Visual Grounding

Luke Ye

52194506006

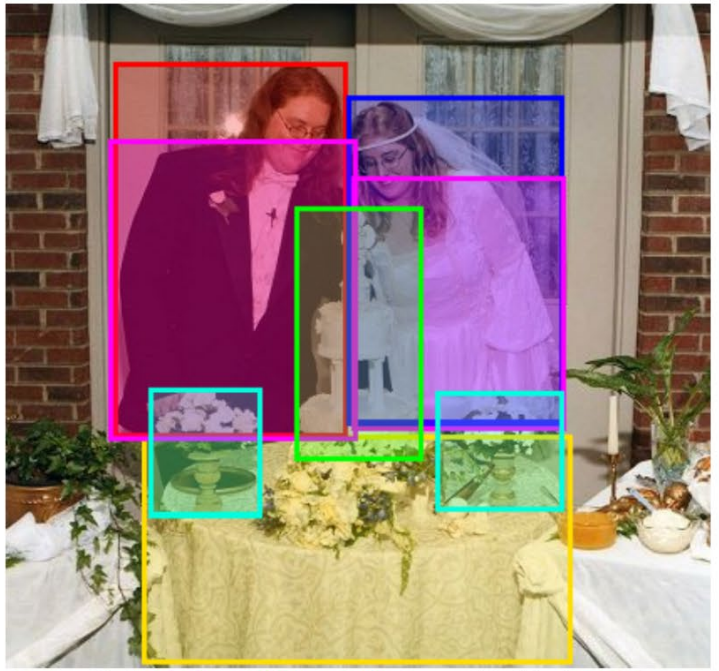
Outline

- 任务概述
- 细粒度监督的Visual Grounding
- 粗粒度（弱）监督的Visual Grounding
- 未来研究路线

任务介绍

A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .

