arXiv:1705.02426v2 [cs.LG] 6 Jul 2017

Analogical Inference for Multi-relational Embeddings

Hanxiao Liu [1] **Yuexin Wu** [1] **Yiming Yang** [1]

Abstract

Large-scale multi-relational embedding refers to the task of learning the latent representations for entities and relations in large knowledge graphs. An effective and scalable solution for this problem is crucial for the true success of knowledge-based inference in a broad range of applications. This paper proposes a novel framework for optimizing the latent representations with respect to the *analogical* properties of the embedded entities and relations. By formulating the learning objective in a differentiable fashion, our model enjoys both theoretical power and computational scalability, and significantly outperformed a large number of representative baseline methods on benchmark datasets. Furthermore, the model offers an elegant unification of several well-known methods in multi-relational embedding, which can be proven to be special instantiations of our framework.

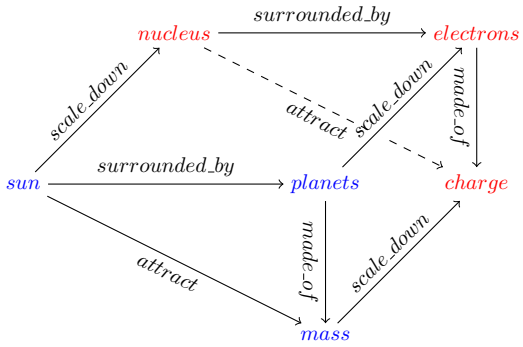

Figure 1. Commutative diagram for the analogy between the Solar System (red) and the Rutherford-Bohr Model (blue) (atom system). By viewing the atom system as a "miniature" of the solar system (via the $scale\_down$ relation), one is able to complete missing facts (triplets) about the latter by mirroring the facts about the former. The analogy is built upon three basic analogical structures (parallelograms): "sun is to $planets$ as $nucleus$ is to $electrons$", "sun is to $mass$ as $nucleus$ is to $charge$" and "$planets$ are to $mass$ as $eletrons$ are to $charge$".

通过类比的关系推断link

1. Introduction

Multi-relational embedding, or knowledge graph embedding, is the task of finding the latent representations of entities and relations for better inference over knowledge graphs. This problem has become increasingly important in recent machine learning due to the broad range of important applications of large-scale knowledge bases, such as Freebase (Bollacker et al., 2008), DBpedia (Auer et al., 2007) and Google's Knowledge Graph (Singhal, 2012), including question-answering (Ferrucci et al., 2010), information retrieval (Dalton et al., 2014) and natural language processing (Gabrilovich & Markovitch, 2009).

A knowledge base (KB) typically stores factual information as subject-relation-object triplets. The collection of such triplets forms a directed graph whose nodes are entities and whose edges are the relations among entities. Real-world knowledge graph is both extremely large and highly incomplete by nature (Min et al., 2013). How can we use the observed triplets in an incomplete graph to induce the unobserved triples in the graph presents a tough challenge for machine learning research.

Various statistical relational learning methods (Getoor, 2007; Nickel et al., 2015) have been proposed for this task, among which vector-space embedding models are most particular due to their advantageous performance and scalability (Bordes et al., 2013). The key idea in those approaches is to find dimensionality reduced representations for both the entities and the relations, and hence force the models to generalize during the course of compression. Representative models of this kind include tensor factorization (Singhal, 2012; Nickel et al., 2011), neural tensor networks (Socher et al., 2013; Chen et al., 2013), translation-based models (Bordes et al., 2013; Wang et al., 2014; Lin et al., 2015b), bilinear models and its variants (Yang et al., 2014; Trouillon et al., 2016), pathwise methods (Guu et al., 2015), embeddings based on holographic representations (Nickel et al., 2016), and product graphs that utilizes additional site information for the predictions of unseen edges in a semi-supervised manner (Liu & Yang, 2015; 2016).

[1]Carnegie Mellon University, Pittsburgh, PA 15213, USA. Correspondence to: Hanxiao Liu <hanxiaol@cs.cmu.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

Learning the embeddings of entities and relations can be viewed as a knowledge induction process, as those induced latent representations can be used to make inference about new triplets that have not been seen before.

Despite the substantial efforts and great successes so far in the research on multi-relational embedding, one important aspect is missing, i.e., to study the solutions of the problem from the *analogical inference* point of view, by which we mean to rigorously define the desirable analogical properties for multi-relational embedding of entities and relations, and to provide algorithmic solution for optimizing the embeddings w.r.t. the analogical properties. We argue that analogical inference is particularly desirable for knowledge base completion, since for instance if system A (a subset of entities and relations) is analogous to system B (another subset of entities and relations), then the unobserved triples in B could be inferred by mirroring their counterparts in A. Figure 1 uses a toy example to illustrate the intuition, where system A corresponds to the solar system with three concepts (entities), and system B corresponds the atom system with another three concepts. An analogy exists between the two systems because B is a "miniature" of A. As a result, knowing how the entities are related to each other in system A allows us to make inference about how the entities are related to each other in system B by analogy.

类比推理的解释

Although *analogical reasoning* was an active research topic in classic AI (artificial intelligence), early computational models mainly focused on non-differentiable rule-based reasoning (Gentner, 1983; Falkenhainer et al., 1989; Turney, 2008), which can hardly scale to very large KBs such as Freebase or Google's Knowledge Graph. How to leverage the intuition of analogical reasoning via statistical inference for automated embedding of very large knowledge graphs has not been studied so far, to our knowledge.

It is worth mentioning that analogical structures have been observed in the output of several word/entity embedding models (Mikolov et al., 2013; Pennington et al., 2014). However, those observations stopped there as merely empirical observations. Can we mathematically formulate the desirable analogical structures and leverage them in our objective functions to improve multi-relational embedding? In this case, can we develop new algorithms for tractable inference for the embedding of very large knowledge graphs? These questions present a fundamental challenge which has not been addressed by existing work, and answering these questions are the main contributions we aim in this paper. We name this open challenge as the *analogical inference* problem, for the distinction from rule-based *analogical reasoning* in classic AI.

