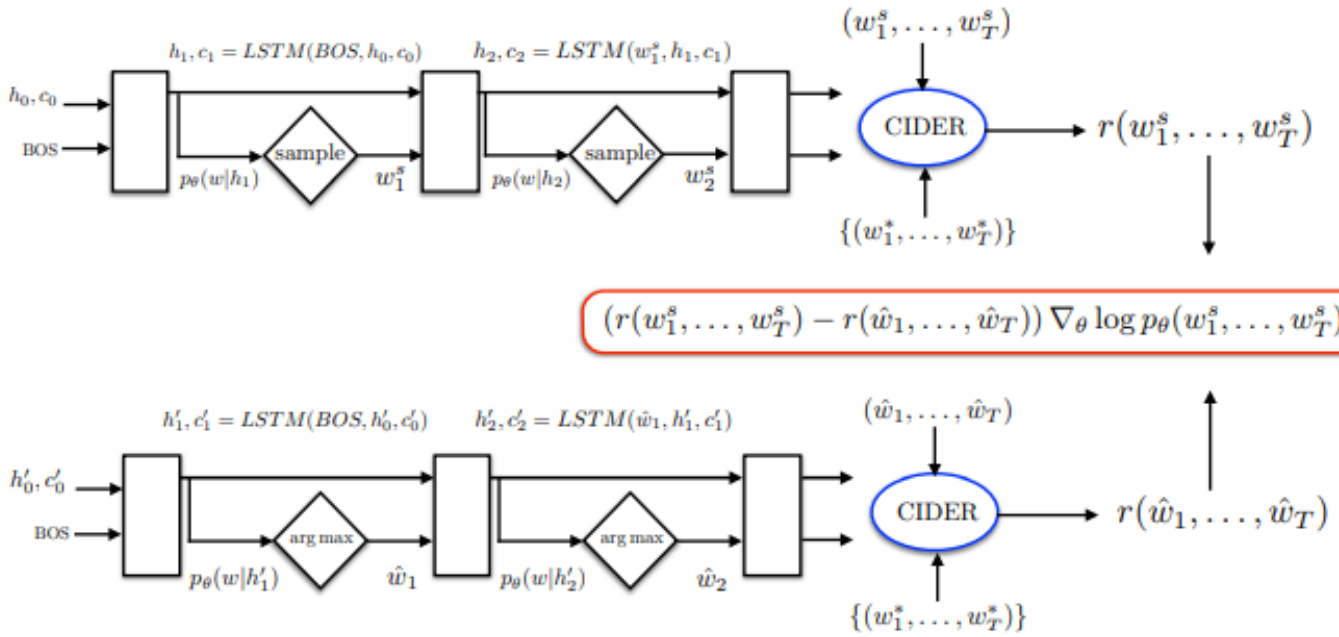
(CVPR 2017)



2015 强化学习 Translation

2017 强化学习 Image Caption

Self-critical sequence training for image captioning (CVPR 2017)



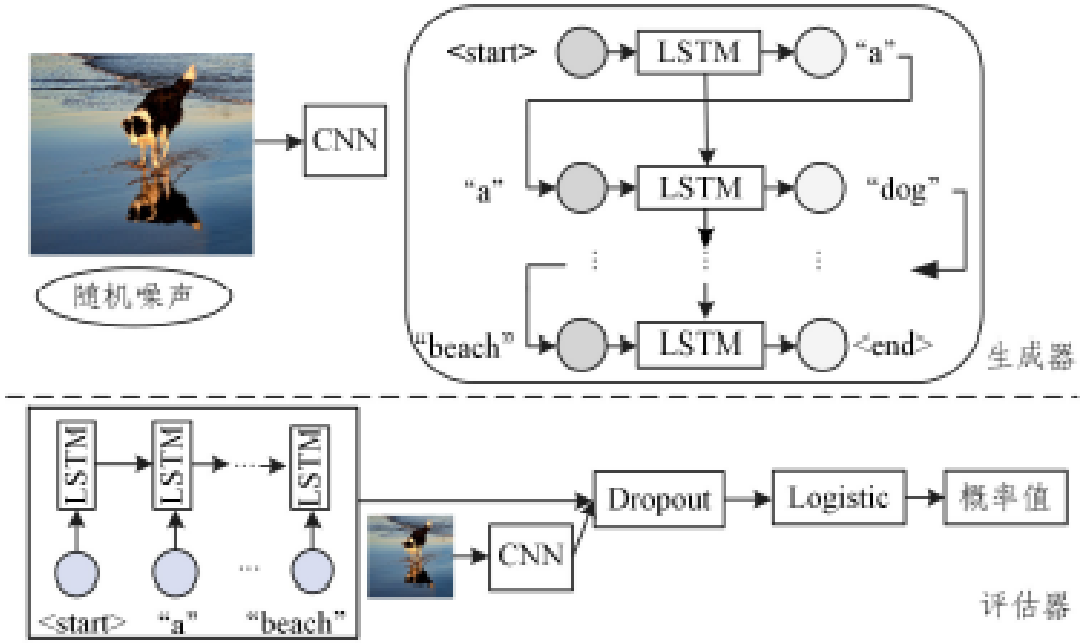
2014 —— GAN

Generative Adversarial Networks

2017 —— Image Caption

Towards Diverse and Natural Image Descriptions via a Conditional GAN

(ICCV 2017)



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

4

Attention编解码器方法

Attention机制-编解码器模型架构

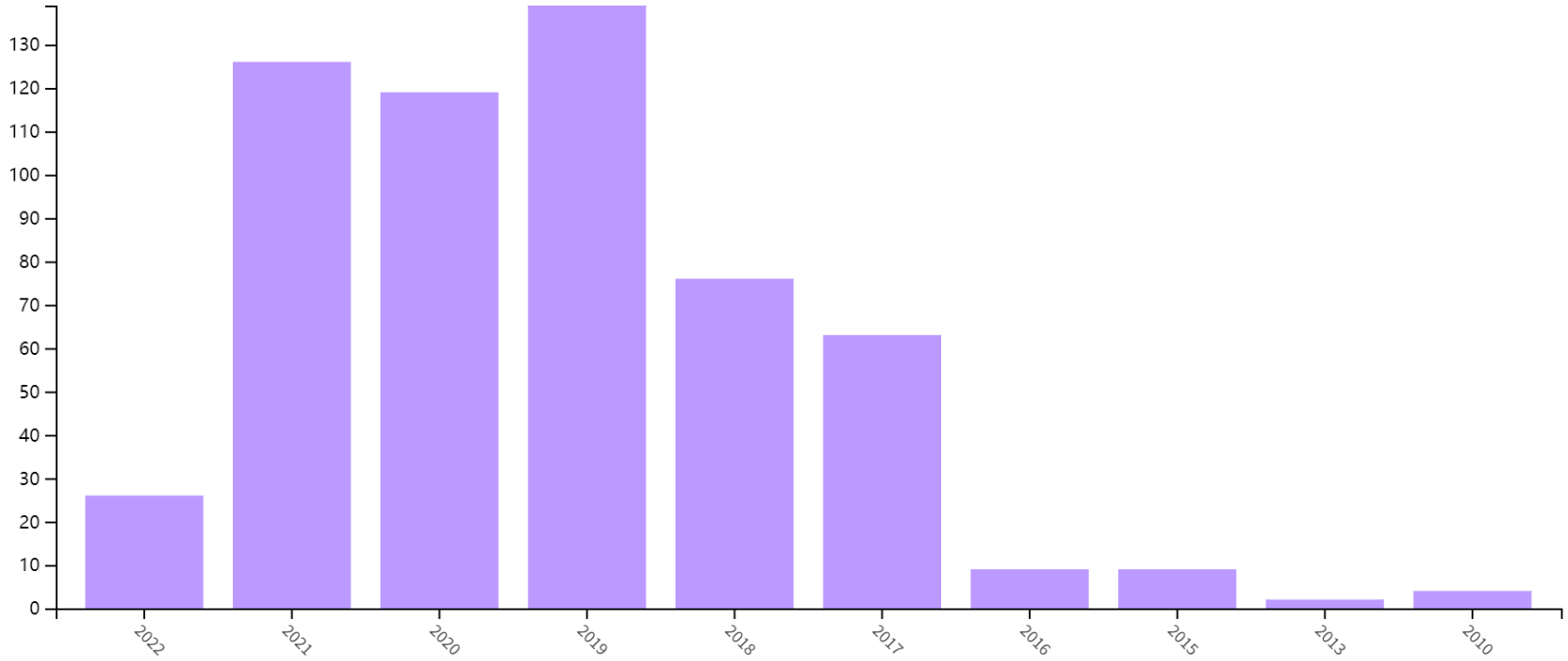
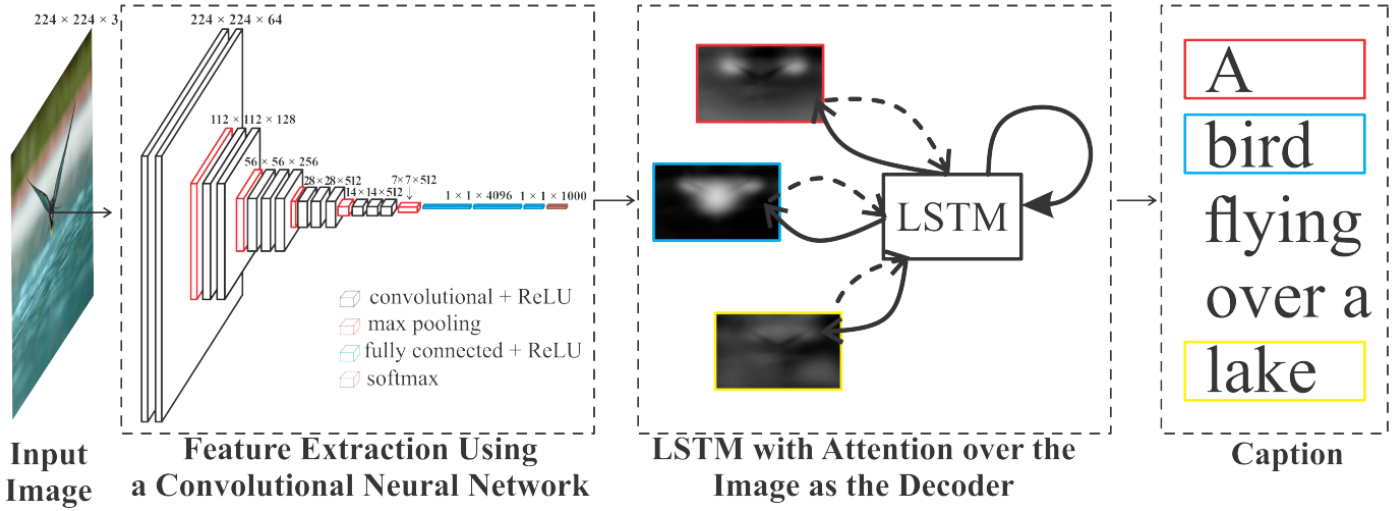


Image caption+Attention

数据来源：web of science

Attention机制-编解码器模型架构



通过CNN做编码器，然后传给RNN，经过RNN对于每张图逐字生成caption向量。

在RNN中添加attention机制

[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention]

发展——Show, Attend and Tell

2015 — Attention + Translation

2016 — Attention + Image Caption

Show and Tell

CNN

RNN

A small white dog laying on a bed

2014 → 2016

Show, Attend and Tell

CNN

RNN + Attention

A dog is standing on a hard wood floor

Attention机制-引入-改进

hard-attention

A man and a woman playing frisbee in a field.

soft-attention

A woman is throwing a frisbee in a park.

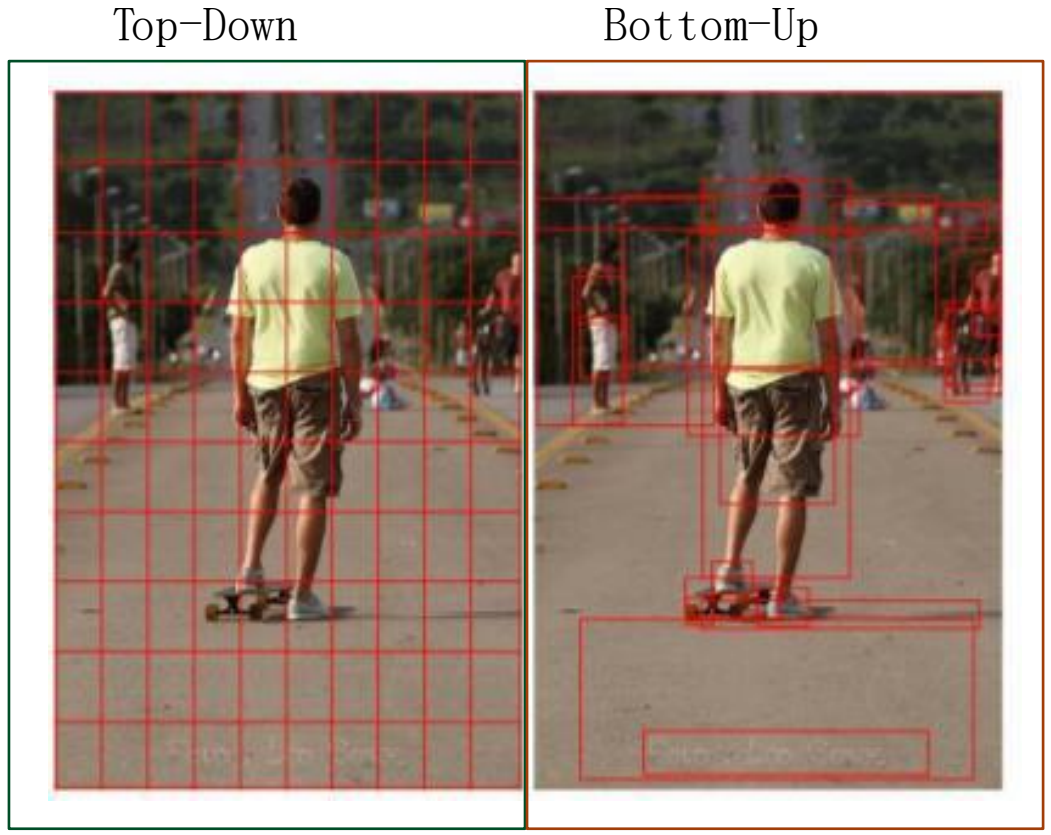
整张图作为输入，并未考虑目标region信息。

[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention]

