Analogical Inference for Multi-Relational Embeddings

Hanxiao Liu, Yuexin Wu, Yiming Yang
Carnegie Mellon University

August 8, 2017

Task Description

Multi-Relational Embeddings:

- Finding latent representations of entities and relations.
- Useful for knowledge base completion (by discovering missing facts), etc.

Novel Contribution:

- Instead of tradition rule-based AI, we impose *analogical structures* in the learning of entity/relation embedding.

Why Analogy? (a toy example)

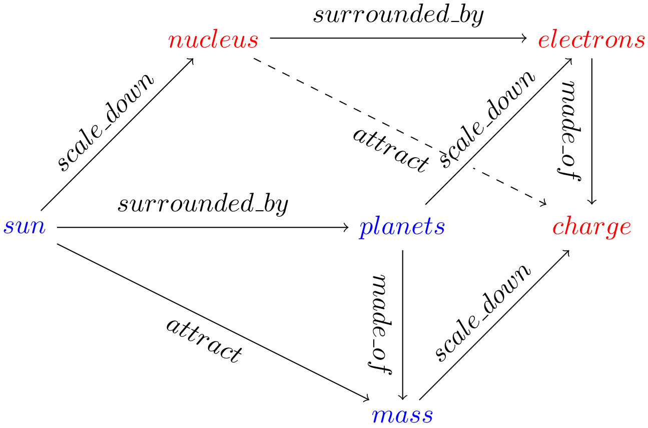

Figure: Solar System (blue) v.s. Atomic System (red).

Knowing the relational structure in one system will help us to understand the other system by analogy.

Basic Formulation

- Denote by vector v_e the embedding of entity e.
- Denote by matrix W_r in the embedding of relation r.
- Assume all valid subject-relation-object (s,r,o) triples approximately satisfy

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

- Define the scoring function of any (s,r,o) triple as:

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

Real Normal Matrices as Desirable

The family of matrices satisfying:

$$W_r^\top W_r = W_r W_r^\top \tag{3}$$

Special cases:

1. Symmetric Matrices
   - $\phi(s, r, o) = \phi(o, r, s)$. E.g. *is_identical*.
2. Skew-symmetric Matrices
   - $\phi(s, r, o) = -\phi(o, r, s)$. E.g. *is_parent_of*.
3. Orthogonal Matrices
   - Useful if r is a bijection (one-to-one mapping).

Commutative Matrices as Necessary

Observation: Analogical structures often imply "parallelograms", e.g.,

"man is to king as woman is to queen"

Or, in an abstract notion:

"a is to b as c is to d"

$$
\begin{tikzcd}
\mathrm{a} \arrow[r, "r"] \arrow[dr, "r'"] & \mathrm{b} \arrow[dr, "r'"] & \\
& \mathrm{c} \arrow[r, "r"] & \mathrm{d}
\end{tikzcd}
$$

Given the parallelogram, if we know $a \xrightarrow{r} b$ and $a \xrightarrow{r'} c$, then $c \xrightarrow{r} d$ and $b \xrightarrow{r'} d$ can be inferred by symmetry.

Commutative Matrices as Necessary (cont'd)

Mathematically, the necessary condition for having an analogical structure is the commutativity of relations:

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

$$\begin{tikzcd}
a \arrow[rr, "r"] \arrow[dr, "r'"'] & & b \arrow[dr, "r'"] & \\
& c \arrow[rr, "r"] & & d
\end{tikzcd}$$

Equivalently, we want the following constraint:

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

Optimization: Straightforward Formulation

Notation: Label $y = +1$ for positive examples and -1 otherwise; Data distribution $\mathcal{D}$; Loss function ℓ.

$$\begin{aligned} \min_{\mathbf{v},\mathbf{W}} \quad & \mathbb{E}_{s,r,o,y\sim\mathcal{D}} \ \ell\left(\phi_{\mathbf{v},\mathbf{W}}(s,r,o), y\right) && (6) \\ \text{s.t.} \quad & W_r W_r^\top = W_r^\top W_r \quad \forall r && (7) \\ & W_r W_{r'} = W_{r'} W_r \quad \forall r, r' && (8) \end{aligned}$$

- (7) follows the definition of normal matrices.
- (8) is for the communicative property.

The OPT is expensive due to (i) W_r's are fully dense matrices (ii) large number of equality constraints.

Optimization: Complexity Reduced Version

Solution $\mathbf{v}^*, \mathbf{W}^*$ for the previous OPT can be exactly recovered by solution $\mathbf{v'}^*, \mathbf{W'}^*$ of the following problem:

$$\min_{\mathbf{v'}, \mathbf{W'}} \mathbb{E}_{s,r,o,y \sim \mathcal{D}} \ \ell\left(\phi_{\mathbf{v'},\mathbf{W'}}(s, r, o), y\right) \tag{9}$$

Most notably,

- We show that any W'_r must be block-diagonal with the diagonal block sizes bounded by 2.
  - $O(m)$ free parameters in the $m \times m$ matrix.
- We now have an unconstrained optimization instead.
  - Efficiently solved using SGD without projection.

