复旦大學
FUDAN UNIVERSITY

自然语言处理国际前沿动态综述 之 多模态

基于视觉的跨模态文本生成

复旦大学 魏忠钰

http://www.sdspeople.fudan.edu.cn/zywei/

第十八届中国计算语言学大会（CCL 2019）
2019年10月20日，云南，昆明

目录

语言 - 视觉的跨模态任务

- 图像的描述自动生成 (Image Captioning)
- 图像集的故事自动生成 (Visual Storytelling)
- 图像的文本自动问答 (Visual Question Answering)
- 图像的对话自动生成 (Visual Dialogue)

- 视觉导航任务 (Visual Navigation)
- 文本到图像的自动生成 (Text-to-Image Synthesis)

研究问题

- 视觉的语义表示
- 视觉-语言的跨模态对齐
- 文本的生成方法

核心问题

- 长文本生成（故事生成）
- 多样化文本生成（一对多的映射）
- 语义控制的文本生成（情绪、个性）

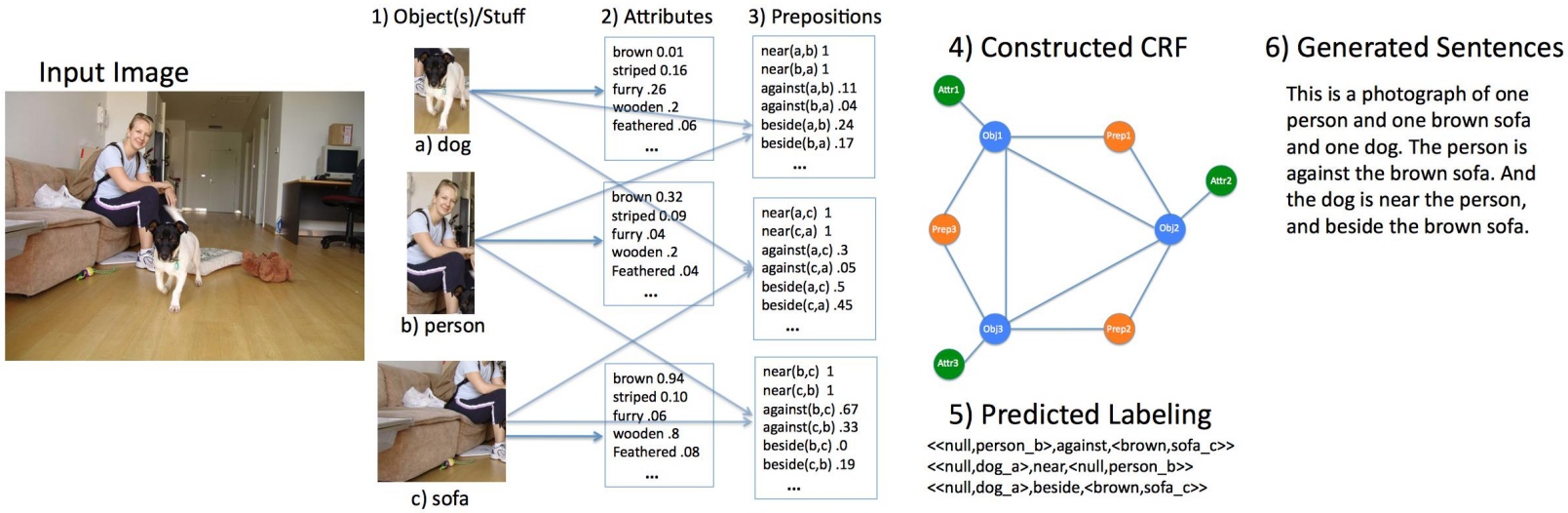
场景相关问题

相关解决方案

- 视觉的语义表示
  - 特征向量（全局、局部）
  - 实体、实体关系
- 文本的生成方法
  - 基于检索的方法
  - 基于模板填充的方法
  - 基于生成的方法（语言模型、神经网络）
- 视觉-语言的跨模态对齐
  - 中间表示：词语，三元组，语义概念，场景图
  - 对齐机制：注意力机制

