Holographic Embeddings of Knowledge Graphs

Maximilian Nickel, Lorenzo Rosasco, Tomaso Poggio
Knowledge Graphs → Digital Assistants

Google

Speak now

Hi! I'm M, your personal assistant in Messenger. You can reach me here anytime!

You can ask things like...

Can you help me order flowers for my mom's birthday?

Where's the best place to go hiking in the Bay Area?

Is there a dog-friendly beach nearby?

Ask me anything
Knowledge graphs provide machine-interpretable data by modeling

\[ \text{knowledge} \approx \text{entities} + \text{their relationships} \]

Facts are represented as binary relations \( R_p(e_s, e_o) \).

vicePresident(Obama, Biden)  memberOf(Obama, Democrats)  \(\implies\)  memberOf(Biden, Democrats)
Relational Knowledge Representation

Knowledge graphs provide machine-interpretable data by modeling 

\[ \text{knowledge} \approx \text{entities} + \text{their relationships} \]

Facts are represented as binary relations \( R_p(e_s, e_o) \).

**Multigraph structure**

- Entity = Node
- Fact = Edge
- Relation type = Edge type

Modern knowledge graphs like Freebase, YAGO, DBpedia are

- **Very large** (FB: 40M entities, 35K relations, 637M facts)
- **Very incomplete** (FB: Nationality for 71% of persons missing)
Learn a statistical model of a knowledge graph

Predict probability of any edge (link prediction)

Applications

- KG Completion
- “Structured” prior for Machine Reading
- Probabilistic QA

Challenges

- Relational nature of data
- Size of modern KGs
Knowledge graph embeddings consist of

**Entity Embeddings + Relation Embeddings + Score Function**

**Goal:** Learn embeddings that best explain the data according to score function

**RESCAL** (Nickel, Tresp, et al., 2011)

\[
\text{score}(\mathcal{R}_p(e_s, e_o)) = e_s^T R_p e_o
\]

- Interpretation as tensor completion
- State-of-the-art results on SRL benchmarks
- Runtime & memory complexity $O(d^2)$
Knowledge graph embeddings consist of

**Entity Embeddings + Relation Embeddings + Score Function**

**Goal**: Learn embeddings that best explain the data according to score function

**TransE** (Bordes et al., 2013):

\[
\text{score}(\mathbf{R}_p(\mathbf{e}_s, \mathbf{e}_o)) = -\|\mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o\|_1
\]

- Inspired by Word2Vec
- Runtime & memory complexity \(O(d)\)
- Less powerful than RESCAL
Holographic Embeddings
Let $\mathcal{E}$ be the set of all entities in a domain

A binary relation $\mathcal{R} \subseteq \mathcal{E} \times \mathcal{E}$ is the subset of all pairs of entities for which the relationship is true

**Characteristic Function of Relations**

$$\phi_p(s, o) = \begin{cases} 1, & (s, o) \in \mathcal{R}_p \\ 0, & \text{otherwise} \end{cases}$$

Observation: this is what we want to learn in link prediction

(Nickel, Rosasco, et al., 2016)
Holographic Embeddings (HOLE)

**Holographic Embeddings**

- model entities as vectors
  \[ e_i \mapsto e_i \in \mathbb{R}^d \]
- model relations types as vectors
  \[ R_k \mapsto r_k \in \mathbb{R}^d \]
- represent pairs of entities as
  \[ (e_s, e_o) \mapsto e_s \star e_o \in \mathbb{R}^d \]

where \( \star : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \) denotes circular correlation

\[
[a \star b]_k = \sum_{i=0}^{d-1} a_i b_{(k+i) \text{ mod } d}.
\]

**Model relationships via the classification of pairs of entities**

\[
\Pr(R_p(e_s, e_o) = 1|\Theta) = \sigma \left( r_p^\top (e_s \star e_o) \right)
\]

where \( \Theta = \{e_i\}_{i=1}^{n_e} \cup \{r_k\}_{k=1}^{n_r} \)

(Nickel, Rosasco, et al., 2016)
Holographic Embeddings (HOLE)

