推理 & 知识推理调研

张记袁

2020-04-05

Outline

- 推理 & 知识推理
- 推理相关的领域
  - KG知识推理（知识补全）
  - 常识推理
  - 自然语言推理
  - 视觉推理
- 其他推理相关benchmark

推理 & 知识推理

- 什么是推理?
  - 古希腊，亚里士多德提出“三段论”，作为现代演绎推理基础
  - Reasoning is “the process of drawing conclusions from the **principles** and **evidence**”. (Wason & Johnson-Laird, 1972)
  - Reasoning: a mechanism that can generate answers to unseen questions by manipulating **existing knowledge** with **inference techniques**. (Zhou et al., 2020)
  - **推理 = 事实证据（knowledge/evidence） + 做出结论**

- 什么是知识推理?
  - **狭义上**指知识图谱中的推理：基于已知事实推出未知事实的计算过程

推理 & 知识推理

Inference vs Reasoning

- inference：推断，得出结论的过程
  - 统计学statistical inference，如贝叶斯推断、近似推断（变分推断）、因果推断
  - ML模型的计算范式，train & inference

- reasoning：推理，根据reason（理由）得出结论
  - 如逻辑推理：
    - 演绎deductive reasoning：逻辑学中的苏格拉底三段论（大前提、小前提->结论）
    - 归纳inductive reasoning：休谟case，“我记得的每一天中太阳会升起”->“太阳明天升起”
    - 溯因abductive reasoning：反绎/溯因，和演绎相反，根据结论找原因/解释

- inference是更高层面概念，某种reasoning就是一个inference的过程?

推理 & 知识推理

- 推理的类型

	类型	说明
归纳&演绎	归纳推理	自底向上，从特殊到一般
	演绎推理	自顶向下，从一般到特殊
是否确定	确定性推理	按照专家规则有完备的推理过程
	不确定推理	概率推理，构建概率模型利用MAP等手段，建模真实世界的不确定性
符号&数值	符号推理	规则（谓词）、知识库（三元组）在关系符号上推理
	数值推理	神经网络、隐含维度空间的向量计算

推理相关的领域

- KG知识推理-知识补全（knowledge reasoning）
- 常识推理（commonsense reasoning）
- 自然语言推理（natural language inference）
- 视觉推理（visual reasoning）

KG知识推理

- 知识推理：
  - **定义：**基于知识图谱已有事实<e1, rel, e2>，推理出新的事实
    - 知识图谱：知识库（二元谓词构成SPO三元组）
    - 如推理出<Melinda, LivesIn, Seatle>
  - **任务：**实体/属性/关系预测

- 应用：
  - 知识补全：完善知识库
  - 问答系统（KBQA；检索问答）
  - 对话系统

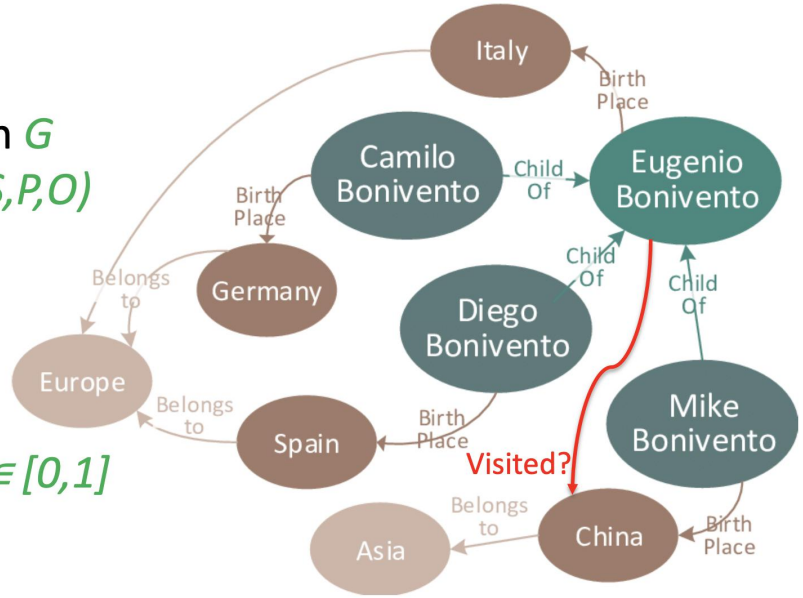
Input:

Knowledge Graph *G*

claim triple *C = (S,P,O)*

Output:

Truth Score $\tau(C) \in [0,1]$



关系补全示例

知识补全常见方法

- 基于路径查找的方法
- 基于强化学习的方法
- 基于知识表示的方法
- 基于图神经网络的方法
- 基于推理规则的方法
- 基于元学习的方法

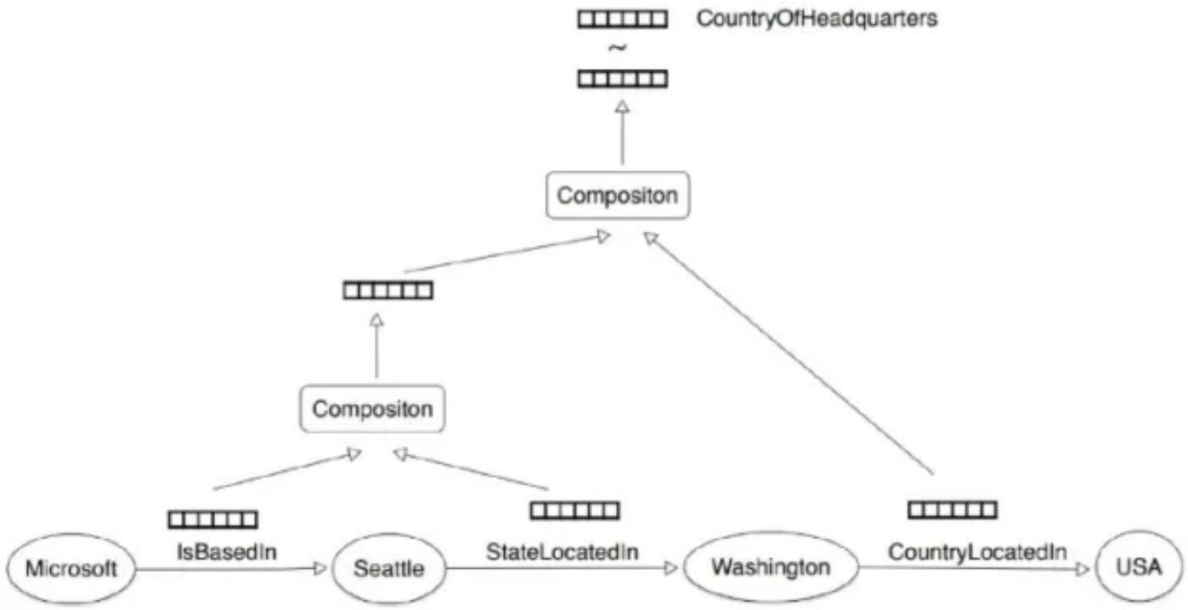
知识补全常见方法——基于路径查找的方法

- 传统的路径查找方法主要是PRA方法（Path Ranking Algorithm）
  - 随机游走获取连接两个实体的所有路径，对路径进行特征构建，训练分类器，预测两个实体之间的关系
  - 优点：直观、解释性好
  - 缺点：很难处理关系稀疏的数据，很难处理低连通度的图，路径特征提取的效率低且耗时