Our specific novel contributions are the following:

论文的贡献

1. A new framework that, for the first time, explicitly models analogical structures in multi-relational embedding, and that improves the state-of-the-art performance on benchmark datasets;
2. The algorithmic solution for conducting analogical inference in a differentiable manner, whose implementation is as scalable as the fastest known relational embedding algorithms;
3. The theoretical insights on how our framework provides a unified view of several representative methods as its special (and restricted) cases, and why the generalization of such cases lead to the advantageous performance of our method as empirically observed.

The rest of this paper is organized as follows: §2 introduces related background where multi-relational embedding is formulated as linear maps. §3 describes our new optimization framework where the desirable analogical structures are rigorously defined by the the commutative property of linear maps. §4 offers an efficient algorithm for scalable inference by exploiting the special structures of commutative linear maps, §5 shows how our framework subsumes several representative approaches in a principled way, and §6 reports our experimental results, followed by the concluding remarks in §7.

2. Related Background

2.1. Notations

Let $\mathcal{E}$ and $\mathcal{R}$ be the space of all entities and their relations. A knowledge base $\mathcal{K}$ is a collection of triplets $(s, r, o) \in \mathcal{K}$ where $s \in \mathcal{E}, o \in \mathcal{E}, r \in \mathcal{R}$ stand for the subject, object and their relation, respectively. Denote by $v \in \mathbb{R}^{|\mathcal{E}| \times m}$ a look-up table where $v_e \in \mathbb{R}^m$ is the vector embedding for entity e, and denote by tensor $W \in \mathbb{R}^{|\mathcal{R}| \times m \times m}$ another look-up table where $W_r \in \mathbb{R}^{m \times m}$ is the matrix embedding for relation r. Both v and W are to be learned from $\mathcal{K}$.

2.2. Relations as Linear Maps

We formulate each relation r as a linear map that, for any given $(s, r, o) \in \mathcal{K}$, transforms the subject s from its original position in the vector space to somewhere near the object o. In other words, we expect the latent representations for any valid (s, r, o) to satisfy

$$v_s^\top W_r \approx v_o^\top \tag{1}$$

The degree of satisfaction in the approximated form of (1) can be quantified using the inner product of $v_s^\top W_r$ and v_o. That is, we define a bilinear score function as:

线性映射的评分函数

$$\phi(s, r, o) = \langle v_s^\top W_r, v_o \rangle = v_s^\top W_r v_o \quad \text{RESCAL} \tag{2}$$

Our goal is to learn v and W such that $\phi(s, r, o)$ gives high scores to valid triples, and low scores to the invalid ones.

In contrast to some previous models (Bordes et al., 2013) where relations are modeled as additive translating operators, namely $v_s + w_r \approx v_o$, the multiplicative formulation in (1) offers a natural analogy to the first-order logic where each relation is treated as a predicate operator over input arguments (subject and object in our case). Clearly, the linear transformation defined by a matrix, a.k.a. a linear map, is a richer operator than the additive transformation defined by a vector. Multiplicative models are also found to substantially outperform additive models empirically (Nickel et al., 2011; Yang et al., 2014).

2.3. Normal Transformations

Instead of allowing arbitrary linear maps to be used for representing relations, a particular family of matrices has been studied for "well-behaved" linear maps. This family is named as the *normal matrices*.

正规矩阵

Definition 2.1 (Normal Matrix)**.** *A real matrix A is normal if and only if $A^\top A = AA^\top$.*

Normal matrices have nice theoretical properties which are often desirable form relational modeling, e.g., they are unitarily diagonalizable and hence can be conveniently analyzed by the spectral theorem (Dunford et al., 1971). Representative members of the normal family include:

- Symmetric Matrices for which $W_r W_r^\top = W_r^\top W_r = W_r^2$. These includes all diagonal matrices and positive semi-definite matrices, and the symmetry implies $\phi(s, r, o) = \phi(o, r, s)$. They are suitable for modeling symmetric relations such as $is\_identical$.

对称关系

- Skew-/Anti-symmetric Matrices for which $W_r W_r^\top = W_r^\top W_r = -W_r^2$, which implies $\phi(s, r, o) = -\phi(o, r, s)$. These matrices are suitable for modeling asymmetric relations such as is_parent_of.

非对称关系

- Rotation Matrices for which $W_r W_r^\top = W_r^\top W_r = I_m$, which suggests that the relation r is invertible as W_r^{-1} always exists. Rotation matrices are suitable for modeling 1-to-1 relationships (bijections).

1对1关系

- Circulant Matrices (Gray et al., 2006), which have been implicitly used in recent work on holographic representations (Nickel et al., 2016). These matrices are usually related to the learning of latent representations in the Fourier domain (see §5 for more details).

In the remaining parts of this paper, we denote all the real normal matrices in $\mathbb{R}^{m\times m}$ as $\mathcal{N}_m(\mathbb{R})$.

3. Proposed Analogical Inference Framework

Analogical reasoning is known to play a central role in human induction about knowledge (Gentner, 1983; Minsky, 1988; Holyoak et al., 1996; Hofstadter, 2001). Here we provide a mathematical formulation of the analogical structures of interest in multi-relational embedding in a latent semantic space, to support algorithmic inference about the embeddings of entities and relations in a knowledge graph.

3.1. Analogical Structures

Consider the famous example in the word embedding literature (Mikolov et al., 2013; Pennington et al., 2014), for the following entities and relations among them:

"man is to $king$ as $woman$ is to $queen$"

In an abstract notion we denote the entities by a (as man) , b (as $king$), c (as $woman$) and d (as $queen$), and the relations by r (as $crown$) and r' (as $male \mapsto female$), respectively. These give us the subject-relation-object triplets as follows:

$$a \xrightarrow{r} b, \quad c \xrightarrow{r} d, \quad a \xrightarrow{r'} c, \quad b \xrightarrow{r'} d \tag{3}$$

For multi-relational embeddings, r and r' are members of $\mathcal{R}$ and are modeled as linear maps in our case.