基于模板填充的生成方法

- 从图片中识别**实体，属性，以及关系**
- 将识别的实体和关系通过**模板填充**的方式产生句子

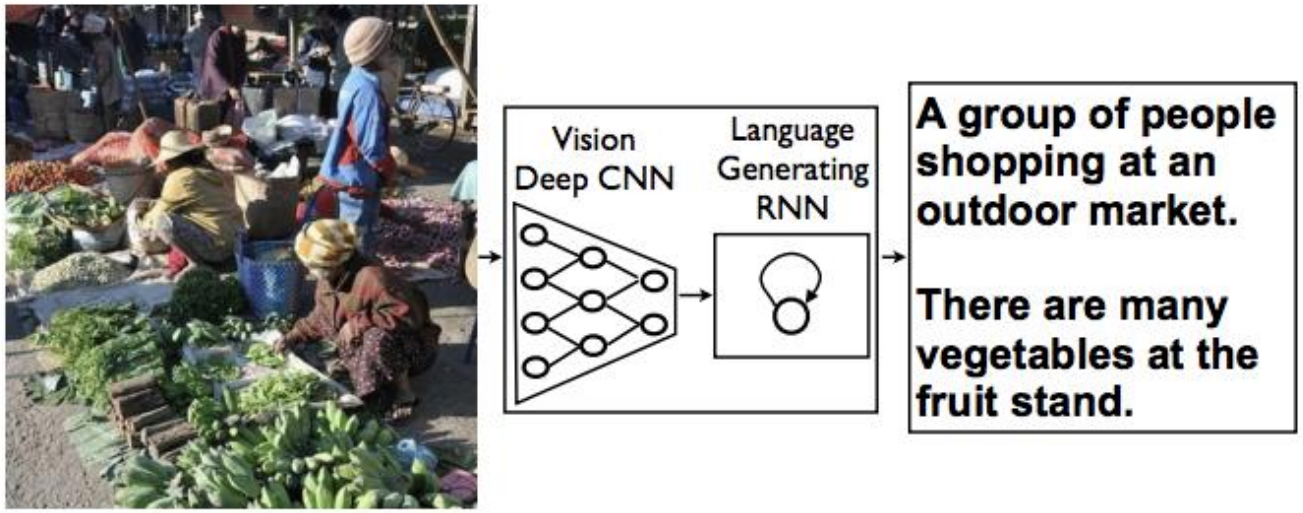


- 缺点：生成句子样式单一；受限于物体识别的种类

Baby talk: Understanding and generating simple image descriptions, CVPR'11

基于神经网络的端到端生成方法

- 利用编码器-解码器框架
  - 编码器：卷积神经网络（CNN）抽取**图像全局特征**
  - 解码器：长短时记忆网路（LSTM）**进行逐字的文本生成**

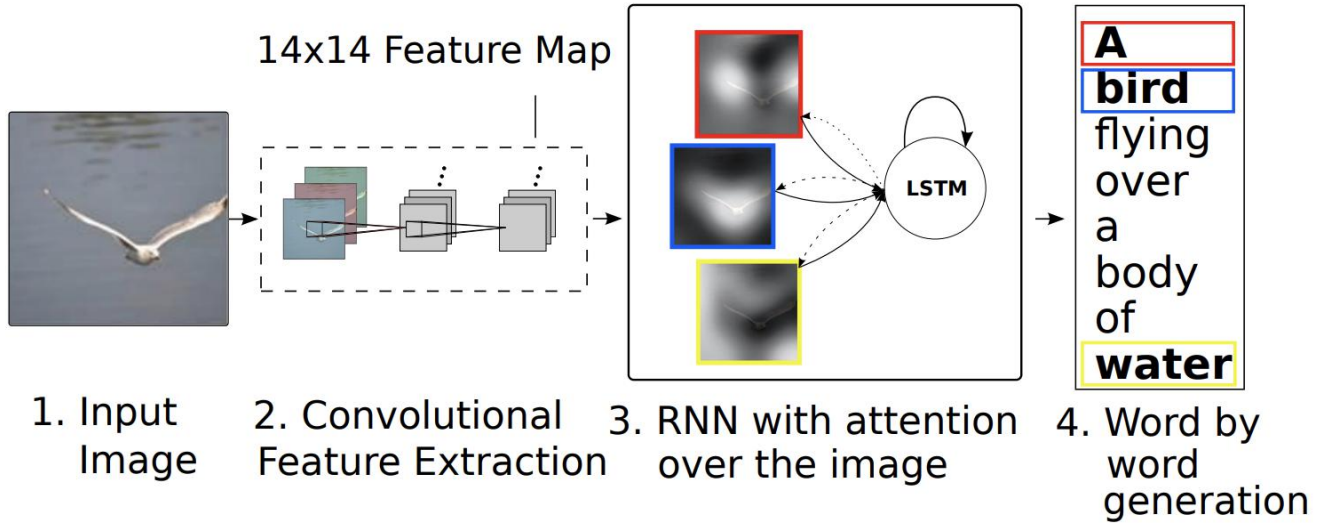


- 缺点: 图片和文本的对齐完全依赖于图片的全局特征。

Show and tell: A neural image caption generation, CVPR'15

关注图像局部特征的端到端生成方法

- 引入**图像的局部信息**（卷积神经网络的中间层特征）
- 通过**注意力机制**建立局部特征和文字生成过程的联系

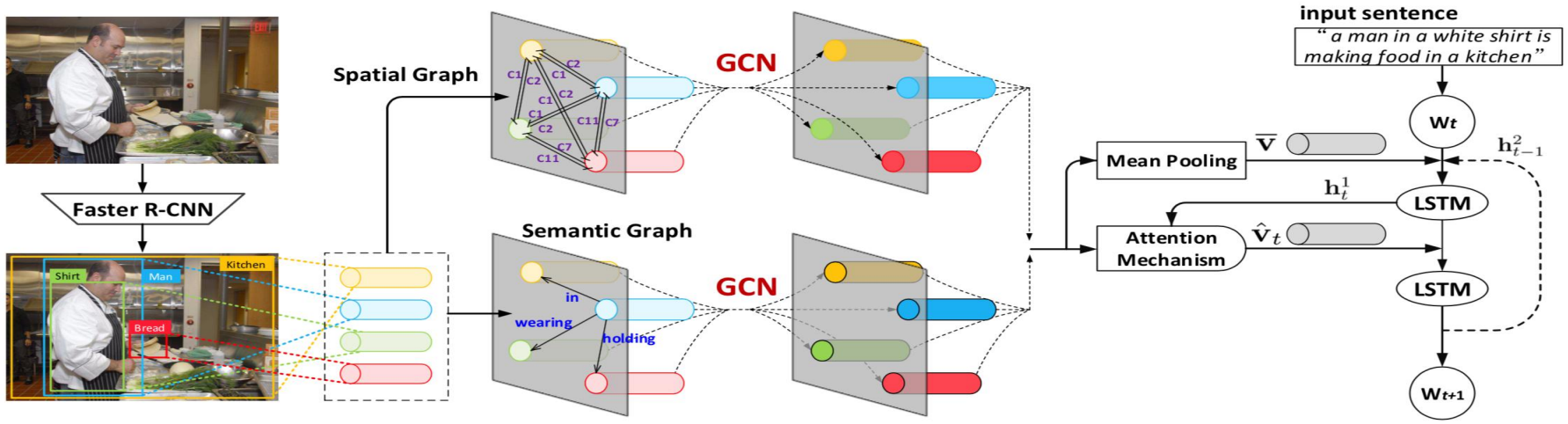


- 限制: 局部信息是预先定义的，没有语义内涵；局部信息的表示相互独立，没有考虑互相之间的联系。

Show, attend and tell: Neural image caption generation with visual attention, ICML'15