论文地址：https://link.springer.com/content/pdf/10.1007/s10994-010-5205-8.pdf

知识补全常见方法——基于路径查找的方法

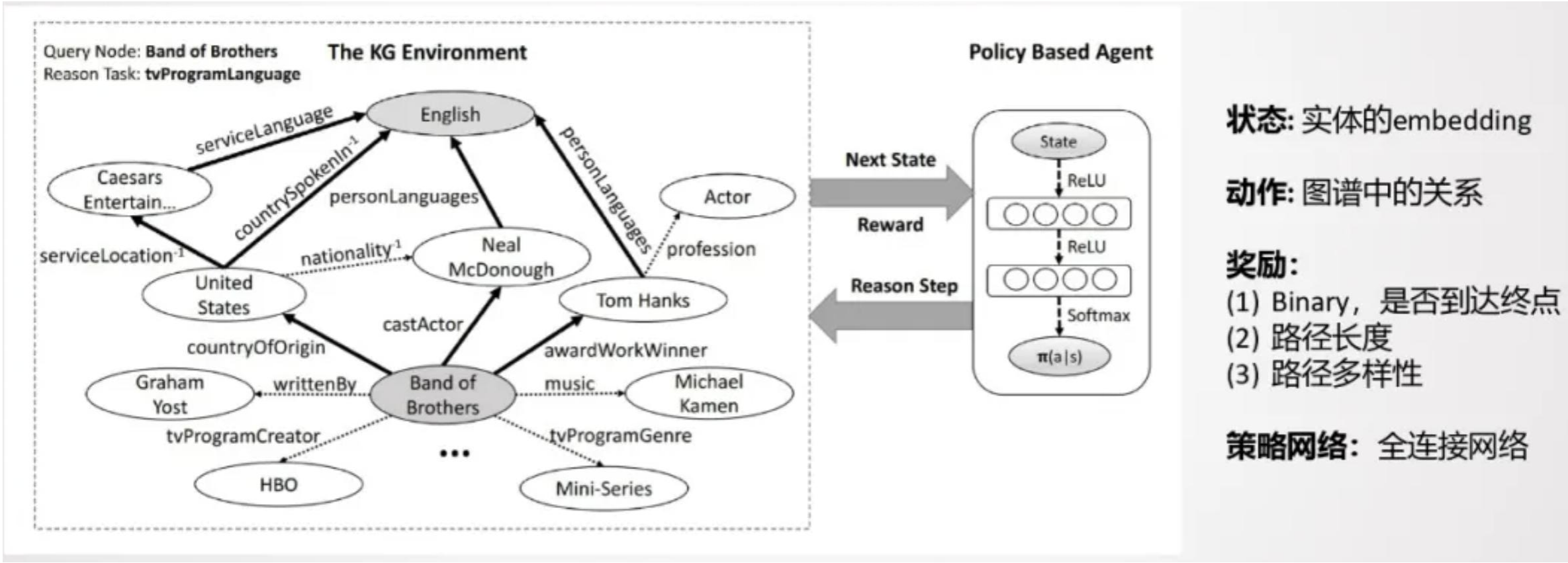
- RNN+PRA
  - 给定实体对集合，利用PRA查找一定数量的路径
  - 使用RNN沿着路径进行向量化建模；
  - 通过比较路径向量与待预测关系向量间的关联度来进行关系补全。



论文地址：https://www.aclweb.org/anthology/P15-1016.pdf

知识补全常见方法——基于强化学习的方法

- DeepPath：将KG推理转化为判断能否找到从h实体到t实体的路径，建模为序列决策问题



知识补全常见方法——基于知识表示的方法

- 知识表示学习：对知识图谱中的实体和关系学习其低维度的嵌入式表示。
- 常见的知识表示学习方法

- TransE

$$f_r(h,t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{L_1/L_2}.$$

- TransH

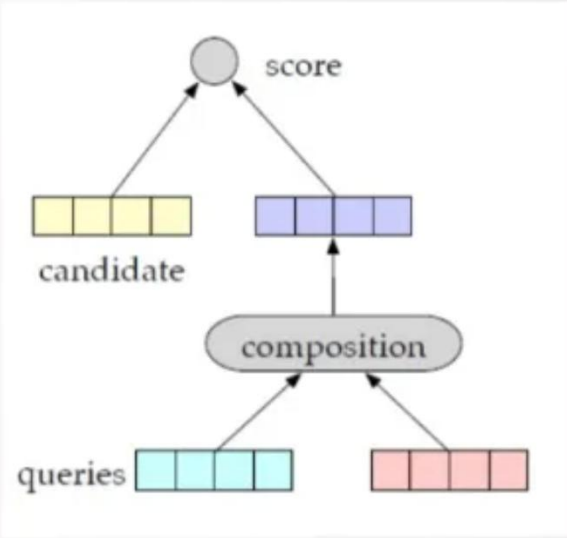
$$f_r(h,t) = -\left\|\left(\mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r\right) + \mathbf{r} - \left(\mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r\right)\right\|_2^2,$$

- TransR

$$f_r(h,t) = -\|\mathbf{M}_r\mathbf{h} + \mathbf{r} - \mathbf{M}_r\mathbf{t}\|_2^2,$$

- TrasnF

$$f_r(h,t) = (\mathbf{h} + \mathbf{r})^\top \mathbf{t} + \mathbf{h}^\top (\mathbf{t} - \mathbf{r}).$$



- 基于实体和关系的表示对缺失三元组进行预测
- 许多前面提到的知识表示方法都可以用在知识图谱补全中

知识补全常见方法——基于知识表示的方法

- RESCAL（2011）：矩阵分解，将所有三元组构成的大张量分解成实体 & 关系向量
  - $f_r(h,t) = h^T M_r t$
- DistMult（2014）：简化RESCAL，但只能处理对称关系
  - $f_r(h,t) = h^T diag(M_r) t$
- ComplEx（2016）：拓展到复数域，解决复杂对称关系
  - $f_r(h,t) = Re(h^T diag(M_r)\bar{t})$
- NTN（2013）：张量神经网络
  - 用bilinear tensor layer代替传统fc，刻画两个实体之间的交互，对三元组的打分函数：

$$g(e_1, R, e_2) = u_R^T f\left(e_1^T W_R^{[1:k]} e_2 + V_R \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_R\right)$$