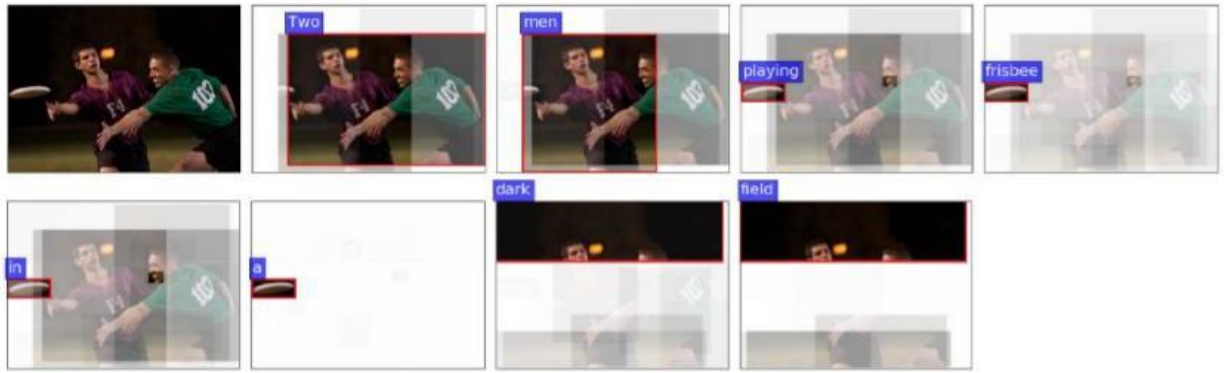
Attention Is All You Need

WHAT IS *ATTENTION* ?

[NeuralPS-2017 Attention Is All You Need]



利用目标检测Faster-Rcnn获得自底向上的注意力特征

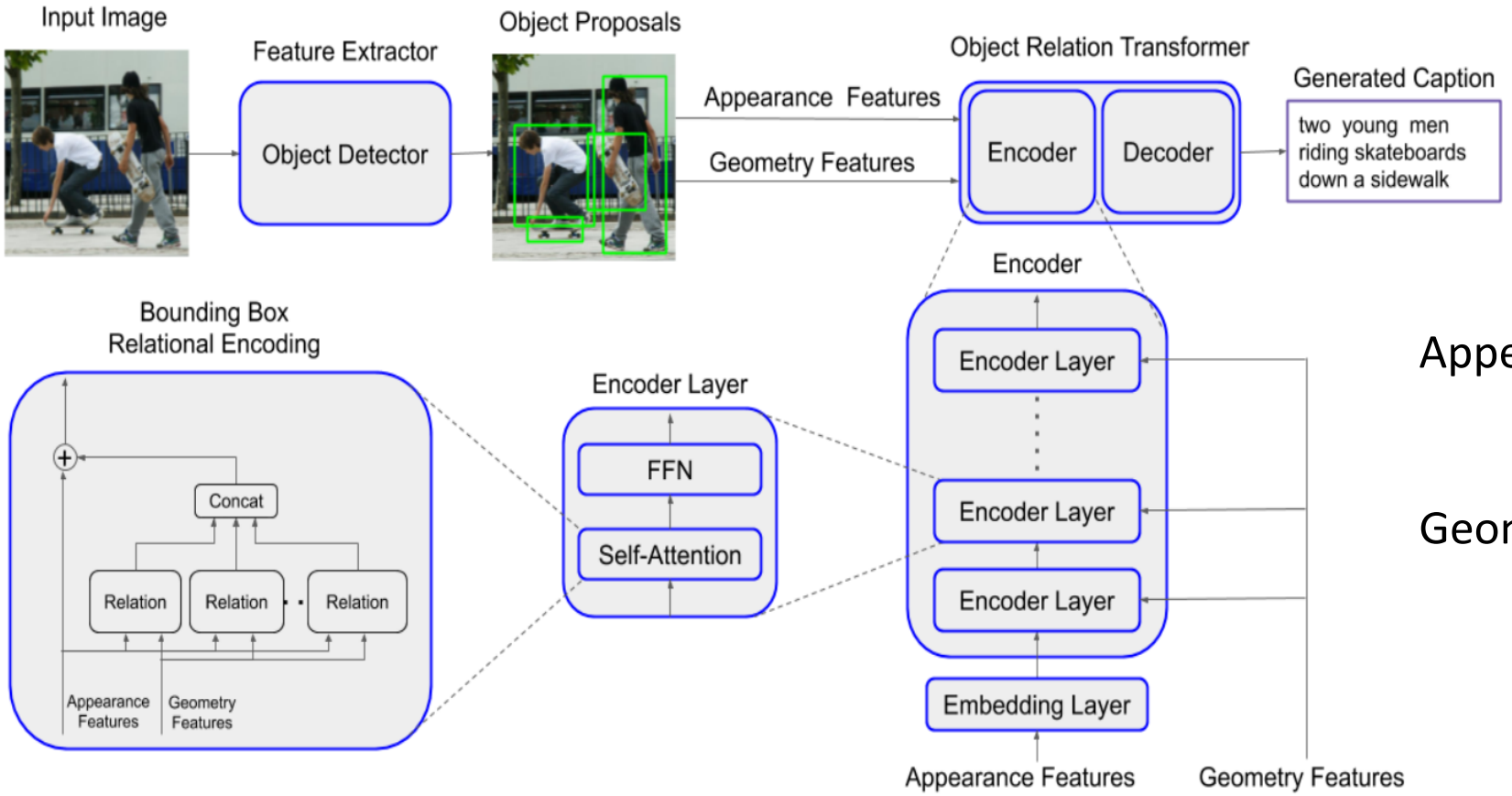


具体来说，bottom-up机制基于Faster R-CNN，即基于目标来计算attention。得到图片中每个目标或显著区域的特征向量表示。

[2018-CVPR Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering]

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

Attention加入位置空间信息



在模型中加入了Geometry 信息，适于模型提取空间位置特征

Appearance Features: Resnet-101

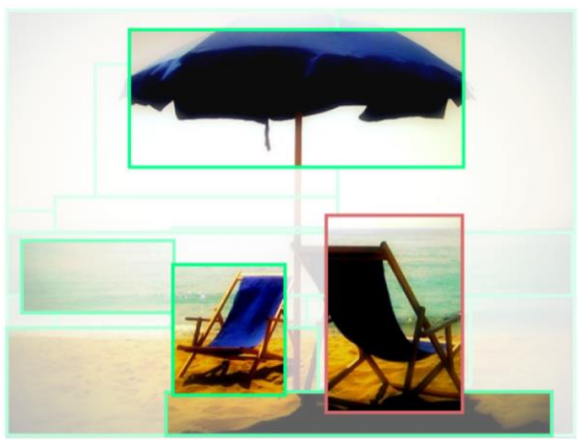
Geometry Features: Faster-RCNN

加入了空间信息，但未考虑多区域多细粒度的交互

[NeurIPS-2019 Transforming Objects into Words]

Object Relation Transformer

Attention加入位置空间信息



two chairs and an umbrella on a beach

two beach chairs under an umbrella on the beach

a group of young men riding skateboards down a sidewalk

two young men riding skateboards down a sidewalk

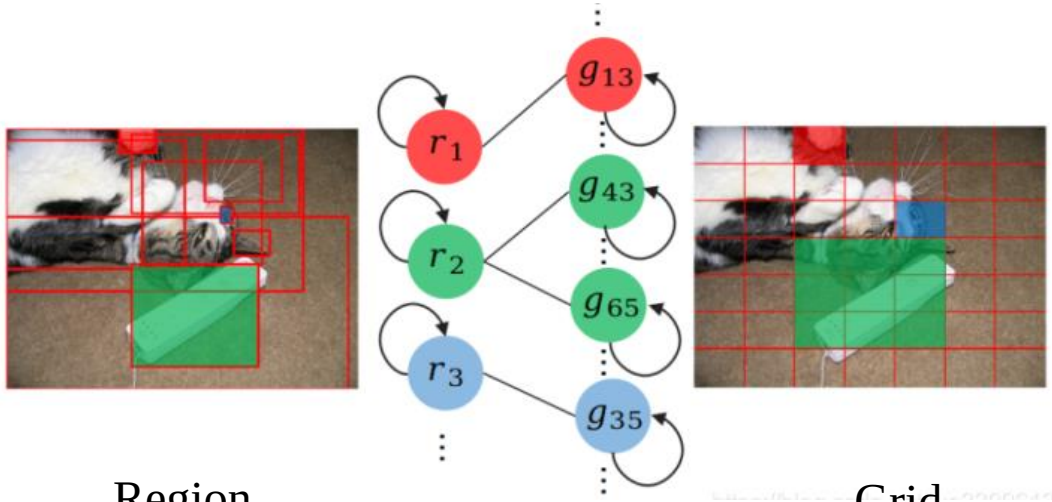
Standard:two chairs and an umbrella on a breach

ORT:two beach charis under an umbrella on the breach

Standard:a group of young men riding skateboards down a sidewalk

ORT:two young men riding skateboards down a sidewalk

[NeuralPS-2019 Transforming Objects into Words]

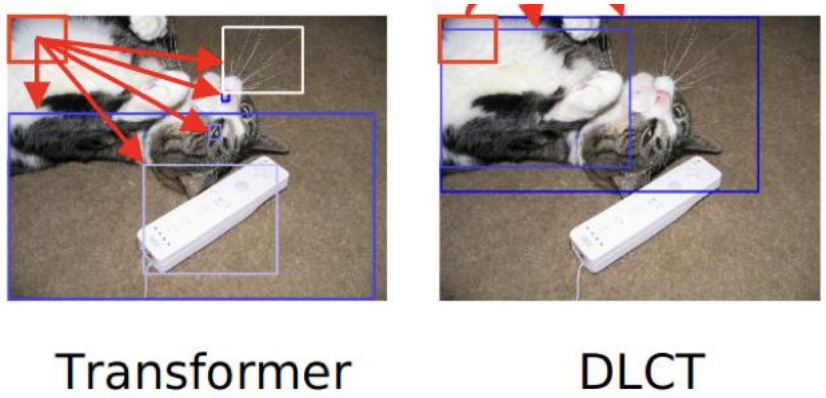


Region　　Grid

顶层区域，交互性好，包含上下文信息，但细粒度低

底层特征，交互性差，不包含上下文信息但是细粒度高

作者提出了位置约束交叉注意模型来模拟区域和网格之间复杂的相互作用，实现层间融合。为了避免引入语义噪声，首先创建一个几何对齐图，所有区域和网格特征都表示为独立的节点，当且仅当网格和区域的有交点时，该网格节点与区域节点才相连。有交集的区域和网格(用相同颜色突出显示)由无向边连接，以消除语义上的不相关信息。注意每个节点都有一个自连接的边。



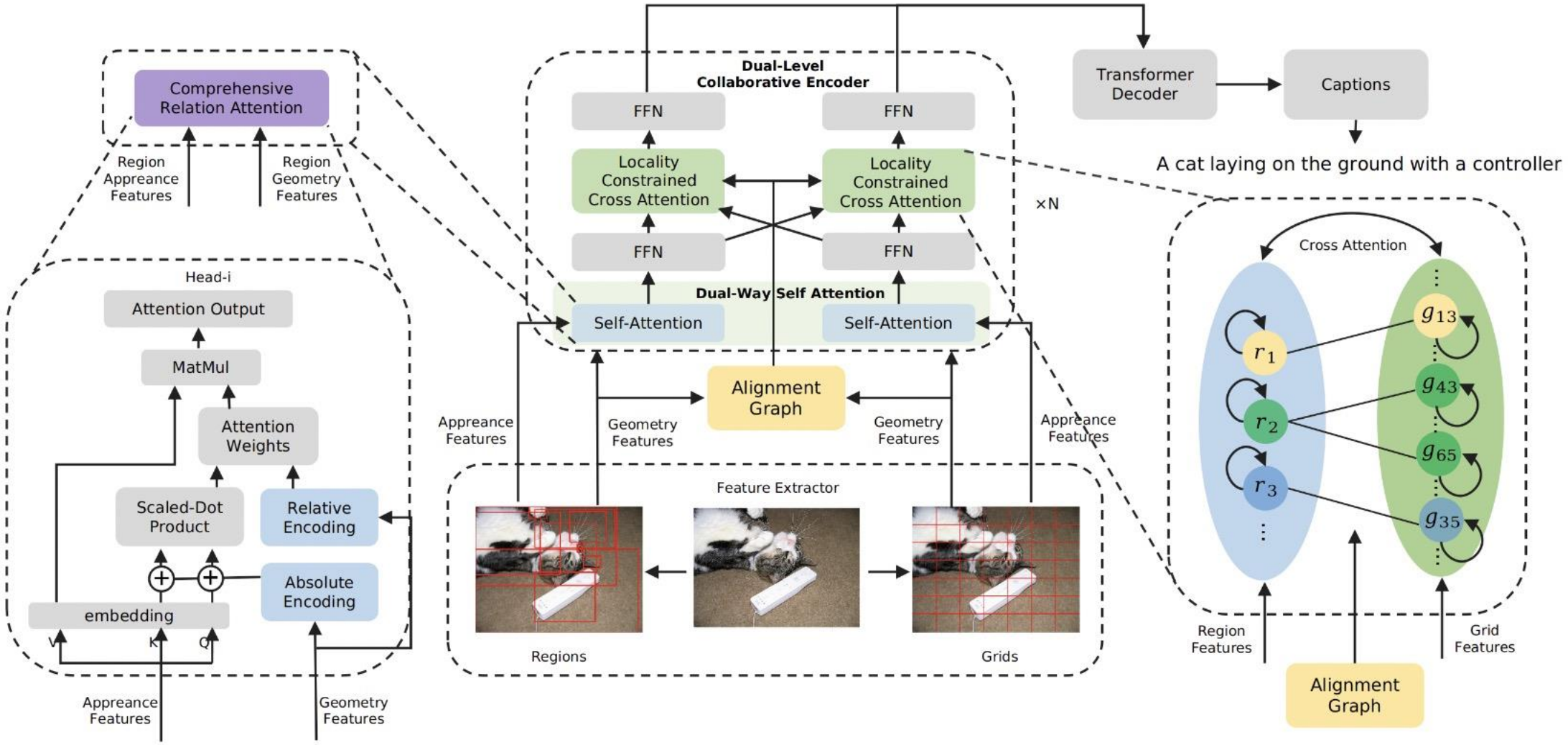
Transformer　　DLCT

Transformer 注意力机制考虑了图像中任意region之间的交互，而DLCT选取特定的rejion和特定grid进行交互。

Transformer: A brown horse grazing in a field

DLCT: Two horses grazing in a field of grass

[2021 Dual-Level Collaborative Transformer for Image Captioning]