Holographic embeddings use circular correlation \( \star : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d \)

\[(e_s, e_o) \approx e_s \star e_o\]

which is defined for \( a, b \in \mathbb{R}^d \) as

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

Compressed Tensor Product

\[
\begin{array}{ccc}
    b_0 & b_1 & b_2 \\
a_0 & & \\
    a_1 & & \\
    a_2 & & \\
\end{array}
\]

(Plate, 1995)
Holographic embeddings use circular correlation $\star : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$

$$(e_s, e_o) \approx e_s \star e_o$$

which is defined for $a, b \in \mathbb{R}^d$ as

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

Compressed Tensor Product

(Plate, 1995)
Holographic Embeddings (HOLE)

Holographic embeddings use circular correlation $\ast : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$

$$(e_s, e_o) \approx e_s \ast e_o$$

which is defined for $a, b \in \mathbb{R}^d$ as

$$[a \ast b]_k = \sum_{i=0}^{d-1} a_i b_{(k+i) \mod d}.$$ 

Compressed Tensor Product

(Plate, 1995)
Circular Correlation as a Compositional Operator

Components of entity embeddings ≈ latent features of entities

Model relation instances via interactions of latent features

e.g., partyOf relation in the US presidents example:

Liberal persons are members of liberal parties
Conservative persons are members of conservative parties

HoLE as a Neural Network

\[
\begin{align*}
\text{Liberal Person} & \land \text{Liberal Party} \\
\text{Conserv. Person} & \land \text{Conserv. Party}
\end{align*}
\]
Components of entity embeddings \( \approx \) latent features of entities

Model relation instances via interactions of latent features

e.g., \textit{partyOf} relation in the US presidents example:
- Liberal persons are members of liberal parties
- Conservative persons are members of conservative parties

\textbf{HoLE as a Neural Network}
**Runtime Complexity:** We can compute circular correlation efficiently via fast Fourier transforms (FFT) in $O(d \log d)$

$$a \ast b = \mathcal{F}^{-1}(\mathcal{F}(a) \odot \mathcal{F}(b))$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the FFT and its inverse.

**Memory Complexity:** Since circular correlation is a function $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, the memory complexity is $O(d)$
Holographic Associative Memory

Let \((a_i, b_i)\) be stimulus-response pairs

- **Storage** \(m \leftarrow \sum_i a_i \ast b_i\)
- **Retrieval** \(b' \leftarrow a \ast m\)
- **Clean-up** \(b \leftarrow \text{arg max}_i b_i^T (a \ast m)\)

Holographic Embeddings

Let \(S_o = \{(s, p) \mid R_p(e_s, e_o) = 1\}\)

- **Storage** \(e_o \leftarrow \sum_{(s, p)} r_p \ast e_s\)
- **Retrieval** \(r' \leftarrow e_s \ast e_o\)
- **Probability** \(\sigma(r_p^T(e_s \ast e_o))\)

**Generalization**, not memorization

(Plate, 1995; Poggio, 1973; Gabor, 1969; Willshaw, 1985)
Experiments
Link Prediction on WordNet

- **WordNet** consists of lexical relationships between words
- **WN18** subset (Bordes et al., 2013)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>40,943</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relation types</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facts</td>
<td>151,442</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Optics**
- **Holography**
- **hypernym**
- **derivational form**
- **optical**

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
<th>Hits@1</th>
<th>Hits@3</th>
<th>Hits@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOLE</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>RESCAL</td>
<td>0.89</td>
<td>0.84</td>
<td>0.9</td>
<td>0.93</td>
</tr>
<tr>
<td>TRANSE</td>
<td>0.5</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRANSR</td>
<td>0.88</td>
<td>0.61</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>ER-MLP</td>
<td>0.71</td>
<td>0.78</td>
<td>0.86</td>
<td></td>
</tr>
</tbody>
</table>
Link Prediction on Freebase

- **Freebase** consists of general facts about the world (e.g., harvested from Wikipedia, MusicBrainz, etc.)