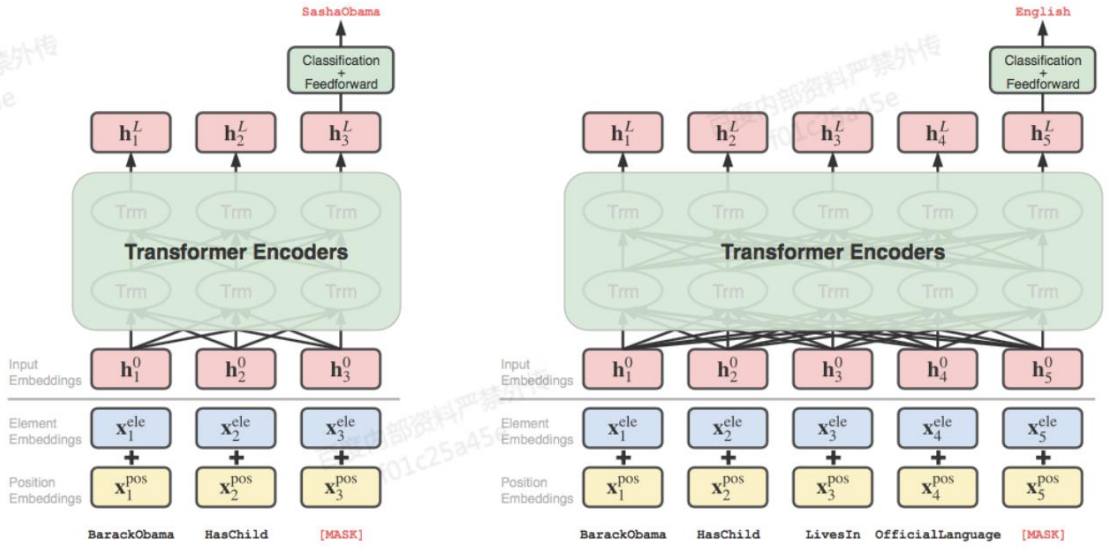
知识补全常见方法——基于知识表示的方法

- 常见KGE在关系预测的结果对比

表示学习方法	WN18					FB15k				
	MR（filter）	MRR（filter）	Hit@1	Hit@3	Hit@10	MR（filter）	MRR（filter）	Hit@1	Hit@3	Hit@10
TransE	251	0.454	0.089	0.823	0.892	125	0.380	0.231	0.472	0.471
TransH	303	—	—	—	0.867	84	—	—	—	0.584
TransR	219	—	—	—	0.920	77	—	—	—	0.687
TransD	212	—	—	—	0.925	67	—	—	—	0.773
DistMult	—	0.822	0.930	0.945	0.936	—	0.654	0.402	0.613	0.824
ComplEx	—	0.941	0.936	0.945	0.947	—	0.692	0.599	0.759	0.840
ANALOGY	—	0.94	0.939	0.944	0.947	—	0.725	0.646	0.785	0.854

知识补全常见方法——基于知识表示的方法

- CoKE：Contextualized Knowledge Graph Embedding
- 改进点：
  1. 本方法采用edges和paths作为input编码实体、关系的embedding
  2. 采用transformer编码信息得到embedding



论文地址：https://arxiv.org/abs/1911.02168

	FB15k				WN18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
Methods that use triples alone								
SimplE (Kazemi and Poole, 2018)	.727	.660	.773	.838	.942	.939	.944	.947
TorusE (Ebisu and Ichise, 2018)	.733	.674	.771	.832	.947	.943	.950	.954
ConvE (Dettmers et al., 2018)	.745	.670	.801	.873	.942	.935	.947	.955
ConvR (Jiang et al., 2019)	.782	.720	.826	.887	.951	.947	**.955**	.958
RotatE (Sun et al., 2019)	.797	.746	.830	.884	.949	.944	.952	.959
HypER (Balažević et al., 2019a)	.790	.734	.829	.885	.951	.947	**.955**	.958
TuckER (Balažević et al., 2019b)	.795	.741	.833	.892	**.953**	**.949**	**.955**	.958
Methods that use graph contexts or rules								
R-GCN+ (Schlichtkrull et al., 2017)	.696	.601	.760	.842	.819	.697	.929	**.964**
KBLRN (Garcia-Duran and Niepert, 2017)	.794	.748	–	.875	–	–	–	–
ComplEx-NNE+AER (Ding et al., 2018)	.803	.761	.831	.874	.943	.940	.945	.948
pLogicNet* (Qu and Tang, 2019)	.844	.812	.862	.902	.945	.939	.947	.958
CoKE (with triples alone)	**.855**	**.826**	**.872**	**.906**	.952	.947	**.955**	.960

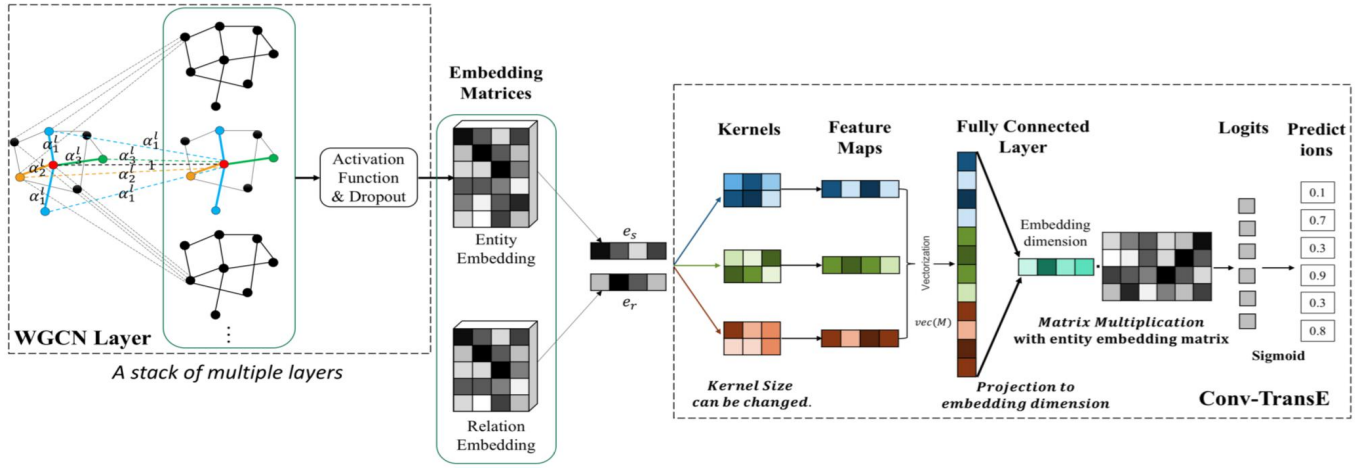
Table 2: Link prediction results on FB15k and WN18. Baseline results are taken from original papers.

	FB15k-237				WN18RR			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
Methods that use triples alone								
ConvE (Dettmers et al., 2018)	.316	.239	.350	.491	.46	.39	.43	.48
ConvR (Jiang et al., 2019)	.350	.261	.385	.528	.475	.443	.489	.537
RotatE (Sun et al., 2019)	.338	.241	.375	.533	.476	.428	.492	**.571**
HypER (Balažević et al., 2019a)	.341	.252	.376	.520	.465	.436	.477	.522
TuckER (Balažević et al., 2019b)	.358	.266	.394	.544	.470	.443	.482	.526
Methods that use graph contexts or rules								
R-GCN+ (Schlichtkrull et al., 2017)	.249	.151	.264	.417	–	–	–	–
KBLRN (Garcia-Duran and Niepert, 2017)	.309	.219	–	.493	–	–	–	–
pLogicNet* (Qu and Tang, 2019)	.332	.237	.367	.524	.441	.398	.446	.537
CoKE (with triples alone)	**.364**	**.272**	**.400**	**.549**	**.484**	**.450**	**.496**	.553

Table 3: Link prediction results on FB15k-237 and WN18RR. Baseline results are taken from original papers.

