基于结构和视觉联合编码的场景图谱补全表示学习方法

Representation Learning for Scene Graph Completion via Jointly Structural and Visual Embedding

罗永豪 (16214343)

工程（软件工程）
数据科学与计算机学院

目录

1 绪论

2 预备知识

3 RLSV 模型

4 实验设计与分析

5 总结与期望

6 附录

目录

1 绪论

2 预备知识

3 RLSV 模型

4 实验设计与分析

5 总结与期望

6 附录

场景图谱 (Scene Graph)

场景图谱是对图片所描述场景的一种基于图的、半结构化的表示形式[1]：

- 视觉三元组（visual triple）的集合：{(头部实体，关系，尾部实体)}
- 有向图：节点表示实体（entity），有向边表示关系（relation）
- 实体的构成：实体类型、实体附带的属性、实体的包围盒

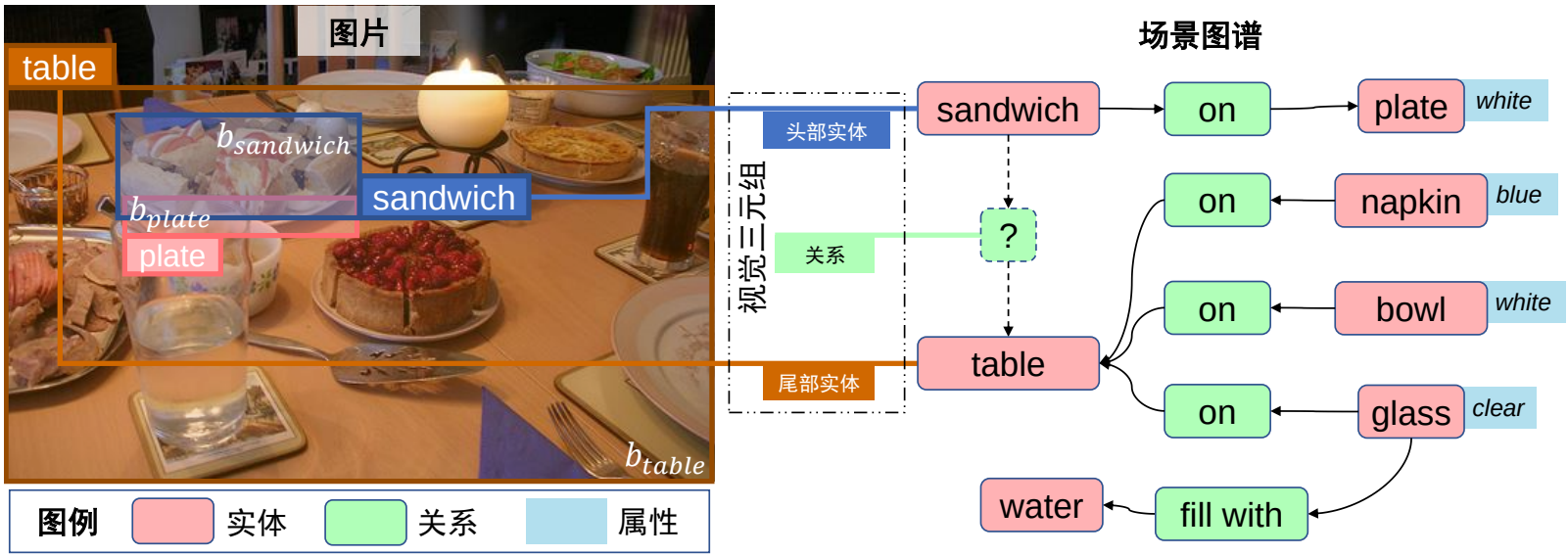


图 1-1: 来自 Visual Genome 场景图谱数据集[2]的一张示例图片和它的场景图谱

场景图谱的应用与产生

场景图谱成为了很多计算机视觉以及人工智能应用的重要资源，例如：

- 图像检索[1] (Johnson *et al.*, 2015)
- 图像的视觉问答[3] (Zhu *et al.*, 2017)

特点：实体向量化、考虑图结构 → 完整的场景图谱可以提供丰富的资源

场景图谱的产生方法：

- 人工标注：众包收集的 Visual Genome 数据集[2] (Krishna *et al.*, 2017)
- 自动生成：物体检测器 + 视觉关系建模，如 VRD[4] (Lu *et al.*, 2016)，VTransE[5] (Zhang *et al.*, 2017a)，PPR-FCN[6] (Zhang *et al.*, 2017b)