建模视觉局部特征关系的生成方法

- 构建图像的**场景图用以建模实体之间的关系**
- 利用图卷积神经网络对场景图中的实体和关系进行表示更新

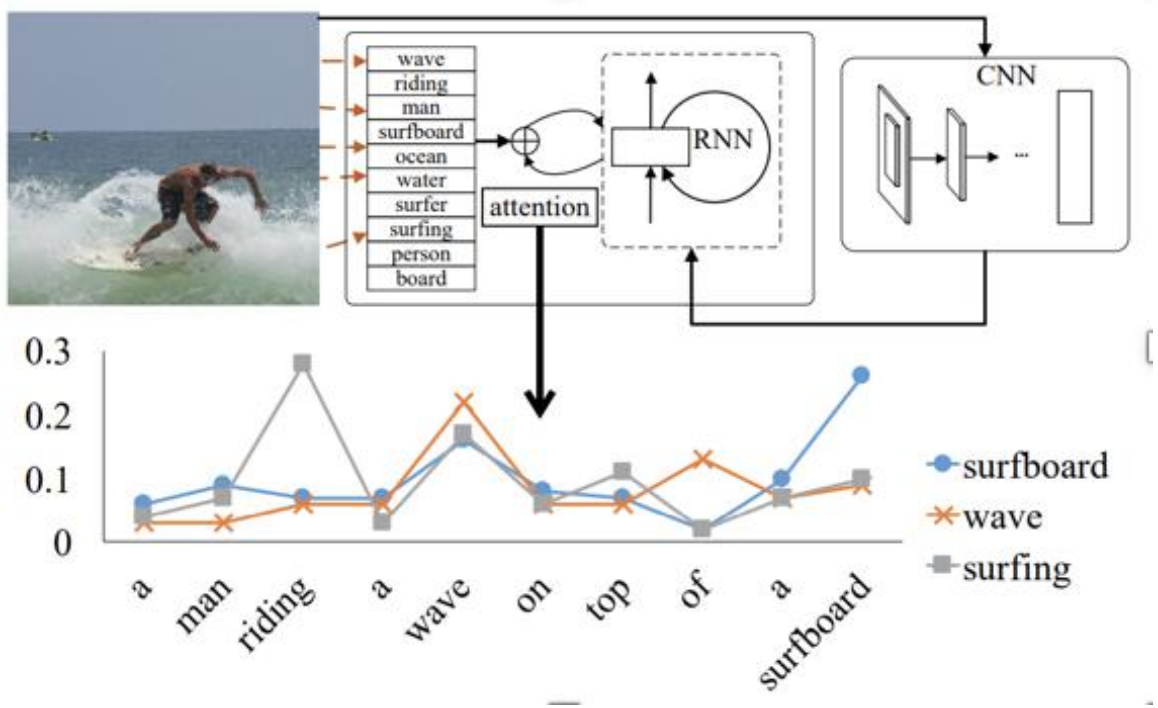


- 限制: 视觉特征建模与文本生成仅仅依靠自顶向下的注意力机制联系，缺少显示的语义关联。

Exploring visual relationship for image captioning, ECCV'18

引入图像语义概念的端到端生成方法

- 引入**语义概念**作为中间表示
- 在解码过程中利用**注意力机制**与语义概念进行对齐

