# Ground 视频描述任务介绍

郑济元
11821029
jiyuanz@zju.edu.cn

2019 年 6 月 15 日

## 摘要

本文主要介绍一个 CVPR2019 新提出的一个任务，Grounding 的视频描述任务。原文名 “Grouding Video Description”, 作者针对目前的视频描述任务不能准确理解视频以及缺少可解释性，在 ActivityNet Captions [2] 基础上构建了需要同时生成 grounding 和视频描述的数据集。然后作者提出了用于该任务的一种模型，实验结果也表明加入 grouding 的监督信息可以更好地生成准确的视频描述，作者并将这种方法拓展到了图像 grouding 数据集上也获得了超过其他基准方法的表现。

## 1 整体介绍

传统的没有 well-grounded 数据集只标注了图像或者视频对应的描述。在这种标注下训练得到的模型很容易产生 bias 以及生成一些图像或者视频中没有的物体，而仅仅是因为这些物体在训练数据中类似的上下文环境中出现了。

作者在文中举了个例子，如图 1，没有 grounding supervision 的视频描述方法生成了“一个男人站在体育馆前”，而实际上这句描述中“男人”是视频中有的，但“体育馆”是视频中推测不出的。而对于图 1，描述中出现了“一个男人”，在这里，将描述中的“男人”和图里的“男人”对应起来是很重要。我们知道“男人”指的是坐着的弹奏钢琴的男人而不是站着的男人，所以我们要求模型也能做到这一点。这一步对于搭建一个可解释的模型或者是做进一步的推理性的任务非常关键，比如进一步提问“他看着我吗?”，理解这句话中的代词指向模型就需要知道是图里的哪个人。

因此，作者希望能去构建一个可以 ground 到图像的系统。这可以利用 grouding 的监督信息帮助更好生成描述，也有助于探索模型的可解释性。

而现有的数据集不能满足这一要求，因此作者先是在 ActivityNet Captions [2] 基础上构建了需要同时生成 grounding 和视频描述的数据集。然后作者提出了用于该任务的一种模型，实验结果也表明加入 grouding 的监督信息可以更好地生成准确的视频描述，作者并将这种方法拓展到了图像 grouding 数据集上也获得了超过其他基准方法的表现。

## 2 相关工作

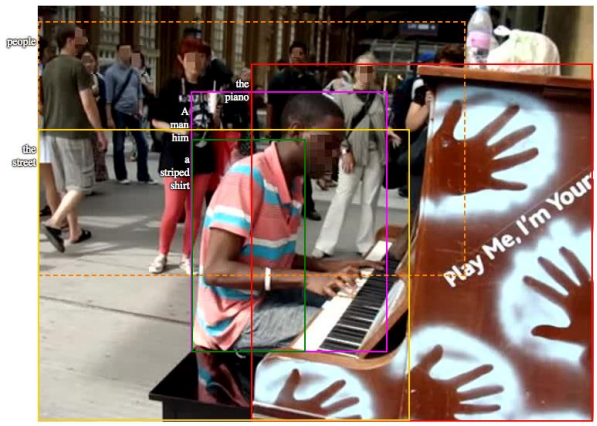
视频描述任务中，基于物体检测的描述生成模型通常把这个任务分解成两步。第一步是直接使用训练好的物体检测器来检测物体得到 proposals，然后再基于这些 proposals 使用 attention 机制去生成描述

A man is seen standing in a room speaking to the camera while holding a bike.

w/o grounding supervision: A man is standing in a gym .
[42]: A man is seen speaking to the camera while holding a piece of exercise equipment.
GT: A man in a room holds a bike and talks to the camera.


A group of people are in a raft down a river.

w/o grounding supervision: A group of people are in a river.
[42]: A large group of people are seen riding down a river and looking off into the distance.
GT: Several people are on a raft in the water.




A man in a striped shirt is playing the piano on the street while people watch him.

或者直接对 proposals 分类得到文本标签再填入预先生成的模板中。但是，第一步中的使用训练好的物体检测器事实上一开始就有着一定的 bias，物体检测器本身训练的 domain 和数据集中物体的 domain 不同，并不一定有效，而根据 domain 重新 fine-tune 又需要大量的标记，这对于视频是更加不可能的。

对于细粒度的描述任务而言，还有一种可能的想法是对 attention 机制进行标注并进行监督学习。一方面实际上一些关于 VQA 任务的工作 [1, 7] 已经指出模型产生的 attention 和人在进行这个任务产生的 attention 并不一样。另一方面，一些在特征图上的 attention 进行监督的学习的方法 [3, 6] 被认为是有效的。

## 3 ActivityNet-Entities 数据集

在图像方面，其实已经有 Flickr30k Entities 这样的 Grounding 数据集，而 YouCook2 数据集则只限制在 cooking 领域，而且只有 val 和 test 上有 bounding box 的标注。ActivityNet Humans 只标了一部分数据中的人。