特点：关系分类器，但准确度较低

场景图谱的不完整问题

尽管是在 Visual Genome[2] 这样质量较高的场景图谱数据集中，我们也很容易发现有两个物体之间的关系没有被标注列出。

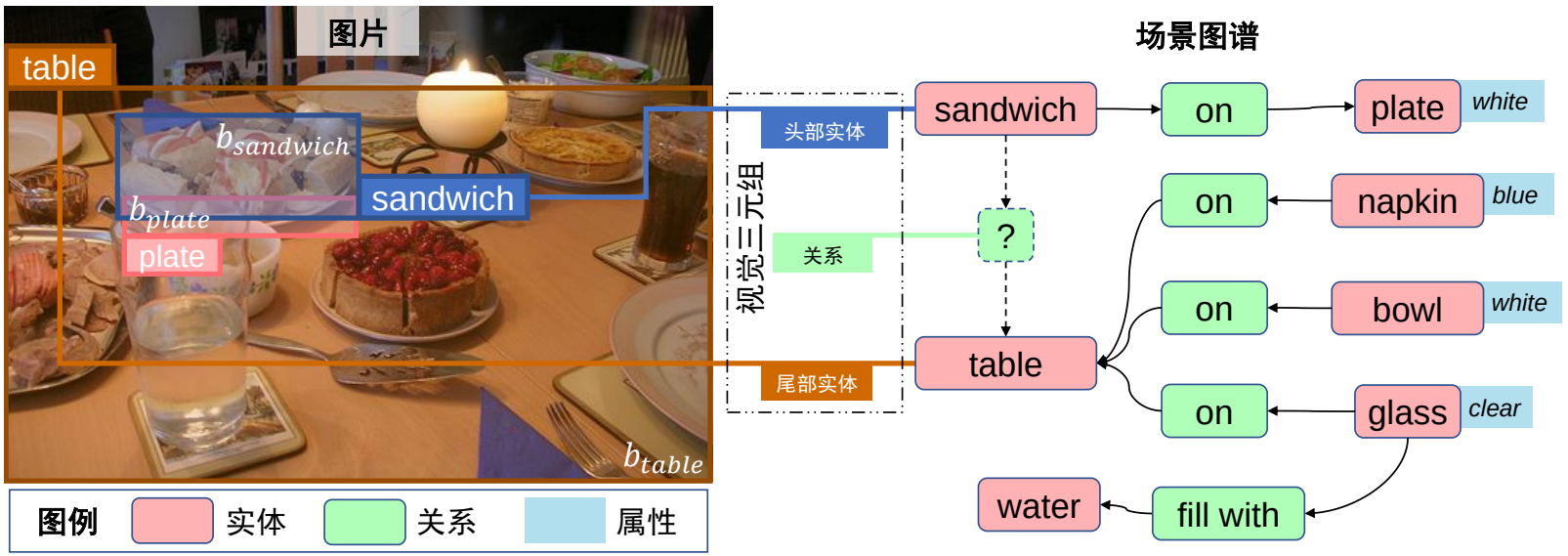


图 1-1: 来自 Visual Genome 场景图谱数据集[2]的一张示例图片和它的场景图谱

场景图谱的不完整问题

尽管是在 Visual Genome[2] 这样质量较高的场景图谱数据集中，我们也很容易发现有两个物体之间的关系没有被标注列出。

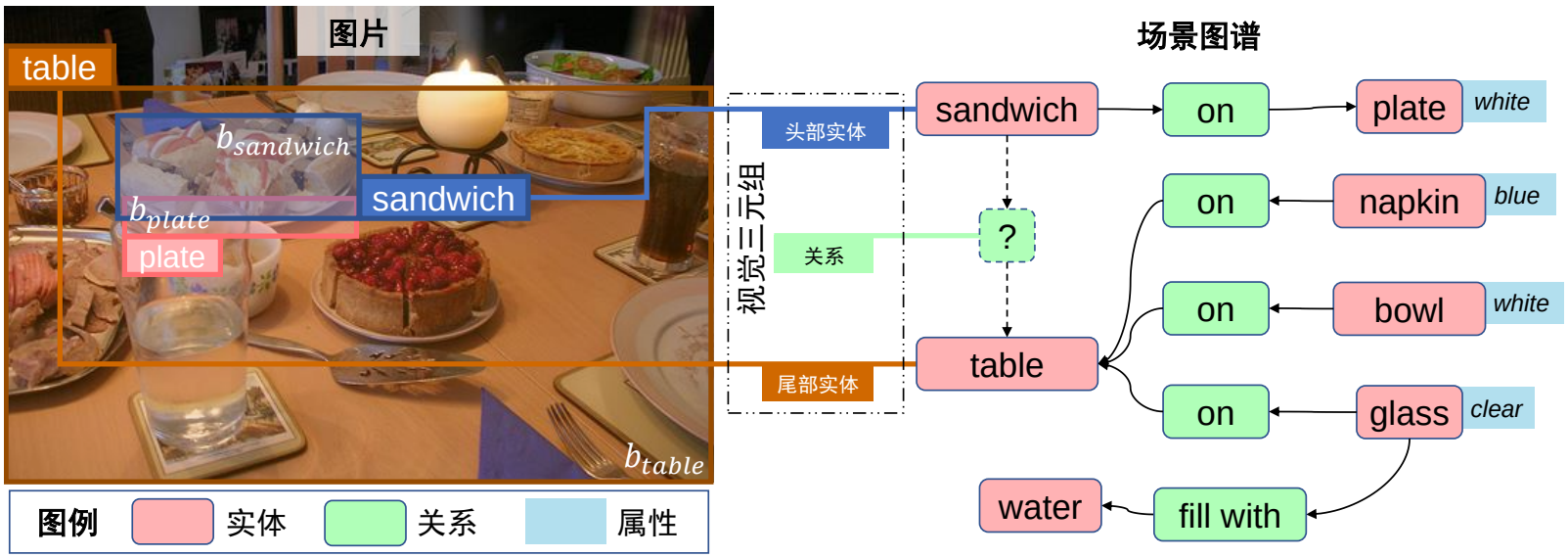


图 1-1: 来自 Visual Genome 场景图谱数据集[2]的一张示例图片和它的场景图谱

结合已有的场景图谱去推测一些缺失的视觉三元组

知识图谱 (Knowledge Graph)