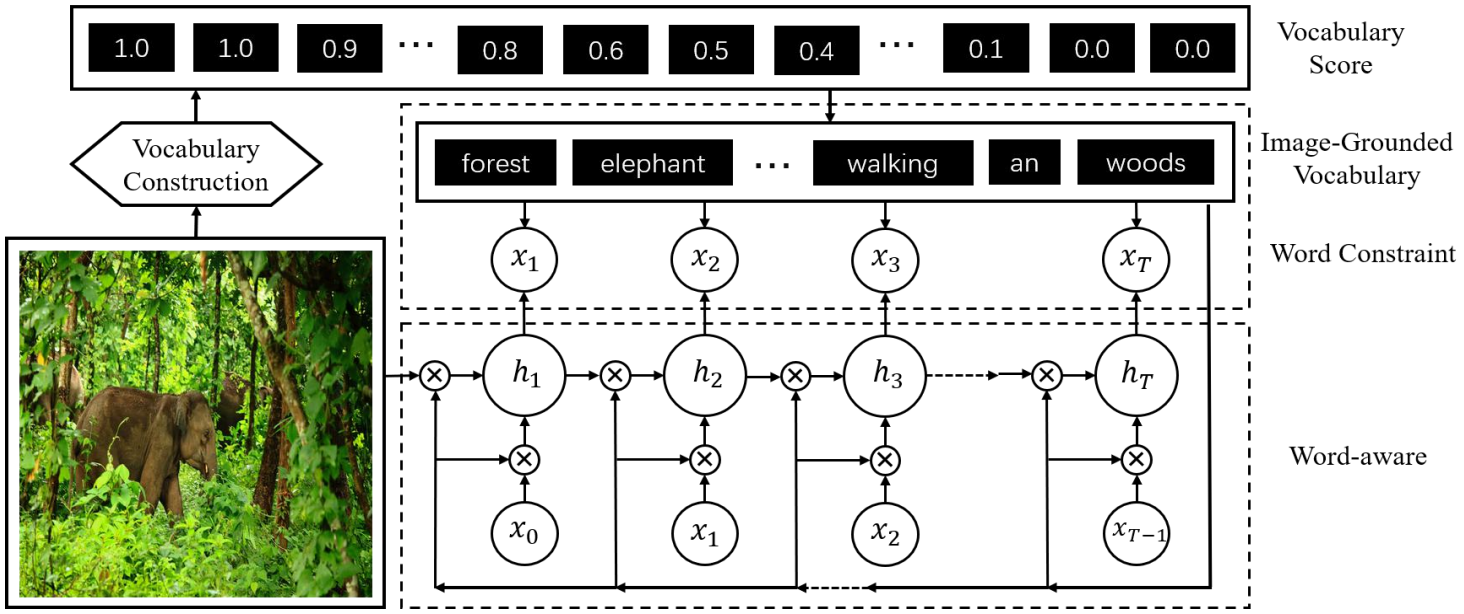


- 限制: 语义信息仅作为外部特征指导文本生成

Image Captioning with Semantic Attention, CVPR'16

基于图像印证的词汇表的生成方法

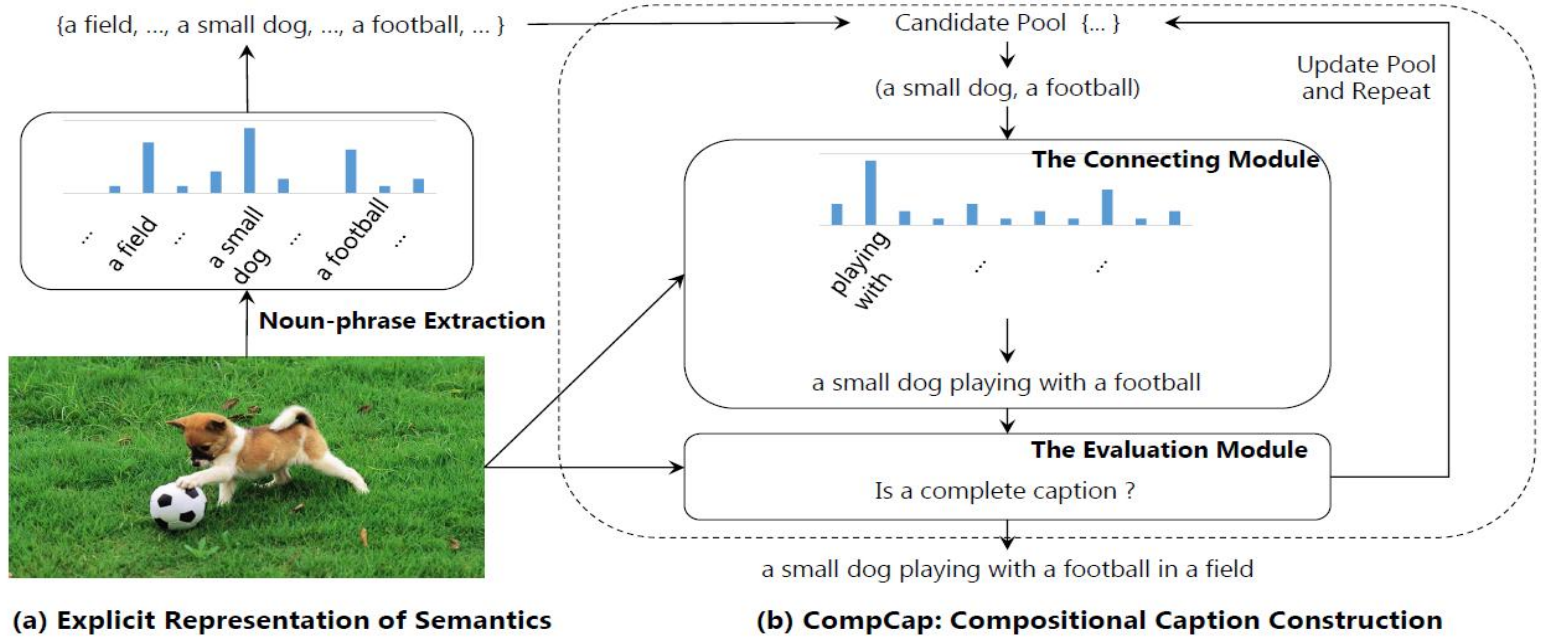
- 在图像中学习**图像印证的词汇表**
- 基于图像印证的词汇表进行文本生成



Bridging by Word: Image-Grounded Vocabulary Construction for Visual Captioning, ACL'19

