ACL 2018 long paper

利用简单约束改进知识图谱的表示学习

Improving Knowledge Graph Embedding Using Simple Constraints

中国科学院信息工程研究所

【摘要】 获得知识图谱中实体和关系的向量化表示，即表示学习，是近几年的一个研究热点，相关技术可以用于大规模知识推理。早期的模型通过定义元组上的简单模型实现这个任务，尽管效率高，但是推理精度有限。近期的一些尝试主要包括在基本模型上利用外部信息以及设计更复杂的元组模型上。在这篇工作中，我们尝试另一种思路，通过对基本模型施加一些简单约束来提升表示学习的效果。我们考虑了对实体添加非负性约束，以及对关系添加蕴含性约束，前者可以帮助我们学到更紧凑，更具有可解释性实体向量；后者将逻辑蕴含结构编码到关系的分布式表示中。这些简单的约束在提升性能的同时，又不会增加基本模型的复杂度。我们在公开数据集FB15K,WN18,DB100K上做了测试，证明这种简单约束可以显著提高基本模型的性能，表现甚至比一些更复杂的模型还要好，此外，我们还对实体和关系向量的可解释性做了一个可视化分析，我们的代码和数据可以通过<https://github.com/iieir-km/ComplEx-NNE_AER>获得。

1. **Introduction**

在过去的几十年中，大规模知识图谱获得了巨大的发展，诞生了诸如WordNet, Freebase，Google Knowledge Vault的项目。知识图谱可以形式化地表示为一个三元组集合，其中是实体，是关系。这种结构化的知识表示方式为自然语言处理和信息检索的诸多下游应用，比如问答，学术检索，机器阅读等都提供了丰富的背景知识，具有非常广阔的应用前景。

近年来，知识图谱的表示学习受到了较多关注，其关键思想是获得实体和关系的向量化表示，这种向量化表示在简化了操作的同时又保持了图的内在结构。早期在该方向上的探索主要是单纯利用元组数据，并在之上设计简单的打分模型。这类基本模型虽然高效，但是推理精度有限。近期的一些工作为了进一步提升推理精度，主要有两种方式，第一种方式是利用外部数据，比如文本，逻辑规则等，但这种方式会依赖于具体的数据，缺乏通用性；第二种是设计更复杂的模型，但这种方式会同时增加计算代价。

在这篇工作中，我们转而尝试对基本模型添加一些简单的约束。特别的，我们对实体添加了非负性约束，对关系添加了近似蕴含性约束。前者要求实体向量处在一个非负的向量空间内；非负性约束在词向量，矩阵分解的领域已经被广泛研究过，非负性可以使得我们获得更加紧凑，更加稀疏的向量表示，使学到的向量更加具有可解释性。后者可以将逻辑蕴含结构编码到向量的表示中，进一步提升推理能力。

我们最终在公开的数据集上测试了方法的性能，在三个数据集上均显著地超过对比方法，甚至比一些复杂的基于深度网络的模型还要好。 同时我们也对学到的实体和关系向量的可解释性做了一个可视化分析，证明添加了简单约束之后，模型具有了更好的可解释性。

1. **Related Work**

相关工作关主要有三类

第一类包含基本表示学习模型，其中有基于位移的模型TransE及其变种，以及基于语义匹配的模型RESCAL及其变种，这些模型都只用到了元组信息。基于这些基本的模型，一些模型利用外部信息来提高表示学习的性能，比如利用关系路径，利用文本信息等。另外，也有一些模型设计更复杂的打分函数来提高性能，比如采用卷积神经网络设计打分函数。

第二类与我们相关的工作是利用逻辑规则的模型，但此类大部分工作都需要实例化逻辑规则，在空间和时间上都比较低效；为了避免实例化，可以采用次序嵌入的方法，我们的方法与此类方法类似，但是之前的工作主要针对实体对进行建模，不能预测未在训练集中出现过的实体对，同时也只考虑了严格的蕴含规则，没有建模近似性。

第三类与我们相关的工作是非负表示学习，非负性约束已经在很多领域都被研究过，之前有研究显示，非负性能够导致学到的向量的稀疏性，同时也能学到更具有可解释性的向量表示。在一些词向量的研究中也表示，施加非负性约束可以学到更加具有可表示性的向量。

1. **Our Approach**

我们首先介绍用来建模元组的基本模型，然后介绍非负性约束以及近似蕴含性约束，最后给出最终的模型

**3.1 基本模型**

我们采用了ComplEx模型，这是目前性能最好的基本表示学习模型之一，ComplEx将实体和关系表示为复数向量，每一个复数向量,然后定义一个打分函数



对于每一个元组，该打分函数计算一个分数来衡量元组成立的可能性，分数越高，成立的可能性也越高。

**3.2 实体非负性约束**

实体的非负性约束要求实体向量处在一个非负的空间内，具体的，我们要求



非负实体向量仅仅包含正向的语义，可以更高效地编码信息，比如对于实体*“巴黎”*

，正向语义仅仅需要编码*“巴黎是法国的首都”,*而负向语义则需要编码*“巴黎不是中国的首都，巴黎不是美国的首都，巴黎不是俄罗斯的首都…”，*事实上会造成信息冗余。此外，非负向量还能够导致稀疏性，这可以使得我们学到更加稀疏的表示，从而获得更有可解释性的实体向量。