知识图谱是一种包含多种关系的数据：

- 事实三元组（fact triple）的集合：{(头部实体，关系，尾部实体)}
- 实体的构成：实体类型

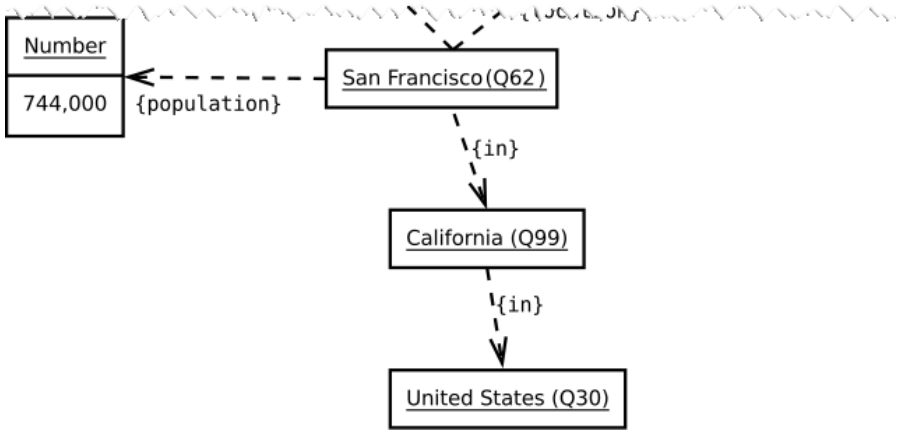


图 1-2: Wikidata 上的一小部分知识图谱示例图

知识表示学习（knowledge representation learning）：

- 为了解决知识图谱的不完整问题
- 实体和关系表示成低维度的稠密向量
- 翻译模型：TransE[7] (Bordes *et al.*, 2013)，TransD[8] (Ji *et al.*, 2015)

场景图谱和知识图谱的不同点

表 1-1: 场景图谱和知识图谱的不同点对照表

	场景图谱	知识图谱
三元组类型	视觉三元组	事实三元组
三元组例子	$(sandwich, on, plate)$	$(San\ Francisco, in, California)$
实体的涵盖面	实体类型、属性、包围盒	实体类型
三元组表示的内容	与具体某张图片相关	现实世界成立的事实
图谱的可合并性	实体类型相同但实例有多个	实体类型相同且实例一致

场景图谱和知识图谱的不同点

表 1-1: 场景图谱和知识图谱的不同点对照表

	场景图谱	知识图谱
三元组类型	视觉三元组	事实三元组
三元组例子	$(sandwich, on, plate)$	$(San\ Francisco, in, California)$
实体的涵盖面	实体类型、属性、包围盒	实体类型
三元组表示的内容	与具体某张图片相关	现实世界成立的事实
图谱的可合并性	实体类型相同但实例有多个	实体类型相同且实例一致

结构信息：实体和关系组成的三元组信息

视觉信息：实体和关系在图像之上的反映

根据场景图谱的特性修改翻译模型

本文工作

1. 对场景图谱以及场景图谱补全任务给出了形式化定义
2. 提出了结构和视觉联合编码表示学习（RLSV）模型
   - 视觉特征提取模块：得到视觉三元组的视觉向量编码
   - 层级投影模块：融合视觉三元组的结构编码和视觉编码
3. 通过场景图谱补全的任务验证了 RLSV 模型的有效性
   - 链接预测
   - 视觉三元组分类

目录

1 绪论

2 预备知识

3 RLSV 模型

4 实验设计与分析

5 总结与期望

6 附录

场景图谱的形式化表示

- $\mathcal{I}$：全体图片组成的集合
- $\mathcal{R}$：全体关系的集合
- $\mathcal{E}_t$：全体实体类型的集合
- $\mathcal{A}$：全体属性的集合

定义 2.1 (场景图谱)

给定图片 $I \in \mathcal{I}$，其场景图谱是一个视觉三元组集合 $\mathcal{T}_I \subseteq \mathcal{E}_I \times \mathcal{R}_I \times \mathcal{E}_I$：

- $\mathcal{E}_I$：图片 I 中的实体集合；
- $\mathcal{R}_I$：图片 I 中的关系集合且 $\mathcal{R}_I \subseteq \mathcal{R}$；
- 实体：$e_{I,k} = (e_{t,I,k}, \mathcal{A}_{I,k}, b_{I,k}) \in \mathcal{E}_I$，包括了实体类型 $e_{t,I,k} \in \mathcal{E}_t$，实体附带的属性集合 $\mathcal{A}_{I,k} \subseteq \mathcal{A}$，包围盒 $b_{I,k}$；
- 视觉三元组：$(h, r, t) \in \mathcal{T}_I$，其中 $h, t \in \mathcal{E}_I$ 且 $r \in \mathcal{R}_I$。

例 2.1

在图 1-1 所示的场景图谱 $\mathcal{T}_I$ 中，有 $|\mathcal{E}_I| = 7$、$|\mathcal{R}_I| = 2$、$|\mathcal{T}_I| = 5$。
视觉三元组 $((sandwich, \{\}, b_{sandwich}), on, (plate, \{white\}, b_{plate}))$，可简写成 $(sandwich, on, plate)$。