基于短语拼接的生成方法

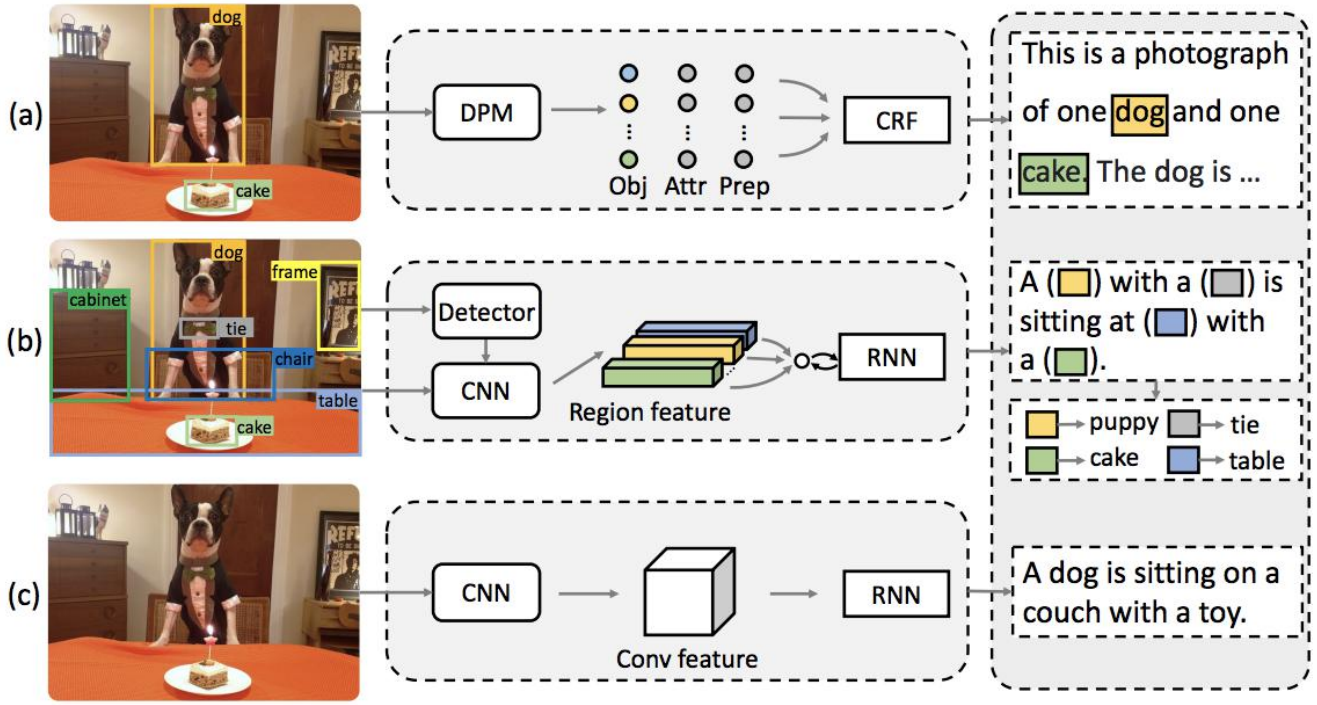
- 在图像中抽取**短语作为图像语义表征**
- 利用预先学习的**短语拼接**模块进行文本生成



