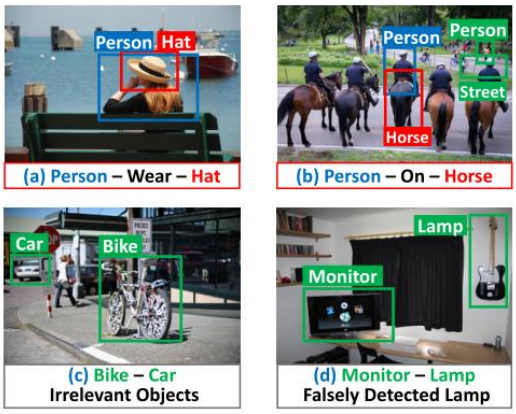
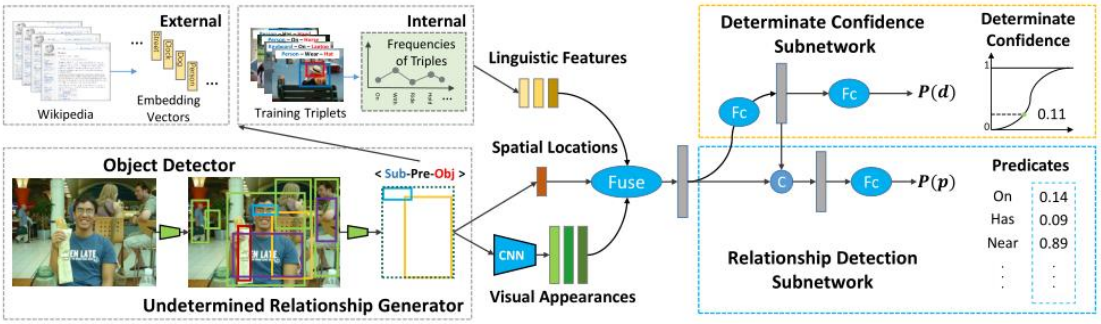
On Exploring Undetermined Relationships for Visual Relationship Detection
视觉关系检测中不确定关系的探索（CVPR-2019）
https://github.com/Atmegal/MFURLN-CVPR-2019-relationship-detection-method

【任务】视觉关系检测
【数据集】VRD、VG
【动机/贡献点】大量的关系未被标注 ==> 利用不确定的关系，改善视觉关系检测的性能
（1）确定关系：人类标记的关系
（2）不确定关系：未标记对象对构成的其他关系（具有关系但未被人类标记的对象对、没有关系的对象对、具有错误检测结果的物体对）



不确定关系可以视作对确定关系的补充？包含否定样本以及人们的不喜欢的偏好（包括不太重要的未标记的关系及异常的表达，如杯子在桌子上）

【方法】提出了一种基于多模式特征的不确定关系学习网络（MF-URLN），来自动生成有用的不确定关系。
（1）使用检测器检测出对象，组成对象对
（2）使用设计的标准将该对象对与人类标记的关系进行比较
（3）提取并融合视觉、空间和语义模态的特征
（4）使用置信度子网络+关系预测子网络，生成不确定关系



$$P(R) = P(p|s,o,d)P(d|s,o)P(s|B_s)P(o|B_o)$$

其中s,p,o,d,R分别代表主语、谓词、宾语、置信度和关系

检测：Faster R-CNN + VGG-16，给出$P(s|B_s)$和$P(o|B_o)$，保留概率大于 0.05 的检测对象
确定关系的标准：主语和宾语的类别与标注一致，且 IoU 均大于 0.5