知识补全常见方法——基于图神经网络的方法

- GNN用于处理图结构的数据，随着信息在节点之间的传播聚合，编码图中节点间的依赖关系。
- KG图结构 + GNN = 学习图结构的实体&关系表示
  - 图神经网络作为encoder学习知识表示，KGE打分函数作为decoder进行关系预测
  - 相关工作：
    - R-GCN + DistMult, ESWC18
    - W-GCN + Conv-TransE, AAAI19
    - GAT + TransE, AAAI19
    - GAT + ConvKB, ACL19



Model	FB15k-237 Hits @10	FB15k-237 Hits @3	FB15k-237 Hits @1	FB15k-237 MRR	WN18RR Hits @10	WN18RR Hits @3	WN18RR Hits @1	WN18RR MRR
DistMult (Yang et al. 2014)	0.42	0.26	0.16	0.24	0.49	0.44	0.39	0.43
ComplEx (Trouillon et al. 2016)	0.43	0.28	0.16	0.25	0.51	0.46	0.41	0.44
R-GCN (Schlichtkrull et al. 2018)	0.42	0.26	0.15	0.25	—	—	—	—
ConvE (Dettmers et al. 2017)	0.49	0.35	0.24	0.32	0.48	0.43	0.39	0.46
Conv-TransE	0.51	0.37	0.24	0.33	0.52	0.47	0.43	0.46
SACN	0.54	0.39	0.26	0.35	**0.54**	**0.48**	**0.43**	**0.47**
SACN using **FB15k-237-Attr**	**0.55**	**0.40**	**0.27**	**0.36**	—	—	—	—
Performance Improvement	**12.2%**	**14.3%**	**12.5%**	**12.5%**	**12.5%**	**11.6%**	**10.3%**	**2.2%**

论文地址：https://arxiv.org/pdf/1811.04441.pdf

知识补全常见方法——基于推理规则的方法

- 基于规则的推理，如传统的AMIE、FOIL存在搜索空间大，导致推理效率低的问题，针对此问题，提出两类优化方案：
  - 推理规则与embedding结合
  - 神经网络模型与传统的推理模型结合

知识补全常见方法——基于推理规则的方法

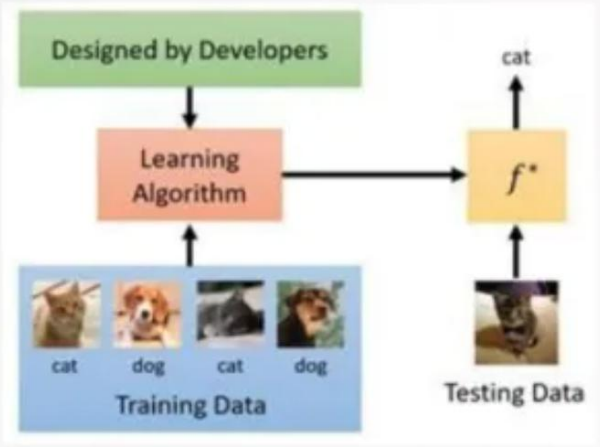
- 神经网络模型与传统的推理模型结合（pLogicNet）
- 使用马尔科夫逻辑网定义三元组的联合分布
- 假设：逻辑规则推理得到的三元组，与基于KG embedding获得的三元组，分布一致。
- 基于以上假设，使用EM算法进行训练：
  - E-step: 限定逻辑规则的权重，基于逻辑规则生成三元组，学习KGE模型权重。
  - M-step: 限定KGE模型，更新逻辑规则的权重

Category	Algorithm	FB15k-237					WN18RR				
		MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10
KGE	TransE [3]	181	0.326	22.9	36.3	52.1	3410	0.223	1.3	40.1	53.1
	DistMult [18]	254	0.241	15.5	26.3	41.9	5110	0.43	39	44	49
	ComplEx [44]	339	0.247	15.8	27.5	42.8	5261	**0.44**	**41**	**46**	51
	ConvE [8]	244	0.325	**23.7**	35.6	50.1	4187	0.43	40	44	52
Rule-based	BLP [7]	1985	0.092	6.2	9.8	15.0	12051	0.254	18.7	31.3	35.8
	MLN [35]	1980	0.098	6.7	10.3	16.0	11549	0.259	19.1	32.2	36.1
Ours	pLogicNet	**173**	0.330	23.1	**36.9**	**52.8**	3436	0.230	1.5	41.1	53.1
	pLogicNet*	**173**	**0.332**	**23.7**	36.7	52.4	**3408**	**0.441**	39.8	44.6	**53.7**

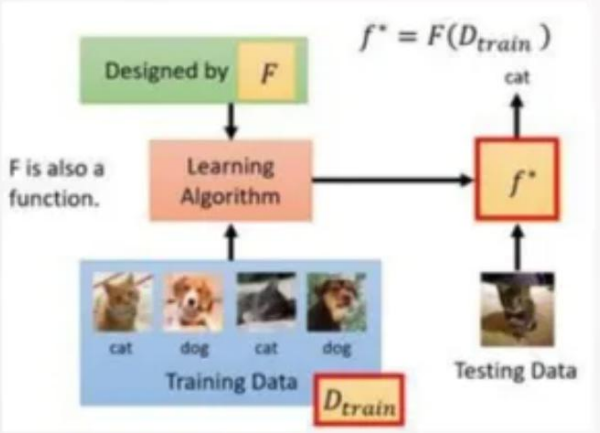
论文地址：https://papers.nips.cc/paper/2019/file/13e5ebb0fa112fe1b31a1067962d74a7-Paper.pdf

知识补全常见方法——基于元学习的方法

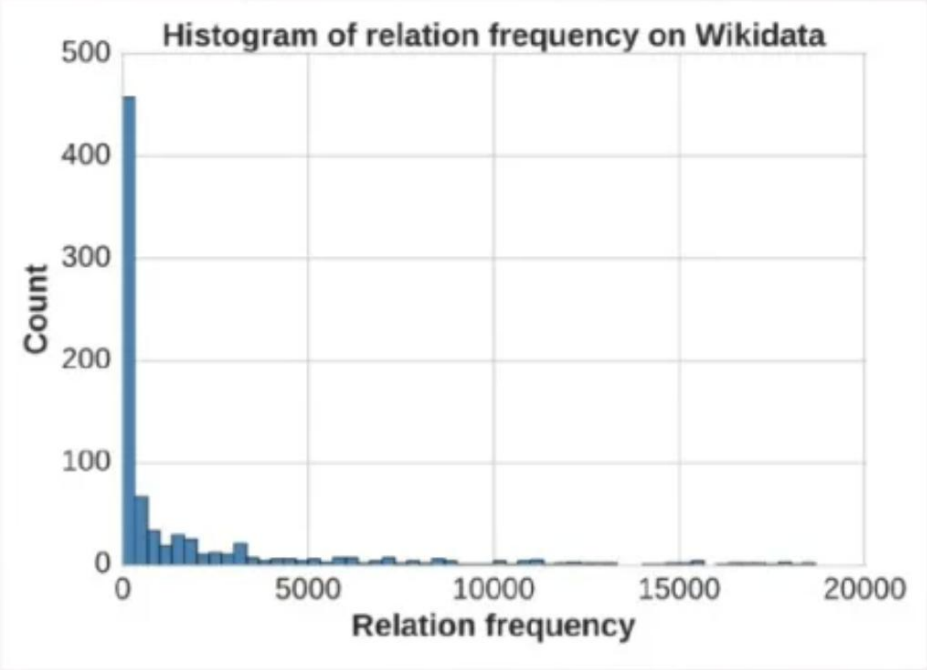
传统机器学习



元学习



- 为什么需要元学习?