A Neural Compositional Paradigm for Image Captioning, NIPS'18

结合模板与循环神经网络的生成方法

- 在图像中抽取实体，属性信息作为**图像词汇**
- 利用端到端生成模型，产生**表达模板**
- 使用槽填充的方法将图像词汇插入表达模板



Neural Baby Talk, CVPR'18

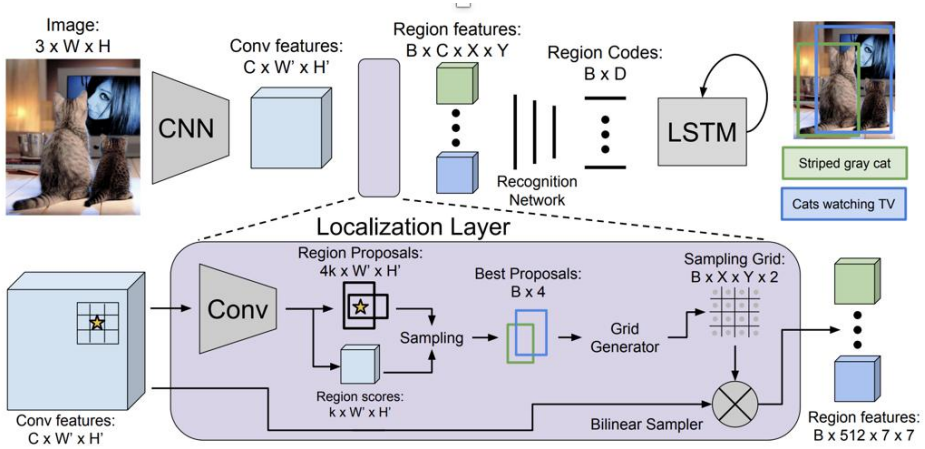
基于视觉的文本生成小结

图像表示	文本生成	跨模态对齐	训练方式	相关文献
实体	模板填充	三元组	pipeline	Baby talk…, CVPR'11 Midge…, EACL'12
实体	模板填充+ 神经生成模型	三元组	pipeline	Neural Baby Talk, CVPR'18
实体+ 实体关联	神经生成模型	场景图	端到端	Visual relationship, ECCV'18
全局特征	神经生成模型	图像特征	端到端	Show and tell, CVPR'15
全局+ 局部特征	神经生成模型	图像特征	端到端	Show, attend and tell,ICML'15
全局	神经生成模型	图像特征+ 语义概念	端到端	Semantic Attention, CVPR'16
全局	短语拼接	语义概念	pipeline	Compositional Paradigm, NIPS'18

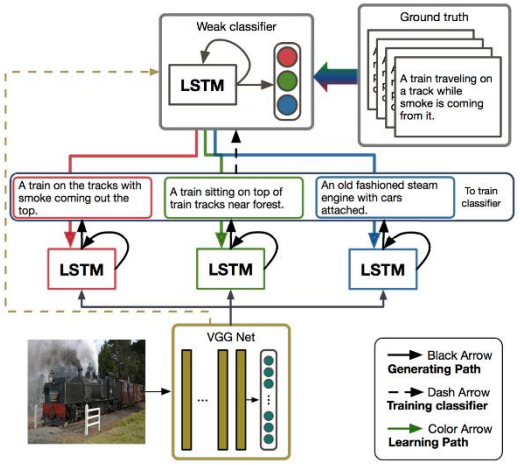
- 图像特征
  - 全局 → 局部
  - 实体 → 实体关联
- 文本生成
  - 模板→生成模型→结合
  - 自回归 → 拼接
- 跨模态对齐
  - 三元组 → 场景图
  - 图像特征 → 语义概念

多样化的视觉文本生成

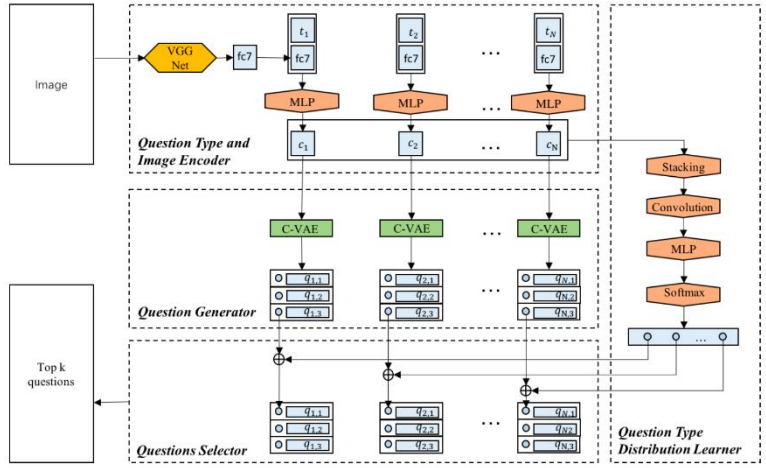
- 编码过程：利用细粒度的图像信息
- 解码过程：利用多个解码器
- 任务信息：利用特定任务信息，如问题类型



Densecap: Fully convolutional localization networks for dense captioning, CVPR'16



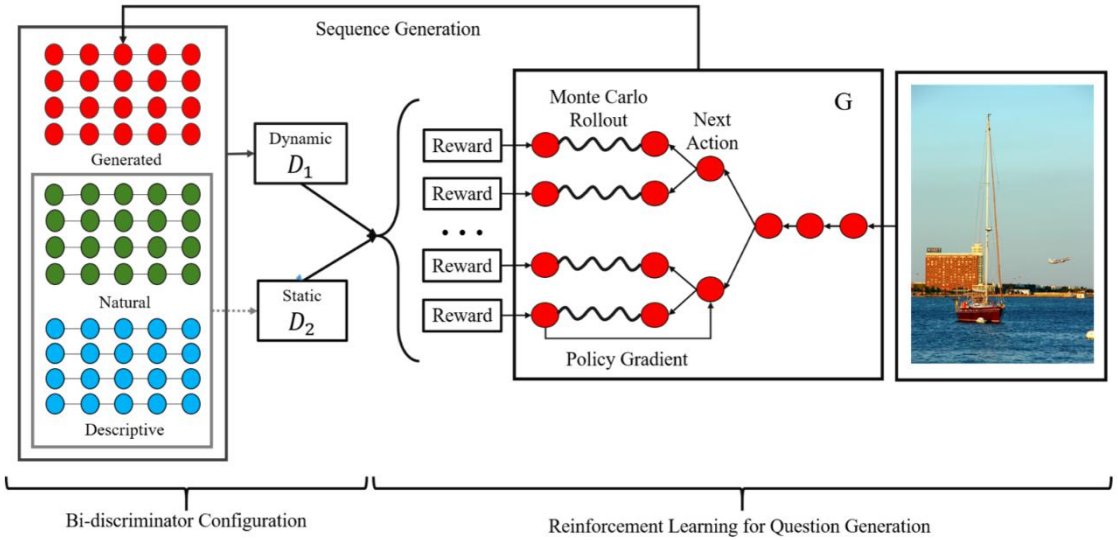
Diverse Image Captioning via GroupTalk, IJCAI'16

