Holographic Embeddings of Knowledge Graphs

Maximilian Nickel, Lorenzo Rosasco, Tomaso A. Poggio. AAAI 2016

Abstract

- HolE:全息嵌入
- 采用**循环相关(circular correlation)**来创建构图表示
- 可以捕获丰富的交互信息但同时仍然有效计算，易于训练，并可扩展到非常大的数据集。

Compositional Representations

Compositional Representations 简介

可能存在的三元组:Rp(subject,object) φp : E × E → {±1}

$$\Pr(\phi_p(\mathsf{s},\mathsf{o}) = 1|\Theta) = \sigma(\eta_{spo}) = \sigma(\boldsymbol{r}_p^\top(\boldsymbol{e}_s \circ \boldsymbol{e}_o)) \qquad (1)$$

σ(关系 (实体 ◦ 实体))

Compositional Representations 简介

xi表示三元组,yi={+ - 1}表示标签,对于一个数据集D(包含正负样本),根据eq1学习最能解释D的表示Θ可以通过最小化eq2的损失.

$$\min_{\Theta} \sum_{i=1}^{m} \log(1 + \exp(-y_i \eta_i)) + \lambda \|\Theta\|_2^2. \qquad (2)$$

Compositional Representations 简介

但是真实世界的知识图谱中,一般只会存放正确的存在故意的错例比较少或者不会被储存.

可以用pairwise ranking loss ,eq3来使得正三元组的概率比负三元组高.

$$\min_{\Theta} \sum_{i \in \mathcal{D}_+} \sum_{j \in \mathcal{D}_-} \max(0, \gamma + \sigma(\eta_j) - \sigma(\eta_i)) \tag{3}$$

important property of compositional models

1. 实体的表示和意义不会随着实体在组合表示中的位置而变化.(作为subject和object的时候是一样的)
2. 由于所有实体和关系的表示是在eq2,eq3中共同学习的,所以模型学习到了三元组之间传播的信息,以及数据的全局依赖关系.

compositional operators

Tensor Product

all pairwise multiplicative interactions between the features of a and b:

$$[\boldsymbol{a} \otimes \boldsymbol{b}]_{ij} = a_i b_j. \tag{4}$$

Intuitively, a feature in the tuple representation a ⊗ b is “on”

(has a high absolute magnitude), if and only if the corresponding features of both entities are “on” (See also fig. 1a).

Concatenation, Projection, and Non-Linearity

Let ⊕ : Rd1 × Rd2 → Rd1+d2 denote concatenation and

ψ : R → R be a non-linear function such as tanh.

The composite tuple representation is then given by a ◦ b = ψ(W (a ⊕ b)) ∈ Rh, such that

$$[\psi(W(\boldsymbol{a} \oplus \boldsymbol{b}))]_i = \psi\left(\sum\nolimits_j w_{ij}^a a_j + \sum\nolimits_j w_{ij}^b b_j\right) \quad (5)$$

the projection matrix W ∈ Rh×2d is learned in combination with the entity and relation embeddings.

Non-compositional Methods

TRANSE models the score of a fact as the distance between relation-specific translations of entity embeddings:

$$score(R_p(s, o)) = -\, dist(e_s + r_p, e_o) \,. \quad (6)$$

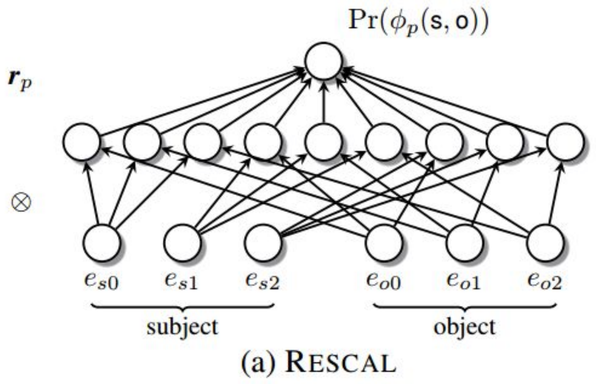


Figure 1: RESCAL and HOLE as neural networks. RESCAL represents pairs of entities via d^2 components (middle layer). In contrast, HOLE requires only d components.