场景图谱补全的形式化表示

场景图谱补全就是利用现有的场景图谱以及图片的信息，向已有的场景图谱添加更多视觉三元组的过程，使得补全后的场景图谱更加完整。

定义 2.2 (场景图谱补全)

给定图片 $I \in \mathcal{I}$ 及其场景图谱 $\mathcal{T}_I$，若 $\exists e_{I,p}, e_{I,q} \in \mathcal{E}_I$ 且不存在一个关系 $r_I \in \mathcal{R}_I$ 使得 $(e_{I,p}, r_I, e_{I,q}) \in \mathcal{T}_I$ 成立，那么，场景图谱补全就是要把 $\mathcal{T}_I$ 扩展成 $\mathcal{T}_I^+ = \mathcal{T}_I \cup \{(e_{I,p}, r_I^+, e_{I,q}) | e_{I,p}, e_{I,q} \in \mathcal{E}_I, r_I^+ \in \mathcal{R}, p \neq q\}$，使得 $|\mathcal{T}_I^+| > |\mathcal{T}_I|$。

例 2.2

图 1-1 所示的场景图谱 $\mathcal{T}_I$ 中，实体 $sandwich$ 与 $table$ 之间存在关系的缺失：

- 以向 $\mathcal{T}_I$ 添加一条新的视觉三元组 $(sandwich, on, table)$；
- 把 $\mathcal{T}_I$ 扩充成 $\mathcal{T}_I^+ = \mathcal{T}_I \cup \{(sandwich, on, table)\}$；
- 视觉三元组的总数增加，即 $|\mathcal{T}_I^+| = 6 > |\mathcal{T}_I| = 5$。

结构信息和视觉信息向量化

知识图谱的结构信息向量化

- 翻译模型：简单有效，参数量适中（如 TransE[7]，TransD[8]等）

结构信息和视觉信息向量化

知识图谱的结构信息向量化

- 翻译模型：简单有效，参数量适中（如 TransE[7]，TransD[8]等）

TransE[7]

对知识图谱 $\Delta = \{(h_t, r, t_t) | h_t, t_t \in \mathcal{E}_t, r \in \mathcal{R}\}$，给定一条三元组 (h_t, r, t_t)，当这条三元组成立时，TransE 希望 $\boldsymbol{h}_t + \boldsymbol{r} \approx \boldsymbol{t}_t$。因此，TransE 的评分函数为：

$$E(h_t, r, t_t) = |\boldsymbol{h}_t + \boldsymbol{r} - \boldsymbol{t}_t|_{L_1/L_2}, \tag{2-1}$$

其中，$\boldsymbol{h}_t, \boldsymbol{t}_t, \boldsymbol{r} \in \mathbb{R}^d$，$d$ 表示维度，L_1/L_2 表示使用向量的一范数或二范数。

结构信息和视觉信息向量化

知识图谱的结构信息向量化

- 翻译模型：简单有效，参数量适中（如 TransE[7]，TransD[8]等）

TransE[7]

对知识图谱 $\Delta = \{(h_t, r, t_t) | h_t, t_t \in \mathcal{E}_t, r \in \mathcal{R}\}$，给定一条三元组 (h_t, r, t_t)，当这条三元组成立时，TransE 希望 $\boldsymbol{h}_t + \boldsymbol{r} \approx \boldsymbol{t}_t$。因此，TransE 的评分函数为：

$$E(h_t, r, t_t) = |\boldsymbol{h}_t + \boldsymbol{r} - \boldsymbol{t}_t|_{L_1/L_2}, \tag{2-1}$$

其中，$\boldsymbol{h}_t, \boldsymbol{t}_t, \boldsymbol{r} \in \mathbb{R}^d$，$d$ 表示维度，L_1/L_2 表示使用向量的一范数或二范数。