The relational maps in (3) can be visualized using a commutative diagram (Adámek et al., 2004; Brown & Porter, 2006) from the Category Theory, as shown in Figure 2, where each node denotes an entity and each edge denotes a linear map that transforms one entity to the other. We also refer to such a diagram as a "parallelogram" to highlight its particular *algebraic structure*[1].

使用例子说明类比关系以及使用类比关系进行推导

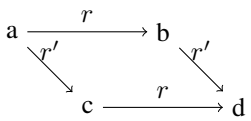

Figure 2. Parallelogram diagram for the analogy of "a is to b as c is to d", where each edge denotes a linear map.

The parallelogram in Figure 2 represents a very basic analogical structure which could be informative for the inference about unknown facts (triplets). To get a sense about why analogies would help in the inference about unobserved facts, we notice that for entities a, b, c, d which form an analogical structure in our example, the parallelogram structure is fully determined by symmetry. This means that if we know $a \xrightarrow{r} b$ and $a \xrightarrow{r'} c$, then we can induce the remaining triplets of $c \xrightarrow{r} d$ and $b \xrightarrow{r'} d$. In other words, understanding the relation between man and $king$ helps us to fill up the unknown relation between $woman$ and $queen$.

[1]Notice that this is different from parallelograms in the geometric sense because each edge here is a linear map instead of the difference between two nodes in the vector space.

Analogical structures are not limited to parallelograms, of course, though parallelograms often serve as the building blocks for more complex analogical structures. As an example, in Figure 1 of §1 we show a compound analogical structure in the form of a triangular prism, for mirroring the correspondent entities/relations between the atom system and the solar system. Formally define the desirable analogical structures in a computationally tractable objective for optimization is the key for solving our problem, which we will introduce next.

3.2. Commutative Constraint for Linear Maps

线性映射还需要满足的约束

Although it is tempting to explore all potentially interesting parallelograms in the modeling of analogical structure, it is computationally intractable to examine the entire powerset of entities as the candidate space of analogical structures. A more reasonable strategy is to identify some desirable properties of the analogical structures we want to model, and use those properties as constraints for reducing the candidate space.

An desirable property of the linear maps we want is that all the directed paths with the same starting node and end node form the *compositional equivalence*. Denoting by "$\circ$" the composition operator between two relations, the parallelogram in Figure 2 contains two equivalent compositions as:

$$r \circ r' = r' \circ r \tag{4}$$

which means that a is connected to d via either path. We call this the *commutativity* property of the linear maps, which is a necessary condition for forming commutative parallelograms and therefore the corresponding analogical structures. Yet another example is given by Figure 1, where sun can traverse to $charge$ along multiple alternative paths of length three, implying the commutativity of relations $surrounded\_by$, $made\_of$, $scale\_down$.

The composition of two relations (linear maps) is naturally implemented via matrix multiplication (Yang et al., 2014; Guu et al., 2015), hence equation (4) indicates

交换性

a从b到d和从c到d是一样的

$$W_{r \circ r'} = W_r W_{r'} = W_{r'} W_r \tag{5}$$

One may further require the commutative constraint (5) to be satisfied for any pair of relations in $\mathcal{R}$ because they may be simultaneously present in the same commutative parallelogram for certain subsets of entities. In this case, we say the relations in $\mathcal{R}$ form a commuting family.

It is worth mentioning that $\mathcal{N}_m(\mathbb{R})$ is not closed under matrix multiplication. As the result, the composition rule in eq. (5) may not always yield a legal new relation—$W_{r \circ r'}$ may no longer be a normal matrix. However, any commuting family in $\mathcal{N}_m(\mathbb{R})$ is indeed closed under multiplication. This explains the necessity of having a commuting family of relations from an alternative perspective.

3.3. The Optimization Objective

The generic goal for multi-relational embedding is to find entity and relation representations such that positive triples labeled as $y = +1$ receive higher score than the negative triples labeled as $y = -1$. This can be formulated as

$$\min_{v,W} \ \mathbb{E}_{s,r,o,y \sim \mathcal{D}} \ \ell\left(\phi_{v,W}(s,r,o), y\right) \tag{6}$$

where $\phi_{v,W}(s,r,o) = v_s^\top W_r v_o$ is our score function based on the embeddings, ℓ is our loss function, and $\mathcal{D}$ is the data distribution constructed based on the training set $\mathcal{K}$.

To impose analogical structures among the representations, we in addition require the linear maps associated with relations to form a commuting family of normal matrices. This gives us the objective function for ANALOGY:

$$\min_{v,W} \ \mathbb{E}_{s,r,o,y \sim \mathcal{D}} \ \ell\left(\phi_{v,W}(s,r,o), y\right) \tag{7}$$

$$\text{s.t.} \ \ W_r W_r^\top = W_r^\top W_r \quad \forall r \in \mathcal{R} \tag{8}$$

$$W_r W_{r'} = W_{r'} W_r \quad \forall r, r' \in \mathcal{R} \tag{9}$$

where constraints (8) and (9) are corresponding to the normality and commutativity requirements, respectively. Such a constrained optimization may appear to be computationally expensive at the first glance. In §4, however, we will recast it as a simple lightweight problem for which each SGD update can be carried out efficiently in $O(m)$ time.

直接在上面的条件下优化计算是困难的
可以进行优化

4. Efficient Inference Algorithm

The constrained optimization (7) is computationally challenging due to the large number of model parameters in tensor W, the matrix normality constraints, and the quadratic number of pairwise commutative constraints in (9).

