Recent VQA Approaches

Wentao Mo[1]

[1]Department of Machine Intelligence
Peking University

2021 年 4 月 28 日

Outline

1. Visual Feature Extractor
   - VC R-CNN
   - Grid Feature
   - VinVL
2. Feature Fusion Methods
   - MFH
   - MCAN
   - TRRNet
   - BERT/Transformer-based

Visual Commonsense R-CNN

提出 VC R-CNN. 使用 causual intervention$P(Y \mid \mathrm{do}(X))$ 代替传统的 lld. 相信应该使用 causual commonsense feature 而不是单纯的 visual feature.
Replace

$$P(Y \mid X) = \sum_z P(Y \mid X, z)\underline{P(z \mid X)} \tag{1}$$

w/ intervention

$$P(Y \mid do(X)) = \sum_z P(Y \mid X, z)\underline{P(z)} \tag{2}$$

提出 proxy task 为预测 local context label of Y. 关于 confounder set Z, 我们保存一些固定数量的 dictionary $N \times d$, N 是数据集中类别的数量 (MSCOCO, 80), 每个 d 维特征都是平均的 RoI 特征. 特征通过 Faster RCNN pretrain.
总 obj. 为 self-classification+contextual-pair-classification loss

$$L(X) = L_{\text{self}}\ (p, x^c) + \frac{1}{K}\sum_i L_{cxl}\left(p_i, y_i^c\right) \tag{3}$$

Visual Commonsense R-CNN

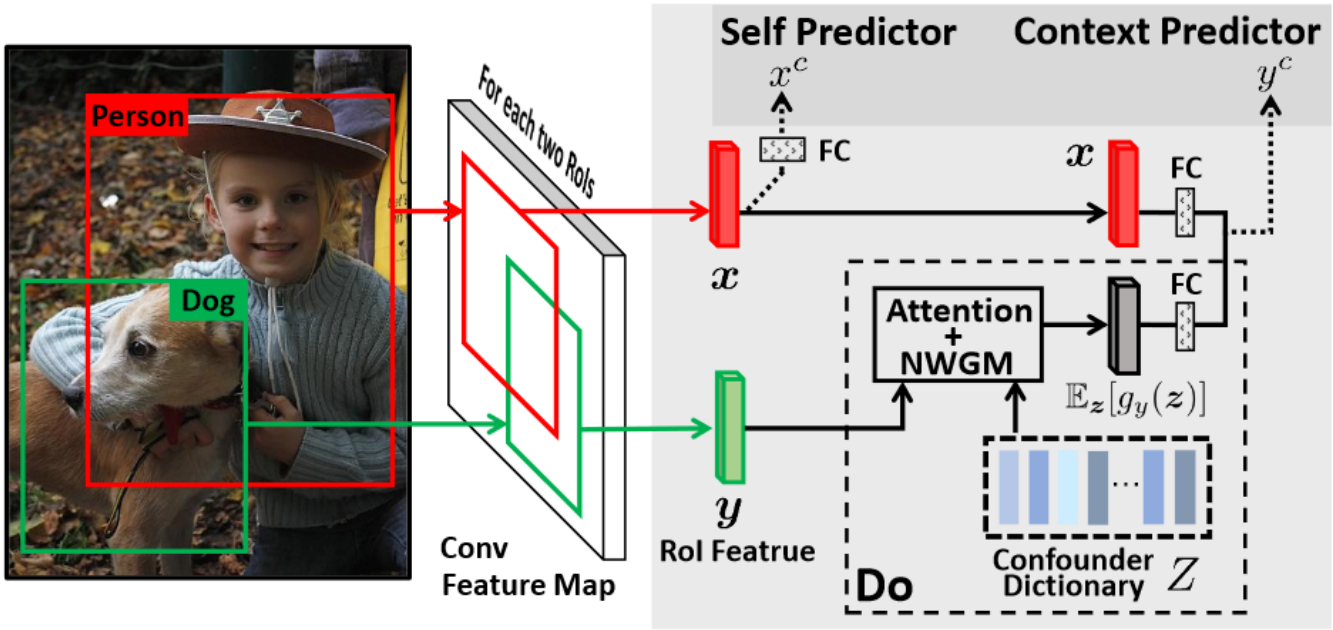

Figure 4. The overview of VC R-CNN. Any R-CNN backbone (*e.g.*, Faster R-CNN [54]) can be used to extract regions of interest (RoI) on the feature map. Each RoI is then fed into two sibling branches: a **Self Predictor** to predict its own class, *e.g.*, x^c, and a **Context Predictor** to predict its context labels, *e.g.*, y^c, with our **Do** calculus. The architecture is trained with a multi-task loss.