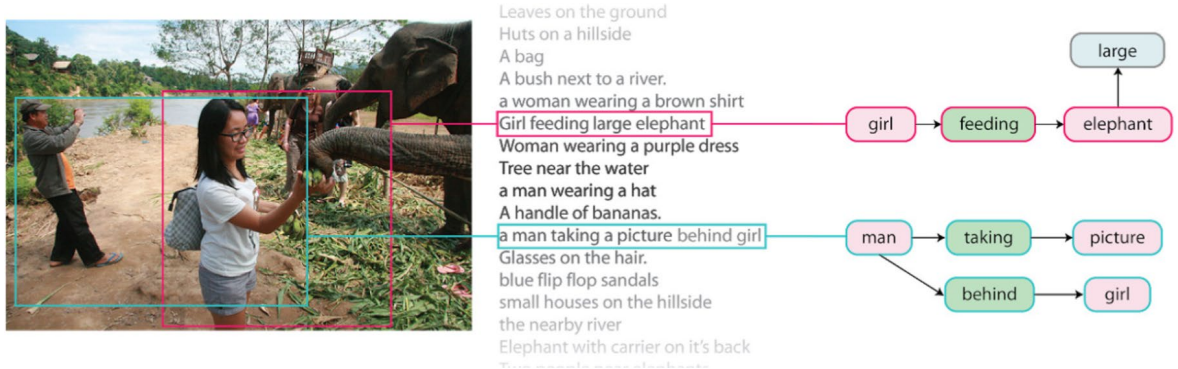


A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

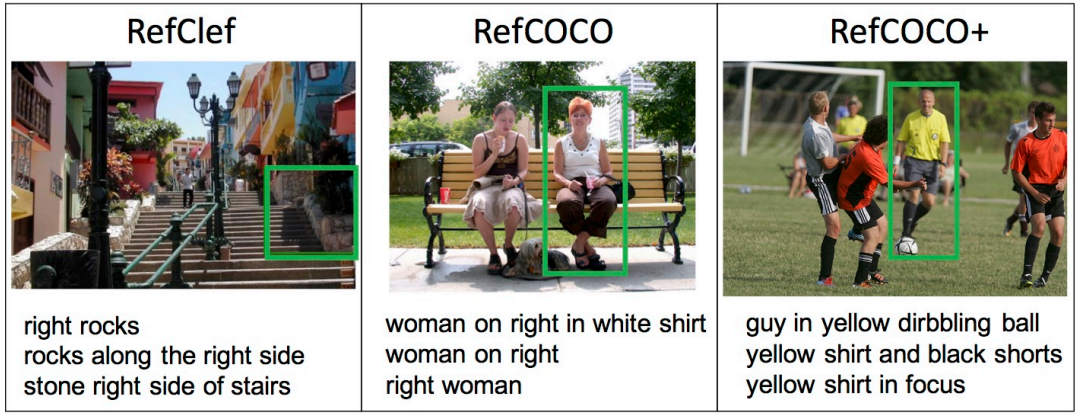
数据集

- Flickr30k Entities

- Visual Genome

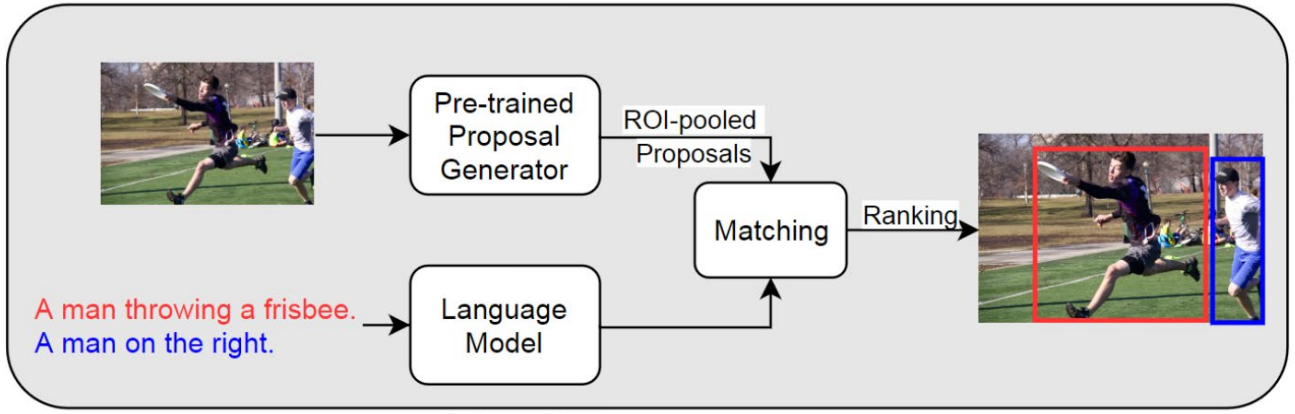


- Refer dataset
  - RefCOCOg
  - RefCOCO/RefCOCO+/RefClef

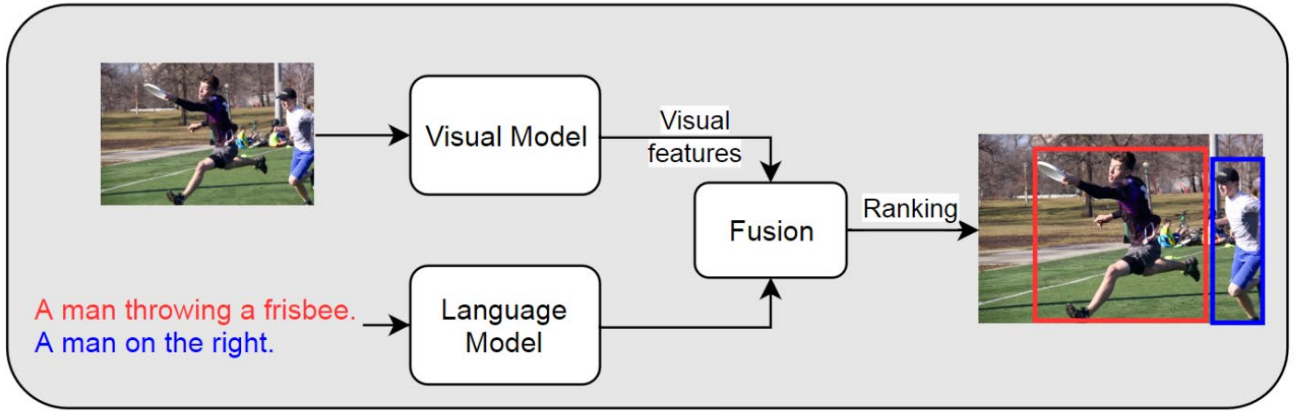


细粒度监督的Visual Grounding

- 方法
  - Two-Stage
    - Faster-RCNN
  - One-Stage
    - YOLO
    - FPN
