What Object Should I Use? - Task Driven Object Detection

我应该使用什么物品？——任务驱动对象检测（CVPR-2019）

https://coco-tasks.github.io/（项目） https://github.com/coco-tasks/dataset （数据集）

https://github.com/yassersouri/task-driven-object-detection （论文）

任务：任务驱动的对象检测

输入是一项**任务**，输出是**最适合**解决该问题的**对象的边界框**

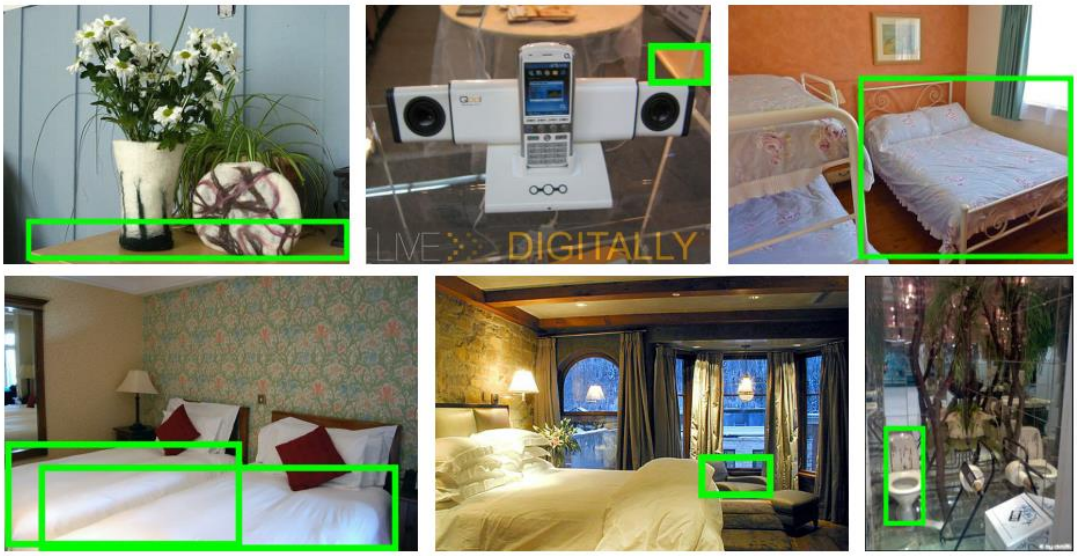
Figure 1. What object in the scene would a human choose to serve wine? In the left image, the wine glass is preferred to other drinking glasses. In the right image, neither a wine glass nor other drinking glasses are present. The cup is therefore chosen by the human.