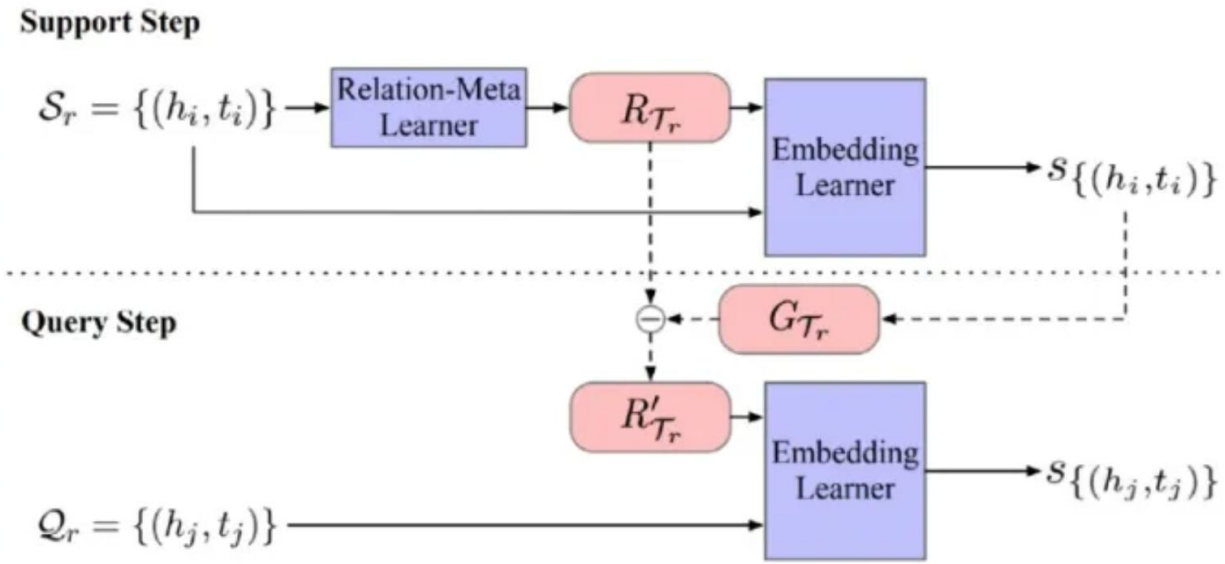
大量关系出现频次很低
越是低频的关系越可能需要补全

知识补全常见方法——基于元学习的方法

- 基于优化的方法：（ MetaR）该方法用来训练一个模型，该模型在很小的数据上可以快速收敛优化达到较好的效果。
- 模型训练的关键点：
  - 使用support set生成gradient meta协助快速学习relation meta
  - 使用query set结果的Loss更新模型参数

	MRR		Hits@10		Hits@5		Hits@1	
NELL-One	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
GMatching_RESCAL	.188	–	.305	–	.243	–	.133	–
GMatching_TransE	.171	–	.255	–	.210	–	.122	–
GMatching_DistMult	.171	–	.301	–	.221	–	.114	–
GMatching_ComplEx	.185	.201	.313	.311	.260	.264	.119	.143
GMatching_Random	.151	–	.252	–	.186	–	.103	–
MetaR (BG:Pre-Train)	.164	.209	.331	.355	.238	.280	.093	.141
MetaR (BG:In-Train)	**.250**	**.261**	**.401**	**.437**	**.336**	**.350**	**.170**	**.168**
Wiki-One	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
GMatching_RESCAL	.139	–	.305	–	.228	–	.061	–
GMatching_TransE	.219	–	.328	–	.269	–	.163	–
GMatching_DistMult	.222	–	.340	–	.271	–	.164	–
GMatching_ComplEx	.200	–	.336	–	.272	–	.120	–
GMatching_Random	.198	–	.299	–	.260	–	.133	–
MetaR (BG:Pre-Train)	**.314**	**.323**	**.404**	**.418**	**.375**	**.385**	**.266**	**.270**
MetaR (BG:In-Train)	.193	.221	.280	.302	.233	.264	.152	.178

论文地址：https://arxiv.org/pdf/1909.01515.pdf



知识补全总结

- 小结
  - 基于知识表示的方法模型简单清晰，效果最好，但是可解释性较差
  - 在知识图谱中进行路径查找可以进行更加复杂的知识推理，重点在于如何缓解大规模的图谱中的路径数量爆炸以及无用信息过多的问题
  - 基于推理的方法将逻辑规则与图谱表示相结合，缓解了稀疏数据的表示学习问题，并且增强了逻辑规则的泛化能力，但当前阶段效果不理想
  - 元学习算法致力于解决知识图谱补全中长尾关系的问题，让模型在极少量训练数据的情况下有快速适配的能力
- 未来方向
  - 概率化逻辑推理与知识表示结合，解决推理过程中的不确定性问题（规则+知识表示方法）
  - 持续强化知识推理的可解释性（规则+知识表示方法）
  - 知识推理的小样本学习（基于元学习的方法）

参考文献：A Survey on Knowledge Graphs- Representation, Acquisition and Applications, 2020

常识推理

- 常识库：
  - CYC（1995）：人工构建的本体规则，最新的ResearchCyc版包含超过700万的常识断言。
  - ConceptNet（2004）：MIT主导众包构建的常识知识库， 主要是人们日常使用词语/短语/概念及其关系组成的语义网。目前版本(ConceptNet5)已经包含有800万节点以及2800万关系描述。
  - WebChild 2.0（2017）：从Web内容中抽取名词-形容词关系构建的常识库，包含超过200万concept和activity，以及超过1800万断言。
  - ATOMIC（2019）：关注if-then形式常识，构建了87万条<事件,关系,事件>的常识图谱，定义了3大类共9种类型的因果关系。

- 常识推理：
  - KG知识推理的子集（MSRA直接把KG推理、常识推理等推理任务直接建模为机器推理问题）
  - 主流方法：基于常识库/知识库，常见的任务形式是QA问答（如问答题、选择题）

常识推理 - 相关Benchmark

- COPA（2011）：
  - **Choice of Plausible Alternatives，1000道常识因果推理的选择题（2选项）**

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.

Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.

COPA数据示例

Table 1: Results of COPA evaluation.

Method	Corpus	Accuracy (%)
Random		50.0
PM I (Roemmele et al., 2011)	Project Gutenberg	58.8
PMI-EX (Gordon et al., 2011)	Personal stories	65.4
CS w/o $MWP_{\lambda=1.0}$ (Luo et al., 2016)	Causal Net	70.2
CS w/o $MWP_{\lambda=0.8}$	ClueWeb12	69.9
CS w/ $MWP_{\lambda=0.7}$	ClueWeb12	**71.2**

Task	Model	Acc. (%) best	mean	std
COPA	Sasaki et al. (2017)	71.2	–	–
	BERT-large	80.8	75.0	3.0
	BERT-SOCIAL IQA	**83.4**	80.1	2.0

leaderboard

常识推理 - 相关Benchmark

- CommonsenseQA
  - 以色列特拉维夫大学 & AllenAI 常识问答任务，12000道选择题（5个候选答案）
  - 数据集构建时已经保证每个候选答案都和问题中的词汇具有语义关联，因此正确回答该数据集中的问题需要有效利用问题和候选答案的相关背景知识。

Where would I not want a fox?
👍 hen house, 👎 england, 👎 mountains,
👎 english hunt, 👎 california

Why do people read gossip magazines?
👍 entertained, 👎 get information, 👎 learn,
👎 improve know how, 👎 lawyer told to