[2021 Dual-Level Collaborative Transformer for Image Captioning]

效果对比

Model	B-1		B-2		B-3		B-4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST (ResNet-101)	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down (ResNet-101)	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
HAN (ResNet-101)	80.4	94.5	63.8	87.7	48.8	78.0	36.5	66.8	27.4	36.1	57.3	71.9	115.2	118.2
GCN-LSTM (ResNet-101)	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE (ResNet-101)	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoA (ResNet-101)	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
HIP (SENet-154)	81.6	95.9	66.2	90.4	51.5	81.6	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
M2 (ResNet-101)	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer (ResNet-101)	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
X-Transformer (SENet-154)	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
DLCT (ResNeXt-101)	82.0	96.2	66.9	91.0	52.3	83.0	40.2	73.2	29.5	39.1	59.4	74.8	131.0	133.4
DLCT (ResNeXt-152)	**82.4**	**96.6**	**67.4**	**91.7**	**52.8**	**83.8**	**40.6**	**74.0**	**29.8**	**39.6**	**59.8**	**75.3**	**133.3**	**135.4**

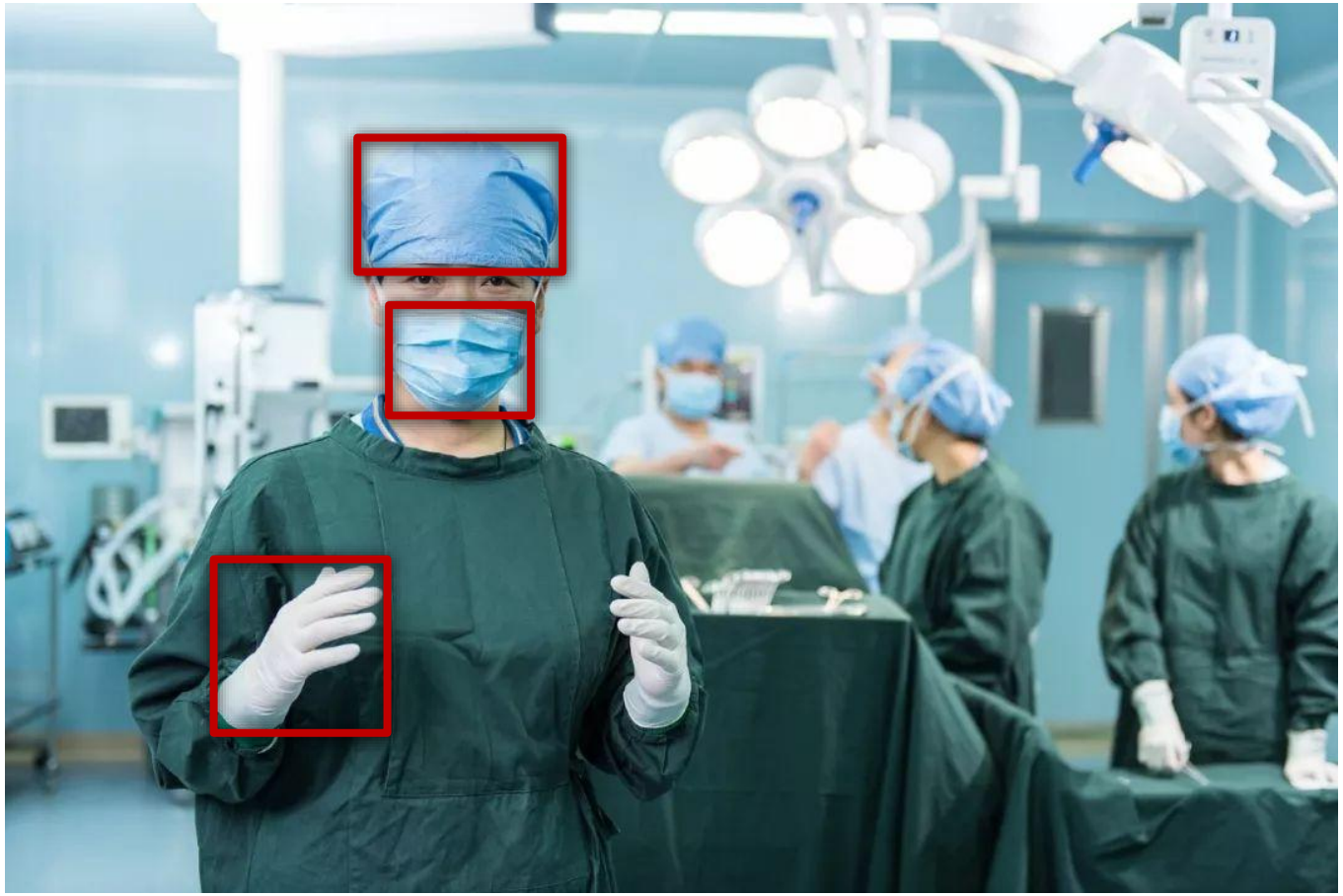
北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

5

基于图、强化学习的方法

汇报人：刘天锐

仅仅end-to-end vs. graph-based

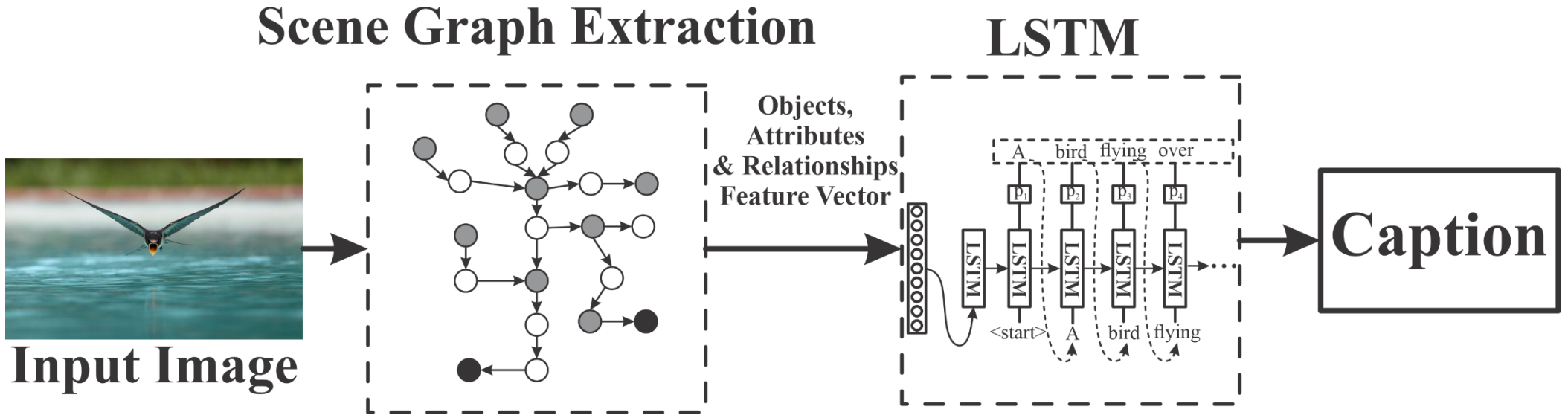


encoder

decoder

A wooden bench sitting net to a tree.

基于图的一般流程

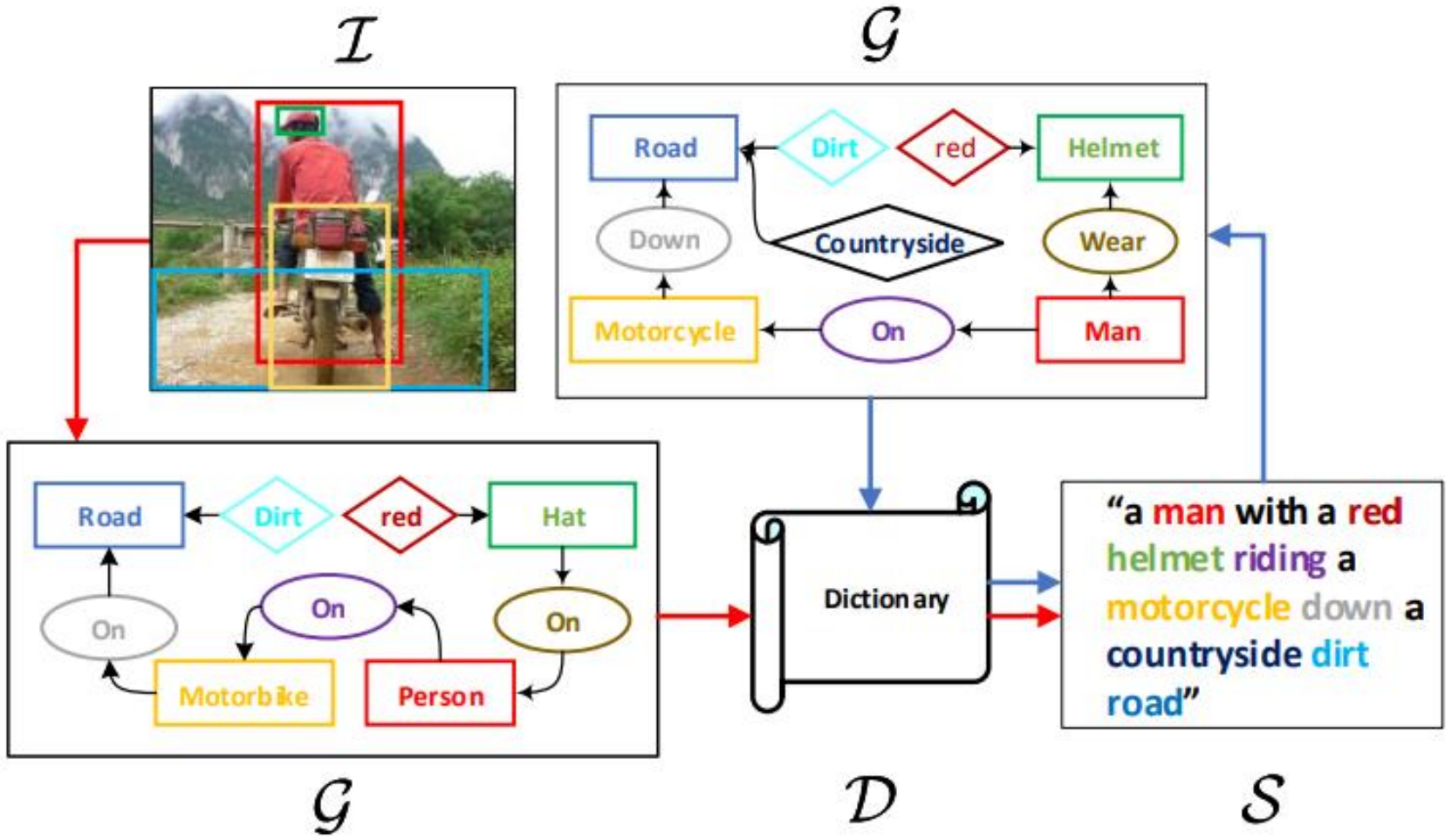


图卷积网络（GCNs）用来编码场景图中的区域和关系
获得的特征向量被传递给LSTM解码器生成标题

场景图（scene graph）

Intuition：
推理互补端到端，
知识→视觉语言
连接实体、属性、关系

编码器：$V \leftarrow I$
映射：$G \leftarrow V$
$V' \leftarrow R(V,G;D)$
解码器：$S \leftarrow V'$