因为视频存在很多类似动作这样的动态信息，为了准确性，作者只对名词性短语进行标注。而且对每帧数据都进行标注的话，代价太大了。所以实际操作上作者选择了根据视频内容进行分段，然后只对每一段中其中一帧标注。

作者主要是在 ActivityNet-Captions 这个数据量比较大的视频描述数据集基础上构建，所有的 bounding box 是精确到实例的，而不是一整类物体。比如1中对视频描述中的“男人”只会对坐着的男人标注。

作者对每一个视频段都随机采样 10 帧并把这些帧和对应的描述提供给众包工人。众包工人先识别出描述中的所有名词性短语，再在这些帧中选取可以完全辨认这些短语的某一帧并绘制 bounding box。然后，众包工人解析描述中可能存在的指代之类的，并同样和 bounding box 相对应起来。

## 4 使用 Grouding 信息的描述生成

作者提出了 grounded 的视频描述结构。这个结构主要包括三个模块：grounding 模块，region attention 模块和语言生成模块。Grounding 模块从视频中检测视觉信息，region attention 模块则动态地 attend 到这些视觉信息形成高层的表示并提供给语言生成模块，具体模型图见 1。

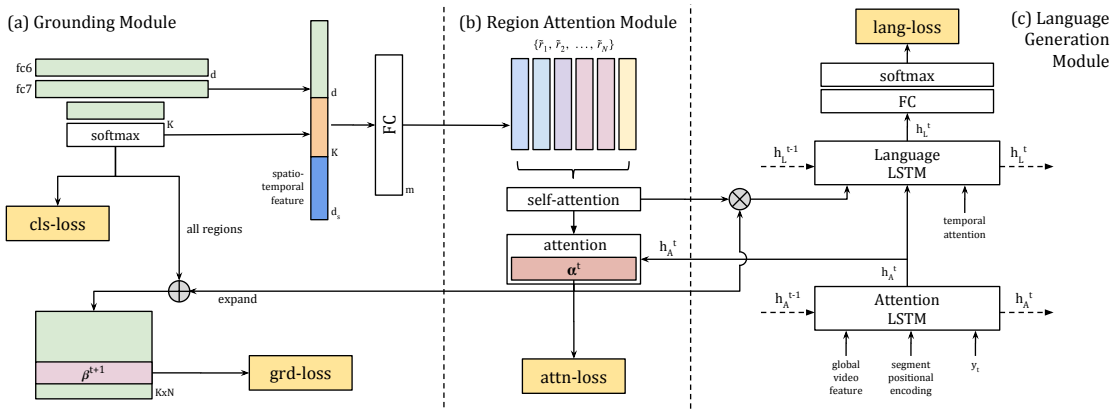



图 1:

作者采用了三种方式来将物体级别的监督结合进去: 区域的分类, 物体的定位以及监督的 attention。

整体来看，作者就是同时优化语言生成任务和 grounding 任务，总的损失函数为 $L = L_{sent} + \lambda_\alpha L_{attn} + \lambda_c L_{cls} + \lambda_\beta L_{grd}$

## 4.1 语言生成模块

语言生成模块的主要是将语言模型迁移过来适合视频输入, 主要由两个 LSTM 组成。第一个 LSTM 将视频信息和单词 embedding 编码到一个 hidden state $h_A^t \in \mathbb{R}^m$，第二个 LSTM 则用于生成语言。语言生成模块会动态地 attend 视频的帧（被称为 temporal attention）以及区域的编码信息（被称为 region attention）来生成单词。

Temporal attention 的部分作者采用了类似 [9] 的方法，只是将 self-attention 的 encoder 换成了 Bi-GRU。

## 4.2 Region Attention 模块

不同于 Temporal attention 主要是在帧层面的信息, 也就是相对 coarse-grained 一点, region attention 则更关注视频中 fine-grained 的信息。

假设我们有一系列关于 grounding 物体的类别标签 $c_1, c_2, \ldots, c_{\|}$。给定一系列物体的区域，grounding 模块就是要对这些区域估计类别的概率。

作者采用了一种简单的方法去得到这组概率:

$$M_s(R) = \text{Softmax}(W_c^T R + B\not{k}^T) \tag{1}$$

其中 $W_c$ 和 $B$ 分别是参数矩阵和参数向量，$M_s$ 就是多个区域的类别的分布组成的矩阵。

一方面，为了能利用预先训练好的物体检测器，作者先是将 grounding 物体的类别标签和预训练数据集中语义相近的标签来近似，并通过在自己的数据集上跑一遍物体检测器来初始化 $W_c$ 和 $B$。另一方面，作者进一步将每个 region 的表示扩展到时间和空间的维度。在原始的 region 表示上加入 label 概率的表示 $M_s(R)$ 以及携带时序信息的帧索引，即

$$\tilde{R} = W_g[R|M_s(R)|M_l] \tag{2}$$

为了进一步建模 region 之间的关系，作者在此基础上又加了一层 self-attention 层，这样我们就得到了 grounding-aware 的 region 表示。

