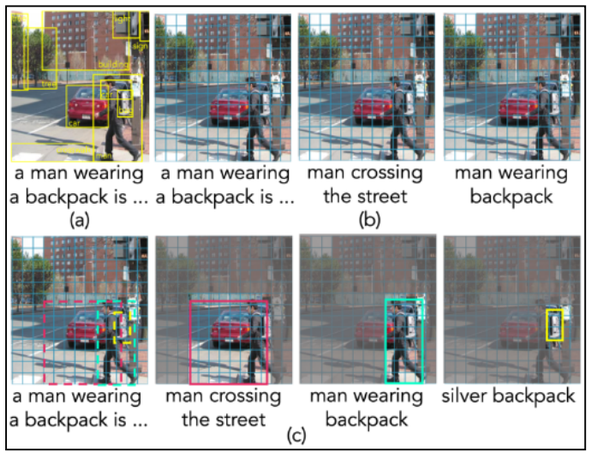
Methods -- old VS. new



(a) Fine-grained Methods Drawbacks

Object Level

1. 被检测的目标可能并非与文本相关
2. 以对象为中心的特征不能表示多个对象关系
3. 很难确定合适的下游任务

(b) Coarse-grained Methods Drawbacks

Image Level

1. 忽视了局部的对象及其特征，只考虑了全局
2. 不适用于如视觉推理等下游任务

(c) Multi-Grained Method (this paper)

Object & Image Level

1. 对于 visual concept 无限制 2. 不局限 level

Visual Concept (Multi-Grained)

(1) an object (2) a region (3) the image itself

Contributions

1. multi-grained 的视觉语言对齐任务
2. 优化模型，给文本定位图像 + 视觉语言对齐
3. 在诸多下游任务中取得很好的成绩

A. X-VLM Model

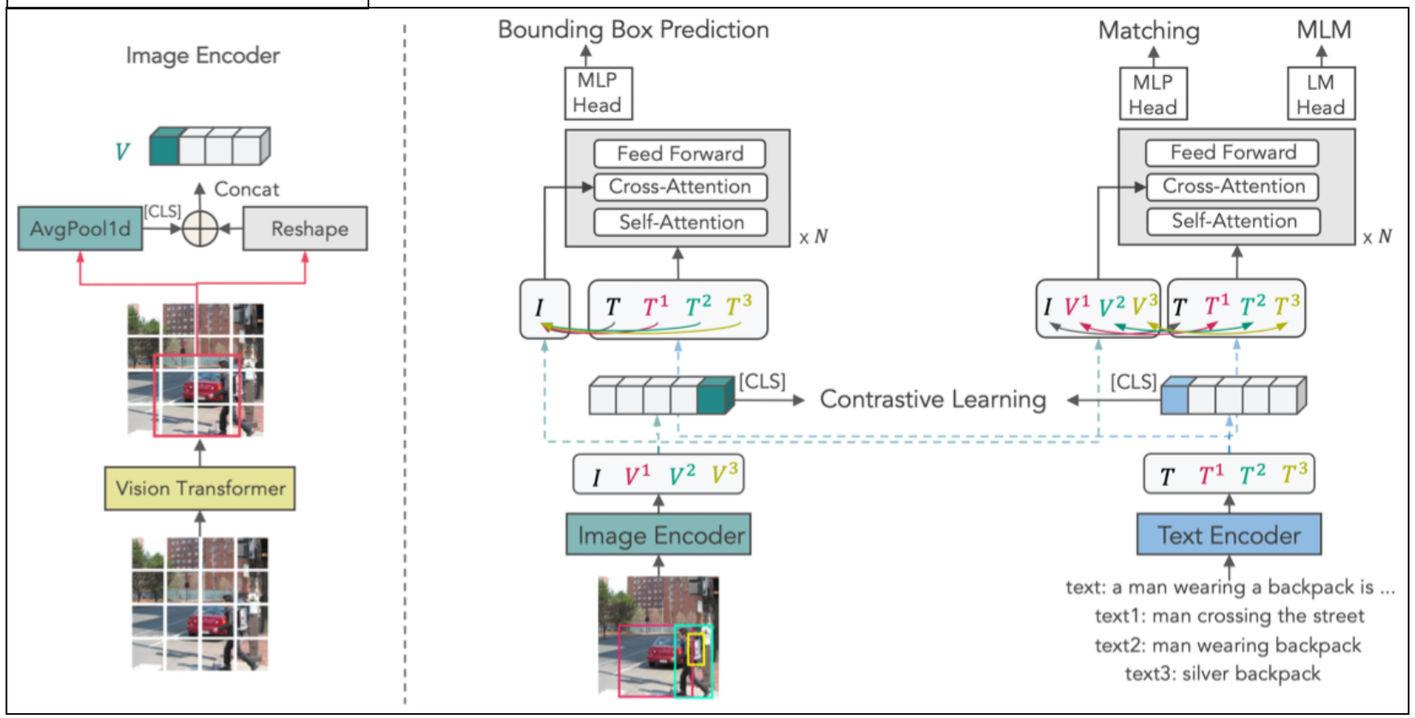


Image Encoder

1. ViT，49个Patch
2. 保证位置信息，reshape
3. 前加特征的均值

Cross-Modal Modeling

1. 给定文本，定位区域
2. 对比学习
3. 匹配预测-视觉语言对齐
4. 掩码语言模型

B. How to optimize

1. Bounding Box Prediction

$$\mathcal{L}_{\text{bbox}} = \mathbb{E}_{(V^j,T^j)\sim I;I\sim D}[\mathcal{L}_{\text{iou}}(\boldsymbol{b}_j,\hat{\boldsymbol{b}}_j) + ||\boldsymbol{b}_j - \hat{\boldsymbol{b}}_j||_1]$$

2. Contrastive Learning

$$\mathcal{L}_{\text{cl}} = \frac{1}{2}\mathbb{E}_{V,T\sim D}\big[\mathrm{H}(\boldsymbol{y}^{\text{v2t}}(V),\boldsymbol{p}^{\text{v2t}}(V)) + \mathrm{H}(\boldsymbol{y}^{\text{t2v}}(T),\boldsymbol{p}^{\text{t2v}}(T))\big]$$

3. Matching Prediction

$$\mathcal{L}_{\text{match}} = \mathbb{E}_{V,T\sim D}\mathrm{H}(\boldsymbol{y}^{\text{match}},\boldsymbol{p}^{\text{match}}(V,T))$$

4. Masked Language Modeling

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{t_j\sim\hat{T};(V,\hat{T})\sim D}\mathrm{H}(\boldsymbol{y}^j,\boldsymbol{p}^j(V,\hat{T}))$$

构建了本文的需要被优化的损失函数

$$\mathcal{L} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{mlm}}$$

C. V+L Downstream Tasks

1. 视觉问答任务 2. 视觉推理等相关任务 3. 视觉定位任务 4. 图像文本生成任务