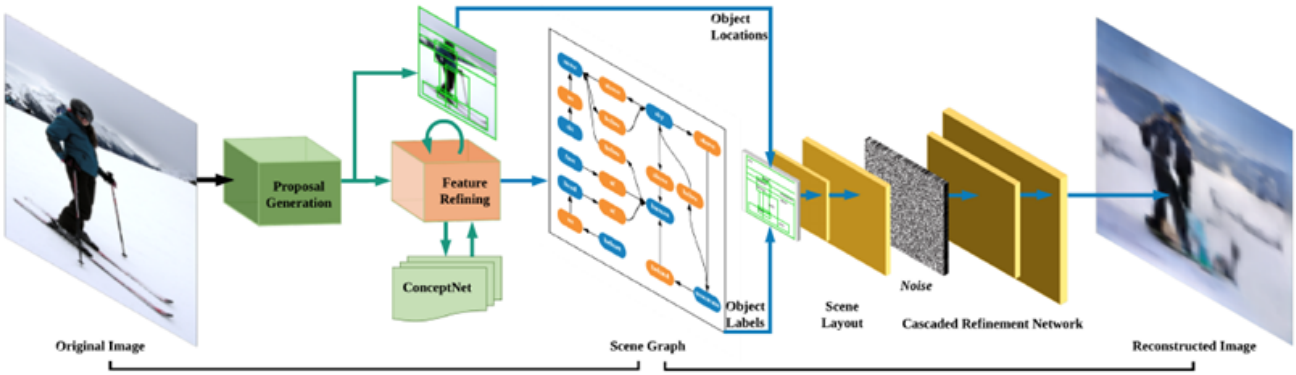
Scene Graph Generation with External Knowledge and Image Reconstruction 论文笔记

整体框架

•作者通过引入外部常识知识库的方式来解决场景图生成所面对的数据集偏差过大，含有噪声的问题。

•作者在提取完图片的场景图以后，同过外部常识库对提取到的结果进行进一步的修正。

•通过根据场景图重建图片，对提取的场景图的准确度进行了进一步的约束。



Proposal 框生成

通过RPN网络提取物体proposal框。

$$[o_0, \cdots, o_{N-1}] = f_{\mathrm{RPN}}(\mathbf{I})$$

为了降低计算量，作者对提取到的框划分子图。特别地，对于一对proposal，子图建议根据两个目标得分的乘积得到的置信度得分来构造为联合bbox。因此一个关系可表示为：

$$\langle o_i, o_j, s_k^i \rangle$$

根据外部知识库进行特征 refine

物体和子图交替refine： 根据每个顶点邻域的信息，更新每个物体的编码。

$$\bar{\mathbf{o}}_i = \mathbf{o}_i + f_{s\to o}\left(\sum_{\mathbf{s}_k^i \in \mathbf{S}^i} \alpha_k^{s\to o} \cdot \mathbf{s}_k^i\right)$$

$$\bar{\mathbf{s}}_k = \mathbf{s}_k + f_{o\to s}\left(\sum_{\mathbf{o}_i^k \in \mathbf{O}^k} \alpha_i^{o\to s} \cdot \mathbf{o}_i^k\right)$$

为了解决数据集中数据集中视觉关系的偏差，作者借助外部知识库提出一种特征refine网络。作者首先预测每一个物体的类别a_i，然后根据a_i从外部知识库中匹配。

$$a_i \xrightarrow{\text{retrieve}} \langle a_i, a^r_{i,j}, a^o_j, w_{i,j} \rangle, j \in [0, K-1]$$

where $a^r_{i,j}$, a^o_j and $w_{i,j}$ are the top-K corresponding relationships, the object entity and the weight, respectively.