图像的视觉信息向量化

- 卷积神经网络（如 VGG[9]，ResNet[10]等）

目录

1 绪论

2 预备知识

3 RLSV 模型

4 实验设计与分析

5 总结与期望

6 附录

基本框架

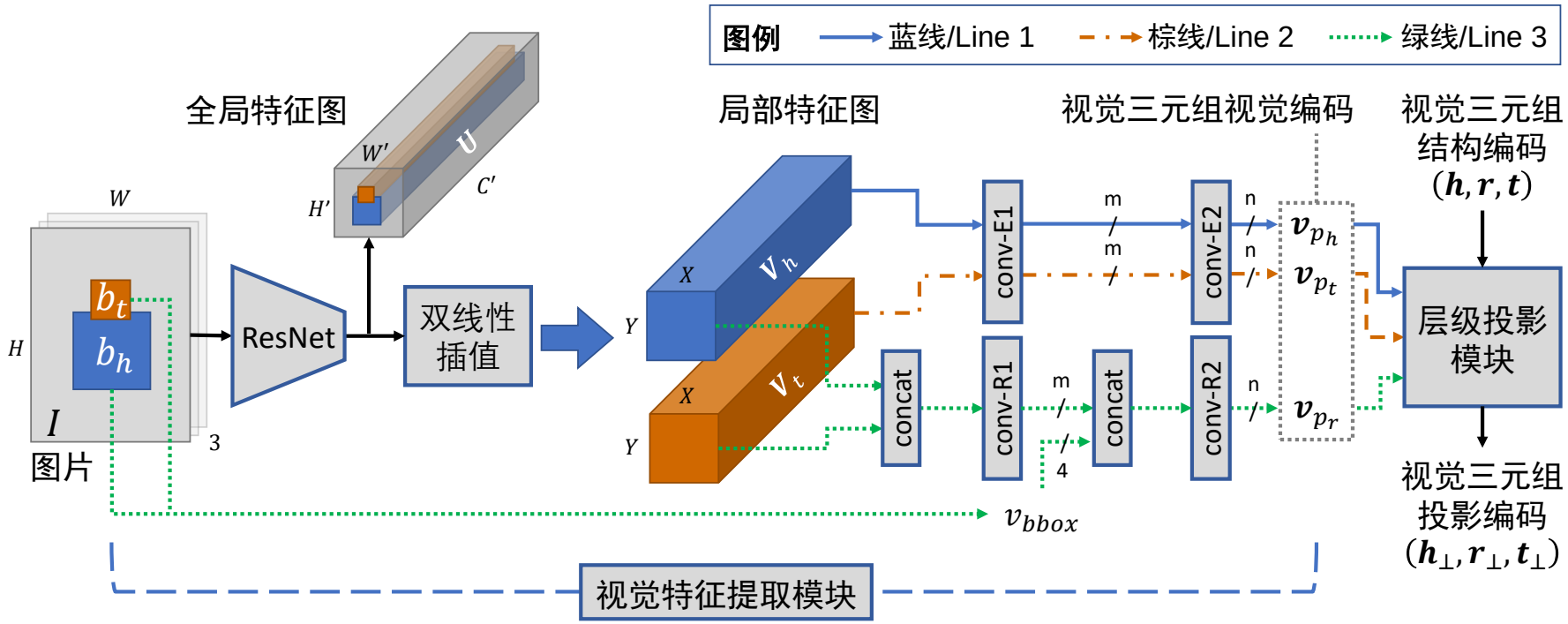


图 3-1: RLSV 模型的基本框架图

根据公式 (2-1)，RLSV 模型的视觉三元组评分函数定义如下：

$$E_I(h, r, t) = |\boldsymbol{h}_\perp + \boldsymbol{r}_\perp - \boldsymbol{t}_\perp|_{L_1/L_2}。 \tag{3-1}$$

- 在以下的讨论中，我们假设所有的向量编码都是一个 n 维的向量

层级投影模块

层级投影

$$\begin{aligned} \boldsymbol{h}_\perp &= \mathbf{M}_h^v \mathbf{M}_h^r \mathbf{M}_h^a \boldsymbol{h}, \\ \boldsymbol{t}_\perp &= \mathbf{M}_t^v \mathbf{M}_t^r \mathbf{M}_t^a \boldsymbol{t}, \\ \boldsymbol{r}_\perp &= \mathbf{M}_r^v \boldsymbol{r}。 \end{aligned} \tag{3-2}$$

参考 TransD 构造动态投影矩阵，设投影向量 $\boldsymbol{e}_p$，$\boldsymbol{r}_p$，$\boldsymbol{a}_p \in \mathbb{R}^n$。

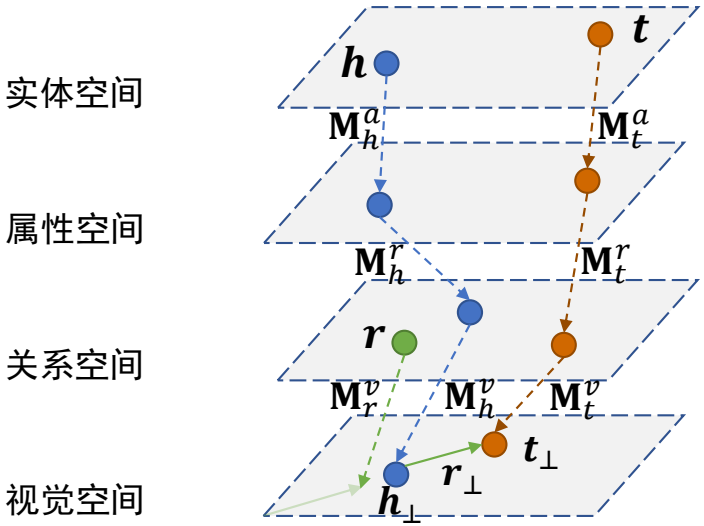


图 3-2: 层级投影模块示意图

层级投影模块

层级投影

$$\begin{aligned} \boldsymbol{h}_\perp &= \mathbf{M}_h^v \mathbf{M}_h^r \mathbf{M}_h^a \boldsymbol{h}, \\ \boldsymbol{t}_\perp &= \mathbf{M}_t^v \mathbf{M}_t^r \mathbf{M}_t^a \boldsymbol{t}, \\ \boldsymbol{r}_\perp &= \mathbf{M}_r^v \boldsymbol{r}。 \end{aligned} \tag{3-2}$$

参考 TransD 构造动态投影矩阵，设投影向量 $\boldsymbol{e}_p$，$\boldsymbol{r}_p$，$\boldsymbol{a}_p \in \mathbb{R}^n$。