Interestingly, by exploiting the special properties of commuting normal matrices, we will show in Corollary 4.2.1 that ANALOGY can be alternatively solved via an another formulation of substantially lower complexity. Our findings are based on the following lemma and theorem:

Lemma 4.1. *(Wilkinson & Wilkinson, 1965) For any real normal matrix A, there exists a real orthogonal matrix Q and a block-diagonal matrix B such that $A = QBQ^\top$, where each diagonal block of B is either (1) A real scalar, or (2) A 2-dimensional real matrix in the form of $\begin{bmatrix} x & -y \\ y & x \end{bmatrix}$, where both x, y are real scalars.*

M_r是块对角矩阵，每个块只能是这两种其中之一

The lemma suggests any real normal matrix can be block-diagonalized into an almost-diagonal canonical form.

Theorem 4.2 (Proof given in the supplementary material)**.** *If a set of real normal matrices $A_1, A_2, \ldots$ form a commuting family, namely $A_i A_j = A_j A_i \ \forall i, j$, then they can be block-diagonalized by the same real orthogonal basis Q.*

块对角矩阵则是由多个对角块组成的矩阵，每个对角块可以有自己的大小和结构

The theorem above implies that the set of dense relational matrices $\{W_r\}_{r\in\mathcal{R}}$, if mutually commutative, can always be *simultaneously block-diagonalized* into another set of sparse almost-diagonal matrices $\{B_r\}_{r\in\mathcal{R}}$.

Corollary 4.2.1 (Alternative formulation for ANALOGY)**.** *For any given solution (v^*, W^*) of optimization (7), there always exists an alternative set of embeddings (u^*, B^*) such that $\phi_{v^*,W^*}(s,r,o) \equiv \phi_{u^*,B^*}(s,r,o)$, $\forall(s,r,o)$, and (u^*, B^*) is given by the solution of:*

$$\min_{u,B} \ \mathbb{E}_{s,r,o,y\sim\mathcal{D}} \ \ell\left(\phi_{u,B}(s,r,o), y\right) \tag{10}$$

$$B_r \in \mathcal{B}_m^n \quad \forall r \in \mathcal{R} \tag{11}$$

where $\mathcal{B}_m^n$ denotes all $m \times m$ almost-diagonal matrices in Lemma 4.1 with $n < m$ real scalars on the diagonal.

proof sketch. With the commutative constraints, there must exist some orthogonal matrix Q, such that $W_r = QB_rQ^\top$, $B_r \in \mathcal{B}_m^n$, $\forall r \in \mathcal{R}$. We can plug-in these expressions into optimization (7) and let $u = vQ$, obtaining

$$\phi_{v,W}(s,r,o) = v_s^\top W_r v_o = v_s^\top Q B_r Q^\top v_o \tag{12}$$

$$= u_s^\top B_r u_o = \phi_{u,B}(s,r,o) \tag{13}$$

In addition, it is not hard to verify that constraints (8) and (9) are automatically satisfied by exploiting the facts that Q is orthogonal and $\mathcal{B}_m^n$ is a commutative normal family. □

Constraints (11) in the alternative optimization problem can be handled by simply binding together the coefficients within each of those 2×2 blocks in B_r. Note that each B_r consists of only m free parameters, allowing the gradient w.r.t. any given triple to be efficiently evaluated in $O(m)$.

使用ANALOGY推导其他方法

5. Unified View of Representative Methods

In the following we provide a unified view of several embedding models (Yang et al., 2014; Trouillon et al., 2016; Nickel et al., 2016), by showing that they are restricted versions under our framework, hence are implicitly imposing analogical properties. This explains their strong empirical performance as compared to other baselines (§6).

5.1. DistMult

DistMult (Yang et al., 2014) embeds both entities and relations as vectors, and defines the score function as

$$\phi(s,r,o) = \langle v_s, v_r, v_o \rangle \tag{14}$$

$$\text{where} \ \ v_s, v_r, v_o \in \mathbb{R}^m, \forall s,r,o \tag{15}$$

where $\langle\cdot,\cdot,\cdot\rangle$ denotes the generalized inner product.

Proposition 5.1. *DistMult embeddings can be fully recovered by ANALOGY embeddings when $n = m$.*

有n个对角块，n=m，则每一个对角块都是实数标量，整个矩阵也就是对角矩阵

Proof. This is trivial to verify as the score function (15) can be rewritten as $\phi(s,r,o) = v_s^\top B_r v_o$ where B_r is a diagonal matrix given by $B_r = \text{diag}(v_r)$. □

Entity analogies are encouraged in DistMult as the diagonal matrices $\text{diag}(v_r)$'s are both normal and mutually commutative. However, DistMult is restricted to model symmetric relations only, since $\phi(s,r,o) \equiv \phi(o,r,s)$.

5.2. Complex Embeddings (ComplEx)

ComplEx (Trouillon et al., 2016) extends the embeddings to the complex domain $\mathbb{C}$, which defines

$$\phi(s,r,o) = \Re\left(\langle v_s, v_r, \overline{v_o} \rangle\right) \tag{16}$$

$$\text{where} \ \ v_s, v_r, v_o \in \mathbb{C}^m, \forall s,r,o \tag{17}$$

where $\overline{x}$ denotes the complex conjugate of x.