一方面，作者先是使用了常规的 attention 的方式，即在每一步生成描述的部分，令语言模块的 hidden state attend 到这些 region 表示即可。另一

| Method | B@1 | B@4 | M | C | S | Attn. | Grd. | $F1_{all}$ | $F1_{loc}$ | Cls. |
|---|---|---|---|---|---|---|---|---|---|---|
| Masked Transformer [9] | 22.9 | **2.41** | 10.6 | **46.1** | 13.7 | – | – | – | – | – |
| Bi-LSTM+TempoAttn [9] | 22.8 | 2.17 | 10.2 | 42.2 | 11.8 | – | – | – | – | – |
| Our Unsup. (w/o SelfAttn) | 23.1 | 2.16 | 10.8 | 44.9 | **14.9** | 16.1 | 22.3 | 3.73 | 11.7 | 6.41 |
| Our Sup. Attn.+Cls. (GVD) | **23.6** | 2.35 | **11.0** | 45.5 | 14.7 | **34.7** | **43.5** | **7.59** | **25.0** | **14.5** |

表 1: Results on ActivityNet-Entities test set.

方面，作者还对 attention 信息进行监督。作者希望模型在生成 groundable 单词的时候 attend 到正确的 region。在 attend 的每一步，对所有的 regions 对应生成一组 indicators$\gamma^t = [\gamma_1^t, \gamma_2^t, \ldots, \gamma_N^t]$，假如 region 和那个单词对应的 bounding box 有超过 0.5IoU，则对应的 indicator 为 1，否则为 0，因此这样情况下的 attention loss 为

$$L_{attn} = -\sum_{i=1}^{N} \gamma_i^t \log \alpha_i^t \tag{3}$$

### 4.3 Grounding 模块

使用 4.2中得到的 region 和类别之间的相似度矩阵 $M_s$，我们首先可以针对 region 进行分类级别的监督，也就是利用每个 region 在类别上的分布即相似度矩阵和 ground truth 计算交叉熵得到对应的分类 loss$L_cls$。

对于一个描述中可以 grounding 的单词，可以同样建立一个关于单词 grouding 的 loss。即把该时刻的相似度矩阵 $M_s$ 看作是每个 region 在不同 label 上的一个 confidence 分数 $\beta$，则对于该时刻该单词的 grounding loss 可以表示为:

$$L_{grd} = -\sum_{i=1}^{N} \beta_i^t log \beta_i^t \tag{4}$$

和 $L_{attn}$ 一样，最后的 loss 是在所有 groundable 单词上的平均，这里和 attention loss 的区别在于，前者是不预先知道生成的单词的，而后者是知道目标物体的类型。

## 5 实验

实验方面，作者主要在自己构建的 ActivityNet-Entities 数据集上进行了主实验以及 ablation study，另外也扩展到了 Flickr30k Entities[4] 数据集上进行了实验。

作者采用了 MaskedTransformer[9] 以及 Bi-LSTM+TempoAttn[9] 模型作为 baseline。在描述生成方面的指标主要有 BLEU@1,BLEU@4,METEOR,CIDEr, 和 SPICE; 在 grounding 的方法，主要参考 [5, 8] 中定位准确度的做法得到 grouding 和 attention 两方面的分数 (*Grd* 和 *Attn*)。

## 6 结果

作者的主要结果在表 1中展现，可以相比于没有 grouding 信息的描述生成 (即表中的 Unsup 版本)，使用了 grounding 信息的模型显著得提升了生成的描述质量。这里提升了 Attn 和 Grd 的分数则是相对显而易见的，毕竟是有监督学习。虽然看到在 BLEU 分数和 CIDEr 分数上虽然没有之前的最好结果好，但是经过人工评测，作者发现之前的方法会生成重复的

句子，实际上质量上不如本文中的好。

# 参考文献

[1] Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding 163* (2017), 90–100.

[2] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Carlos Niebles, J. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 706–715.

[3] Liu, C., Mao, J., Sha, F., and Yuille, A. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence* (2017).

[4] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 2641–2649.

[5] Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., and Schiele, B. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision* (2016), Springer, pp. 817–834.

[6] Yu, Y., Choi, J., Kim, Y., Yoo, K., Lee, S.-H., and Kim, G. Supervising neural attention models for video captioning by human gaze data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 490–498.

[7] Zhang, Y., Niebles, J. C., and Soto, A. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), IEEE, pp. 349–357.

[8] Zhou, L., Louis, N., and Corso, J. J. Weakly-supervised video object grounding from text by loss weighting and object interaction. *arXiv preprint arXiv:1805.02834* (2018).

[9] Zhou, L., Zhou, Y., Corso, J. J., Socher, R., and Xiong, C. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8739–8748.