比起RESCAL,HoLE需要的参数少很多.

(a) Compositional Representations

Operator	$\circ$	**Memory** $\boldsymbol{r}_p$	**Runtime** $\boldsymbol{r}_p^\top(\boldsymbol{e}_s \circ \boldsymbol{e}_o)$
Tensor Product	$\otimes$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2)$
Circular Correlation	$\star$	$\mathcal{O}(d)$	$\mathcal{O}(d \log d)$

张量积与循环相关的关系内存占用量以及时间复杂度比较

Holographic Embeddings

Combine tensor product and TransE

the circular correlation of vectors to represent pairs of entities,用循环相关的向量表示实体对.

we use the compositional operator:

$$\boldsymbol{a} \circ \boldsymbol{b} = \boldsymbol{a} \star \boldsymbol{b}, \tag{7}$$

where $\star : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ denotes *circular correlation:*[1]

$$[\boldsymbol{a} \star \boldsymbol{b}]_k = \sum_{i=0}^{d-1} a_i b_{(k+i) \bmod d}. \tag{8}$$

全息嵌入(holographic embeddings)模型

the probability of triple:

Hence, we model the probability of a triple as

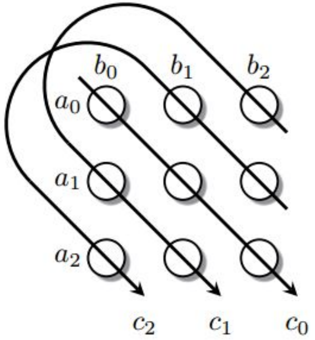
$$\Pr(\phi_p(\mathsf{s},\mathsf{o}) = 1|\Theta) = \sigma(\boldsymbol{r}_p^\top(\boldsymbol{e}_s \star \boldsymbol{e}_o)). \tag{9}$$

circular correlation

可以解释为张量积的压缩.

the tensor product assigns a separate component cij = aibj for each pairwise interaction of entity features, in correlation each component corresponds to a sum over a fixed partition of pairwise interactions.

节点表示张量积中的元素,箭头表示求和



$$\boldsymbol{c} = \boldsymbol{a} \star \boldsymbol{b}$$

$$c_0 = a_0 b_0 + a_1 b_1 + a_2 b_2$$
$$c_1 = a_0 b_2 + a_1 b_0 + a_2 b_1$$
$$c_2 = a_0 b_1 + a_1 b_2 + a_2 b_0$$

Figure 2: Circular correlation as compression of the tensor product. Arrows indicate summation patterns, nodes indicate elements in the tensor product. Adapted from (Plate 1995).

circular correlation

1. 在语义相近交互的部分可以共享一个权重.(例子:当建立partyof 模型的时候,知道subject和object是自由党人+自由党,或者是保守党人+保守党是很有用的)
2. 这些interactions会被分在同一组.
3. a subset of latent features are relevant to model relational patterns.
4. 然后可以将不相关的交互分组在相同的分区中,并在rp中分配了一个小权重。
5. 分区不是学习,,而是相关操作的预先准备.

circular correlation computation

$$a \star b = \mathcal{F}^{-1}\left(\overline{\mathcal{F}(a)} \odot \mathcal{F}(b)\right)$$

F(·) and F −1(·) denote the fast Fourier transform (FFT) and its inverse.[快速傅里叶变化]

F(a)带上标 denotes the complex conjugate of F(a) ∈ Cd

[https://zh.wikipedia.org/wiki/%E5%85%B1%E8%BD%AD%E5%A4%8D%E6%95%B0]

den denotes the Hadamard (entrywise) product.[例:矩阵每个对应位置元素乘]

快速傅里叶变换

快速傅里叶变换（英语：**Fast Fourier Transform, FFT**），是快速计算序列的离散傅里叶变换（DFT）或其逆变换的方法[1]。傅里叶分析将信号从原始域（通常是时间或空间）转换到频域的表示或者逆过来转换。FFT会通过把DFT矩阵分解为稀疏（大多为零）因子之积来快速计算此类变换。[2] 因此，它能够将计算DFT的复杂度从只用DFT定义计算需要的 {\displaystyle O(n^{2})}，降低到 {\displaystyle O(n\log n)}，其中 {\displaystyle n} 为数据大小。

Circular convolution

$$[\boldsymbol{a} * \boldsymbol{b}]_k = \sum_{i=0}^{d-1} a_i b_{(k-i) \bmod d}. \tag{10}$$

在组合运算中,相关与起循环卷积有两个好处

非交换:可以体现出关系的非对称性

Non Commutative Correlation, unlike convolution, is not commutative, i.e., $\boldsymbol{a} \star \boldsymbol{b} \neq \boldsymbol{b} \star \boldsymbol{a}$. Non-commutativity is necessary to model asymmetric relations (directed graphs) with compositional representations.