Proposition 5.2. *ComplEx embeddings of embedding size m can be fully recovered by ANALOGY embeddings of embedding size $2m$ when $n = 0$.*

Proof. Let $\Re(x)$ and $\Im(x)$ be the real and imaginary parts of any complex vector x. We recast ϕ in (16) as

$$\phi(r,s,o) = +\left\langle \Re(v_r), \Re(v_s), \Re(v_o) \right\rangle \tag{18}$$

$$+\left\langle \Re(v_r), \Im(v_s), \Im(v_o) \right\rangle \tag{19}$$

$$+\left\langle \Im(v_r), \Re(v_s), \Im(v_o) \right\rangle \tag{20}$$

$$-\left\langle \Im(v_r), \Im(v_s), \Re(v_o) \right\rangle = {v'_s}^\top B_r v'_o \tag{21}$$

The last equality is obtained via a change of variables: For any complex entity embedding $v \in \mathbb{C}^m$, we define a new real embedding $v' \in \mathbb{R}^{2m}$ such that

$$\begin{cases} (v')_{2k} & = \Re(v)_k \\ (v')_{2k-1} & = \Im(v)_k \end{cases} \quad \forall k = 1, 2, \ldots m \tag{22}$$

The corresponding B_r is a block-diagonal matrix in $\mathcal{B}_{2m}^0$ with its k-th block given by $\begin{bmatrix} \Re(v_r)_k & -\Im(v_r)_k \\ \Im(v_r)_k & \Re(v_r)_k \end{bmatrix}$. □

第i个对角块

5.3. Holographic Embeddings (HolE)

HolE (Nickel et al., 2016) defines the score function as

$$\phi(s,r,o) = \langle v_r, v_s * v_o \rangle \tag{23}$$

$$\text{where} \ \ v_s, v_r, v_o \in \mathbb{R}^m, \forall s,r,o \tag{24}$$

where the association of s and o is implemented via circular correlation denoted by $*$. This formulation is motivated by the holographic reduced representation (Plate, 2003).

To relate HolE with ANALOGY, we rewrite (24) in a bilinear form with a circulant matrix $C(v_r)$ in the middle

$$\phi(r,s,o) = v_s^\top C(v_r) v_o \tag{25}$$

where entries of a circulant matrix are defined as

$$C(x) = \begin{bmatrix} x_1 & x_m & \cdots & x_3 & x_2 \\ x_2 & x_1 & x_m & & x_3 \\ \vdots & x_2 & x_1 & \ddots & \vdots \\ x_{m-1} & & \ddots & \ddots & x_m \\ x_m & x_{m-1} & \cdots & x_2 & x_1 \end{bmatrix} \tag{26}$$

It is not hard to verify that circulant matrices are normal and commute (Gray et al., 2006), hence entity analogies are encouraged in HolE, for which optimization (7) reduces to an unconstrained problem as equalities (8) and (9) are automatically satisfied when all W_r's are circulant.

The next proposition further reveals that HolE is equivalent to ComplEx with minor relaxation.

Proposition 5.3. *HolE embeddings can be equivalently obtained using the following score function*

$$\phi(s, r, o) = \Re\left(\langle v_s, v_r, \overline{v_o}\rangle\right) \tag{27}$$

$$\textit{where } \; v_s, v_r, v_o \in \mathfrak{F}(\mathbb{R}^m), \forall s, r, o \tag{28}$$

where $\mathfrak{F}(\mathbb{R}^m)$ denotes the image of $\mathbb{R}^m$ in $\mathbb{C}^m$ through the Discrete Fourier Transform (DFT). In particular, the above reduces to ComplEx by relaxing $\mathfrak{F}(\mathbb{R}^m)$ to $\mathbb{C}^m$.

Proof. Let $\mathfrak{F}$ be the DFT operator defined by $\mathfrak{F}(x) = Fx$ where $F \in \mathbb{C}^{m\times m}$ is called the Fourier basis of DFT. A well-known property for circulant matrices is that any $C(x)$ can always be diagonalized by F, and its eigenvalues are given by Fx (Gray et al., 2006).

Hence the score function can be further recast as

$$\phi(r, s, o) = v_s^\top F^{-1} \operatorname{diag}(Fv_r) Fv_o \tag{29}$$

$$= \frac{1}{m} \overline{(Fv_s)}^\top \operatorname{diag}(Fv_r)(Fv_o) \tag{30}$$

$$= \frac{1}{m} \langle \overline{\mathfrak{F}(v_s)}, \mathfrak{F}(v_r), \mathfrak{F}(v_o)\rangle \tag{31}$$

$$= \Re\left[\frac{1}{m} \langle \overline{\mathfrak{F}(v_s)}, \mathfrak{F}(v_r), \mathfrak{F}(v_o)\rangle\right] \tag{32}$$

Let $v'_s = \overline{\mathfrak{F}(v_s)}, v'_o = \overline{\mathfrak{F}(v_o)}$ and $v'_r = \frac{1}{m}\mathfrak{F}(v_r)$, we obtain exactly the same score function as used in ComplEx

$$\phi(s, r, o) = \Re\left(\langle v'_s, v'_r, \overline{v'_o}\rangle\right) \tag{33}$$

(33) is equivalent to (16) apart from an additional constraint that v'_s, v'_r, v'_o are the image of $\mathbb{R}$ in the Fourier domain. □

6. Experiments

6.1. Datasets

We evaluate ANALOGY and the baselines over two benchmark datasets for multi-relational embedding released by previous work (Bordes et al., 2013), namely a subset of Freebase (FB15K) for generic facts and WordNet (WN18) for lexical relationships between words.

The dataset statistics are summarized in Table 1.

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	#train	#valid	#test
FB15K	14,951	1,345	483,142	50,000	59,071
WN18	40,943	18	141,442	5,000	5,000

Table 1. Dataset statistics for FB15K and WN18.