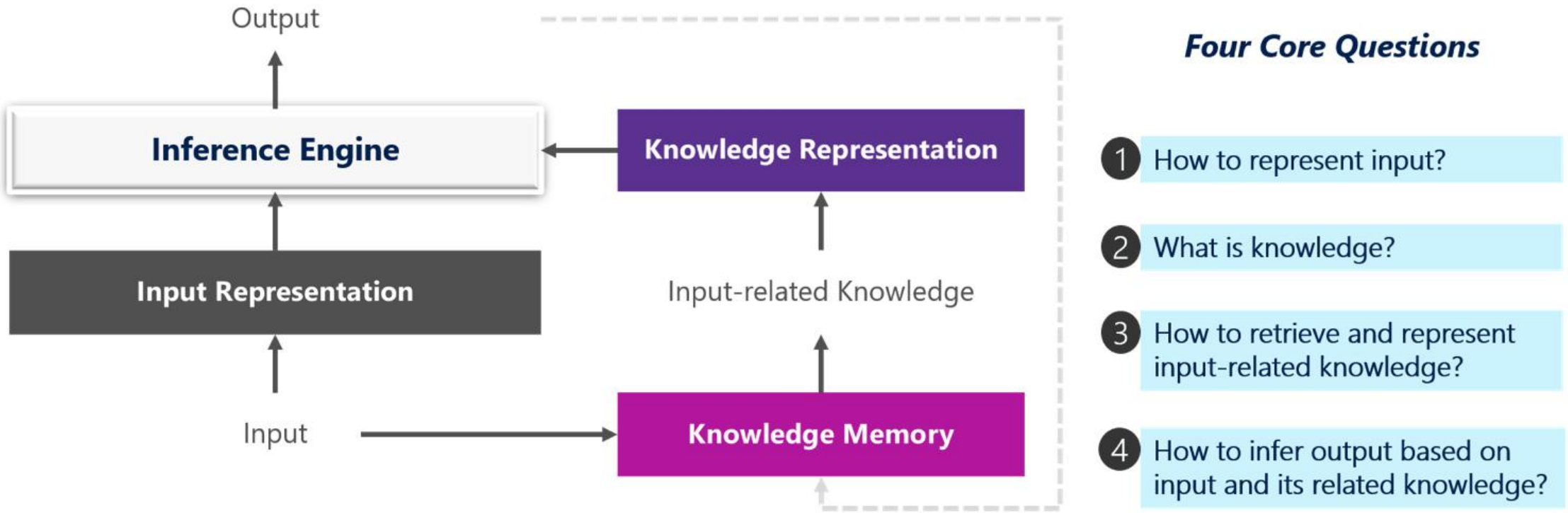
What do all humans want to experience in their own home?
👍 feel comfortable, 👎 work hard, 👎 fall in love,
👎 lay eggs, 👎 live forever

Model	Affiliation	Date	Accuracy	Accuracy (*Used ConceptNet)
Human		03/10/2019	88.9	
Albert + KCR(knowledge chosen by relations, single model)	ITNLP (Harbin Institute of Technology)	07/12/2020		79.5
UnifiedQA (single model)	Allen Institute for AI	04/23/2020	79.1	
Albert + PathGenerator (ensemble model)	USC MOWGLI / INK Lab	05/14/2020		78.2
T5 (single model)	Allen Institute for AI	04/23/2020	78.1	
TeGBERT (single model)	anonymous	07/22/2020		76.8
ALBERT (ensemble model)	Zhiyan Technology	12/18/2019	76.5	

常识推理 - 相关Benchmark

- 其他常识推理相关评测
  - Abductive Natural Language Inference (aNLI)，2020：反绎（溯因）推理，根据观测句选择正确的原因/解释
  - ARC-Easy、ARC: AI2 Reasoning Challenge，2020：7787道小学水平的科学选择题（如问人体什么携带氧气，选择红细胞），需要常识推理+问答，
  - Quoref，2020：给定篇章，回答包含指代的问题，47000篇Wikipedia的段落及24000个问题
  - Social IQa: Commonsense Reasoning about Social Interactions, EMNLP 2019：社交场景的常识推理题，37000个QA对来验证模型对日常事件/场景的社交推理能力
  - Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning，EMNLP 2019：35.6万道需要常识阅读理解的选择题

常识推理 - MSRA研究



机器推理整体框架

Ming Zhou, Nan Duan, Shujie Liu, Heung-Yeung Shum. Progress in Neural NLP: Modeling, Learning and Reasoning. Engineering, 2019.

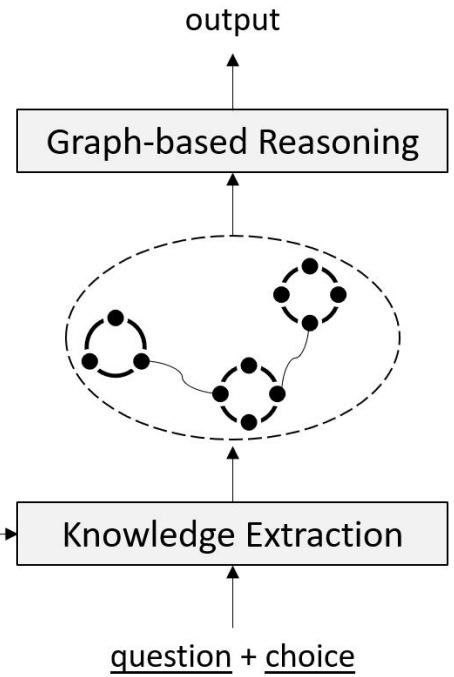
常识推理 - MSRA研究

- reasoning = a mechanism that can generate answers to unseen questions by manipulating existing knowledge with inference techniques. 基于这个定义，一个推理系统包括两个部分：知识+推理引擎
  - 知识：
    - 如何定义知识?
      - KG、常识、规则、文本中提取的断言等；
      - 预训练好的模型也是一种知识（预训练本质是将每个单词在海量文本中的上下文存在模型中）
    - 如何提取&表示知识?
      - 实体链接 + 知识embedding；预训练模型对输入的编码也是知识提取
  - 推理引擎：
    - 如何针对输入，根据知识进行推理、产生回答?
      - 根据输入，召回相关的知识/证据
      - 对输入、知识/证据、候选结果进行联合建模（如预训练模型、GNN）

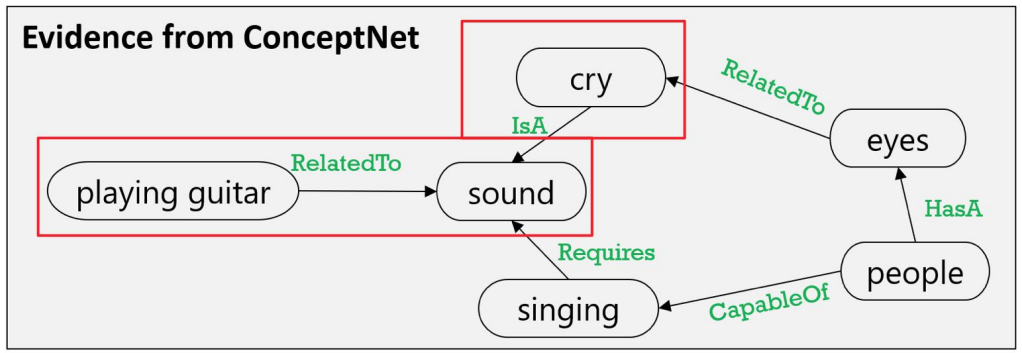
常识推理 - MSRA研究

以AAAI 2020论文为例，提出的Graph-based Reasoning方法，处理CommonsenseQA任务（得分79.3，Top1 79.5）

Question: What do people typically do while playing guitar ?
A. cry **B**. hear sounds **C**. singing (√) **D**. anthritis **E**. making music