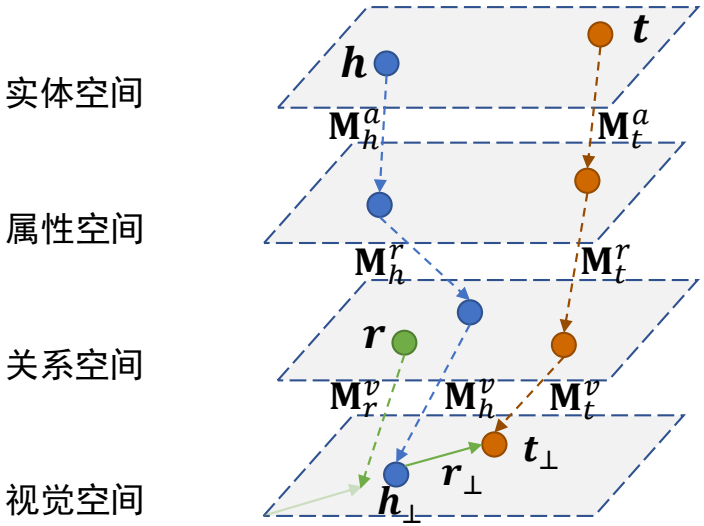


图 3-2: 层级投影模块示意图

属性空间

$$\mathbf{M}_e^a = \left(\sum_{i=1}^{N_a} \beta_i \boldsymbol{a}_{ip}\right) \boldsymbol{e}_p^{\mathrm{T}} + \mathbf{I}^{n\times n}, \tag{3-3}$$

$$\varepsilon_i = \boldsymbol{w}_b^T \tanh(\mathbf{W}_v \boldsymbol{v}_{p_e} + \mathbf{W}_a \boldsymbol{a}_{ip}), \tag{3-4}$$

$$\beta_i = \mathrm{softmax}(\varepsilon_i)。 \tag{3-5}$$

关系空间

$$\begin{aligned} \mathbf{M}_h^r &= \boldsymbol{r}_p \boldsymbol{v}_{p_h}^{\mathrm{T}} + \mathbf{I}^{n\times n}, \\ \mathbf{M}_t^r &= \boldsymbol{r}_p \boldsymbol{v}_{p_t}^{\mathrm{T}} + \mathbf{I}^{n\times n}。 \end{aligned} \tag{3-6}$$

视觉空间

$$\begin{aligned} \mathbf{M}_h^v &= \boldsymbol{v}_{p_h} \boldsymbol{h}_p^{\mathrm{T}} + \mathbf{I}^{n\times n}, \\ \mathbf{M}_t^v &= \boldsymbol{v}_{p_t} \boldsymbol{t}_p^{\mathrm{T}} + \mathbf{I}^{n\times n}, \\ \mathbf{M}_r^v &= \boldsymbol{v}_{p_r} \boldsymbol{r}_p^{\mathrm{T}} + \mathbf{I}^{n\times n}。 \end{aligned} \tag{3-7}$$

目标函数

根据公式 (3-1)，我们希望一条成立的视觉三元组 (h, r, t) 的分数尽量低。因此，目标函数是一个最大化间隔函数（max-margin function）。对于单张图片 I 及其场景图谱 $\mathcal{T}_I$ 有：

$$L_\theta(I, \mathcal{T}_I) = \sum_{(h,r,t)\in\mathcal{T}_I} \sum_{(h',r',t')\in\mathcal{T}'_I} [E_I(h, r, t) + \gamma - E_I(h', r', t')]_+, \quad (3\text{-}8)$$

其中 $[x]_+ \triangleq \max(0, x)$， γ 是间隔距离超参数，θ 表示所有可学习参数。

$\mathcal{T}'_I$ 则表示从正样本集合 $\mathcal{T}_I$ 通过负采样（negative sampling）得到的视觉三元组的负样本集合，其定义如下：

$$\begin{aligned}\mathcal{T}'_I = &\{(h', r, t)|h' \in \mathcal{E}_I\} \cup \{(h, r, t')|t' \in \mathcal{E}_I\} \\ &\cup \{(h, r', t)|r' \in \mathcal{R}\}, \quad (h, r, t) \in \mathcal{T}_I。\end{aligned} \quad (3\text{-}9)$$

目录

1 绪论

2 预备知识

3 RLSV 模型

4 实验设计与分析

5 总结与期望

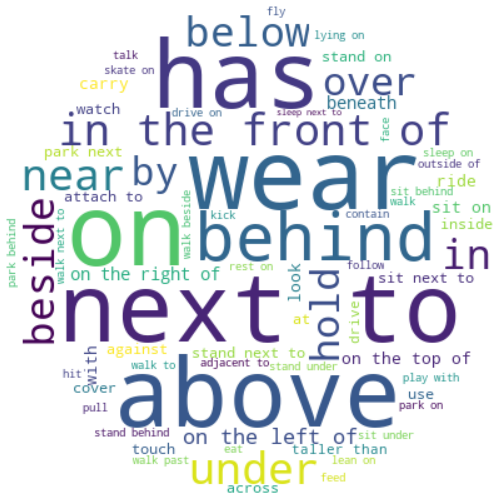
6 附录

VRD 与 VG 数据集

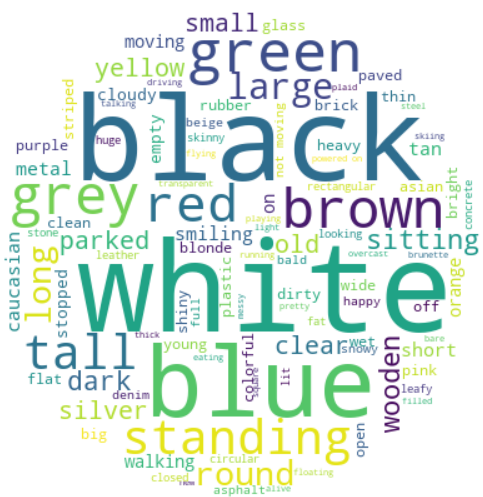
表 4-1: VRD [4] 和 VG [2] 数据集的统计数据表

数据集	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{E}_t\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{A}\rvert$	#Train	#Valid	#Test
VRD	5,000	100	70	100	26,057	5,641	5,641
VG	99,993	200	100	100	657,404	141,627	141,627