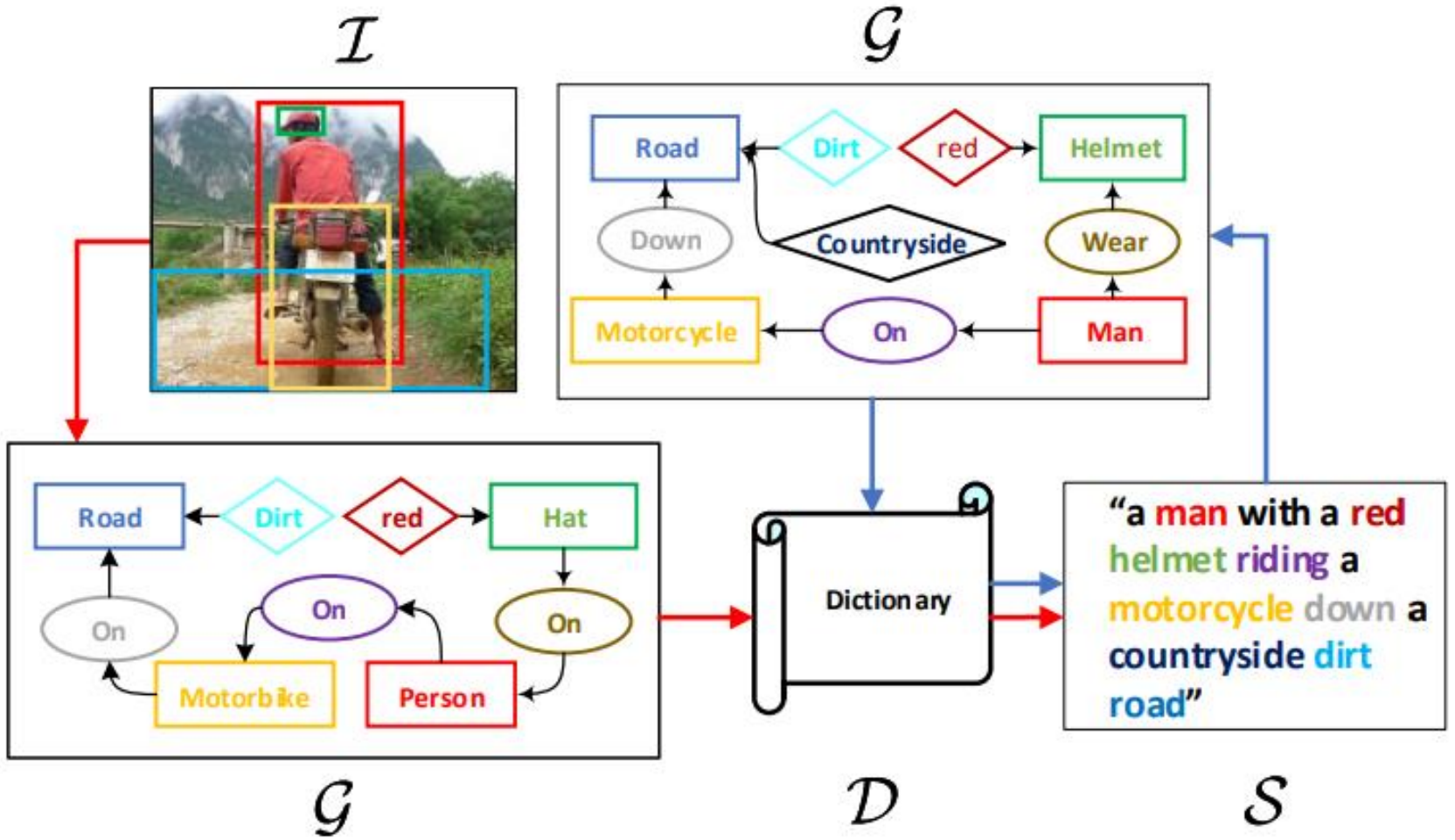


Auto-encoding scene graphs for image captioning. 2019 CVPR

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

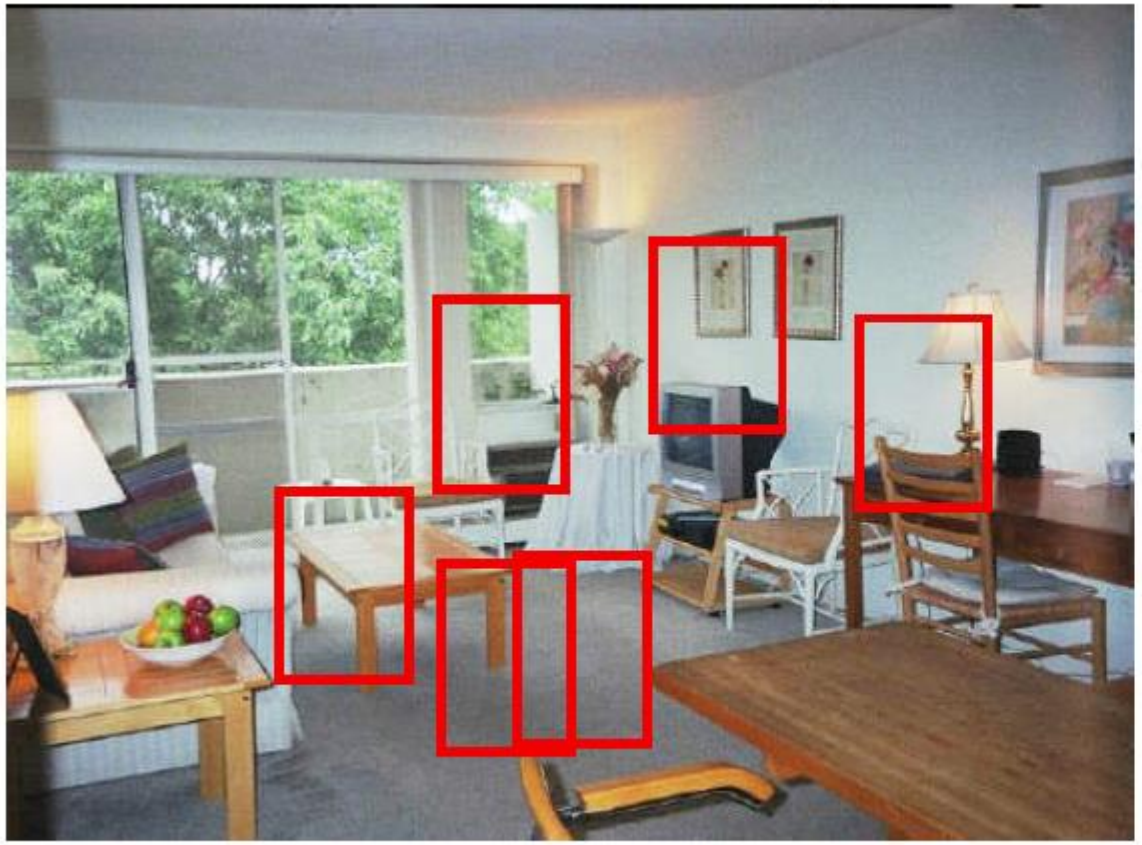
Contribution：

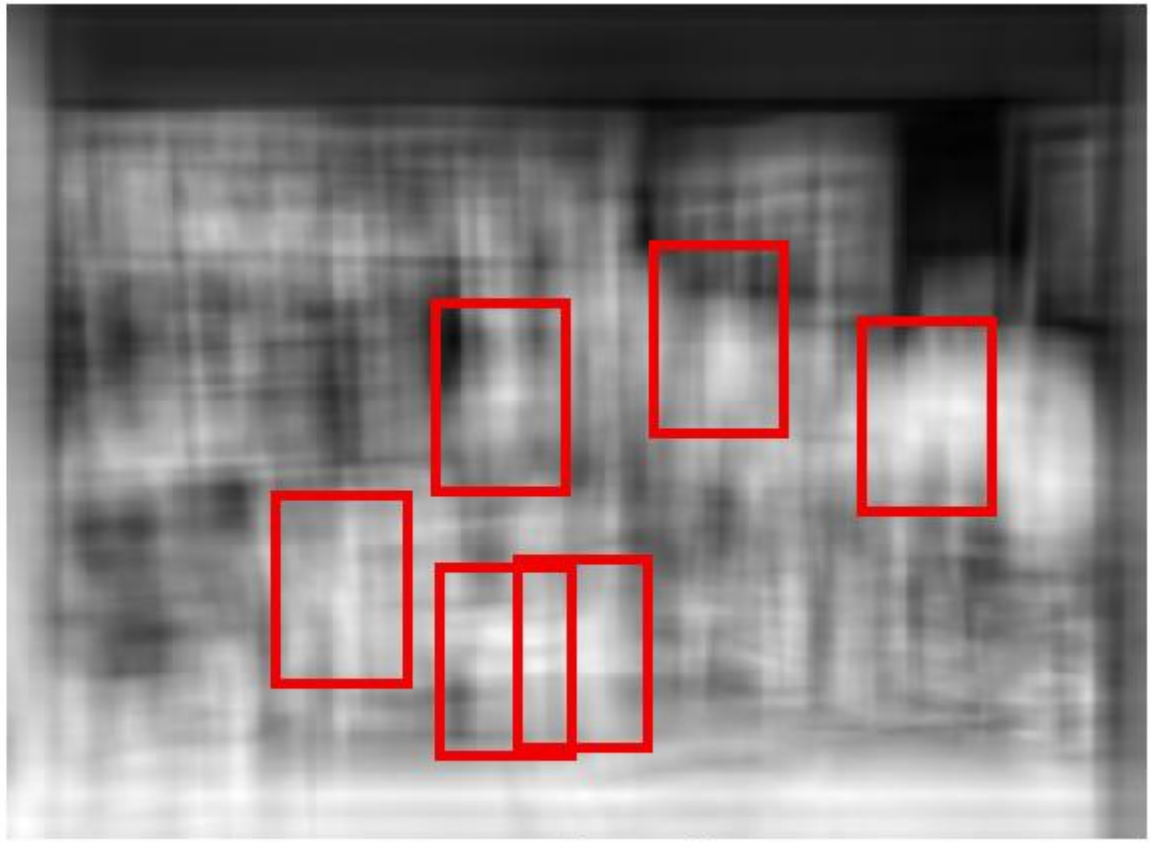
- SGAE（右半部分）
- 多模态图卷积网络
  （补充视觉特征）
- 共享词典D
  （指导生成）