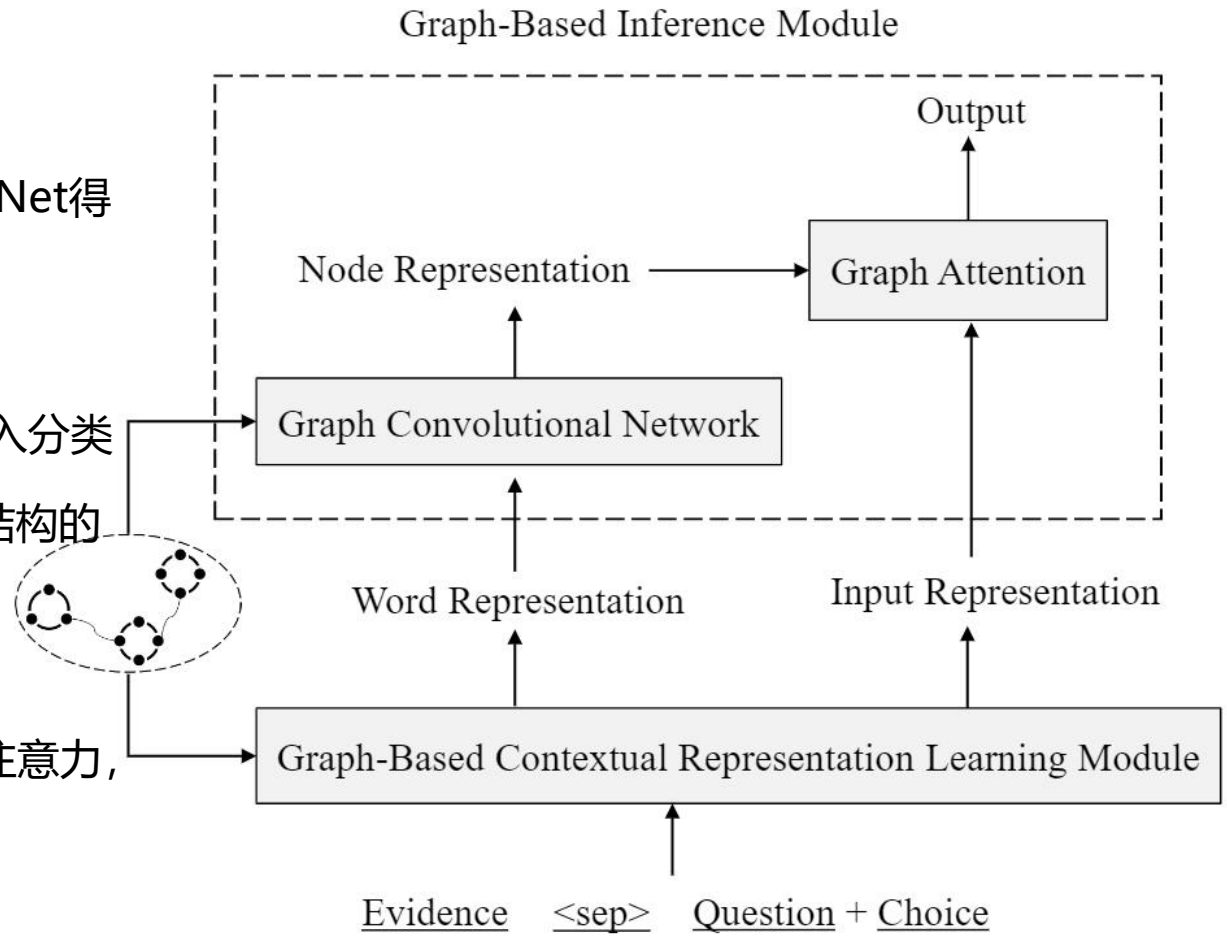
- 知识提取模块：
  - 根据**问题&每个选项的组合**，从多源知识（ConceptNet+Wikipedia）中提取相关知识
  - 证据构建成图结构的形式，如“playing guitar & cry”，从ConceptNet获取的证据如下



** Graph-based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering, AAAI 2020*

常识推理 - MSRA研究

- Graph-based推理模块
  - Graph-based Contextual表示学习
    - 图结构的证据经过拓扑排序得到序列
    - <多源证据序列, 问题, 选项>输入预训练模型XLNet得到Contextual表示
    - 实现多源常识的融合
  - GCN + Graph Attention结合证据对Contextual输入分类
    - 图结构提供更多证据的语义信息，GCN实现图结构的证据聚合，得到图节点表示
    - 图节点表示=XLNet词表示+GCN节点表示
    - Graph Attention用input表示对节点表示分配注意力，得到图表示
    - concat图表示和input表示+MLP计算q-a得分



自然语言推理 – Natural Language Inference(NLI)

- **定义：**给定两个句子（premise和hypothesis），判断它们的关系（蕴含/矛盾/中性）。
  - **如果premise -> hypothesis，则称“premise蕴含hypothesis”**（等同一阶谓词逻辑的蕴含）

	ID	sentence	label
Premise		A dog jumping for a Frisbee in the snow.	
Hypothesis	Example 1	An animal is outside in the cold weather, playing with a plastic toy.	*entailment*
	Example 2	A cat washed his face and whiskers with his front paw.	*contradiction*
	Example 3	A pet is enjoying a game of fetch with his owner.	*neutral*

- **常用方法：**传统方法（模板+分类器）、预训练模型+分类
- 常见数据集/评测Benchmark
  - SNLI (2015)：斯坦福NLI评测数据，600k句子pair
  - MultiNLI (2017)：引入不同来源的数据，如电话对话等，后来被包含在GLUE (2018)评测
  - 其他：RTE-1~7 (2005~2011)、SICK (2014, 10k句子pair)、SciTail (2018, 27k句子pair)、SherLIic (2019)

视觉推理 - 相关Benchmark

- 多模(CV+NLP)+推理任务
  - 基于给定图像，回答自然语言问题

- GQA (2019)：
  - 斯坦福Manning组发布的视觉推理问答数据集，2000万条问题。
  - 数据涉及多种推理技巧、 multi-hop推理问题，模型需要推理能力

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → **Relate:** girl,holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food,on → **Choose** color: yellow | brown

Rank	Participant team	Binary	Open	Consistency	Plausibility	Validity	Distribution	Accur
1	Human Performance (human)	91.20	87.40	98.40	97.20	98.90	0.00	89.30
2	DREAM+Unicoder-VL (MSRA)	84.46	68.60	91.47	83.75	96.42	3.68	76.04
3	TRRNet (Ensemble)	82.12	66.89	89.00	83.58	96.76	1.29	74.03
4	MIL-nbgao	80.80	67.64	91.76	83.90	96.73	1.70	73.81
5	Kakao Brain	79.68	67.73	77.02	83.70	96.36	2.46	73.33

视觉推理 - 相关Benchmark

- VCR (2019)：Visual Commonsense Reasoning，AllenAI提出的视觉常识推理任务
  - 包含约29万个问题、答案和解释pair，涵盖超过11万个不重复的电影场景。