**3.3 关系近似蕴含性约束**

除了非负性约束，我们考虑关系的近似蕴含性约束，这里可以记作，近似蕴含性约束是说对于两个关系，前者可以近似蕴含后者。比如表示我们在知道了某人的出生地（）, 以大概率能推断出这个人的国籍（）。每一条近似蕴含都有一个权重，代表该蕴含成立的置信度，权重越大，代表置信度越大。这种带置信度的蕴含约束可以通过规则挖掘系统自动地从元组中挖掘出来。

为了建模近似蕴含约束，我们首先建模蕴含约束，在知识图谱中，我们可以通过如下的方式建模蕴含



该式子的含义是说，对于任何实体对，如果为真并取得较高的打分，则总是为真并会取得更高的打分。但是这样的约束并不利于我们求解，因为需要遍历所有的实体对，所以我们采用次序嵌入的方式将其改写，当满足非负性约束的条件下，则上面公式的一个充分条件就是



这样，对与每一条蕴含约束，只涉及两个关系向量的计算，计算量大大减少。

最后，我们考虑建模蕴含的近似性，这里引入松弛变量建模近似性，可以得到最终的近似蕴含约束如下，以上的具体推导细节，可以参考原论文的附录部分。



* 1. **最终模型**

结合基本模型，我们的目标函数可以定义如下：



其中，目标函数中第一项是对元组建模，这里使用了logistic loss,第二项是近似蕴含约束中的实部的约束，是一个hinge loss，直观的理解是当，则视为违反约束，大的越多，违反的程度也越大，如果,则没有违反约束，不施加任何惩罚。第三项是一个平方项，用于建模近似蕴含约束的虚部约束，迫使两个虚部相等，这两项是通过一些变形去掉了近似蕴含约束中的松弛变量获得，相关的推导细节可以参考原论文。最后一项是正则，用于防止过拟合。

* 1. **复杂度分析**

带非负与蕴含的模型的复杂度是,其中是维度，是原来ComplEx模型用于建模元组的复杂度；是近似蕴含约束的复杂度，是近似蕴含约束的个数，该数目远远小于知识图谱中的元组数目，可以近似认为是一个常数，是非负性约束带来的复杂度，其中是一个mini-batch中中平均的实体个数，可以看到，带简单约束的模型与基本模型复杂度同阶，是一个很高效的算法。

1. **Experiment**

我们的实验分为两个部分，第一部分是链接预测实验，这个在之前很多相关工作中都会做，用来测试表示学习的性能，第二部分是针对实体关系向量可解释性的可视化分析。

**4.1 数据集**

****我们在三个数据集上做了实验，分别是WN18,FB15K,DB100K，WN18和FB15K是WordNet是Freebase的一个子集，最早在这篇工作中公布，DB100K是我们在DBPedia中抽取的的一个子集。各数据集的统计数据如下,其中#Ent,#Rel,#Train/Valid/Test,#Cons分别表示实体数，关系数，训练/验证/测试集的元组数目以及近似蕴含约束。

蕴含约束通过规则自动抽取软件AMIE+获得，在各数据集上（从上到下，依次是WN18,FB15K,DB100k）的近似蕴含约束样例如下



**4.2 链接预测**

链接预测任务是给定一个缺失头实体或尾实体的元组，预测出相应的头尾实体。在评测中，给定测试元组集合，对于其中每一个元组的头（尾）实体，将其随机地替换为除了测试元组之外没有在知识图谱中出现过的实体，然后计算这些元组的的打分，获得对应测试元组的排序。根据这个排序，计算两个指标， 是平均倒排值，,其中，表示正确元组排进前N的比例。

我们选取的比较方法有三类，第一类是基本模型，这类模型只利用元组信息，同时打分函数也是简单的

第二类模型是结合了逻辑规则的模型，

第三类模型是近期取得最好结果的一些比较复杂的模型，包含基于深度神经网络的模型，以及，是利用了大量的负例来提升性能。

从表中我们可以看到，ComplEx-NNE（添加了实体非负性约束的ComplEx模型）以及ComplEx-NNE+AER（添加了实体非负性和关系蕴含性约束的模型）都能够明显地提升ComplEx的性能，其中ComplEx-NNE-AER能显著好于对比方法，这其中包含了基本模型，利用逻辑规则的模型，以及一些复杂的模型，其中包括的基于深度神经网络的模型。由此可见，这种简单的约束在保证计算效率的同时，又能够显著提升性能。