MF-URLN 的不确定关系学习网络包括两部分：**多模态特征提取网络**和**关系学习网络**

（一）多模态特征提取网络：视觉特征+空间特征+语义特征

（1）视觉特征：主语和宾语分别的边界框，边界框的并集

（2）空间特征：利用三组边界框

$(x^s_{min}, y^s_{min}, x^s_{max}, y^s_{max})$ 、 $(x^o_{min}, y^o_{min}, x^o_{max}, y^o_{max})$ 和 $(x^u_{min}, y^u_{min}, x^u_{max}, y^u_{max})$

$$[\frac{x^s_{min}-x^u_{min}}{x^u_{max}-x^u_{min}}, \frac{y^s_{min}-y^u_{min}}{y^u_{max}-y^u_{min}}, \frac{x^s_{max}-x^u_{max}}{x^u_{max}-x^u_{min}}, \frac{y^s_{max}-y^u_{max}}{y^u_{max}-y^u_{min}}, \frac{x^o_{min}-x^u_{min}}{x^u_{max}-x^u_{min}}, \frac{y^o_{min}-y^u_{min}}{y^u_{max}-y^u_{min}}, \frac{x^o_{max}-x^u_{max}}{x^u_{max}-x^u_{min}}, \frac{y^o_{max}-y^u_{max}}{y^u_{max}-y^u_{min}}]$$

分别表示主语和宾语两个对象相对于整个并集框的位置

（3）语义特征：外部+内部

外部：Wikipedia 2014 上预训练 word2vec（噪声可能较大）

内部：朴素贝叶斯+拉普拉斯平滑（为了适应 zero-shot），对训练集所有的关系三元组频率进行统计，转换为主语和宾语的概率分布

（4）特征融合：

由于高维特征（例如 4096 维视觉特征）会轻易淹没低维特征（例如 8 维空间特征），因此先将特征转换为相同的维度，然后再进行多模态特征的融合。

（二）关系学习网络：置信度子网络+关系预测子网络

（1）置信度子网络

确定对象对的确定置信度，反映了对象对被人工选择和标注的可能性（即具有确定关系的概率）

交叉熵损失：$$\mathrm{CE}(p,y)=\begin{cases}-\log(p) & \text{if} \quad y=1\\ -\log(1-p) & \text{otherwise.}\end{cases}$$

使用上标d和i表示确定和不确定的关系，确定的关系 y=1，不确定的关系 y=0

对于确定的关系，$L^d_{det} = \mathrm{CE}(P(d^d|s^d,o^d),1)$

对于不确定的关系，$L^i_{det} = \mathrm{CE}(P(d^i|s^i,o^i),0)$

总的置信度损失函数：$L_{det}=L^d_{det}+\alpha L^i_{det},$（文中 α =1，认为两者同等重要）

（2）关系预测子网络

确定的关系：$$L^d_{rel}=\sum_{k=1}^{M}\mathrm{CE}(P(p^d_k|s^d,o^d,d^d),y_k)$$ 第 k 个谓词，M 是谓词的数量

不确定的关系，视为没有谓词：$$L^i_{rel}=\sum_{k=1}^{M}\mathrm{CE}(P(p^i_k|s^i,o^i,d^i),0).$$

*当前没有可靠的方法来自动标记这些不确定的关系

总的关系预测损失函数：$L_{rel}=L^d_{rel}+\lambda_1 L^i_{rel}.$

==> 联合损失函数：$L=L_{rel}+\lambda_2 L_{det}.$ ==> $L=L^d_{rel}+\lambda_1 L^i_{rel}+\lambda_2 L^d_{det}+\lambda_2 L^i_{det}.$

【实验】

数据集：VRD（5000 张图像，100 个对象类型，70 个谓词；3700 训练+300 验证+1000 测试，1169 个关系三元组只出现在测试集）、VG（划分：99652 张图像，200 个对象类别，100 个谓词；68794 训练+5000 验证+25858 测试）

任务：谓词检测，短语检测和关系检测

指标：Recall @ N（R_N）

Table 1. Performance comparison of visual relationship detection methods on the VRD dataset. Pre., Phr., and Rel. represent predication detection, phrase detection, and relation detection, respectively. "-" denotes that the result is unavailable.

