Grounded Language-Image Pre-training

动机：scale up dataset

visual grounding：

有张图片，再给你一些文本，要求你在图片中框出这些文本描述的物体

损失函数：

$$S_{ground}$$

指的是region-word alignment scoress

$$O = Enc_I(Img), P = Enc_L(Prompt), S_{ground} = OP^T$$

O是region proposal，P是text embeddings

object detection：

损失函数：

$$L = L_{cls} + L_{loc}$$

loc是bounding box的位置；cls是bounding box的类别

$$O = Enc_I(Img), S_{cls} = OW^T, L_{cls} = loss(S_{cls}; T)$$

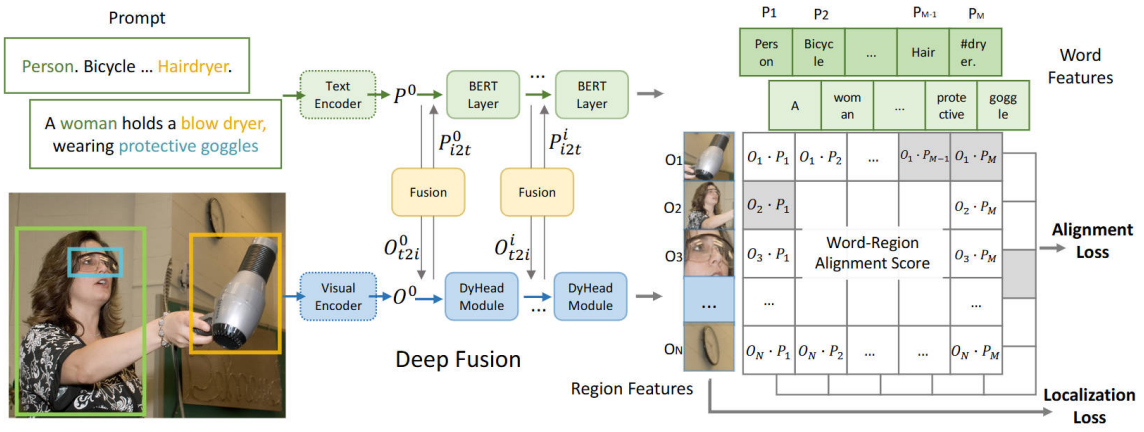
O是region proposal，Scls是类别的logits，Lcls是cross entropy loss，T是标签

Approach：

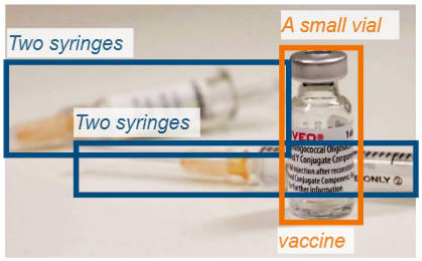
将visual grounding和object detection合并成一个任务（先在小的数据集上实验），用supervision fashion（Detection数据集--FourODs和Grounding数据集--GoldG上），再用self-training fashion（数据集--Cap24M），即伪标签来实现scale up

deep fusion方法

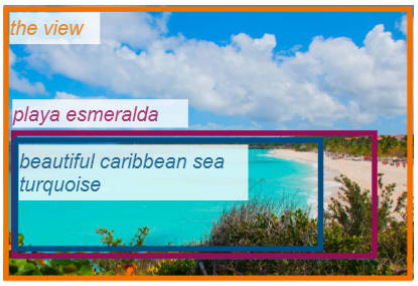
Architecture：



`Inference:`



Two syringes and a small vial of vaccine.



playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

GLIPv2: Unifying Localization and VL Understanding

`合并更多的任务（Object Detection、Instance Segmentation、VL Grounding、Visual Question Answering、Image Caption），以继续scale up dataset`