6.2. Baselines

We compare the performance of ANALOGY against a variety types of multi-relational embedding models developed in recent years. Those models can be categorized as:

- Translation-based models where relations are modeled as translation operators in the embedding space, including TransE (Bordes et al., 2013) and its variants TransH (Wang et al., 2014), TransR (Lin et al., 2015b), TransD (Ji et al., 2015), STransE (Nguyen et al., 2016) and RTransE (Garcia-Duran et al., 2015).
- Multi-relational latent factor models including LFM (Jenatton et al., 2012) and RESCAL (Nickel et al., 2011) based collective matrix factorization.
- Models involving neural network components such as neural tensor networks (Socher et al., 2013) and PTransE-RNN (Lin et al., 2015b), where RNN stands for recurrent neural networks.
- Pathwise models including three different variants of PTransE (Lin et al., 2015a) which extend TransE by explicitly taking into account indirect connections (relational paths) between entities.
- Models subsumed under our proposed framework (§5), including DistMult (Yang et al., 2014) based simple multiplicative interactions, ComplEx (Trouillon et al., 2016) using complex coefficients and HolE (Nickel et al., 2016) based on holographic representations. Those models are implicitly leveraging analogical structures per our previous analysis.
- Models enhanced by external side information. We use Node+LinkFeat (NLF) (Toutanova & Chen, 2015) as a representative example, which leverages textual mentions derived from the ClueWeb corpus.

6.3. Evaluation Metrics

Following the literature of multi-relational embedding, we use the conventional metrics of Hits@k and Mean Reciprocal Rank (MRR) which evaluate each system-produced

ranked list for each test instance and average the scores over all ranked lists for the entire test set of instances.

The two metrics would be flawed for the *negative instances* created in the test phase as a ranked list may contain some positive instances in the training and validation sets (Bordes et al., 2013). A recommended remedy, which we followed, is to remove all training- and validation-set triples from all ranked lists during testing. We use "filt." and "raw" to indicate the evaluation metrics with or without filtering, respectively.

In the first set of our experiments, we used on Hits@k with k=10, which has been reported for most methods in the literature. We also provide additional results of ANALOGY and a subset of representative baseline methods using MRR, Hits@1 and Hits@3, to enable the comparison with the methods whose published results are in those metrics.

6.4. Implementation Details

6.4.1. Loss Function

We use the logistic loss for ANALOGY throughout all experiments, namely $\ell(\phi(s,r,o),y) = -\log\sigma(y\phi(s,r,o))$, where σ is the sigmoid activation function. We empirically found this simple loss function to perform reasonably well as compared to more sophisticated ranking loss functions.

6.4.2. Asynchronous AdaGrad

Our C++ implementation[2] runs over a CPU, as ANALOGY only requires lightweight linear algebra routines. We use asynchronous stochastic gradient descent (SGD) for optimization, where the gradients with respect to different mini-batches are simultaneously evaluated in multiple threads, and the gradient updates for the shared model parameters are carried out without synchronization. Asynchronous SGD is highly efficient, and causes little performance drop when parameters associated with different mini-batches are mutually disjoint with a high probability (Recht et al., 2011). We adapt the learning rate based on historical gradients using AdaGrad (Duchi et al., 2011).

6.4.3. Creation of Negative Samples

Since only valid triples (positive instances) are explicitly given in the training set, invalid triples (negative instances) need to be artificially created. Specifically, for every positive example (s,r,o), we generate three negative instances (s',r,o), (s,r',o), (s,r,o') by corrupting s, r, o with random entities/relations $s' \in \mathcal{E}$, $r' \in \mathcal{R}$, $o' \in \mathcal{E}$. The union of all positive and negative instances defines our data distribution $\mathcal{D}$ for SGD updates.

[2]Code available at https://github.com/quark0/ANALOGY.

Table 2. Hits@10 (filt.) of all models on WN18 and FB15K categories into three groups: (i) 19 baselines without modeling analogies; (ii) 3 baselines and our proposed ANALOGY which implicitly or explicitly enforce analogical properties over the induced embeddings (see §5); (iii) One baseline relying on large external data resources in addition to the provided training set.

Models	WN18	FB15K
Unstructured (Bordes et al., 2013)	38.2	6.3
RESCAL (Nickel et al., 2011)	52.8	44.1
NTN (Socher et al., 2013)	66.1	41.4
SME (Bordes et al., 2012)	74.1	41.3
SE (Bordes et al., 2011)	80.5	39.8
LFM (Jenatton et al., 2012)	81.6	33.1
TransH (Wang et al., 2014)	86.7	64.4
TransE (Bordes et al., 2013)	89.2	47.1
TransR (Lin et al., 2015b)	92.0	68.7
TKRL (Xie et al., 2016)	–	73.4
RTransE (Garcia-Duran et al., 2015)	–	76.2
TransD (Ji et al., 2015)	92.2	77.3
CTransR (Lin et al., 2015b)	92.3	70.2
KG2E (He et al., 2015)	93.2	74.0
STransE (Nguyen et al., 2016)	93.4	79.7
DistMult (Yang et al., 2014)	93.6	82.4
TransSparse (Ji et al., 2016)	93.9	78.3
PTransE-MUL (Lin et al., 2015a)	–	77.7
PTransE-RNN (Lin et al., 2015a)	–	82.2
PTransE-ADD (Lin et al., 2015a)	–	84.6
NLF (with external corpus) (Toutanova & Chen, 2015)	*94.3*	*87.0*
ComplEx (Trouillon et al., 2016)	**94.7**	84.0
HolE (Nickel et al., 2016)	**94.9**	73.9
Our ANALOGY	**94.7**	**85.4**

6.4.4. Model Selection

We conducted a grid search to find the hyperparameters of ANALOGY which maximize the filtered MRR on the validation set, by enumerating all combinations of the embedding size $m \in \{100, 150, 200\}$, ℓ_2 weight decay factor $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}\}$ of model coefficients v and W, and the ratio of negative over positive samples $\alpha \in \{3, 6\}$. The resulting hyperparameters for the WN18 dataset are $m = 200, \lambda = 10^{-2}, \alpha = 3$, and those for the FB15K dataset are $m = 200, \lambda = 10^{-3}, \alpha = 6$. The number of scalars on the diagonal of each B_r is always set to be $\frac{m}{2}$. We set the initial learning rate to be 0.1 for both datasets and adjust it using AdaGrad during optimization. All models are trained for 500 epochs.