		Pre.	Phr.		Rel.	
		$R_{50/100}$	R_{50}	R_{100}	R_{50}	R_{100}
语义知识	VRD-Full [23]	47.9	16.2	17.0	13.9	14.7
端到端网络	VTransE [37]	44.8	19.4	22.4	14.1	15.2
端到端网络	VIP-CNN [20]	-	22.8	27.9	17.3	20.0
其它方法	Weak-S [26]	52.6	17.9	19.5	15.8	17.1
其它方法	PPRFCN [38]	47.4	19.6	23.2	14.4	15.7
语义知识	LKD:S [34]	47.5	19.2	20.0	16.6	17.7
语义知识	LKD:T [34]	54.1	22.5	23.6	18.6	20.6
语义知识	LKD:S+T [34]	55.2	23.1	24.0	19.2	21.3
端到端网络	DVSRL [22]	-	21.4	22.6	18.2	20.8
端到端网络	TFR [15]	52.3	17.4	19.1	15.2	16.8
深度结构学习	DSL [41]	-	22.7	24.0	17.4	18.3
其它方法	STA [32]	48.0	-	-	-	-
其它方法	Zoom-Net [33]	50.7	24.8	28.1	18.9	21.4
其它方法	CAI+SCA-M [33]	56.0	25.2	28.9	19.5	22.4
其它方法	VSA [12]	49.2	19.1	21.7	16.0	17.7
	MF-URLN	**58.2**	**31.5**	**36.1**	**23.9**	**26.8**

Table 2. Performance comparison of six methods on the VG dataset. "-" denotes that the result is unavailable.

	Pre.		Phr.		Rel.	
	R_{50}	R_{100}	R_{50}	R_{100}	R_{50}	R_{100}
VTransE [37]	62.6	62.9	9.5	10.5	5.5	6.0
PPRFCN [38]	64.2	64.9	10.6	11.1	6.0	6.9
DSL [41]	-	-	13.1	15.6	6.8	8.0
STA [32]	62.7	62.9	-	-	-	-
VSA [12]	64.4	64.5	9.7	10.0	6.0	6.3
MF-URLN	**71.9**	**72.2**	**26.6**	**32.1**	**14.4**	**16.5**

Table 3. Performance comparison on the zero-shot set of the VRD dataset. "-" denotes that the result is unavailable.

	Pre.	Phr.		Rel.	
	$R_{50/100}$	R_{50}	R_{100}	R_{50}	R_{100}
VRD-Full [23]	12.3	5.1	5.7	4.8	5.4
VTransE [37]	-	2.7	3.5	1.7	2.1
Weak-S [26]	21.6	6.8	7.8	6.4	7.4
LKD:S [34]	17.0	**10.4**	**10.9**	**8.9**	**9.1**
LKD:T [34]	8.8	6.5	6.7	6.1	6.4
DVSRL [22]	-	9.2	10.3	7.9	8.5
TFR [15]	17.3	5.8	7.1	5.3	6.5
STA [32]	20.6	-	-	-	-
MF-URLN	26.9	5.9	7.9	4.3	5.5
MF-URLN-IM	**27.2**	6.2	9.2	4.5	6.4

Zero-shot 效果不好：一些 unseen 的确定关系被错误地分类为不确定，仍然需要更好的策略来产生和利用不确定的关系

消融实验：V 视觉，S 空间，$L_{ex,in}$ 外部和内部语义

Table 4. R_{50} predicate detection and relation detection of the MF-URLN and its eight variants on the VRD dataset.

	Transforming		Concatenating	
	Pre.	Rel.	Pre.	Rel.
Baseline: V	52.29	22.64	53.01	22.85
Baseline: $L_{ex,in}$	53.39	18.49	53.94	18.07
Baseline: S	43.43	17.94	43.44	17.95
V+S	54.66	23.15	52.36	22.75
V+$L_{ex,in}$	57.27	23.21	55.45	22.62
$L_{ex,in}$+S	57.10	23.67	**56.04**	**23.29**
V+S+L_{in}	56.87	23.15	53.25	22.51
V+S+L_{ex}	57.69	23.50	55.29	22.83
MF-URLN	**58.22**	**23.89**	55.77	22.61

需要考虑更好的多特征融合策略