    - Retina Net
- 改进点
  - 空间位置编码
  - 图像文本注意力
    - Co-Attention
    - Graph-Attention
  - 多区域匹配



Two-Stage approach



One-Stage approach

粗粒度（弱）监督的Visual Grounding

RPN

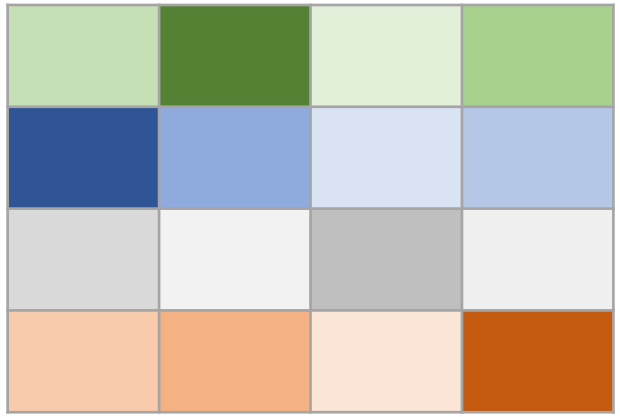
Parser

[The man]

[bat]

[the pitch]

[the umpire]



Aggregator

$Sim(Image, Caption)$

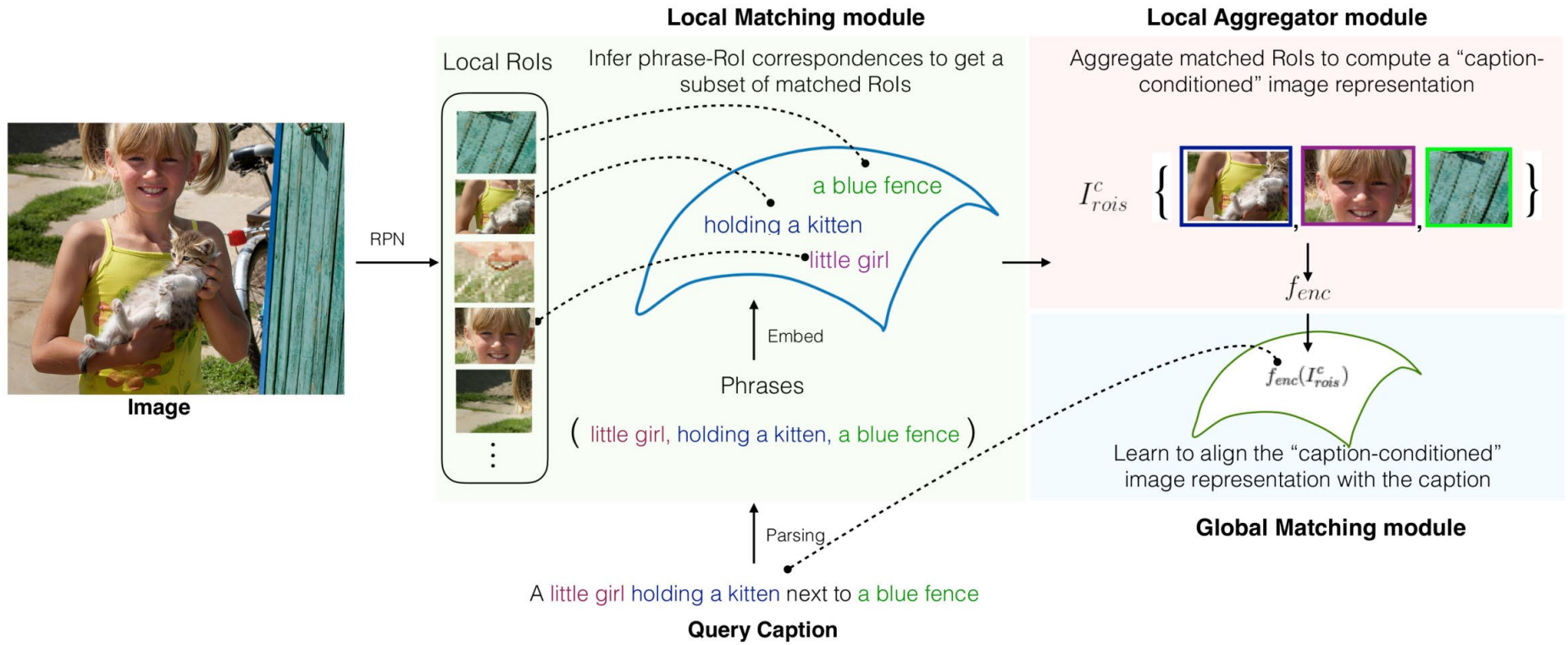
The man at bat readies to swing at the pitch while the umpire looks on