提出数据集：COCO-Tasks（14 个任务，40000 张图像）

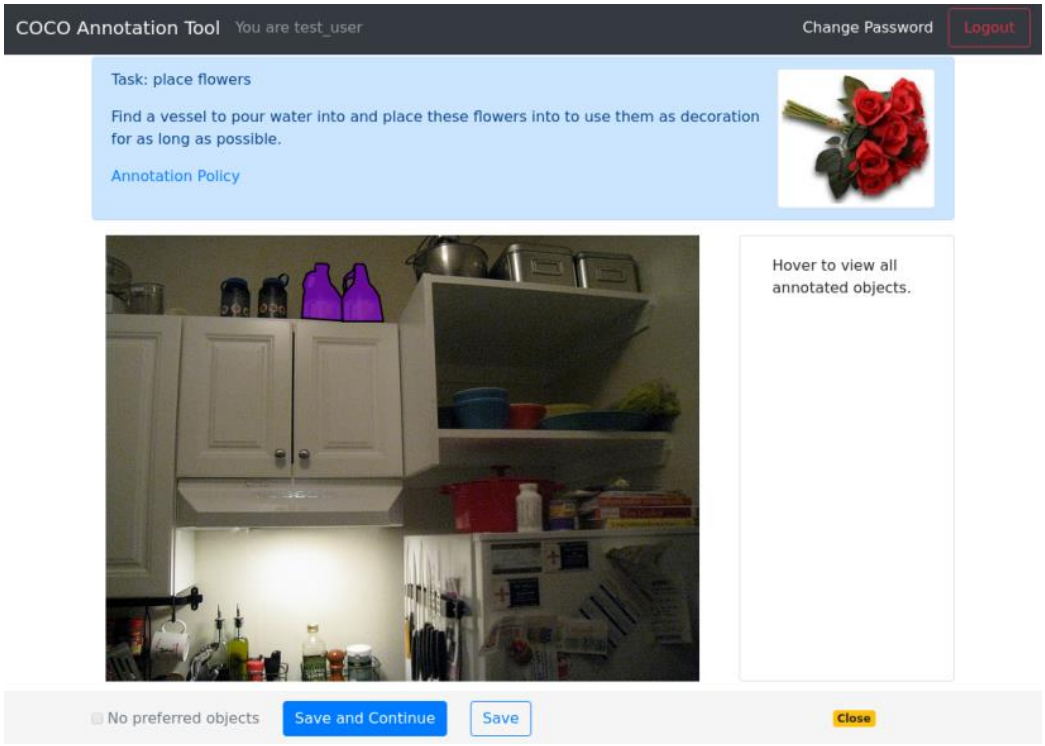
对于每个任务，从 COCO train2014 中选取 3600 张作为训练，COCO val2014 中选取 900 张作为测试（训练集：30,229 张图像，测试集：9,495 张图像）

Task #	Task	COCO supercategories
1	step on something	furniture
2	sit comfortably	furniture
3	place flowers	kitchen, outdoor
4	get potatoes out of fire	sports, kitchen, outdoor
5	water plant	kitchen, indoor
6	get lemon out of tea	kitchen
7	dig hole	sports, kitchen, indoor
8	open bottle of beer	furniture, kitchen, indoor
9	open parcel	kitchen, indoor
10	serve wine	kitchen
11	pour sugar	kitchen
12	smear butter	kitchen
13	extinguish fire	kitchen, indoor
14	pound carpet	sports

确保 40%的图像包含来自这些超级类别的多个类别，40%的图像包含来自这些超级类别的一个类别但有多个实例，10%的图像恰好包含一个类别的一个实例，10%随机抽样。