Auto-encoding scene graphs for image captioning. 2019 CVPR

存在的问题：grounded captioning

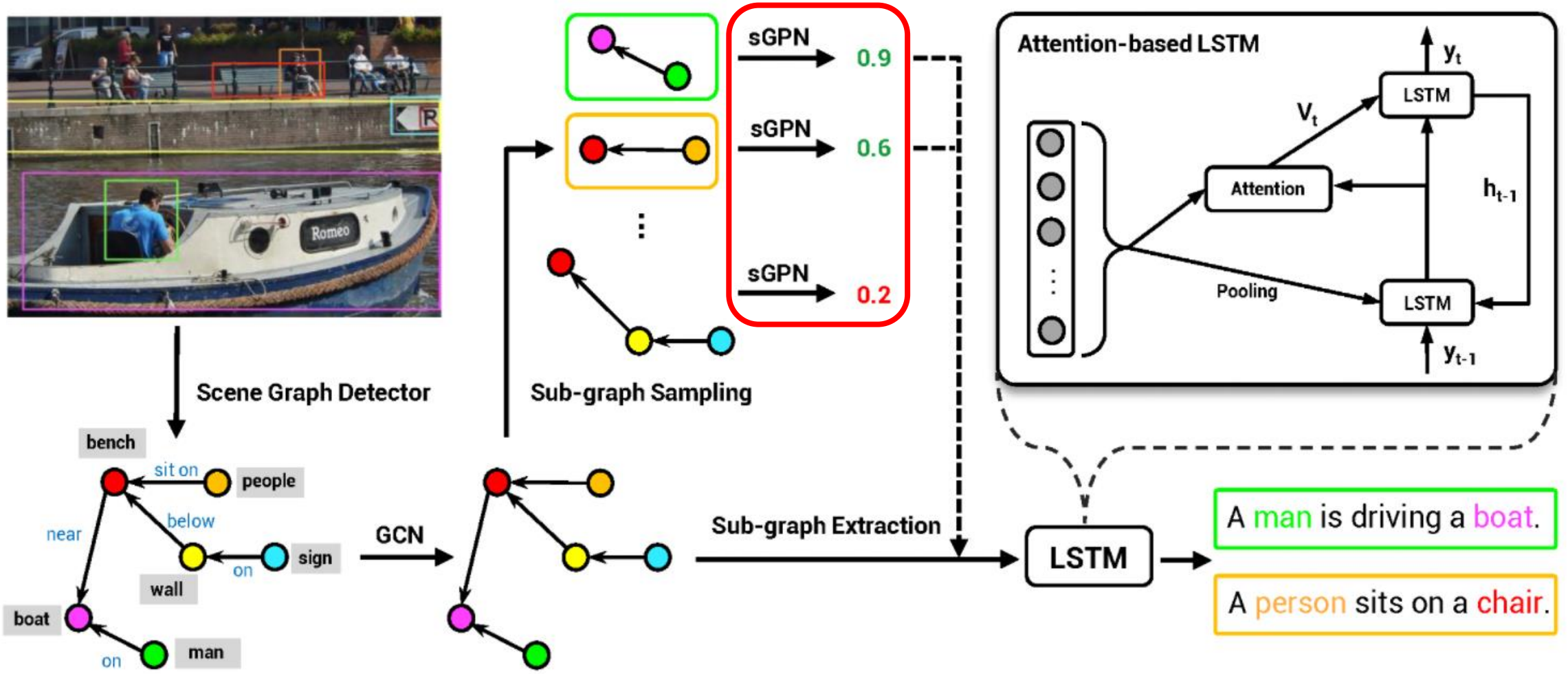




Pretty much garbage

图片源自北京理工大学付莹“计算机视觉”PPT

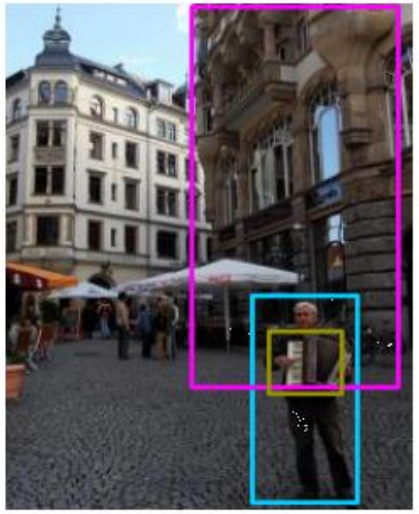
Grounded captioning



Comprehensive image captioning via scene graph decomposition. 2020 ECCV

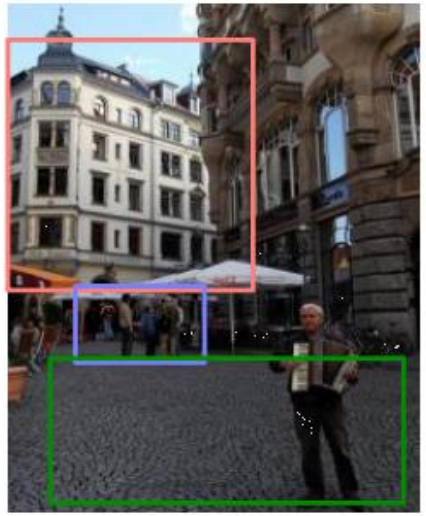
Grounded captioning

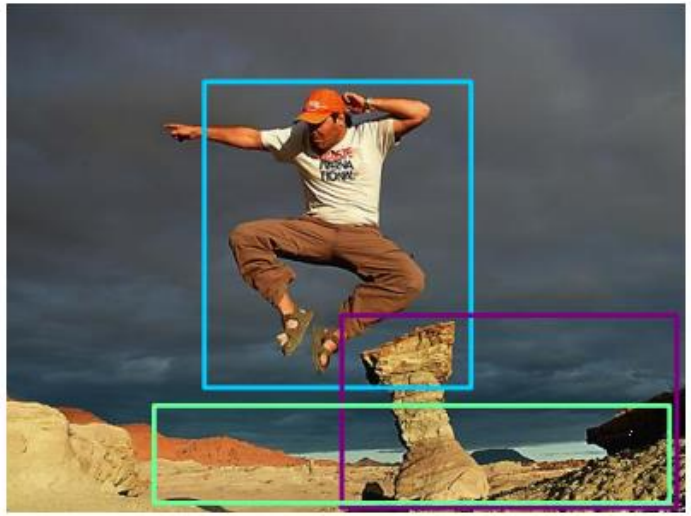
A man is playing a accordionin front of a building.



A man in a orange hat and brown pants is jumping off a rock.

People walking down a street in a city.





A man is jumping off a rock in a rocky area.

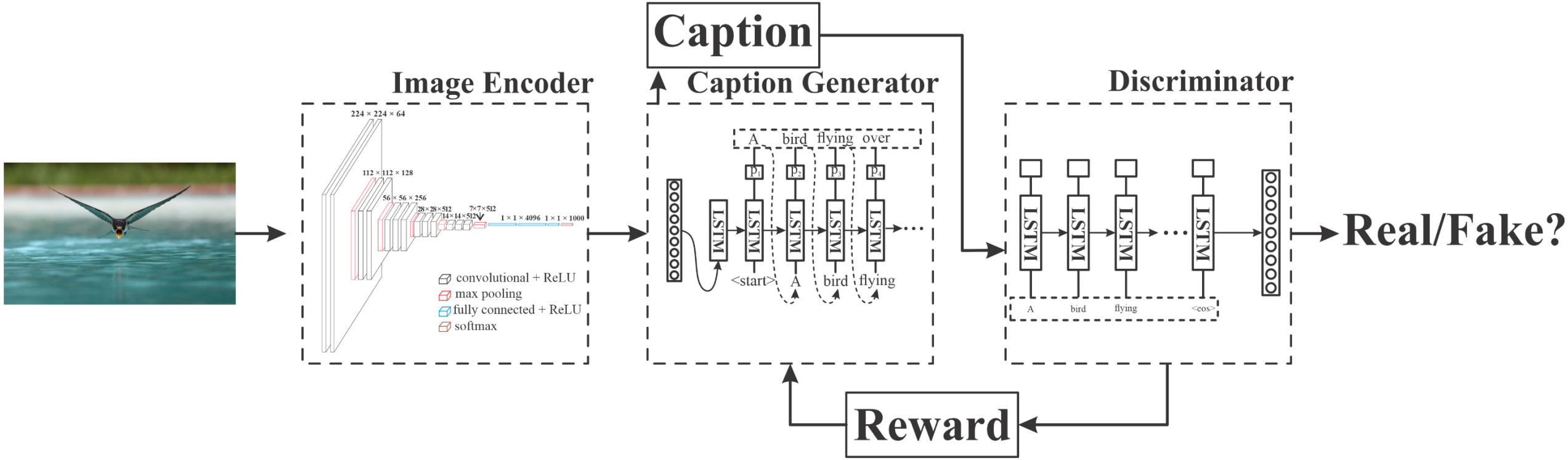
基于无监督+强化学习的方法

ImageNet： 1400W+张样例图片，27大类、2W+小类

Open Images： 900万张图像， 7800种类别

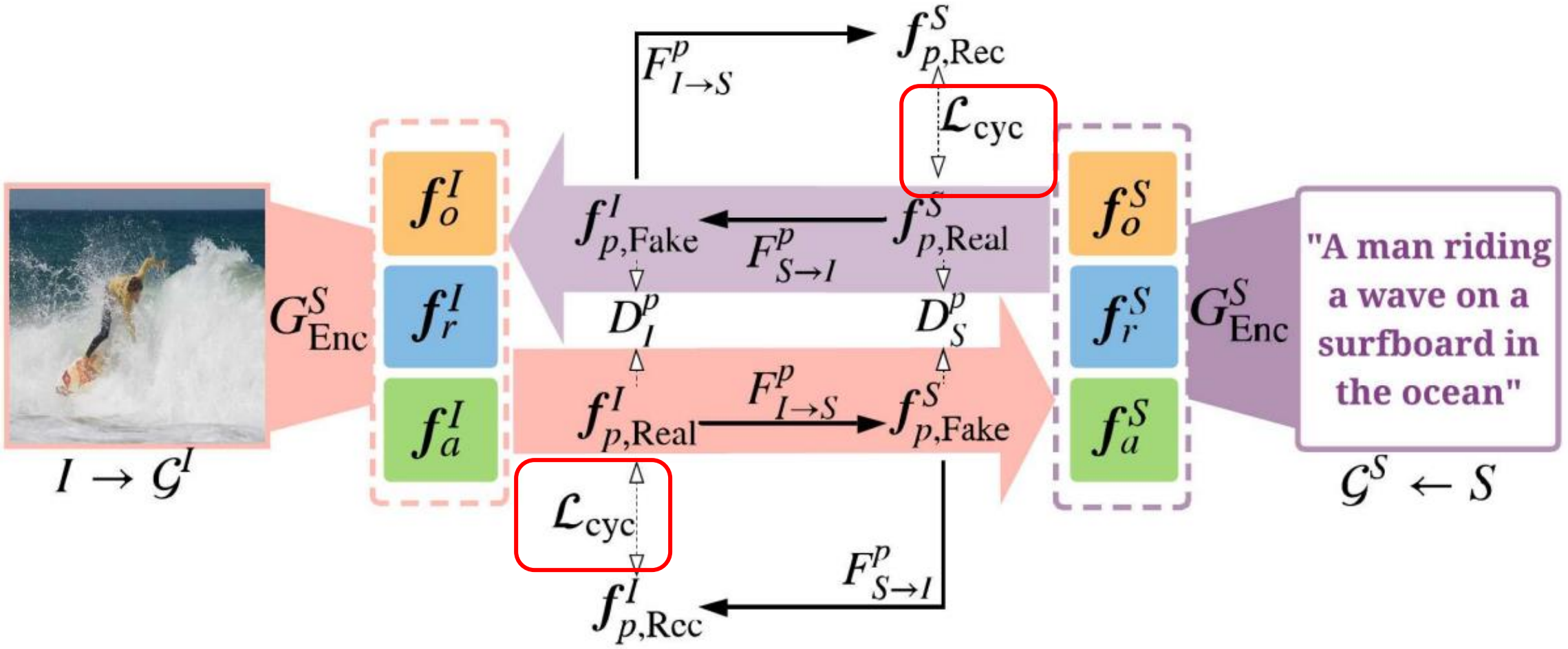
Microsoft COCO： 100类，训练11w+，验证5k和测试4w+，总计20w+ **（很小+多种语言标题+新类别图片 →耗时耗力）**

基于无监督+强化学习的一般流程



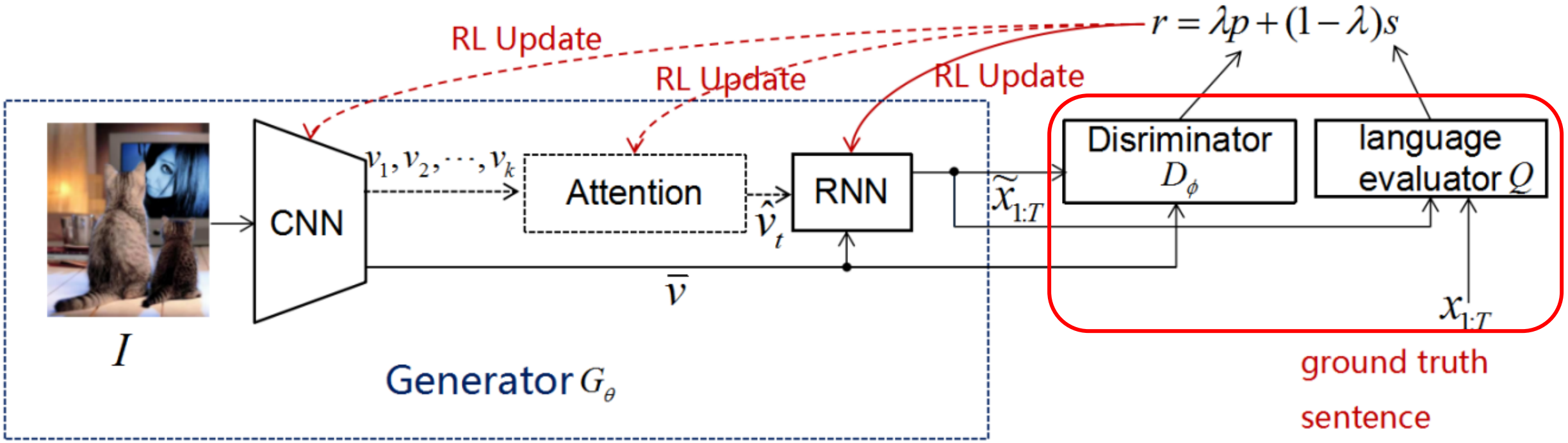
VGGNet 图像编码器，LSTM 字幕生成器，鉴别器也是LSTM，鉴别器对生成器进行相应的奖励，决定给定的标题是真实的还是由模型生成的。

训练时无监督（cycleGAN）



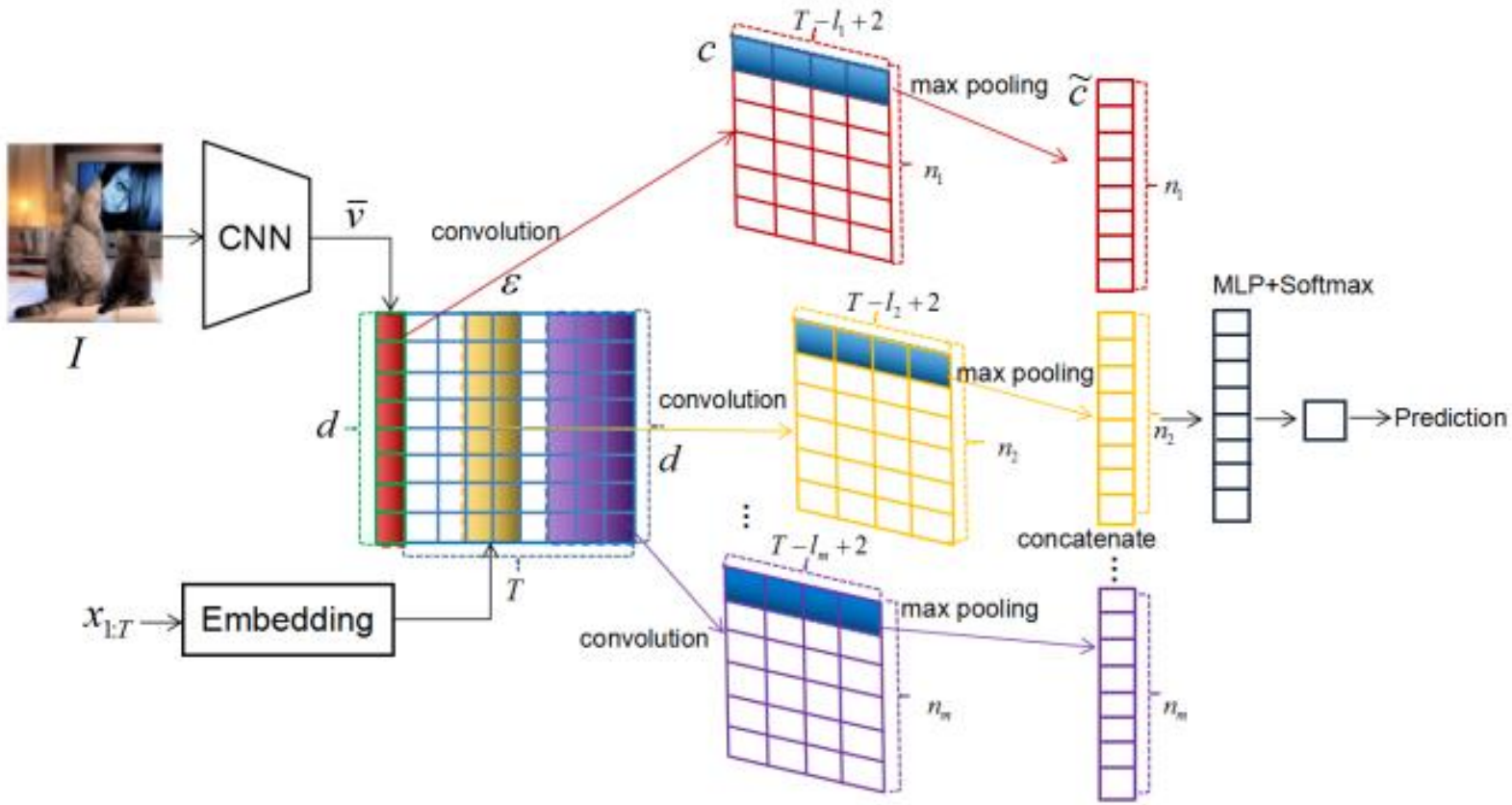
Unpaired image captioning via scene graph alignments. 2019 ICCV