(a) 实体类型词云



(b) 关系词云



(c) 属性词云

图 4-1: VRD 数据集的词云图

链接预测

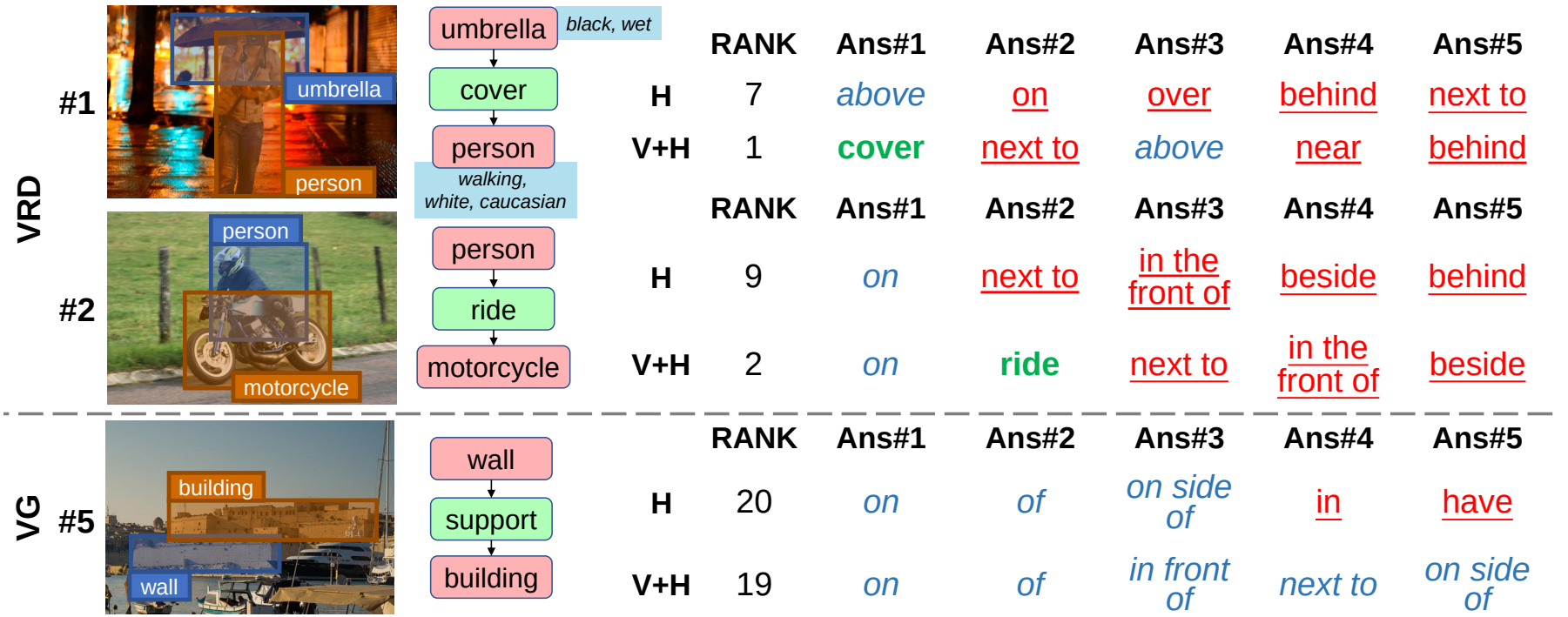
- 给定一条残缺的视觉三元组 $(h, ?, t)$，补全 r。
- 测试方法：对于图片 I 的一条测试视觉三元组 (h, r, t)，
  - “raw”设定：用 $\mathcal{R}$ 中的所有关系去替换 r，分数低表示排名高；
  - “filt”设定：剔除在训练集或验证集中出现过的三元组包含的关系。

表 4-2: 在 VRD 数据集上的关系链接预测实验结果

评价指标	rAVG		Hits@1 (%)		Hits@5 (%)		Hits@10 (%)	
	raw	filt	raw	filt	raw	filt	raw	filt
Rand	35.71	35.59	1.63	1.63	6.98	7.02	14.02	14.06
TransE	5.34	5.22	28.01	29.39	66.85	67.56	89.45	89.81
TransD	5.22	5.10	29.80	31.04	68.94	69.42	85.37	85.94
VTransE	5.17	4.97	38.29	44.09	75.57	76.58	87.32	87.79
PPR-FCN	4.96	4.75	34.05	38.77	76.32	77.75	87.70	88.14
RLSV-V	7.07	6.96	28.62	29.27	62.44	63.00	80.16	80.62
RLSV-H	3.66	3.54	48.69	50.97	82.18	83.09	93.26	93.51
RLSV-V+H	**3.59**	**3.46**	**49.32**	**51.68**	**83.23**	**84.15**	**93.33**	**93.62**