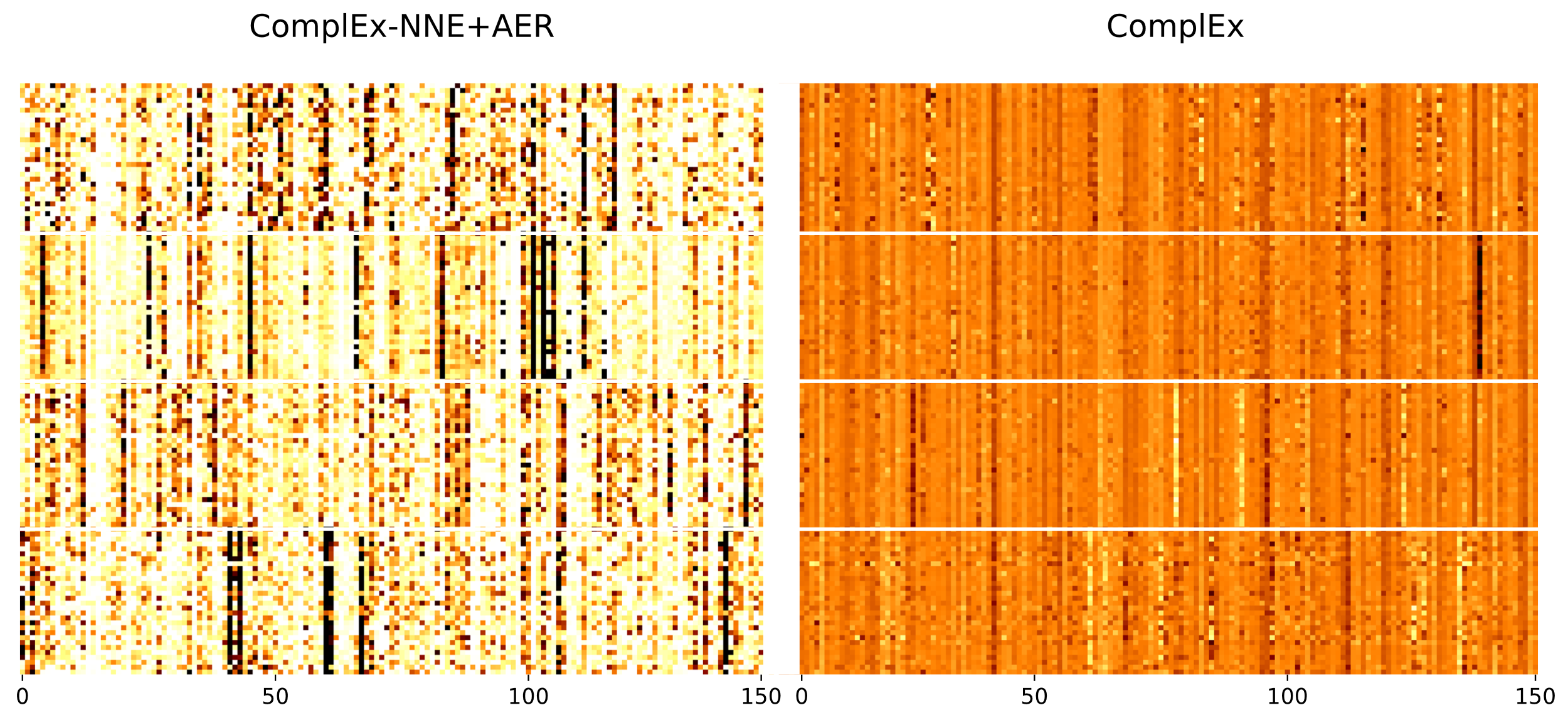


**4.2 可视化分析**

接下来，我们对实体和关系的可解释性做一个可视化的分析。

**4.2.1 实体语义分布热图**

我们从四个类别（从上到下依此reptile,wine\_region,species,programming\_language）中分别随机选取了30个实体，将这些实体向量堆叠起来（横轴代表维度），然后绘制热图，图中的颜色越深的地方代表值越大，我们可以从图中看出，带约束的模型所学到的向量（左边），相比不带约束的模型学到的向量（右边）更加稀疏；同时，左边图中，同一类型的实体，也能够激活一致地相同的维度，说明有关这些类型的语义也集中在了这些维度

S

**4.2.2 实体语义纯净度曲线**

语义纯净度是指对于一个维度，该维度取值较大的实体都应该尽可能的属于同一类别，我们用熵来衡量语义纯净度；熵越低，语义纯净度越高。图中绘制了ComplEx，ComplEx-NNE，ComplEx-NNE-AER每个维度的平均语义纯净度，可以看到ComplEx-NNE，ComplEx-NNE-AER具有更高的语义纯净度。

图片包含 天空, 地图

已生成高可信度的说明

**4.2.3 关系可视化分析**

我们选取了三组蕴含关系（每一组两个，处在上方的关系是蕴含式的左部，下方的是蕴含式的右部）分别是等价，逆反，一般蕴含；等价和逆反是一般蕴含的特例，等价要求对于任意两个实体对，存在，而逆反要求。而对于他们的充分条件，等价要求两个关系满足,，逆反要求满足,一般蕴含只需满足，详细推导可参见论文附录。我们可以看到，图中的关系对均能满足上面的约束。