条件生成对抗网



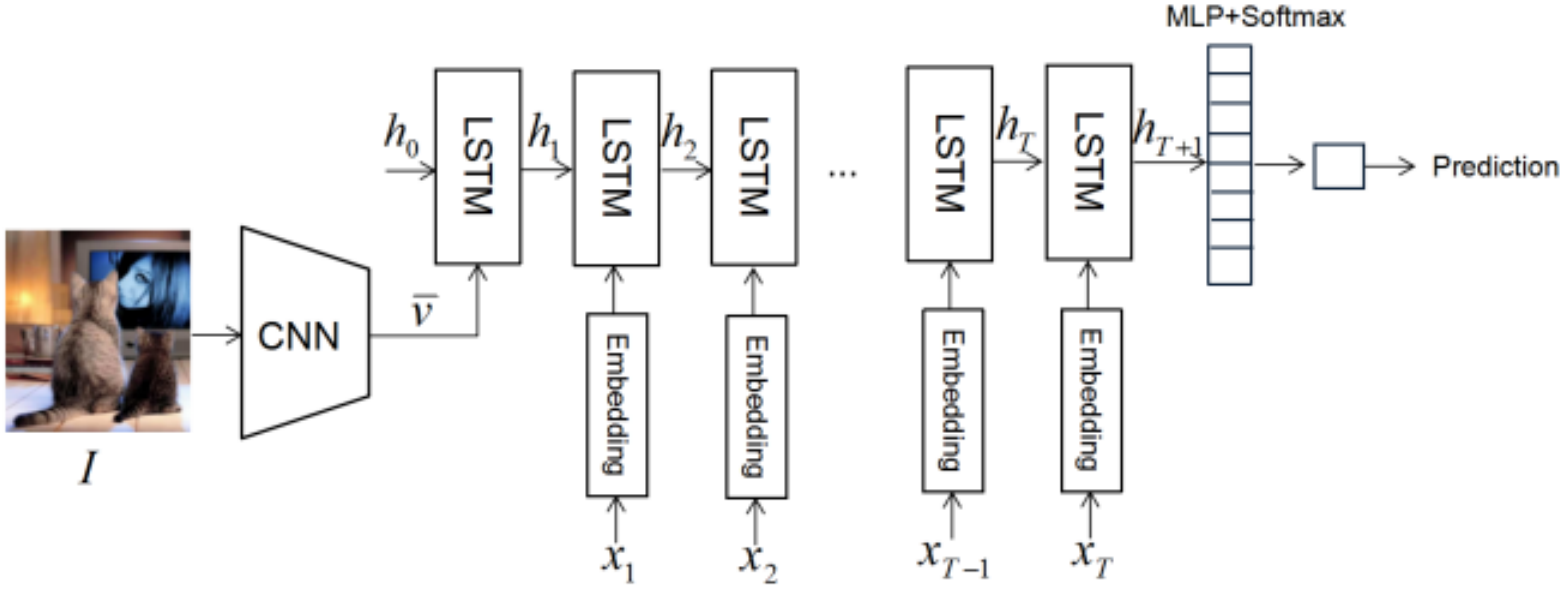
Improving image captioning with conditional generative adversarial nets. AAAI 2019

