Extracting and Modeling Relations with Graph Convolutional Networks

Ivan Titov

with Diego Marcheggiani, Michael Schlichtkrull, Thomas Kipf, Max Welling, Rianne van den Berg and Peter Bloem
Inferring missing facts in knowledge bases: link prediction

- Vaganova Academy
- Mikhail Baryshnikov
- St. Petersburg

- studied_at
- located_in
- lived_in?
Mikhail Baryshnikov studied at Vaganova Academy and lived in St. Petersburg. He danced for the Mariinsky Theatre, which was based in what was then Leningrad (now St. Petersburg).
After a promising start in Mariinsky ballet, Baryshnikov defected to Canada in 1974 ...
KBC: it is natural to represent both sentences and KB with graphs.

For sentences, the graphs encode beliefs about their linguistic structure.

After a promising start in Mariinsky ballet, Baryshnikov defected to Canada in 1974...

How can we model (and exploit) these graphs with graph neural networks?
Outline

Graph Convolutional Networks (GCNs)

Link Prediction with Graph Neural Networks
  - Relational GCNs
  - Denoising Graph Autoencoders for Link Prediction

Extracting Semantic Relations: Semantic Role Labeling
  - Syntactic GCNs
  - Semantic Role Labeling Model
Graph Convolutional Networks: Neural Message Passing
Graph Convolutional Networks: message passing

Undirected graph

Update for node v

Graph Convolutional Networks: message passing

Undirected graph

Update for node $v$

\[ h_v = \text{ReLU}(W_{\text{loop}}h_v + \sum_{u \in \mathcal{N}(v)} Wh_u) \]

GCNs: multilayer convolution operation

Initial feature representations of nodes

Parallelizable computation, can be made quite efficient (e.g., Hamilton, Ying and Leskovec (2017)).
GCNs: multilayer convolution operation

Initial feature representations of nodes

\[ X = H^{(0)} \]

Hidden layer

\[ H^{(1)} \]

Hidden layer

\[ H^{(2)} \]

Hidden layer

\[ H^{(N)} \]

Output

\[ Z = H^{(N)} \]

Parallelizable computation, can be made quite efficient (e.g., Hamilton, Ying and Leskovec (2017)).

Representations informed by node neighbourhoods
Graph Convolutional Networks: Previous work

Shown very effective on a range of problems - citations graphs, chemistry, ...

Mostly:
- Unlabeled and undirected graphs
- Node labeling in a single large graph (transductive setting)
- Classification of graphlets

How to apply GCNs to graphs we have in knowledge based completion / construction?

See Bronstein et al. (Signal Processing, 2017) for an overview
Link Prediction with Graph Neural Networks
Link Prediction

- Vaganova Academy
- Mikhail Baryshnikov
- Mariinsky Theatre
- St. Petersburg

- studied_at
- danced_for
- located_in
- lived_in?
Link Prediction

Vaganova Academy studied_at Mikhail Baryshnikov danced_for Mariinsky Theatre

located_in St. Petersburg

lived_in ? located_in
Link Prediction

Vaganova Academy → studied_at → Mikhail Baryshnikov → danced_for → Mariinsky Theatre

located_in:
- St. Petersburg
KB Factorization

Vaganova Academy \(\xrightarrow{studied\_at}\) Mikhail Baryshnikov \(\xrightarrow{danced\_for}\) Mariinsky Theatre

located_in

St. Petersburg

lived_in ?
A scoring function is used to predict whether a relation holds:

\[
\text{Baryshnikov} \times \text{lived_in} \times \text{St. Petersburg} \]

**KB Factorization**

- **studied_at**: Vaganova Academy
- **danced_for**: Mariinsky Theatre
- **located_in**: St. Petersburg

**RESCAL** (Nickel et al., 2011)
A scoring function is used to predict whether a relation holds:

$\text{DistMult}$

(Yang et al., 2014)
A scoring function is used to predict whether a relation holds:

\[ \text{DistMult} \quad \text{(Yang et al., 2014)} \]

Relies on SGD to propagate information across the graph
Relational GCNs

A scoring function is used to predict whether a relation holds:

\[ \text{DistMult} \] (Yang et al., 2014)

Use the same scoring function but with GCN node representations rather than parameter vectors