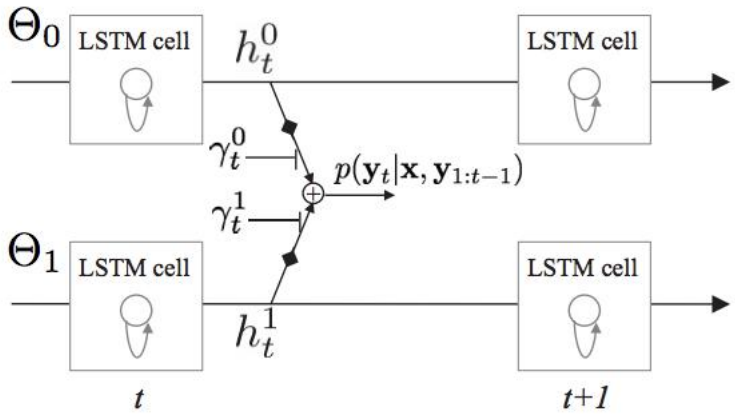


A Question Type Driven Framework to Diversify Visual Question Generation, IJCAI'18

带语义控制的视觉文本生成

- 将额外的语义作为附加特征与图像信息进行混合
- 在生成对抗框架下，构建额外的判别器引导特定的语义
- 多为数据集合驱动的研究



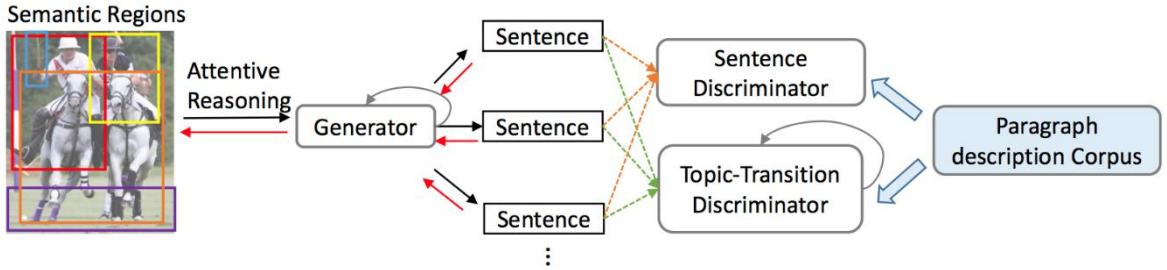


A Reinforcement Learning Framework for Natural Question Generation using Bi-discriminators , COLING'18

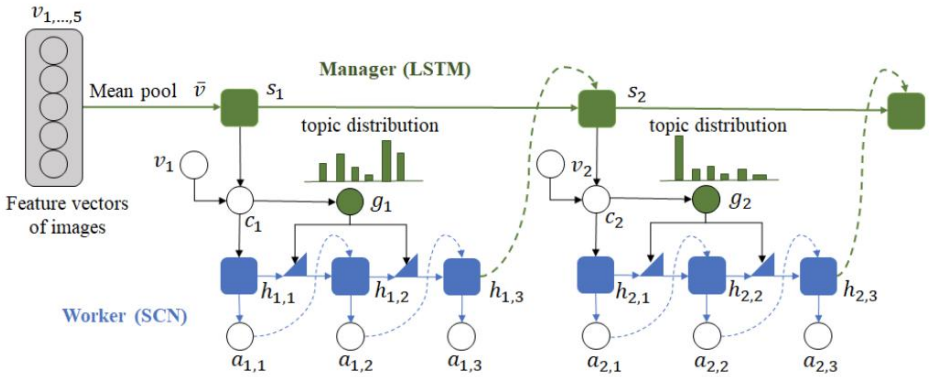
SentiCap: Generating Image Descriptions with Sentiments, AAAI'16

基于视觉的长文生成

- 在长文本生成中考虑话题转移
  - 利用判别器判断话题转移
  - 基于分层次强化学习建模话题转移
- 缺少对多个图像内容交互的建模
- 缺少对生成文本一致性的建模



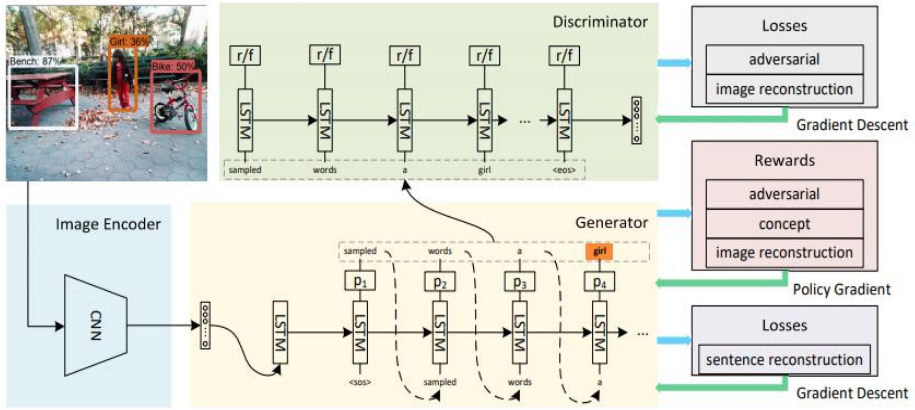
Recurrent Topic-Transition GAN for Visual Paragraph Generation, CVPR'17