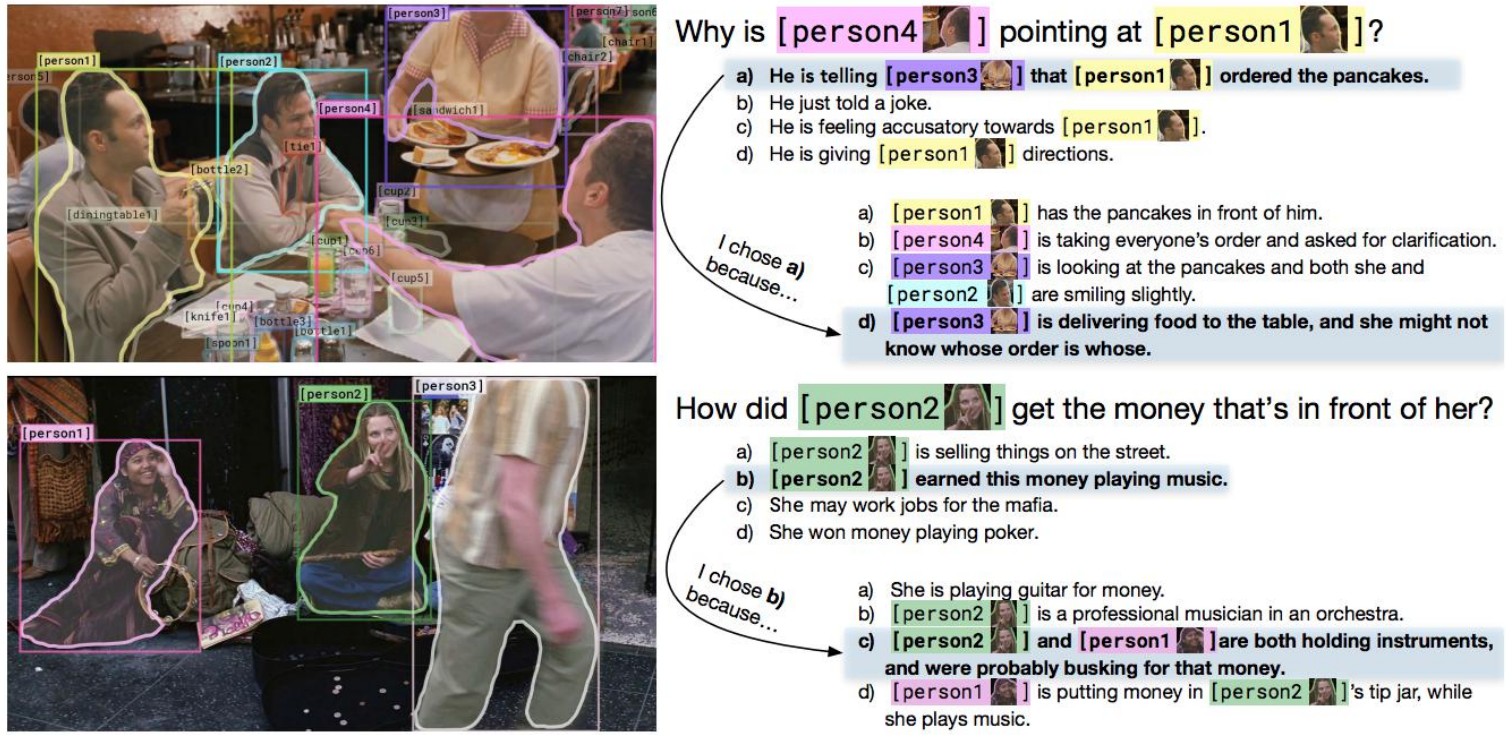
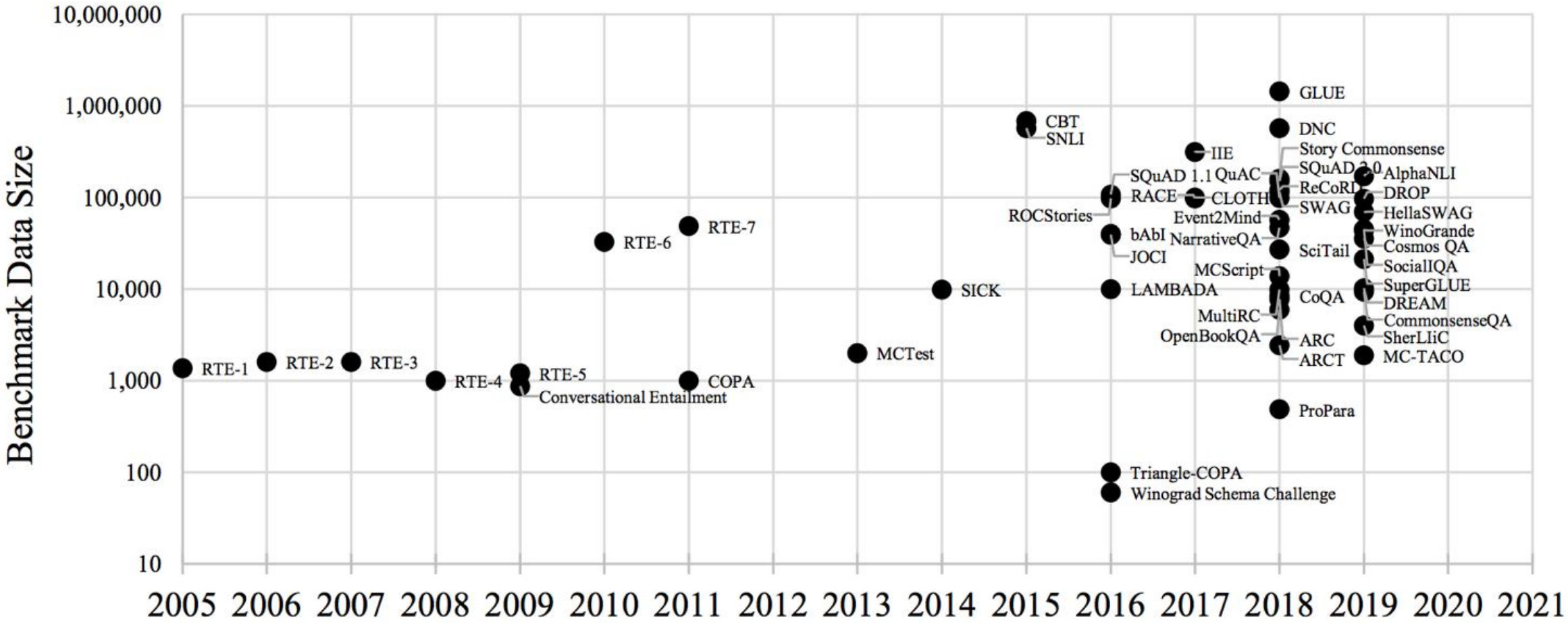


Figure 1: **VCR**: Given an image, a list of regions, and a question, a model must answer the question and provide a *rationale* explaining why its answer is right. Our questions challenge computer vision systems to go beyond recognition-level understanding, towards a higher-order cognitive and commonsense understanding of the world depicted by the image.

Rank	Model	Q->A	QA->R	Q->AR
	Human Performance *University of Washington* (Zellers et al. '18)	91.0	93.0	85.0
November 19, 2020	BLENDER (single model) *WeSee AI team, Tencent*	**81.6**	**86.4**	**70.8**
2 June 24, 2020	ERNIE-ViL-large(ensemble of 15 models) *ERNIE-team - Baidu* https://arxiv.org/abs/2006.16934	**81.6**	86.1	70.5
3 October 28, 2020	MMCNet (ensemble of 4 models) *UC Berkeley*	80.0	83.1	66.9
4 September 30, 2019	UNITER-large (ensemble of 10 models) *MS D365 AI* https://arxiv.org/abs/1909.11740	79.8	83.4	66.8
5 June 24, 2020	ERNIE-ViL-large(single model) *ERNIE-team - Baidu* https://arxiv.org/abs/2006.16934	79.2	83.5	66.3

更多其他推理相关的benchmark

- 涵盖问答CoQA、社区常识问答SocialIQA、阅读理解SQuAD等推理相关的评测任务



Q&A

Thank you!