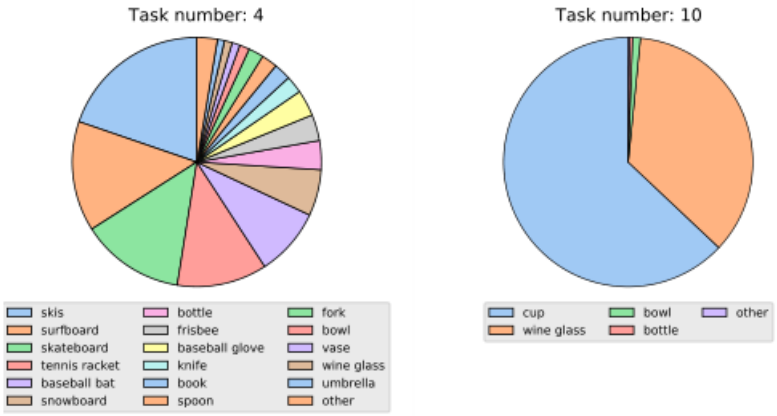


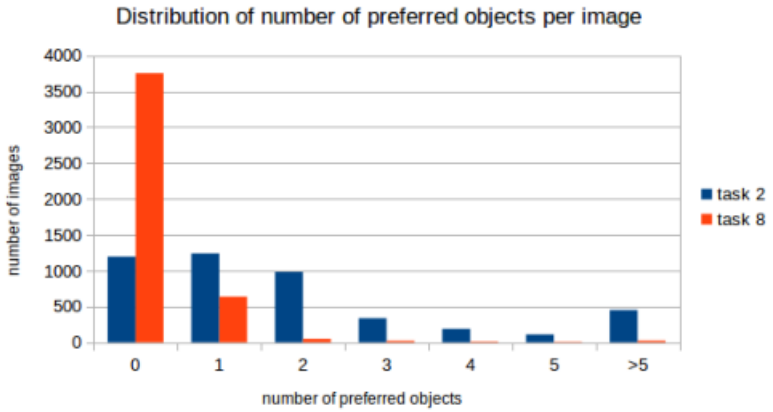
Task1 / Task2 更多的例子



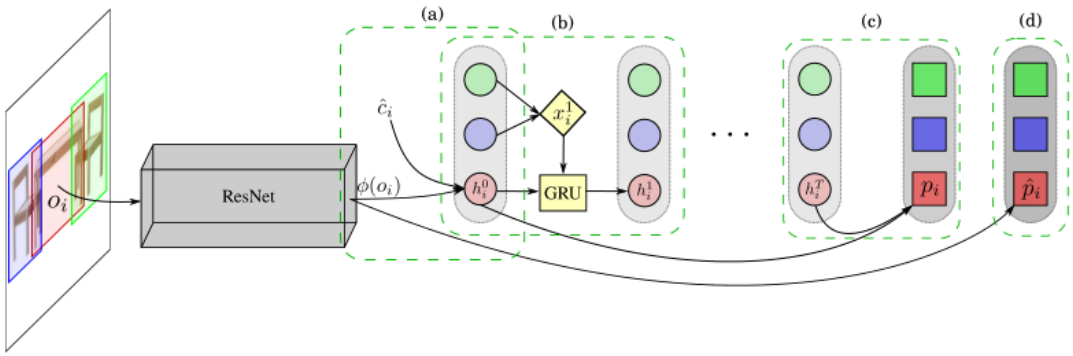
人工标注界面

	Task	Selected object categories	Objects of all selected categories	Objects chosen by humans	Intra class differentiations	Annotation consistency
1	step on something	12	30214	5783	964	0.927
2	sit comfortably	12	31392	9870	1004	0.938
3	place flowers	10	14732	3737	734	0.925
4	get potatoes out of fire	30	32775	6889	525	0.921
5	water plant	13	19050	4043	760	0.918
6	get lemon out of tea	15	22386	4707	661	0.873
7	dig hole	29	34015	6857	402	0.922
8	open bottle of beer	12	18177	1105	373	0.921
9	open parcel	7	7172	1759	160	0.921
10	serve wine	6	19209	3778	566	0.963
11	pour sugar	11	20596	5739	944	0.863
12	smear butter	9	17489	1819	270	0.896
13	extinguish fire	8	14821	2535	272	0.940
14	pound carpet	14	34160	7176	432	0.941





方法：



使用预训练的 ResNet-101，构建门控图神经网络 GGNN（每个节点都是图上的一个对象，每个节点都与所有其他节点相连）。

图像 ***I*** 上有 N 个被检测对象：o_i，bounding box：b_i，检测分数：d_i（使用 COCO 自带的检测框均设置为 1），预测类别：c_i（one-hot 编码），被该任务选取的可能性：p_i，由 ResNet 提取出的特征：$\phi(o_i)$

（a）初始隐藏值：$h_i^0 = g(W_c\hat{c}_i) \odot g(W_\phi\phi(o_i))$

其中，$g(.)$ 为 ReLU，$\odot$ 为逐元素乘法

（b）汇总图中所有其他节点的信息：$x_i^t = \sum_{j,j\neq i} W_p d_j h_j^{t-1} + b_p$

GRU 更新规则（参考其他文献）：

$$
\begin{aligned}
z_i^t &= \sigma(W_z x_i^t + U_z h_i^{t-1} + b_z) \\
r_i^t &= \sigma(W_r x_i^t + U_r h_i^{t-1} + b_r) \\
\hat{h}_i^t &= tanh(W_h x_i^t + U_h(r_i^t \odot h_i^{t-1}) + b_h) \\
h_i^t &= (1 - z_i^t) \odot h_i^{t-1} + z_i^t \odot \hat{h}_i^t
\end{aligned}
$$