Schlichtkrull et al., 2017
Relational GCNs

Info about St. Petersburg reached here

Vaganova Academy \( \xrightarrow{studied\_at} \) Mikhail Baryshnikov \( \xrightarrow{danced\_for} \) Mariinsky Theatre

located_in \( \downarrow \)

St. Petersburg

A scoring function is used to predict whether a relation holds:

DistMult (Yang et al., 2014)

Use the same scoring function but with GCN node representations rather than parameter vectors

Schlichtkrull et al., 2017
A scoring function is used to predict whether a relation holds:

\[
\text{DistMult (Yang et al., 2014)}
\]

Use the same scoring function but with GCN node representations rather than parameter vectors

Schlichtkrull et al., 2017
Relational GCNs
Relational GCNs

How do we train Relational GCNs?

How do we compactly parameterize Relational GCNs?
GCN Denoising Autoencoders

Take the training graph

Schlichtkrull et al (2017)
Produce a noisy version: drop some random edges
Use this graph for encoding nodes with GCNs

Schlichtkrull et al., 2017
GCN Denoising Autoencoders

Force the model to reconstruct the original graph (including dropped edges) (a ranking loss on edges)

Schlichtkrull et al., 2017
Training

Classic DistMult

- Edge loss
- DistMult
- Input

(e.g., node embeddings)

Our R-GCN

- Edge loss
- DistMult
- R-GCN
- Input

(e.g., node embeddings)

Schlichtkrull et al., 2017
Instead of denoising AEs, we can use variational AEs to train R-GCNs

VAE R-GCN can be regarded as an inference network performing amortized variational inference

**Intuition:**

R-GCN AEs are amortized versions of factorization models
Relational GCN

\[ h_v = ReLU\left(\frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} W_{r(u,v)} h_u \right) \]

There are too many relations in realistic KBs, we cannot use full rank matrices \( W_r \).
Relational GCN

Naive logic:

We score with a diagonal matrix (DistMul), let’s use a diagonal one in GCN

\[ W_r = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \]
Relational GCN

**Block diagonal assumption:**

Latent features can be grouped into sets of tightly inter-related features, modeling dependencies across the sets is less important.

\[ W_r = \]

\[ \begin{bmatrix}
  \cdot & \cdot & \cdot \\
  \cdot & \cdot & \cdot \\
  \cdot & \cdot & \cdot \\
\end{bmatrix} \]
Relational GCN

Basis / Dictionary learning:

Represent every KB relation as a linear combination of basis transformations

$$W_r = \sum_{b=1}^{B} a_{rb} V_b$$

coefficients
basis
transformations
Results on FB15k-237 (hits@10)

Our R-GCN relies on DistMult in the decoder: DistMult is its natural baseline. See other results and metrics in the paper. Results for ComplEX, TransE and HolE from code of Trouillon et al. (2016). Results for HolE using code by Nickel et al. (2015).
Relational GCNs

**Fast** and simple approach to Link Prediction

**Captures multiple paths** without the need to explicitly marginalize over them

Unlike factorizations, can be applied to **subgraphs unseen in training**

**FUTURE WORK:**

R-GCNs can be used in combination with **more powerful factorizations / decoders**

Objectives favouring **recovery of paths** rather than edges

**Gates** and memory may be effective
Extracting Semantic Relations
Semantic Role Labeling

Closely related to the relation extraction task
Discovering the predicate-argument structure of a sentence

Sequa makes and repairs jet engines
Semantic Role Labeling

Closely related to the relation extraction task

Discovering the predicate-argument structure of a sentence
- Discover predicates

Sequa makes and repairs jet engines
Semantic Role Labeling

Closely related to the relation extraction task

Discovering the predicate-argument structure of a sentence
- Discover predicates
- Identify arguments and label them with their semantic roles

Sequa makes and repairs jet engines

creator makes repairer

creation

entity repaired
Some syntactic dependencies are mirrored in the semantic graph.
Syntax/semantics interaction

Some syntactic dependencies are **mirrored** in the semantic graph

... but not all of them – the syntax-semantics interface is far from trivial

GCNs provide a flexible framework for capturing interactions between the graphs
Syntactic GCNs: directionality and labels

Direction opposite of syntactic edges

Along syntactic edges

\[ W_{\text{out}} \quad W_{\text{loop}} \quad W_{\text{in}} \]
Syntactic GCNs: directionality and labels