CNN判别（更优）



(a) CNN-based discriminator

RNN判别（更快）



(b) RNN-based discriminator

论文投稿趋势

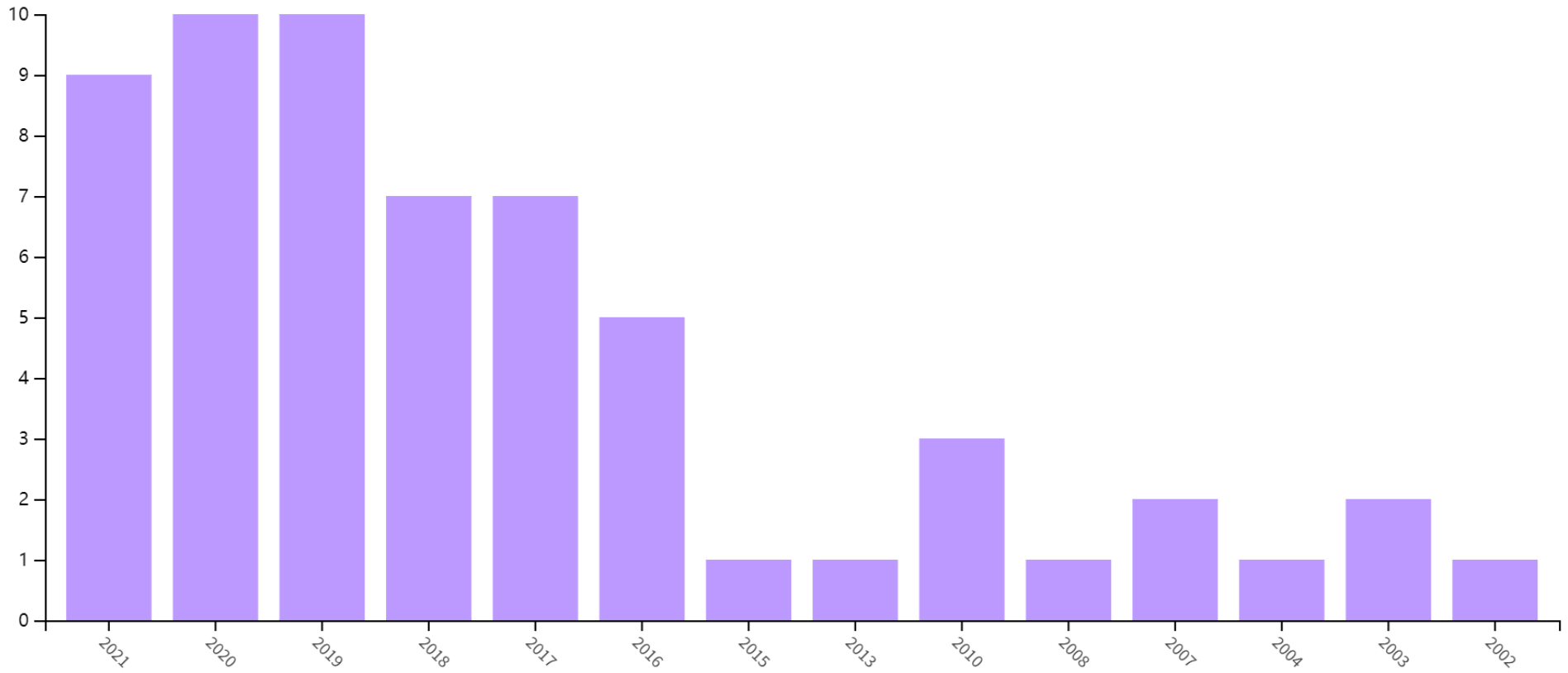


Image caption+无监督

数据来源：web of science

论文投稿趋势

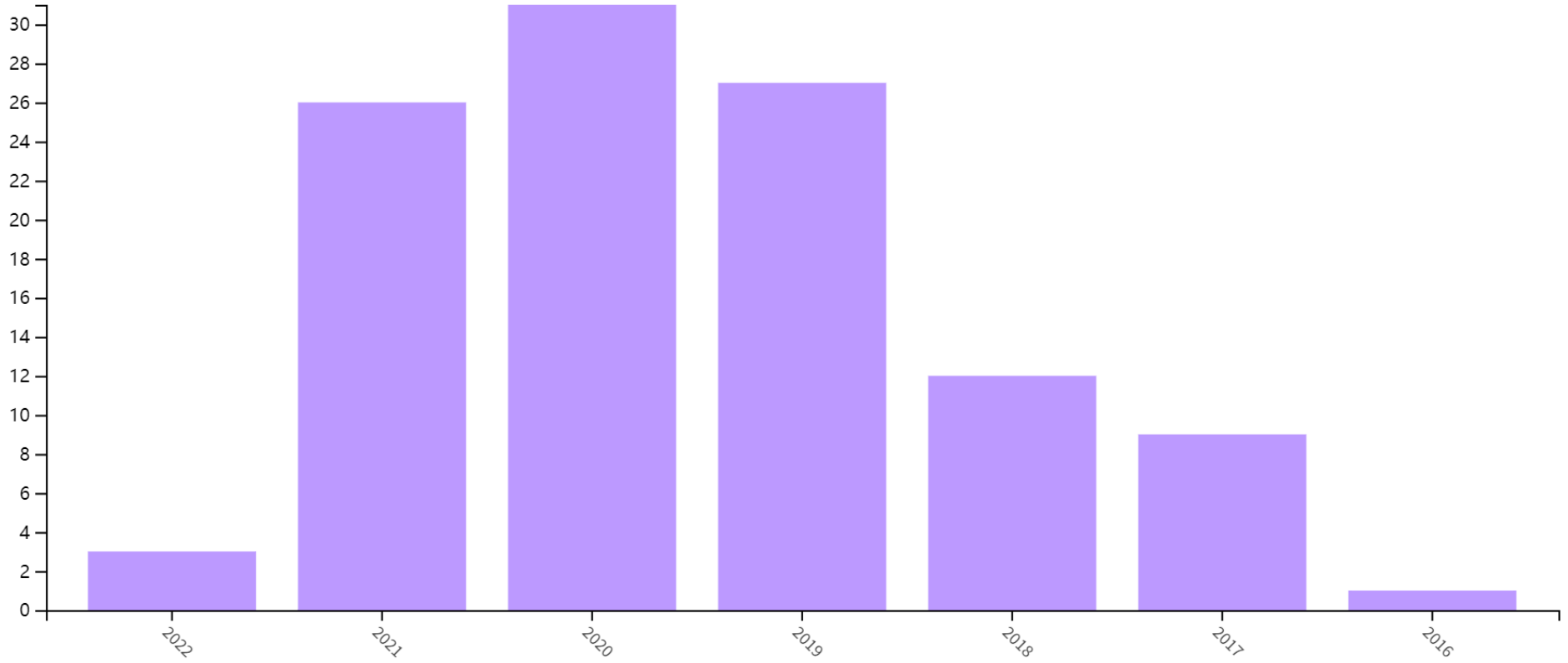


Image caption+强化学习

数据来源：web of science

论文投稿趋势

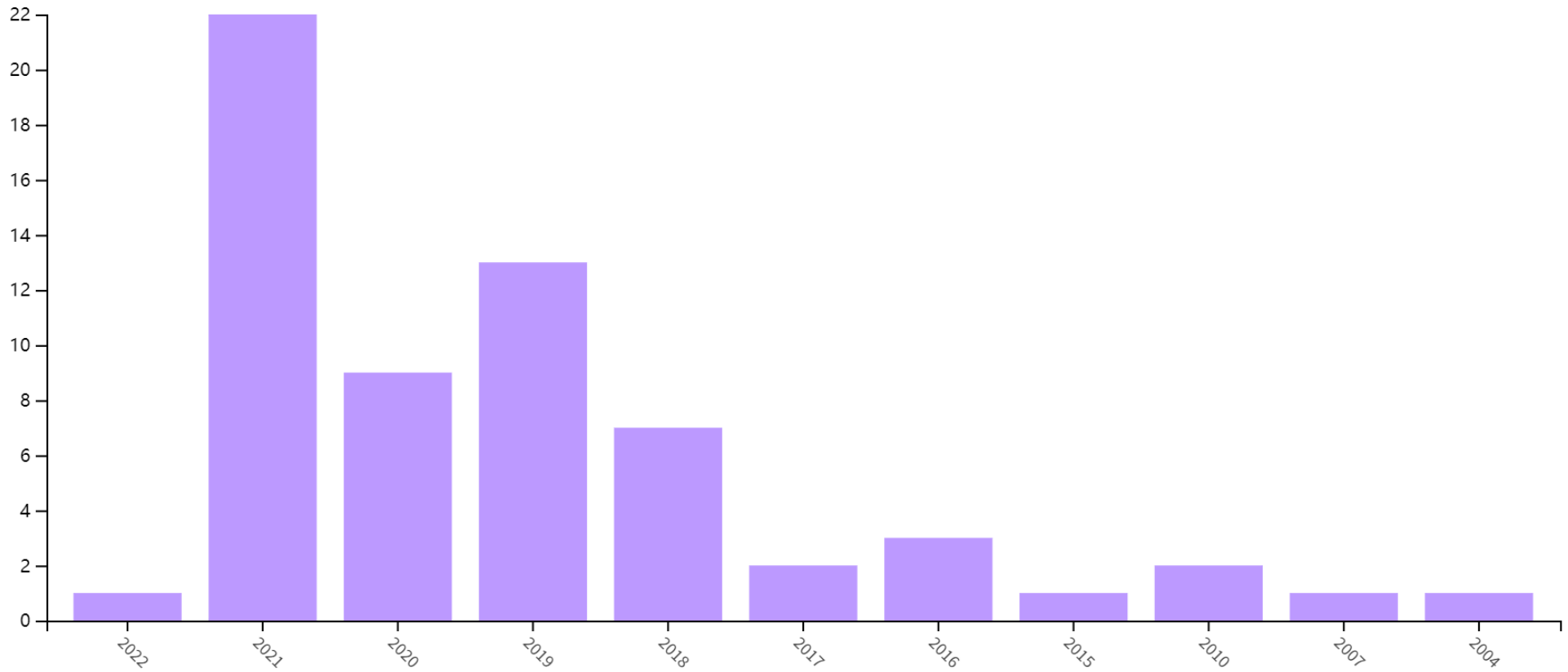


Image caption+场景图

数据来源：web of science

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

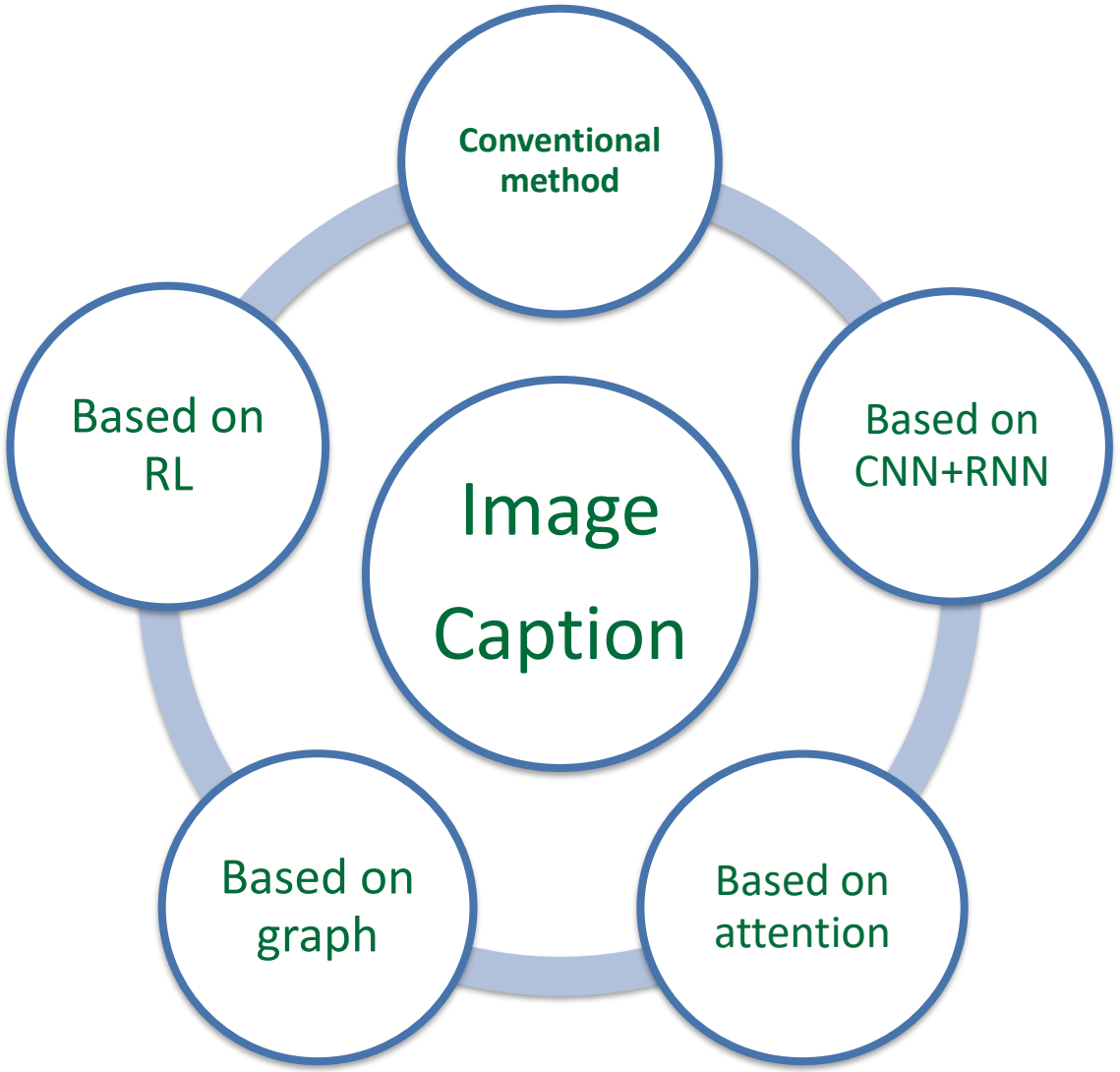
6

技术展示

汇报人：王中琦

当我们从技术复现角度去看待这个课题?

- 复现环境配置：
  - **Windows 11**
  - **Geforce RTX 3080*1**
  - **torch1.10.2+cu113**



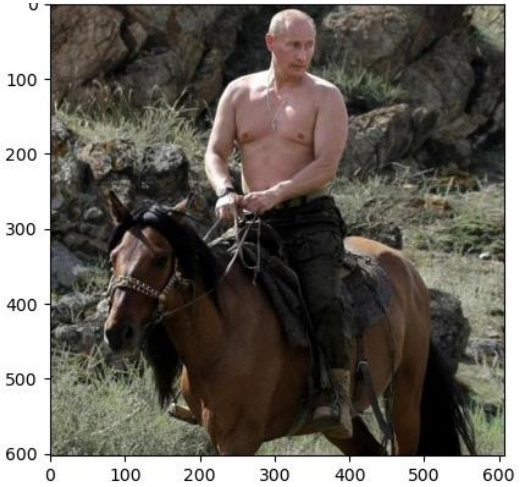
传统方法（2014年以前）

- 没有开源代码
- 国内关于这方面的技术博客几乎为0（百度、必应）
- 所创立的社区目前仍然很活跃
  - WordsEye(http://www.wordseye.com)
- **Github上以“Image Caption”为话题，2014年以前仅创建代码库114个**

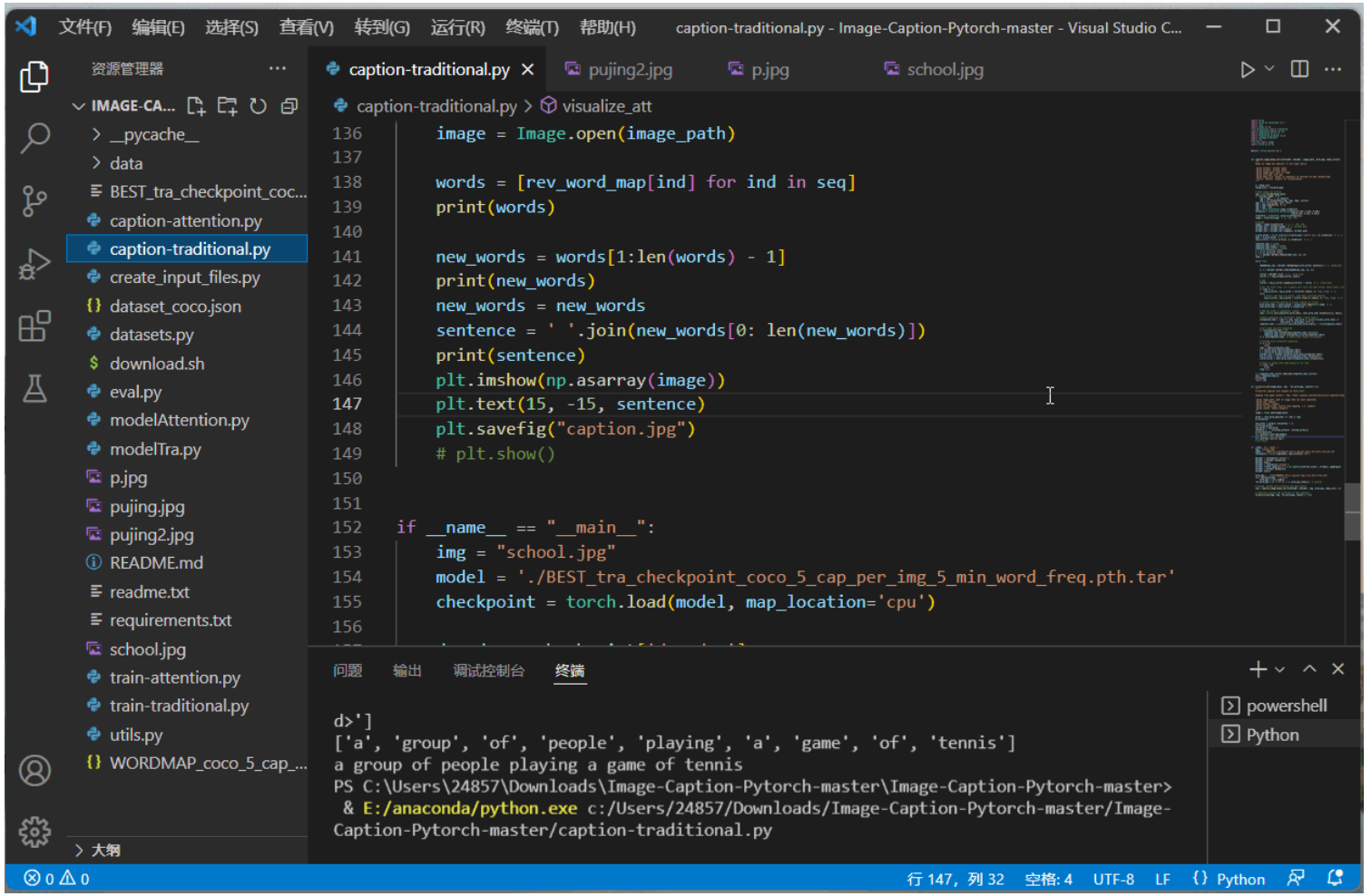
- “encoder-decoder“结构，代码量偏少
- 实验环境比较古老：python2.x、torch0.2、caffe、Lua……
- 最热门的几个开源代码库：
  - karpathy/neuraltalk2: Efficient Image Captioning code in Torch, runs on GPU (github.com)
  - karpathy/neuraltalk: NeuralTalk is a Python+numpy project for learning Multimodal Recurrent Neural Networks that describe images with sentences. (github.com)
  - KranthiGV/Pretrained-Show-and-Tell-model: This repository contains pretrained Show and Tell: A Neural Image Caption Generator implemented in Tensorflow. (github.com)

早期深度学习方法复现

a group of people standing next to each other



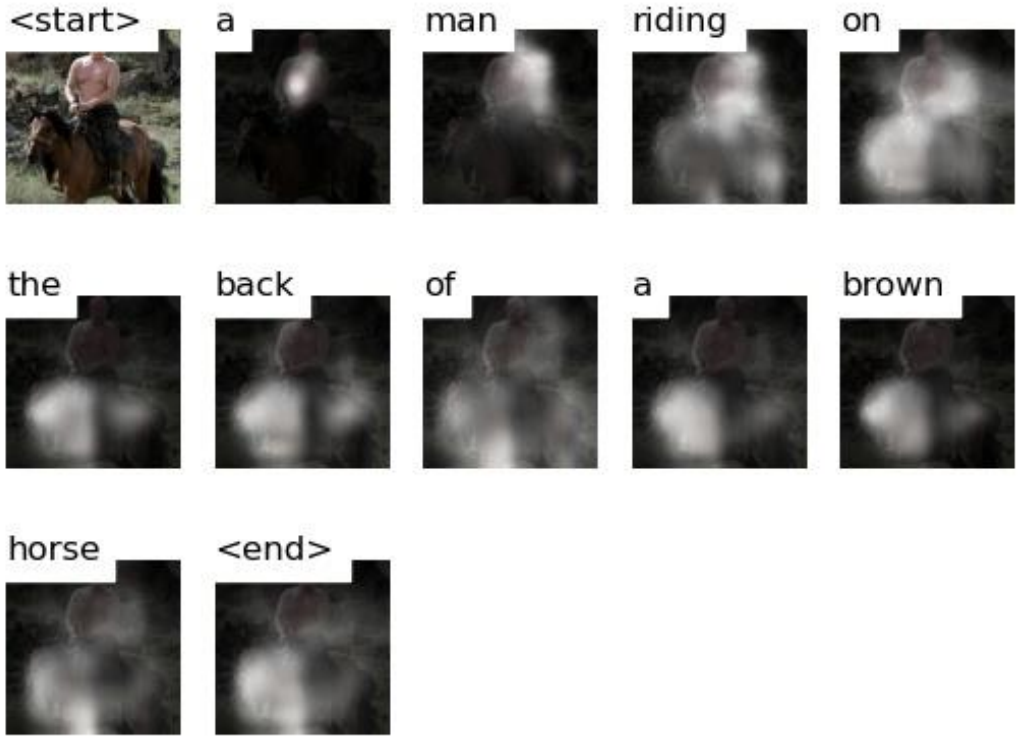
a man is standing next to a horse



Code from：https://github.com/tqxg2018/Image-Caption-Pytorch
Paper: "Show and Tell: A Neural Image Caption Generator." (CVPR 2014)