6.5. Results

Table 2 compares the Hits@10 score of ANALOGY with that of 23 competing methods using the published scores

Table 3. MRR and Hits@{1,3} of a subset of representative models on WN18 and FB15K. The performance scores of TransE and REACAL are cf. the results published in (Trouillon et al., 2016) and (Nickel et al., 2016), respectively.

Models	WN18				FB15			
	MRR (filt.)	MRR (raw)	Hits@1 (filt.)	Hits@3 (filt.)	MRR (filt.)	MRR (raw)	Hits@1 (filt.)	Hits@3 (filt.)
RESCAL (Nickel et al., 2011)	89.0	60.3	84.2	90.4	35.4	18.9	23.5	40.9
TransE (Bordes et al., 2013)	45.4	33.5	8.9	82.3	38.0	22.1	23.1	47.2
DistMult (Yang et al., 2014)	82.2	53.2	72.8	91.4	65.4	24.2	54.6	73.3
HolE (Nickel et al., 2016)	93.8	61.6	**93.0**	**94.5**	52.4	23.2	40.2	61.3
ComplEx (Trouillon et al., 2016)	94.1	58.7	**93.6**	**94.5**	69.2	24.2	59.9	75.9
Our ANALOGY	**94.2**	**65.7**	**93.9**	**94.4**	**72.5**	**25.3**	**64.6**	**78.5**

for these methods in the literature on the WN18 and FB15K datasets. For the methods not having both scores, the missing slots are indicated by "–". The best score on each dataset is marked in the bold face; if the differences among the top second or third scores are not statistically significant from the top one, then these scores are also bold faced. We used one-sample proportion test (Yang & Liu, 1999) at the 5% p-value level for testing the statistical significances[3].

Table 3 compares the methods (including ours) whose results in additional metrics are available. The usage of the bold faces is the same as those in Table 2.

In both tables, ANALOGY performs either the best or the 2nd best which is in the equivalent class with the best score in each case according statistical significance test. Specifically, on the harder FB15K dataset in Table 2, which has a very large number of relations, our model outperforms all baseline methods. These results provide a good evidence for the effective modeling of analogical structures in our approach. We are pleased to see in Table 3 that ANALOGY outperforms DistMult, ComplEx and HolE in all the metrics, as the latter three can be viewed as more constrained versions of our method (as discussed in (§5)). Furthermore, our assertion on HolE for being a special case of ComplEx (§5) is justified in the same table by the fact that the performance of HolE is dominated by ComplEx.

In Figure 3 we show the empirical scalability of ANALOGY, which not only completes one epoch in a few seconds on both datasets, but also scales linearly in the size of the embedding problem. As compared to single-threaded AdaGrad, our asynchronous AdaGrad over 16 CPU threads offers 11.4x and 8.3x speedup on FB15K and WN18, respectively, on a single commercial desktop.

[3]Notice that proportion tests only apply to performance scores as proportions, including Hits@k, but are not applicable to non-proportional scores such as MRR. Hence we only conducted the proportion tests on the Hits@k scores.

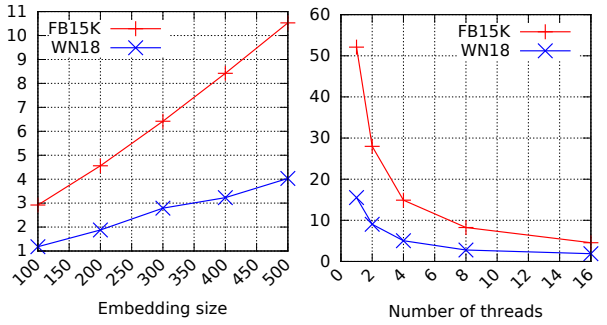

Figure 3. CPU run time per epoch (secs) of ANALOGY. The figure on the left shows the run time over increasing embedding sizes with 16 CPU threads; Figure on the right shows the run time over increasing number of CPU threads with embedding size 200.

7. Conclusion

We presented a novel framework for explicitly modeling analogical structures in multi-relational embedding, along with a differentiable objective function and a linear-time inference algorithm for large-scale embedding of knowledge graphs. The proposed approach obtains the state-of-the-art results on two popular benchmark datasets, outperforming a large number of strong baselines in most cases.

Although we only focused on the multi-relational inference for knowledge-base embedding, we believe that analogical structures exist in many other machine learning problems beyond the scope of this paper. We hope this work shed light on a broad range of important problems where scalable inference for analogical analysis would make an impact, such as machine translation and image captioning (both problems require modeling cross-domain analogies). We leave these interesting topics as our future work.

Acknowledgments

We thank the reviewers for their helpful comments. This work is supported in part by the National Science Founda-

tion (NSF) under grant IIS-1546329.

References

Adámek, Jiří, Herrlich, Horst, and Strecker, George E. Abstract and concrete categories. the joy of cats. 2004.

Auer, Sören, Bizer, Christian, Kobilarov, Georgi, Lehmann, Jens, Cyganiak, Richard, and Ives, Zachary. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pp. 722–735. Springer, 2007.

Bollacker, Kurt, Evans, Colin, Paritosh, Praveen, Sturge, Tim, and Taylor, Jamie. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250. AcM, 2008.

Bordes, Antoine, Weston, Jason, Collobert, Ronan, and Bengio, Yoshua. Learning structured embeddings of knowledge bases. In *Conference on artificial intelligence*, number EPFL-CONF-192344, 2011.

Bordes, Antoine, Glorot, Xavier, Weston, Jason, and Bengio, Yoshua. Joint learning of words and meaning representations for open-text semantic parsing. In *AISTATS*, volume 22, pp. 127–135, 2012.

Bordes, Antoine, Usunier, Nicolas, Garcia-Duran, Alberto, Weston, Jason, and Yakhnenko, Oksana. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pp. 2787–2795, 2013.

Brown, Ronald and Porter, Tim. Category theory: an abstract setting for analogy and comparison. In *What is category theory*, volume 3, pp. 257–274, 2006.