相似成份:

Similiarity Component In the correlation $\boldsymbol{a} \star \boldsymbol{b}$, a single component $[\boldsymbol{a} \star \boldsymbol{b}]_0 = \sum_i a_i b_i$ corresponds to the dot product $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$. The existence of such a component can be helpful to model relations in which the similarity of entities is important. No such component exists in the convolution $\boldsymbol{a} * \boldsymbol{b}$ (see also fig. 1 in the supplementary material).

To compute the representations for entities and relations, we minimize either eq. (2) or (3) via stochastic gradient descent (SGD). Let $\theta \in \{\boldsymbol{e}_i\}_{i=1}^{n_e} \cup \{\boldsymbol{r}_k\}_{k=1}^{n_r}$ denote the embedding of a single entity or relation and let $f_{spo} = \sigma(\boldsymbol{r}_p^\top(\boldsymbol{e}_s \star \boldsymbol{e}_o))$. The gradients of eq. (9) are then given by

$$\frac{\partial f_{spo}}{\partial \theta} = \frac{\partial f_{spo}}{\partial \eta_{spo}} \frac{\partial \eta_{spo}}{\partial \theta},$$

where

$$\frac{\partial \eta_{spo}}{\partial \boldsymbol{r}_p} = \boldsymbol{e}_s \star \boldsymbol{e}_o, \quad \frac{\partial \eta_{spo}}{\partial \boldsymbol{e}_s} = \boldsymbol{r}_p \star \boldsymbol{e}_o, \quad \frac{\partial \eta_{spo}}{\partial \boldsymbol{e}_o} = \boldsymbol{r}_p * \boldsymbol{e}_s. \tag{11}$$

The partial gradients in eq. (11) follow directly from

$$\boldsymbol{r}_p^\top(\boldsymbol{e}_s \star \boldsymbol{e}_o) = \boldsymbol{e}_s^\top(\boldsymbol{r}_p \star \boldsymbol{e}_o) = \boldsymbol{e}_o^\top(\boldsymbol{r}_p * \boldsymbol{e}_s) \tag{12}$$

and standard vector calculus. Equation (12) can be derived as follows: First we rewrite correlation in terms of convolution:

$$\boldsymbol{a} \star \boldsymbol{b} = \widetilde{\boldsymbol{a}} * \boldsymbol{b}$$

where $\widetilde{\boldsymbol{a}}$ denotes the *involution* of $\boldsymbol{a}$, meaning that $\widetilde{\boldsymbol{a}}$ is the mirror image of $\boldsymbol{a}$ such that $\widetilde{a}_i = a_{-i \bmod d}$ (Schönemann 1987, eq. 2.4). Equation (12) follows then from the following identities in convolution algebra (Plate 1995):

$$\boldsymbol{c}^\top(\widetilde{\boldsymbol{a}} * \boldsymbol{b}) = \boldsymbol{a}^\top(\widetilde{\boldsymbol{c}} * \boldsymbol{b}); \qquad \boldsymbol{c}^\top(\widetilde{\boldsymbol{a}} * \boldsymbol{b}) = \boldsymbol{b}^\top(\boldsymbol{a} * \boldsymbol{c}).$$

Similar to correlation, the circular convolution in eq. (11) can be computed efficiently via $\boldsymbol{a} * \boldsymbol{b} = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{a}) \odot \mathcal{F}(\boldsymbol{b}))$.

SGD

Associative Memory

Experiments

数据集

WN18 WordNet is a KG that groups words into synonyms and provides lexical relationships between words. The WN18 dataset consists of a subset of WordNet, containing

40,943 entities, 18 relation types, and 151,442 triples.

