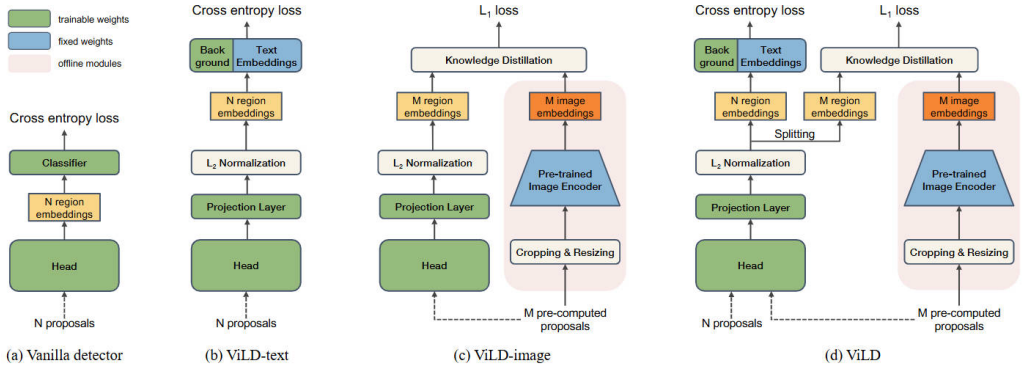
Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation

引言写法：

首先给出一张Figure，然后提个问题自己回答，便直接提出paper的写作动机（scale up the number of classes与zero-shot transfer）

Approach

Architecture	本来Object Detection要分为两个步骤：1）如何定位？Bounding Box画得准不准确；2）如何分类？Bounding Box中的物体判断得对不对；下面的方法只涉及到第二个步骤。	
(a) Vanilla Detector	基准模型，Mask-R-CNN，proposals经过detector heads，得到region embeddings，经过classifier得到bounding box对应的是什么类别。C_B	
(b) ViLD-text	C_B 全给到Text Embeddings：Background Embeddingss；之后将两者分别和Region Embedding做相似度计算。	
(c) ViLD-image	C_B, C_N 蒸馏学习 加快训练的trick——时间和内存	
(d) ViLD	C_B, C_N 右侧粉红底色的模型只在训练时有用到，推理时不需要	
(e) ViLD-ensemble		



proposals是指region proposals。region embeddings对应的是图像特征。

Region Embeddings

$$e_r = R(\phi(I), r)$$

Logits

$$z(r) = [sim(e_r, e_{bg}), sim(e_r, t_1), \ldots, sim(e_r, t_{|C_B|})]$$

region embeddings、background embeddings和text embeddings之间的similarity

Loss

$$L_{ViLD-text} = \frac{1}{N} \sum_{r \in P} L_{CE}(softmax(\frac{z(r)}{\tau}), y_r)$$

和ground truth做cross entropy

$$L_{ViLD-image} = \frac{1}{M} \sum_{\tilde{r} \in \tilde{P}} ||V(crop(I, \tilde{r}_{\{1\times, 1.5\times\}})) - R(\phi(I), \tilde{r})||_1$$

模型Overview