Visual Commonsense R-CNN

具体上, $P(Y \mid do(X)) = \sum_z P\left(y^c \mid \boldsymbol{x}, \boldsymbol{z}\right) P(\boldsymbol{z})$, 使用

$$P(Y \mid do(X)) := \mathbb{E}_{\boldsymbol{z}}\left[\text{Softmax}\left(f_y(\boldsymbol{x}, \boldsymbol{z})\right)\right] \overset{\text{NWGM}}{\approx} \text{Softmax}\left(\mathbb{E}_{\boldsymbol{z}}\left[f_y(\boldsymbol{x}, \boldsymbol{z})\right]\right) \tag{4}$$

并且使用 NWGM(Normalized Weighted Geometric Mean) 来估计上述期望
f 使用线性模型 $f_y(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{W}_2 \cdot g_y(\boldsymbol{z})$, where $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{N \times d}$ 代表了 FC 层. 那么有

$$\mathbb{E}_{\boldsymbol{z}}\left[f_y(\boldsymbol{x}, \boldsymbol{z})\right] = \boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{W}_2 \cdot \mathbb{E}_{\boldsymbol{z}}\left[g_y(\boldsymbol{z})\right] \tag{5}$$

建模 $g_y(\cdot)$ 为 scaled 点积注意力, 具体上有

$$\mathbb{E}_{\boldsymbol{z}}\left[g_y(\boldsymbol{z})\right] = \sum_z \left[\text{Softmax}\left(\boldsymbol{q}^T \boldsymbol{K} / \sqrt{\sigma}\right) \odot \boldsymbol{Z}\right] P(\boldsymbol{z}) \tag{6}$$

使用 NCC 去除 $x \to z$ 的样本.

In Defense of Grid Features for VQA

最近基于 region 的方法逐渐流行并超过了基于 grid 的方法. 但是相容实验发现主要影响性能的是 pre-training 的数据集质量 + 输入图像的高分辨率, grid/region 只是小问题.
传统上, 一般使用 Faster R-CNN, 在 cleaned version VG 上训练. 对于这些方法, 要获得自底向上注意力特征, 进行如下两步

1. Region Selection. 通过一个 Region Proposal Net., 提出候选 region(Regions of Interst, RoIs), 接着通过一个 score comp., 选择 top-N 的区域, 并且两步都使用 NMS.
2. 给出了上述步骤的 regions, 使用 RoIPool 来得到 region-feature.

由于 VG 数据集的复杂性和 Faster R-CNN, 这两步计算上都很昂贵.
具体的 VQA 模型 (特征融合) 使用的是 MFH.