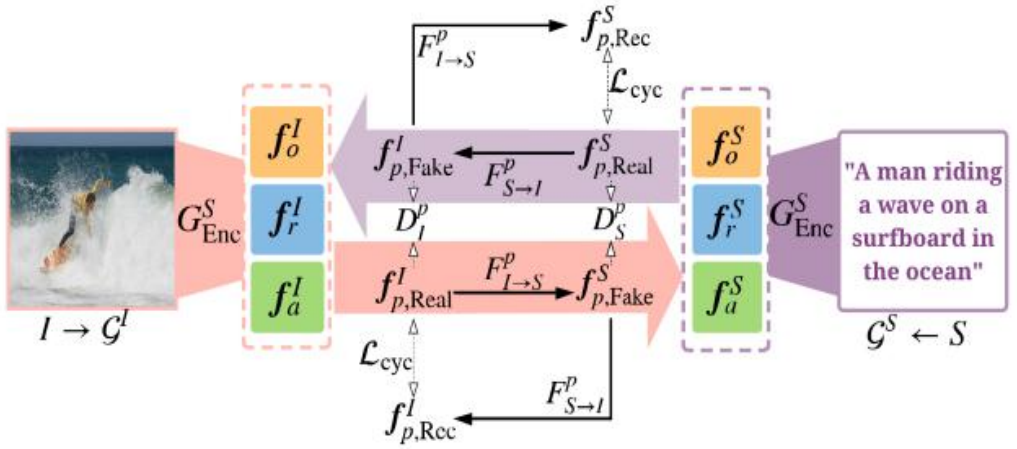
Hierarchically Structured Reinforcement Learning for Topically Coherent Visual Story Generation, AAAI'19

非监督的图像文本生成

- 任务定义：在没有配对图像和文本的情况下进行模型训练
- 预训练模型 + back-translation 进行语义对齐
- 采用场景图作为对齐方式



Unsupervised Image Captioning, CVPR'19



Unpaired Image Captioning via Scene Graph Alignments, ICCV'19

一点自己的想法

- 核心问题驱动的研究应该关注视觉-文本的跨模态语义对齐
  - 场景图会是一个很好的桥接视觉-文本的形式
  - 图像特征到文本之间的对齐失位，图像包含比单个句子更丰富的语义信息，如何解决？
- 场景问题驱动的研究依赖于任务的设定，新语料集合的提出
  - 评价方式是一个很好的探究点，对于多样性等的要求
  - 借助视觉信息，是一个长文本生成的很好的研究场景
- NLPer 需要深入到视觉信息处理的模型中，才有可能提出更好的跨模态建模方式。

参考文献

- Tadas Baltrusaltis, et al., Multimodal machine learning: a survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019
- Xiaodong He and Li Deng, Deep learning for image-to-text Generation: A technical overview, IEEE Signal processing Magazine, 2017
- Raffaella Bernardi, et al., Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures, JAIR, 2016
- Oriol Vinyals, et al., Show and tell: A neural image caption generation, CVPR'15
- Aishwarya Agrawal, et al., VQA: visual question answering, ICCV 2015
- Ting-Hao Huang, et al., Visual story telling, NAACL 2016
- Abhishek Das, et al., Visual Dialogue, CVPR 2017
- Hao Fang, et al., From caption to visual concept and back, CVPR 2015

参考文献

- Girish Kulkarni , et al., Baby Talk: Understanding and Generating Image Descriptions, CVPR 2011
- Kelvin Xu, et al., Show, attend and tell: Neural image caption generation with visual attention, ICML 2015
- Zichao Yang, et al., Stacked Attention Net, CVPR 2016
- Peter Anderson, et al., Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, CVPR 2017
- Ting Yao, Yingwei Pan, Yehao Li, Tao Mei, Exploring Visual Relationship for Image Captioning, ECCV 2018
- Quanzeng You, et, al., Image Captioning with Semantic Attention, CVPR 2016
- Marcella Cornia, et, al., Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions, Arxiv, 2019
- Generating natural questions about an image, ACL'16

参考文献

- Zhihao Fan, et al., Bridging by Word: Image Grounded Vocabulary Construction for Visual Captioning, ACL 2019
- Zhihao Fan, et al., A Reinforcement Learning Framework for Natural Question Generation using Bi-discriminators, COLING 2018
- Zhihao Fan, et al., A Question Type Driven Framework to Diversify Visual Question Generation, IJCAI 2018
- Alexander Mathews, et al., SentiCap: Generating Image Descriptions with Sentiments, AAAI 2016
- Yang Feng, et al., Unsupervised Image Captioning, CVPR 2019
- Jiuxiang Gu, et al., Unpaired Image Captioning via Scene Graph Alignments, ICCV 2019

谢谢