存在问题

- 训练目标和评估目标不一致
- Matching任务提供的监督较弱，模型难以训练
- 在多实例场景下，存在匹配不充分的情况

目标不一致

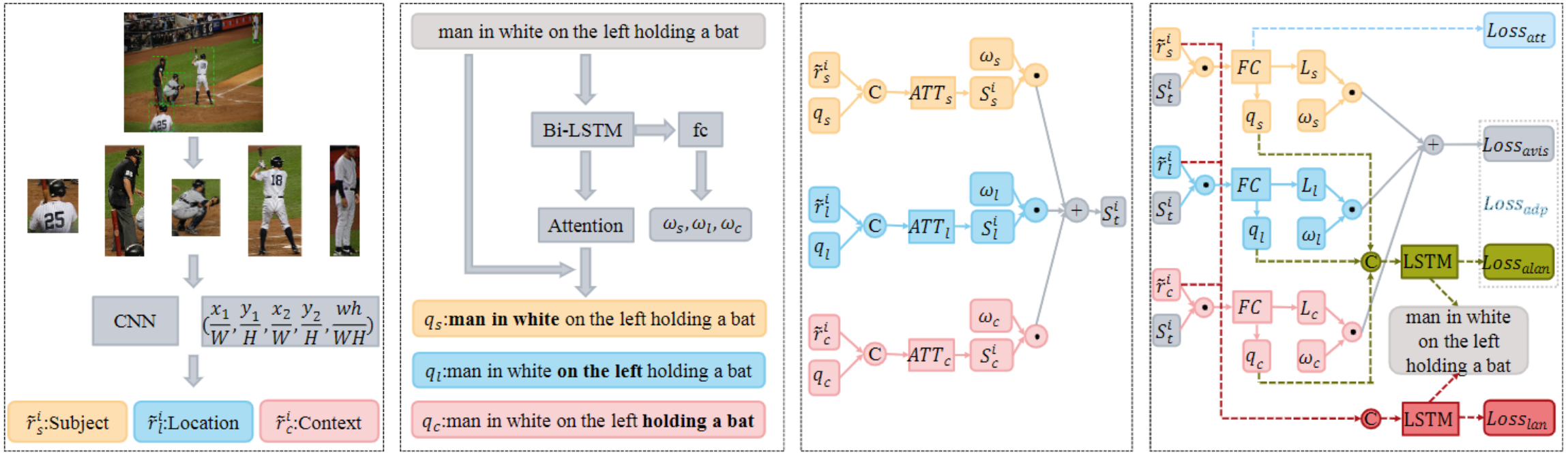
优化Aggregator



ICCV 2019, Align2ground: Weakly supervised phrase grounding guided by image-caption alignment