A Unified View of Existing Work

We explain the strong empirical performance of

- DistMult (Yang et al., ICLR 2015)
- ComplEx (Trouillon et al., ICML 2016)
- HolE (Nickel et al., AAAI 2016)

by showing that they are implicitly imposing analogical structures and are restricted cases of ours.

Connections to Existing Work

Multiplicative Embeddings (DistMult)

$$\phi(s, r, o) = \langle v_s, v_r, v_o \rangle \tag{10}$$

$$\text{where } v_s, v_r, v_o \in \mathbb{R}^m, \forall s, r, o \tag{11}$$

DistMult embeddings of size m can be fully recovered by ANALOGY embeddings of size m.

Intuition: v_r can be viewed as a diagonal $W_r \overset{def}{=} \text{diag}(v_r)$. Diagonal matrices are always commutative.

Connections to Existing Work

Complex Embeddings (ComplEx)

$$\phi(s, r, o) = \Re\left(\langle v_s, v_r, \overline{v_o} \rangle\right) \tag{12}$$

$$\text{where } \; v_s, v_r, v_o \in \mathbb{C}^m, \forall s, r, o \tag{13}$$

ComplEx embeddings of size m can be fully recovered by ANALOGY embeddings of size $2m$.

Intuition: there exists a bijection between any $a + bj \in \mathbb{C}$ and $\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \in \mathbb{R}^{2\times 2}$.

Connections to Existing Work

Holographic Embeddings (HolE)

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

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

HolE embeddings can be equivalently obtained via

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

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

Hence is a restricted case of ComplEx and ANALOGY.

Intuition: Circular convolution $*$ can be converted into element-wise product after Fourier transform.

Experiments

Implementation Details

- Use logistic loss:

$$\ell(\phi(s,r,o),y) = -\log \sigma(y\phi(s,r,o)) \tag{18}$$

- Optimization: Asynchronous AdaGrad (HogWild!)
- For each valid (s,r,o), generate negative examples (s',r,o), (s,r',o), (s,r,o') by corrupting s, r, o.

Evaluation

- Hits and Mean Reciprocal Rank (MRR)
- Benchmark datasets FreeBase-15K and WordNet-18.

Results – Hits@10 (filt.)

Models	WN18	FB15K
Unstructured	38.2	6.3
RESCAL	52.8	44.1
NTN	66.1	41.4
SME	74.1	41.3
SE	80.5	39.8
LFM	81.6	33.1
TransH	86.7	64.4
TransE	89.2	47.1
TransR	92.0	68.7
TKRL	–	73.4
RTransE	–	76.2
TransD	92.2	77.3
CTransR	92.3	70.2
KG2E	93.2	74.0
STransE	93.4	79.7
DistMult	93.6	82.4
TransSparse	93.9	78.3
PTransE-MUL	–	77.7
PTransE-RNN	–	82.2
PTransE-ADD	–	84.6
NLF (+external data)	*94.3*	*87.0*
ComplEx	94.7	84.0
HolE	94.9	73.9
Our ANALOGY	94.7	85.4

Results – Hits@{1,3} & MRR

	WN18				FB15K			
Models	MRR (filt.)	MRR (raw)	Hits@1 (filt.)	Hits@3 (filt.)	MRR (filt.)	MRR (raw)	Hits@1 (filt.)	Hits@3 (filt.)
RESCAL	89.0	60.3	84.2	90.4	35.4	18.9	23.5	40.9
TransE	45.4	33.5	8.9	82.3	38.0	22.1	23.1	47.2
DistMult	82.2	53.2	72.8	91.4	65.4	24.2	54.6	73.3
HolE	93.8	61.6	93.0	94.5	52.4	23.2	40.2	61.3
ComplEx	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

Scalability

The algorithm scales linearly over the embedding size.

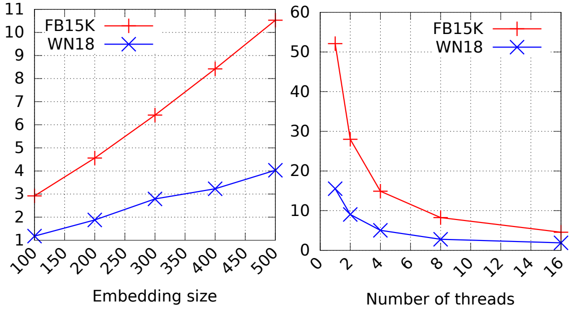

Figure: CPU run time per epoch (secs) of ANALOGY.

Intuition: $O(m)$ for almost-diagonal matrices instead of $O(m^2)$ for dense matrices.

Conclusion

Contributions:

- A new framework that *explicitly* exploit analogy in a differentiable manner.
- Fast algorithm of linear scalability.
- Unified view of several representative works.

Future work: Other applications where analogies might be useful (Machine Translation, Image Captioning, etc.).

Poster #51

Code: https://github.com/quark0/ANALOGY

Thank You!