Chen, Danqi, Socher, Richard, Manning, Christopher D, and Ng, Andrew Y. Learning new facts from knowledge bases with neural tensor networks and semantic word vectors. *arXiv preprint arXiv:1301.3618*, 2013.

Dalton, Jeffrey, Dietz, Laura, and Allan, James. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 365–374. ACM, 2014.

Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Dunford, Nelson, Schwartz, Jacob T, Bade, William G, and Bartle, Robert G. *Linear operators*. Wiley-interscience New York, 1971.

Falkenhainer, Brian, Forbus, Kenneth D, and Gentner, Dedre. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63, 1989.

Ferrucci, David, Brown, Eric, Chu-Carroll, Jennifer, Fan, James, Gondek, David, Kalyanpur, Aditya A, Lally, Adam, Murdock, J William, Nyberg, Eric, Prager, John, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

Gabrilovich, Evgeniy and Markovitch, Shaul. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34: 443–498, 2009.

Garcia-Duran, Alberto, Bordes, Antoine, and Usunier, Nicolas. *Composing relationships with translations*. PhD thesis, CNRS, Heudiasyc, 2015.

Gentner, Dedre. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

Getoor, Lise. *Introduction to statistical relational learning*. MIT press, 2007.

Gray, Robert M et al. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006.

Guu, Kelvin, Miller, John, and Liang, Percy. Traversing knowledge graphs in vector space. *arXiv preprint arXiv:1506.01094*, 2015.

He, Shizhu, Liu, Kang, Ji, Guoliang, and Zhao, Jun. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 623–632. ACM, 2015.

Hofstadter, Douglas R. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pp. 499–538, 2001.

Holyoak, Keith J, Holyoak, Keith James, and Thagard, Paul. *Mental leaps: Analogy in creative thought*. MIT press, 1996.

Jenatton, Rodolphe, Roux, Nicolas L, Bordes, Antoine, and Obozinski, Guillaume R. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems*, pp. 3167–3175, 2012.

Ji, Guoliang, He, Shizhu, Xu, Liheng, Liu, Kang, and Zhao, Jun. Knowledge graph embedding via dynamic mapping matrix. In *ACL (1)*, pp. 687–696, 2015.

Ji, Guoliang, Liu, Kang, He, Shizhu, and Zhao, Jun. Knowledge graph completion with adaptive sparse transfer matrix. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pp. 985–991, 2016. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11982.

Lin, Yankai, Liu, Zhiyuan, Luan, Huanbo, Sun, Maosong, Rao, Siwei, and Liu, Song. Modeling relation paths for representation learning of knowledge bases. *arXiv preprint arXiv:1506.00379*, 2015a.

Lin, Yankai, Liu, Zhiyuan, Sun, Maosong, Liu, Yang, and Zhu, Xuan. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pp. 2181–2187, 2015b.

Liu, Hanxiao and Yang, Yiming. Bipartite edge prediction via transductive learning over product graphs. In *ICML*, pp. 1880–1888, 2015.

Liu, Hanxiao and Yang, Yiming. Cross-graph learning of multi-relational associations. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2235–2243, 2016.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Min, Bonan, Grishman, Ralph, Wan, Li, Wang, Chang, and Gondek, David. Distant supervision for relation extraction with an incomplete knowledge base. In *HLT-NAACL*, pp. 777–782, 2013.

Minsky, Marvin. *Society of mind*. Simon and Schuster, 1988.

Nguyen, Dat Quoc, Sirts, Kairit, Qu, Lizhen, and Johnson, Mark. Stranse: a novel embedding model of entities and relationships in knowledge bases. *arXiv preprint arXiv:1606.08140*, 2016.

Nickel, Maximilian, Tresp, Volker, and Kriegel, Hans-Peter. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 809–816, 2011.

Nickel, Maximilian, Murphy, Kevin, Tresp, Volker, and Gabrilovich, Evgeniy. A review of relational machine learning for knowledge graphs. *arXiv preprint arXiv:1503.00759*, 2015.

Nickel, Maximilian, Rosasco, Lorenzo, and Poggio, Tomaso A. Holographic embeddings of knowledge graphs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pp. 1955–1961, 2016. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12484.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–1543, 2014.

Plate, Tony A. Holographic reduced representation: Distributed representation for cognitive structures. 2003.

Recht, Benjamin, Re, Christopher, Wright, Stephen, and Niu, Feng. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 693–701, 2011.

Singhal, Amit. Introducing the knowledge graph: things, not strings. *Official google blog*, 2012.

Socher, Richard, Chen, Danqi, Manning, Christopher D, and Ng, Andrew. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pp. 926–934, 2013.

Toutanova, Kristina and Chen, Danqi. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015.

Trouillon, Théo, Welbl, Johannes, Riedel, Sebastian, Gaussier, Éric, and Bouchard, Guillaume. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 2071–2080, 2016. URL http://jmlr.org/proceedings/papers/v48/trouillon16.html.

Turney, Peter D. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655, 2008.

Wang, Zhen, Zhang, Jianwen, Feng, Jianlin, and Chen, Zheng. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pp. 1112–1119. Citeseer, 2014.

Wilkinson, James Hardy and Wilkinson, James Hardy. *The algebraic eigenvalue problem*, volume 87. Clarendon Press Oxford, 1965.

Xie, Ruobing, Liu, Zhiyuan, and Sun, Maosong. Representation learning of knowledge graphs with hierarchical types. In *Proceedings of the Twenty-Fifth International*

Joint Conference on Artificial Intelligence, pp. 2965–2971, 2016.

Yang, Bishan, Yih, Wen-tau, He, Xiaodong, Gao, Jianfeng, and Deng, Li. Embedding entities and relations for learning and inference in knowledge bases. *CoRR*, abs/1412.6575, 2014. URL http://arxiv.org/abs/1412.6575.

Yang, Yiming and Liu, Xin. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42–49. ACM, 1999.