In Defense of Grid Features for VQA

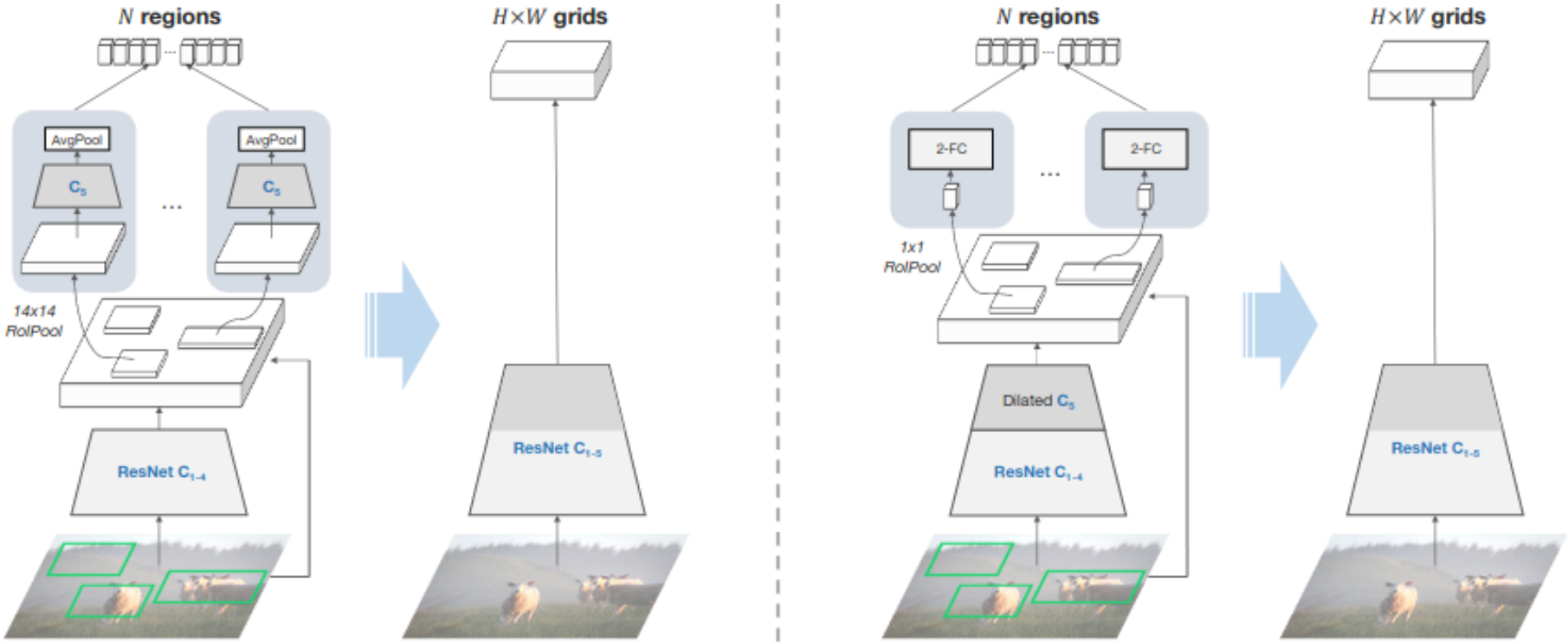

Figure 2: From regions to grids. Left: We convert the original region feature extractor used by bottom-up attention [2] back to the ResNet [15] grid feature extractor for the *same* layer (see Sec. 3.2, weights in blue are transferred), and find it works surprisingly well for VQA [11]. **Right**: We build a detector based on 1×1 `RoIPool` while keeping the output architecture *fixed* for grid features (see Sec. 3.3), and the resulting grid features consistently perform at-par with region features.

In Defense of Grid Features for VQA

Faster R-CNN 是 c4 模型 w/ 属性分类分支的变种. 首先使用 ResNet 的 C_4blocks 来得到 feature map, 接着 per-region feature 先 14×14 RoIPool, 再应用 C_5, 最后 avg-pool 来得到每个 region 的 F. 我们直接使用 C_5 在 grids 来得到特征.

这意味着用一个一维向量来表示一个 region, 而不是 Faster RCNN 里的 HWC 三维. 使用 1×1 RoIPool 会降低物体检测的性能, 对于 VQA, 这要求这个特征尽可能单独地编码信息. 由于预训练的 C_5 输入不适用, 使用最近的直接使用整个 C_5 的 ResNet 的工作[1].

Ablation Study 发现主要影响性能的是 pre-training 的数据集质量 + 输入图像的高分辨率, grid/region 只是小问题. 使用 Res-NeXt 改进了性能.

发现使用更大的图像, 更高的精度. 不同的预训练任务上, detection w/ attr. $>$ detection w/o attr. $>$ classification w/ tag $>$ cls. w/ label.

[1] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. De-formable convnets v2: More deformable, better results. InCVPR, 2019.

VinVL

MFH: Multimodal Factorized High-order Attention

MCAN: Deep Modular Co-Attention Networks

TRRNet: Tiered Relation Reasoning for Compositional VQA

OSCAR & UNITER: Transformer-like Fusion