- **FB15k** subset (Bordes et al., 2013)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>14,951</td>
<td>Relation types</td>
<td>1345</td>
</tr>
<tr>
<td>Facts</td>
<td>592,213</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

The figure shows the performance of different link prediction models on the FB15k dataset. The models include HOLE, RESCAL, TRANSE, TRANSR, and ER-MLP. The metrics include MRR and Hits at 1, 3, and 10. The values for Hits at 1, 3, and 10 for each model are as follows:

- **HOLE**:
  - Hits@1: 0.52
  - Hits@3: 0.4
  - Hits@10: 0.4

- **RESCAL**:
  - Hits@1: 0.35
  - Hits@3: 0.24
  - Hits@10: 0.22

- **TRANSE**:
  - Hits@1: 0.46
  - Hits@3: 0.3
  - Hits@10: 0.35

- **TRANSR**:
  - Hits@1: 0.35
  - Hits@3: 0.4
  - Hits@10: 0.58

- **ER-MLP**:
  - Hits@1: 0.29
  - Hits@3: 0.17
  - Hits@10: 0.32

---

The diagram illustrates the relationships between entities such as Barack Obama, Democratic Party, vicePresident, and Joe Biden.
MRR vs Number of Parameters

FB15k

- HoLE
- TRANSE
- REscAL
- ER-MLP
- TRANSR

MRR

Number of Parameters in Millions

0 5 10 15 20 25 30 35

0.3 0.4 0.5 0.6
Summary

- HoLE combines state-of-the-art relational learning and high scalability in a single model
- Enables complex models of knowledge graphs
- Interpretation in terms of associative memory

Future Work

Since circular correlation is a function $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$

Tuple is vector of same size as entity

Essential property to create recursive representations

Nested Facts

believes(Tom,loves(John,Mary))

Higher-arity Relations

taughtAt(Tom,AI,MIT)
Thank you

Software

• Open-Source Library for Knowledge Graph Embeddings
  http://github.com/mnick/scikit-kge

• Experiments for this Paper
  https://github.com/mnick/holographic-embeddings

Recent Review Article
Simple Reasoning

- **Task**: Predict the region of countries
- **Setting**: 10-fold cross validation over countries

(Nickel et al., 2015)
• **Task**: Predict the region of countries
• **Setting**: 10-fold cross validation over countries

\[ S_1 : \text{partOf}(c, s) \land \text{partOf}(s, r) \Rightarrow \text{partOf}(c, r) \]

(Nickel et al., 2015)
Simple Reasoning

- **Task**: Predict the region of countries
- **Setting**: 10-fold cross validation over countries

\[
S_2 : \text{neighbors}(c_1, c_2) \land \text{partOf}(c_2, r) \Rightarrow \text{partOf}(c_1, r)
\]

(Nickel et al., 2015)
Simple Reasoning

- **Task**: Predict the region of countries
- **Setting**: 10-fold cross validation over countries

```
<table>
<thead>
<tr>
<th>Test Country</th>
<th>Region</th>
<th>Subregion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neighbors</td>
<td>partOf</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neighbors</td>
<td>partOf</td>
<td></td>
</tr>
</tbody>
</table>
```

![Graph showing relationships between countries, regions, and subregions]

\[
S_3 : \text{neighbors}(c_1, c_2) \land \text{partOf}(c_2, s) \land \text{partOf}(s, r) \Rightarrow \text{partOf}(c_1, r)
\]

(Nickel et al., 2015)
Holographic Embeddings keep excellent performance on SRL benchmark datasets

Other knowledge graph embedding models perform worse

(AUC-PR)

(Kinships, Nations, UMLS)

(Nickel, Tresp, et al., 2011; Garcia-Duran et al., 2015)
MAP estimates for $\Theta = \{e_i\}_{i=1}^n \cup \{r_k\}_{k=1}^m$ for the joint distribution

$$\Pr(Y|\Theta) = \prod_{s=1}^n \prod_{p=1}^m \prod_{o=1}^n \Pr(y_{spo} = 1|\sigma(r_p^T(e_s * e_o)))$$

Shared representations enable relational learning

- Entities have same embeddings as subjects, objects, and over all relations
- Embeddings are learned jointly: allows to propagate information between triples
- **Decoupling effect**
  - Known parameters: local computation
  - Parameter learning: global dependencies
- Holds for many compositional models