- Direction opposite of syntactic edges
- Along syntactic edges

Weight matrix for each direction:
\( W_{\text{out}}, W_{\text{in}}, W_{\text{loop}} \)

Bias for each label + direction, e.g. \( b_{\text{in-subj}} \)

\[
h_v = \text{ReLU} \left( \sum_{u \in \mathcal{N}(v)} W_{\text{dir}(u,v)} h_u + b_{\text{lab}(u,v)} \right)\]
Syntactic GCNs: edge-wise gating

Not all edges are equally informative for the downstream task or reliable

\[ g_{u,v} = \sigma \left( h_u \cdot \hat{w}_{dir}(u,v) + \hat{b}_{lab}(u,v) \right) \]

\[ h_v = \text{ReLU} \left( \sum_{u \in \mathcal{N}(v)} g_{u,v} \left( W_{dir}(u,v) \cdot h_u + b_{lab}(u,v) \right) \right) \]

We use parsers to predict syntax

The gate weights the message

Marcheggiani et al., EMNLP 2017
Graph Convolutional Encoders

Sequa makes and repairs jet engines

Encoder (BiRNN, CNN, ..)
Graph Convolutional Encoders

Sequa makes and repairs jet engines

Encoder (BiRNN, CNN, ..)

GCN layer 1
Graph Convolutional Encoders

Sequa makes and repairs jet engines

Encoder (BiRNN, CNN, ..)

GCN layer 1
Graph Convolutional Encoders

Sequa makes and repairs jet engines

Encoder (BiRNN, CNN, ..)
Graph Convolutional Encoders

Sequa makes and repairs jet engines

Encoder (BiRNN, CNN, ..)

GCN layer 1

GCN layer 2

GCN layer 3

W_out

W_out

W_out

W_out

W_out

W_out

W_out

W_out
How do we construct a GCN-based semantic role labeler?
GCNs for Semantic Role Labeling

Repairer

Sequa makes and repairs jet engines

subj - coord - conj - obj - nmod

Marcheggiani et al., EMNLP 2017
Sequa makes and repairs jet engines

Marcheggiani et al., EMNLP 2017
Sequa makes and repairs jet engines

**NULL**

Marcheggiani et al., EMNLP 2017
GCNs for Semantic Role Labeling

Sequa makes and repairs jet engines

Semantic Role Labeler
GCN layer(s)
BiRNN

Marcheggiani et al., EMNLP 2017
GCNs for Semantic Role Labeling

Sequa makes and repairs jet engines

NULL

Semantic Role Labeler

GCN layer(s)

BiRNN

Marcheggiani et al., EMNLP 2017
GCNs for Semantic Role Labeling

Entity Repaired

Sequa makes and repairs jet engines

Semantic Role Labeler
GCN layer(s)
BiRNN

Marcheggiani et al., EMNLP 2017
Results (F1) on Chinese (CoNLL-2009, dev set)

Marcheggiani & Titov (EMNLP, 2017)

Predicate disambiguation is excluded from the F1 metric.
Results (F1) on Chinese (CoNLL-2009, test set)

- Zhao et al. (2009): 77.7
- Björkelund et al. (2010): 78.6
- Roth and Lapata (2016): 79.4
- Ours (GCN): 82.5

Marcheggiani & Titov (EMNLP, 2017)
Results (F1) on English (CoNLL-2009)

- Marcheggiani & Titov (EMNLP, 2017)
- Ensemble of 3

Our model: 89.1
Ensemble (x3): Ours (GCN)
Flexibility of GCN encoders

Simple and fast approach to integrating linguistic structure into encoders

In principle we can exploit almost any kind of linguistic structure:

- Semantic role labeling structure
- Co-reference chains
- AMR semantic graphs
- Their combination
Other applications of syntactic GCN encoders

We also showed them effective as encoders in Neural Machine Translation

Others recently applied them to NER

Bastings et al. (EMNLP, 2017)

Cetoli et al. (arXiv:1709.10053)
Conclusions

GCNs are in subtasks of KBC (and in NLP beyond KBC):

- Semantic Roles: we proposed GCNs for encoding linguistic knowledge
- Link prediction: GCNs for link prediction (and entity classification) in multi-relational knowledge bases

Code available

We are hiring! (PhD students / postdocs)
- Improvement across the board, especially in the middle of the range
Effect of Distance between Argument and Predicate (English)