权重的计算是依靠conceptNet 来评估每个关系的常见性得到的。然后将每一个$< a_i, a^r_{i,j}, a^o_j >$匹配到一个句子序列$< X^0, X^1 \dots X^{T_a-1} >$。然后$x_t = W_e X_t$得到每个单词的编码向量，最后通过双向GRU得到关系的编码。

$$\mathbf{h}^t_k = \mathbf{RNN}_{\text{fact}}(\mathbf{x}^t_k, \mathbf{h}^{t-1}_k),\ t \in [0, T_a - 1] \quad .$$

基于注意力机制的知识融合：对于N 个物体，每个物体对应知识库中K个关系向量。作者提出了一种升级版DMN框架。

$$\begin{aligned}
\mathbf{q} &= \tanh(\mathbf{W}_q \bar{\mathbf{o}} + \mathbf{b}_q) \\
\mathbf{z}^t &= [\mathbf{F} \circ \mathbf{q}; \mathbf{F} \circ \mathbf{m}^{t-1}; |\mathbf{F} - \mathbf{q}|; |\mathbf{F} - \mathbf{m}^{t-1}|] \\
\mathbf{g}^t &= \text{softmax}(\mathbf{W}_1 \tanh(\mathbf{W}_2 \mathbf{z}^t + \mathbf{b}_2) + \mathbf{b}_1) \\
\mathbf{e}^t &= \text{AGRU}(\mathbf{F}, \mathbf{g}^t)
\end{aligned}$$

其中F是知识集合f_k，AGRU为带注意力机制的双向GRU。

$$\mathbf{e}^t_k = g^t_k \text{GRU}(\mathbf{f}_k, \mathbf{e}^t_{k-1}) + (1 - g^t_k)\mathbf{e}^t_{k-1}$$

更新记忆：

$$\mathbf{m}^t = \text{ReLU}(\mathbf{W}_m[\mathbf{m}^{t-1}; \mathbf{e}^t_K; \mathbf{q}] + \mathbf{b}_m).$$

得到最终的向量：

$$\tilde{\mathbf{o}} = \text{ReLU}(\mathbf{W}_c[\bar{\mathbf{o}}; \mathbf{m}^{T_m - 1}] + \mathbf{b}_c)$$

最后完成预测：

$$\begin{aligned}
\mathbf{P}_{i,j} &\sim \text{softmax}(f_{\text{rel}}([\tilde{\mathbf{o}}_i \otimes \bar{\mathbf{s}}_k; \tilde{\mathbf{o}}_j \otimes \bar{\mathbf{s}}_k; \bar{\mathbf{s}}_k])) \\
\mathbf{V}_i &\sim \text{softmax}(f_{\text{node}}(\tilde{\mathbf{o}}_i))
\end{aligned}$$

场景图级别的监督：

$$\mathcal{L}_{\text{im2sg}} = \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{reg}}\mathbf{1}_{u \geq 1}\mathcal{L}_{\text{reg}}$$

图片级别的监督：

将每一个物体的向量扩展为8*8维度，然后将所有物体的向量加在一起求和。通过一个cascaded refinement network生成图片。损失函数包括l1重建损失和对抗损失。

$$\mathcal{L}_{D_i} = \mathbb{E}_{I \sim p_{\text{real}}}[\log D_i(\mathbf{I})]$$
$$\mathcal{L}_{G_i} = \mathbb{E}_{\hat{I} \sim p_{\text{G}}}[\log(1 - D_i(\hat{\mathbf{I}})] + \lambda_p \mathcal{L}_{\text{pixel}}$$

实验结果

Table 5: Comparison with existing methods on **PhrDet** and **SGGen**.

Dataset	Model	PhrDet		SGGen	
		Rec@50	Rec@100	Rec@50	Rec@100
VRD [29]	ViP-CNN [27]	22.78	27.91	17.32	20.01
	DR-Net [5]	19.93	23.45	17.73	20.88
	U+W+SF+LK: T+S [45]	26.32	29.43	19.17	21.34
	Factorizable Net [25]	26.03	30.77	18.32	21.20
	KB-GAN	**27.39**	**34.38**	**20.31**	**25.01**
VG-MSDN [26]	ISGG [41]	15.87	19.45	8.23	10.88
	MSDN [26]	19.95	24.93	10.72	14.22
	Graph R-CNN [42]	–	–	11.40	13.70
	Factorizable Net [25]	22.84	28.57	13.06	16.47
	KB-GAN	**23.51**	**30.04**	**13.65**	**17.57**