更新执行 T 次，本文中 T=3

（c）根据初始和最终隐藏状态计算概率：$p_i = \sigma(f([h_i^0; h_i^T]))$

其中，$f(.)$ 是 2 层的 MLP（多层感知器）

（d）为了使 ResNet 所学习的特征具有判别力，还强制网络仅从单个对象的视觉特征估算适用性得分：$\hat{p}_i = \sigma(\hat{f}(\phi(o_i)))$

最后使用p_i 和 $\hat{p}_i$ 的平均值作为最终概率

实验

Comparison to Baselines mAP@0.5			
	gt bbox	Faster-RCNN detections	Yolo detections
object detector	-	0.206	-
pick best class	0.386	0.141	-
ranker	0.564	0.091	-
classification	0.616	0.288	0.291
proposed + fusion	**0.742**	**0.326**	**0.332**

Comparison to Baselines on Faster-RCNN detections, mAP@0.5															
task #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**avg.**
object detector	0.281	0.258	0.301	0.220	0.305	0.117	0.308	0.00	0.051	0.334	0.097	0.061	0.246	0.309	0.206
pick best class	0.229	0.181	0.198	0.150	0.213	0.058	0.204	0.039	0.033	0.220	0.111	0.05	0.125	0.156	0.141
detection stats	0.246	0.195	0.196	0.142	0.150	0.066	0.162	0.040	0.077	0.218	0.112	0.132	0.093	0.142	0.141
ranker	0.107	0.104	0.115	0.116	0.118	0.033	0.150	0.024	0.046	0.105	0.052	0.050	0.083	0.172	0.091
classification	0.331	0.267	0.368	0.329	0.354	0.146	0.403	0.144	0.176	0.384	0.171	0.245	0.332	0.381	0.288
proposed + fusion	**0.366**	**0.298**	**0.405**	**0.376**	**0.410**	**0.172**	**0.436**	**0.179**	**0.210**	**0.406**	**0.223**	**0.284**	**0.391**	**0.407**	**0.326**
Comparison to Baselines on ground truth bounding boxes, mAP@0.5															
task #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**avg.**
pick best class	0.473	0.637	0.411	0.524	0.392	0.338	0.517	0.098	0.343	0.340	0.453	0.151	0.138	0.585	0.386
detection stats	0.691	0.850	0.638	0.860	0.594	0.660	0.906	0.243	0.679	0.743	0.608	0.575	0.498	0.906	0.675
ranker	0.502	0.687	0.554	0.706	0.604	0.334	0.784	0.215	0.629	0.473	0.404	0.617	0.581	0.812	0.564
classification	0.676	0.762	0.610	0.800	0.549	0.497	0.871	0.265	0.458	0.728	0.435	0.562	0.539	0.870	0.616
proposed + fusion	**0.810**	**0.847**	**0.702**	**0.914**	**0.668**	**0.640**	**0.951**	**0.385**	**0.727**	**0.790**	**0.590**	**0.747**	**0.672**	**0.945**	**0.742**

消融实验：

Ablation experiment results, mAP@0.5		
	gt bbox	Faster-RCNN detections
classifier	0.616	0.288
(a) joint classifier	0.647	0.302
(b) joint classifier + class	0.719	0.301
(c) joint GGNN + class	0.763	0.293
(d) joint GGNN + class + w. aggreg.	-	0.303
(e) proposed	**0.771**	0.318
(f) proposed + fusion	0.742	**0.326**
(g) no visual input	0.589	0.237
(h) no visual input + bounding box	0.412	0.152