Matching任务监督信息弱

引入文本信息重构任务

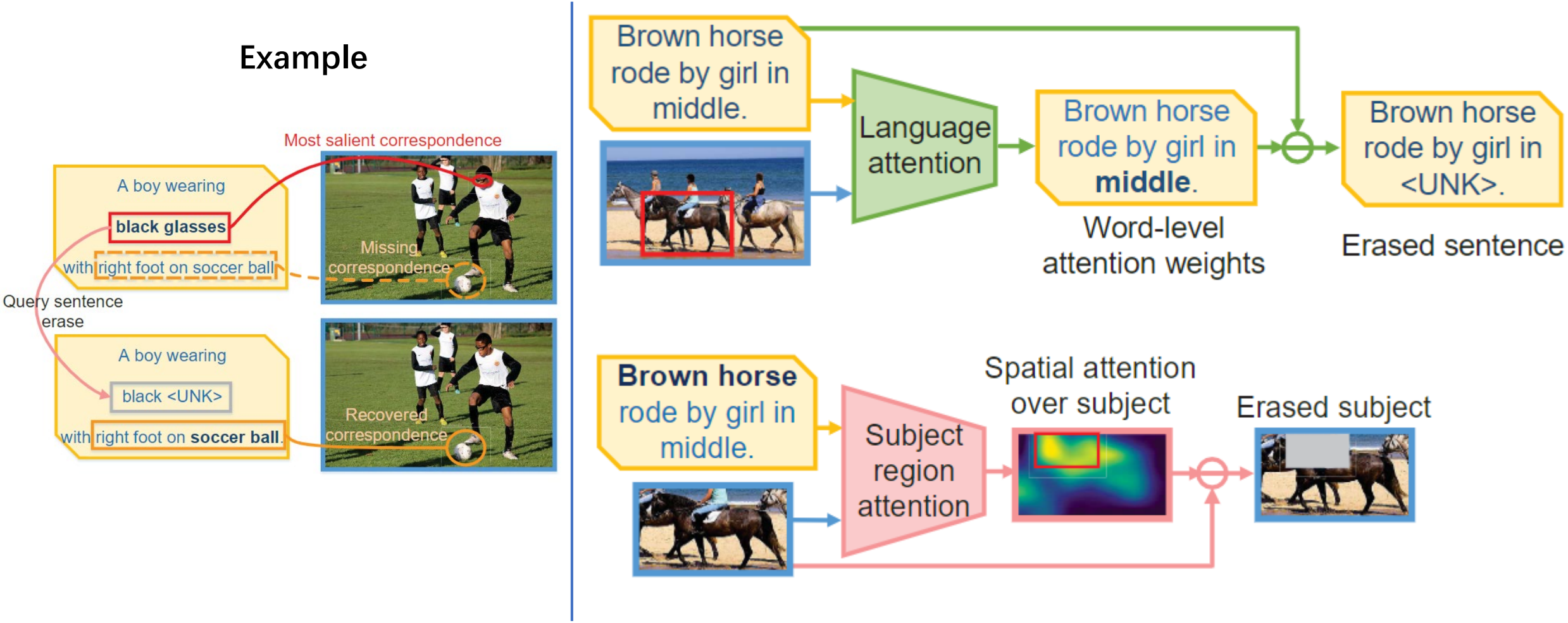


(a) Visual feature encoding. (b) Language feature encoding. (c) Adaptive grounding. (d) Collaborative Reconstruction.

ICCV 2019, Adaptive reconstruction network for weakly supervised referring expression grounding

匹配不充分

引入擦除的增强样本



CVPR 2019, Improving referring expression grounding with cross-modal attention-guided erasing

未来研究路线

- 对抗样本生成
  - 从实体/属性/位置的词组中生成文本对抗样本
  - 使用对抗模型检索具有近似语义的图片或图片块作为图像对抗样本

- 用外部知识辅助约束多Region的Grounding