- 实验环境多样，python2.x/python3.x，torch/tenserflow
- 相关资源最丰富
- 项目维护得最好，常作为教程被引用
  - **pytorch**官方教程sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning: Show, Attend, and Tell | a PyTorch Tutorial to Image Captioning (github.com)
  - 陈云《深度学习框架**Pytorch**：入门与实践》pytorch-book/chapter10-image_caption at master · chenyuntc/pytorch-book (github.com)

基于attention机制复现



A men riding on the back of a brown horse

Code from: https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning
Paper: “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”(CVPR 2016)

基于attention机制复现

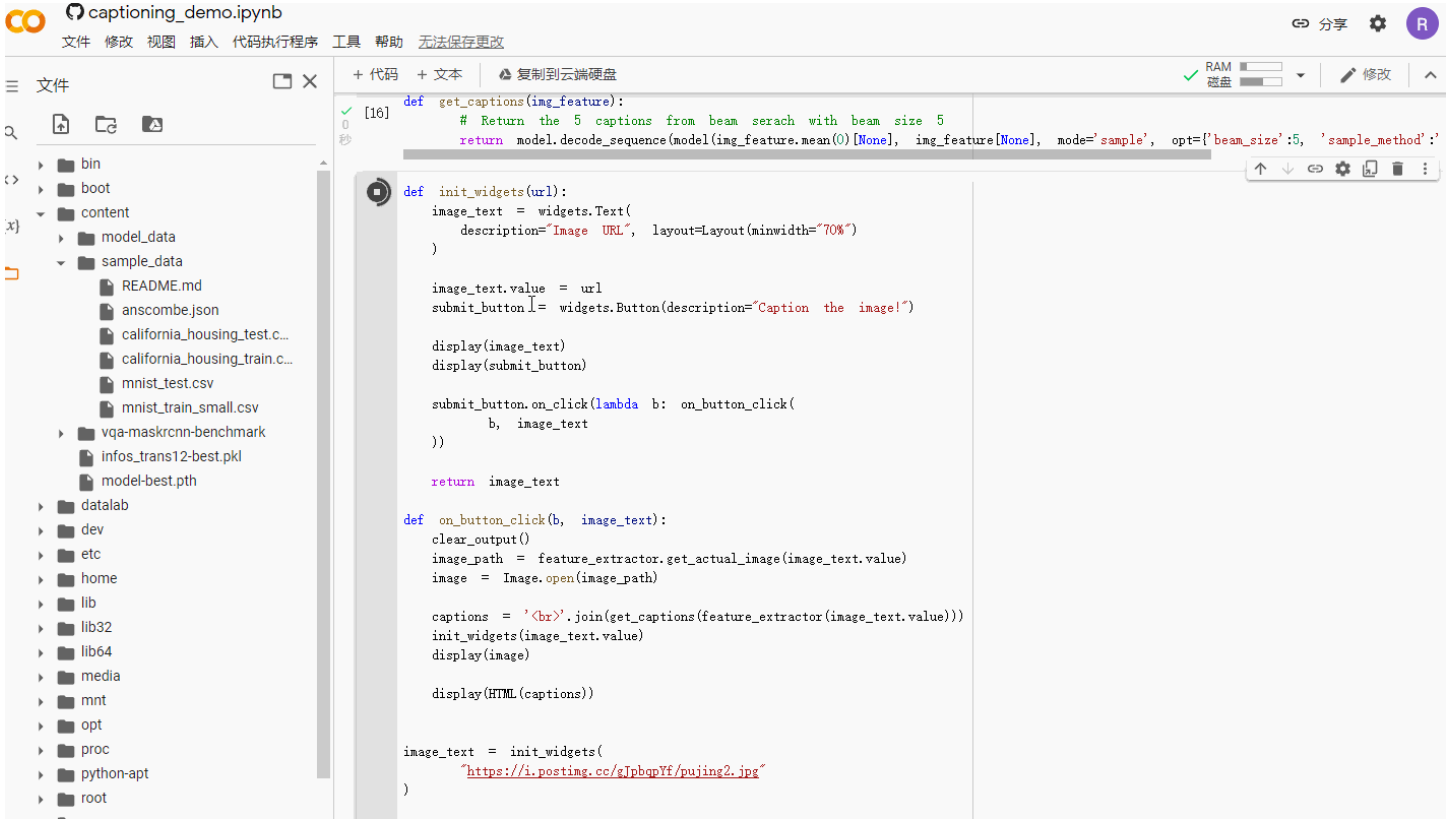
A close up of a person wearing a suit and tie

Code from: https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning
Paper: “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”(CVPR 2016)

- 近四年出现的方法，但开源数目相对少，且复现难度较大
- 环境基本定型：**pytorch1.3+、python3、GPU**
- 诞生了目前最具影响力的代码（祖传）
  - ruotianluo/self-critical.pytorch: Unofficial pytorch implementation for Self-critical Sequence Training for Image Captioning. and others. (github.com)
- 范式：**MSCOCO2014 + Bottom up feature + Karpathy's split**

基于强化学习复现

a man riding on the back of a brown horse
a man riding a horse in the grass
a man riding a brown horse in a field
a man riding a horse in a grassy field
a man riding a brown horse through a field



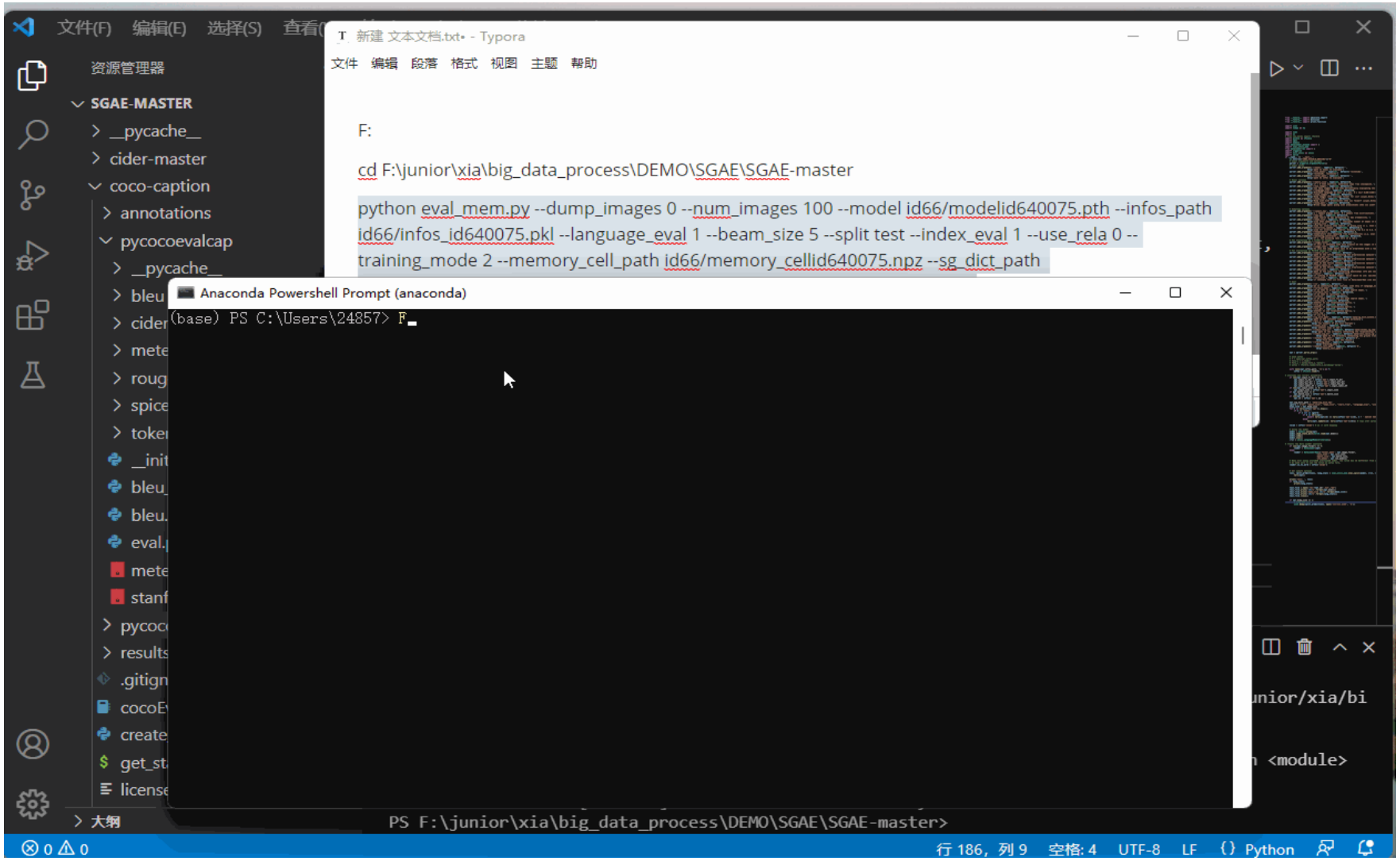
By Google colab

Code from: https://github.com/ruotianluo/self-critical.pytorch
Paper: "Self-critical Sequence Training for Image Captioning" (CVPR 2017)

- 同样，近年诞生的方法，开源生态逐渐丰富，对环境配置要求较高，复现难度较大

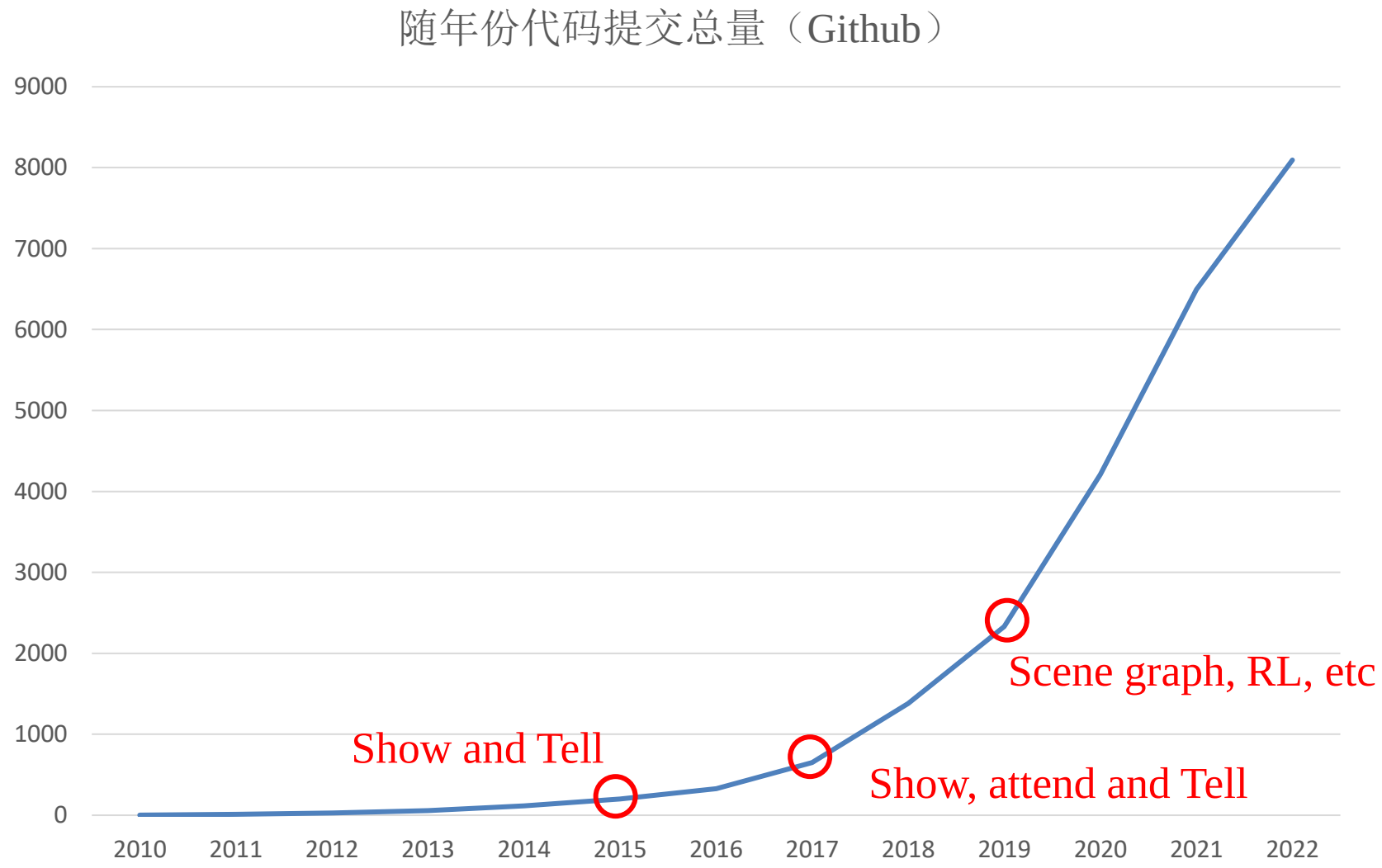
- 维护非常好的代码库：
  - microsoft/scene_graph_benchmark: image scene graph generation benchmark (github.com)
  - google/sg2im: Code for "Image Generation from Scene Graphs", Johnson et al, CVPR 2018 (github.com)
  - KaihuaTang/Scene-Graph-Benchmark.pytorch: A new codebase for popular Scene Graph Generation methods (2020). Visualization & Scene Graph Extraction on custom images/datasets are provided. It's also a PyTorch implementation of paper “Unbiased Scene Graph Generation from Biased Training CVPR 2020” (github.com)

基于场景图



Code from: https://github.com/yangxuntu/SGAE
Paper: "Auto-Encoding Scene Graphs for Image Captioning" (CVPR 2019)

Image Caption提交代码总览



现场演示

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

谢谢观看

答辩人：刘云皓，李桐，邱小尧，
刘天锐，王中琦
时间：2021-3-8