- rAVG：所有正确答案的平均排名，排名越前（低）越优；
- Hits@k：正确答案在前 k 名的比例，比例越高越优，$k \in \{1, 5, 10\}$。

链接预测



		RANK	Ans#1	Ans#2	Ans#3	Ans#4	Ans#5
VRD #1	H	7	*above*	on	over	behind	next to
	V+H	1	**cover**	next to	*above*	near	behind
VRD #2	H	9	*on*	next to	in the front of	beside	behind
	V+H	2	*on*	**ride**	next to	in the front of	beside
VG #5	H	20	*on*	*of*	*on side of*	in	have
	V+H	19	*on*	*of*	*in front of*	*next to*	*on side of*

图 4-2: 关系链接预测的部分视觉三元组样本

RLSV-V+H 作为一个结构信息和视觉信息联合编码的模型表现更佳

RLSV-V+H/RLSV-H 对于难测例能给出合理的关系

视觉三元组分类

- 根据给定的分类阈值，判断视觉三元组为正确或错误。
- 测试方法：
  - 对视觉三元组的关系做随机替换生成负样本；
  - 根据 $E_I(h, r, t)$ 最大化验证集的分类准确率，得到一个关系相关的阈值 σ_r，对于测试集的测例，低于阈值则认为正确。

表 4-3: 在 VRD 和 VG 数据集上的视觉三元组分类实验结果

评价指标	准确率 (%)	
数据集	VRD	VG
VTransE	87.64	91.62
PPR-FCN	86.53	90.23
RLSV-V	86.17	87.58
RLSV-V+H	**90.31**	**93.05**

通过基于翻译的目标函数，RLSV-V+H 可把握相似的关系，区分不相似的关系

目录

1 绪论

2 预备知识

3 RLSV 模型

4 实验设计与分析

5 总结与期望

6 附录

总结与期望

研究背景总结

1. 基于场景图谱的应用也逐渐增多；
2. 场景图谱不完整的问题；
3. 翻译模型在知识图谱上成功运用，但场景图谱与知识图谱存在不同。

总结与期望

研究背景总结

1. 基于场景图谱的应用也逐渐增多；
2. 场景图谱不完整的问题；
3. 翻译模型在知识图谱上成功运用，但场景图谱与知识图谱存在不同。

本文工作总结

1. 形式化定义了场景图谱和场景图谱补全任务；
2. 考虑了场景图谱本身的结构信息和视觉信息，提出了RLSV 模型；
3. 利用现有的大型场景图谱数据集，用链接预测和视觉三元组分类验证了 RLSV 模型的有效性。

总结与期望

研究背景总结

1. 基于场景图谱的应用也逐渐增多；
2. 场景图谱不完整的问题；
3. 翻译模型在知识图谱上成功运用，但场景图谱与知识图谱存在不同。

本文工作总结

1. 形式化定义了场景图谱和场景图谱补全任务；
2. 考虑了场景图谱本身的结构信息和视觉信息，提出了RLSV 模型；
3. 利用现有的大型场景图谱数据集，用链接预测和视觉三元组分类验证了 RLSV 模型的有效性。

研究展望

1. 难样本挖掘的方法加强学习难以拟合的关系；
2. 将现有知识图谱的工作引入到场景图谱当中（如视觉问答）。

目录

1 绪论

2 预备知识

3 RLSV 模型

4 实验设计与分析

5 总结与期望

6 附录

攻读硕士学位期间科研成果

1. Representation Learning for Scene Graph Completion via Jointly Structural and Visual Embedding [C], In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), Sweden, Stockholm, July, 2018（学生第一作者）;

2. 中山大学. 一种基于图像场景图谱对齐的图像查询回答方法: 中国, 201810226645.5 [P] （学生第一作者）。

参考文献 I

[1] Johnson J, Krishna R, Stark M, *et al*. Image Retrieval Using Scene Graphs [C]. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, June, 2015: 3668–3678.

[2] Krishna R, Zhu Y, Groth O, *et al*. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations [J]. International Journal of Computer Vision, 2017, 123 (1): 32–73.

[3] Zhu Y, Lim J J, Fei-Fei L. Knowledge Acquisition for Visual Question Answering via Iterative Querying [C]. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July, 2017: 1154–1163.

[4] Lu C, Krishna R, Bernstein M S, *et al*. Visual Relationship Detection with Language Priors [C]. In Proceedings of European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, October, 2016: 852–869.

[5] Zhang H, Kyaw Z, Chang S, *et al*. Visual Translation Embedding Network for Visual Relation Detection [C]. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July, 2017: 3107–3115.

[6] Zhang H, Kyaw Z, Yu J, *et al*. PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN [C]. In Proceedings of IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October, 2017: 4243–4251.

[7] Bordes A, Usunier N, García-Durán A, *et al*. Translating Embeddings for Modeling Multi-relational Data [C]. In Proceedings of Annual Conference on Neural Information Processing Systems (NIPS), December, 2013: 2787–2795.

参考文献 II

[8] Ji G, He S, Xu L, *et al*. Knowledge Graph Embedding via Dynamic Mapping Matrix [C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), Beijing, China, July, 2015: 687–696.

[9] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [C/OL]. In Proceedings of 3rd International Conference on Learning Representations (ICLR), San Diego, USA, May, 2015. `http://arxiv.org/abs/1409.1556`.

[10] He K, Zhang X, Ren S, *et al*. Deep Residual Learning for Image Recognition [C]. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June, 2016: 770–778.