FB15k Freebase is a large knowledge graph that stores general facts about the world (e.g., harvested from Wikipedia,MusicBrainz, etc.). The FB15k dataset consists of a subset of Freebase, containing

14,951 entities, 1345 relation types, and 592,213 triples.

国家数据集上的任务

locatedIn（e1，e2）和neighborOf（e1，e2）。

在实验中任务是预测locateIn（c，r），其中c包含所有国家，r包括所有地区数据。

80%训练集,10%验证集,10%测试集,使用3个不同的设置

国家数据集上的任务

- S1:验证与测试集中的locatedin(c,r)缺失,正确关系由下式预测
    - locatedIn(c, s) ∧ locatedIn(s, r) ⇒ locatedIn(c, r)
    - s是国家的subregion.
- S2:在S1的基础设置上,使得对于国家c 的locatedIn(c,s)也缺失,正确关系由下式预测:
    - neighborOf(c1, c2) ∧ locatedIn(c2, r) ⇒ locatedIn(c1, r)
- S3:在S1\S2的基础上,所有国家的locatedIn(n,r)缺失,正确关系由下式预测:
    - neighborOf(c1, c2) ∧ locatedIn(c2, s) ∧locatedIn(s, r) ⇒ locatedIn(c1, r)

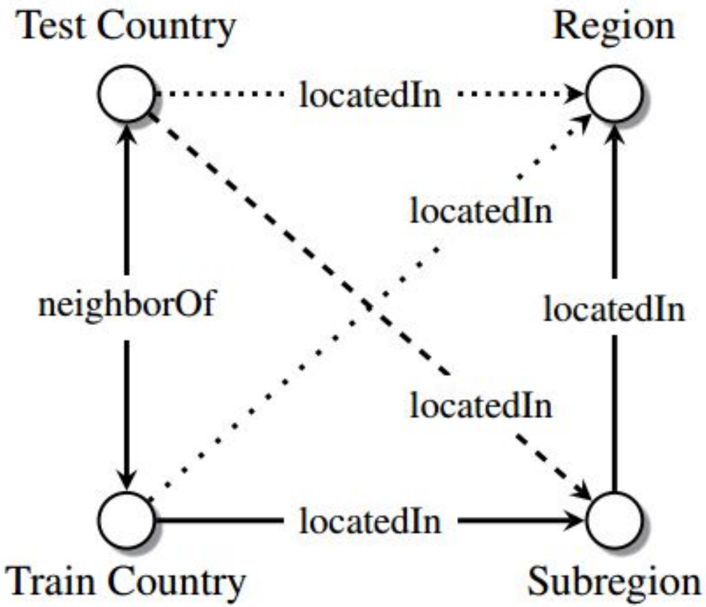


Figure 3: Removed edges in countries experiment: S1) dotted S2) dotted and dashed S3) dotted, dashed and loosely dotted.

Table 2: Results for link prediction on WordNet (WN18), Freebase (FB15k) and Countries data.

(a)

	WN18					FB15k				
	MRR		Hits at			MRR		Hits at		
Method	Filter	Raw	1	3	10	Filter	Raw	1	3	10
TRANSE	0.495	0.351	11.3	88.8	94.3	0.463	0.222	29.7	57.8	**74.9**
TRANSR	0.605	0.427	33.5	87.6	94.0	0.346	0.198	21.8	40.4	58.2
ER-MLP	0.712	0.528	62.6	77.5	86.3	0.288	0.155	17.3	31.7	50.1
RESCAL	0.890	0.603	84.2	90.4	92.8	0.354	0.189	23.5	40.9	58.7
HOLE	**0.938**	**0.616**	**93.0**	**94.5**	**94.9**	**0.524**	**0.232**	**40.2**	**61.3**	73.9

(b)

	Countries		
	AUC-PR		
Method	S1	S2	S3
Random	0.323	0.323	0.323
Frequency	0.323	0.323	0.308
ER-MLP	0.960	0.734	0.652
RESCAL	**0.997**	0.745	0.650
HOLE	**0.997**	**0.772**	**0.697**

在WN18\FB15k中的链路预测结果比较.以及在国家数据中的准确率比较.