Relational Stacked Denoising Autoencoder for Tag Recommendation

Hao Wang

Dept. of Computer Science and Engineering
Hong Kong University of Science and Technology

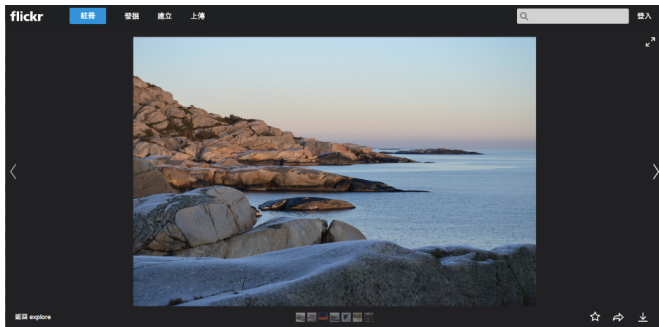Joint work with Xingjian Shi and Dit-Yan Yeung
To appear in AAAI 2015

Outline

Outline

https://www.flickr.com

Tag Recommendation: CiteULike

✓ **An Algorithmic Framework for Performing Collaborative Filtering**
In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1999), pp.
by Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, John Riedl
posted to collaborative_filtering recommendation by wangxinxi on 2014-11-24 14:12:37 ★★ along with 32 people and 6 groups
■ Abstract

✓ **Google News Personalization: Scalable Online Collaborative Filtering**
In Proceedings of the 16th International Conference on World Wide Web (2007), pp. 271-280, doi:10.1145/1242572.1242610
by Abhinandan S. Das, Mayur Datar, Ashutosh Garg, Shyam Rajaram
posted to collaborative_filtering news recommendation by wangxinxi on 2014-11-24 14:08:35 ★★ along with 74 people and 11 groups
■ Abstract

✓ **Probabilistic Matrix Factorization**
In NIPS '08 (2008)
by Ruslan Salakhutdinov, Andriy Mnih
posted to collaborative_filtering matrix_factorization probabilistic recommendation by nguyenthaibinh on 2014-09-14 20:29:47 ✓ along with
■ Abstract

✓ **Active Learning in Collaborative Filtering Recommender Systems**
In E-Commerce and Web Technologies, Vol. 188 (2014), pp. 113-124, doi:10.1007/978-3-319-10491-1_12
by Mehdi Elahi, Francesco Ricci, Neil Rubens
edited by Martin Hepp, Yigal Hoffner
posted to active_learning collaborative_filtering matrix_factorization by wangxinxi on 2014-08-24 09:28:27 ★★★★
■ Abstract

Google news personalization: scalable online collaborative filtering
In Proceedings of the 16th international conference on World Wide Web (2007), pp. 271-280, doi:10.1145/1242572.1242610
by Abhinandan S. Das, Mayur Datar, Ashutosh Garg, Shyam Rajaram
posted to collaborative_filtering by eustache_diemert on 2014-06-06 07:32:15 ★★
■ Abstract

http://www.citeulike.org

Related Work

Content-based:

1. Chen et al., 2008
2. Chen et al., 2010
3. Shen and Fan, 2010

Co-occurrence based:

1. Garg and Weber, 2008
2. Weinberger et al., 2008
3. Rendle and Schmidt-Thieme, 2010

Hybrid:

1. Wu et al., 2009
2. Wang and Blei, 2011
3. Yang et al., 2013

Content-based

1. Chen et al., 2008
2. Chen et al., 2010
3. Shen and Fan, 2010
4. . . .

Pros:
1. Tag independence
2. Interpretability
3. No New-item problem

Cons:
1. Need domain knowledge

Co-occurrence based

1. Garg and Weber, 2008
2. Weinberger et al., 2008
3. Rendle and Schmidt-Thieme, 2010
4. $\cdots$

Pros:

1. No domain knowledge needed

Cons:

1. Requires some form of rating feedback (co-occurrence matrix)
2. New-tag problem and new-item problem

1. Wu et al., 2009
2. Wang and Blei, 2011
3. Yang et al., 2013
4. ...

BEST OF BOTH WORLDS

Collaborative Topic Regression (CTR)
(Wang and Blei, KDD 2011)

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. Dempster, N. M. Laird and D. B. Rubin

Harvard University and Educational Testing Service

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

Summary

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

matrix factorization		topic modeling	
■ (red)	???????????	■ (red)	estimate estimates likelihood maximum estimated missing
■ (blue)	???????????	■ (blue)	algorithm signal input signals output exact performs music
■ (magenta)	???????????	■ (magenta)	distribution random probability distributions sampling stochastic

Collaborative Topic Regression (CTR)
(Wang and Blei, KDD 2011)

Article representation in different methods
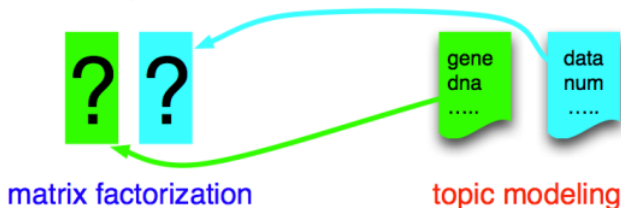
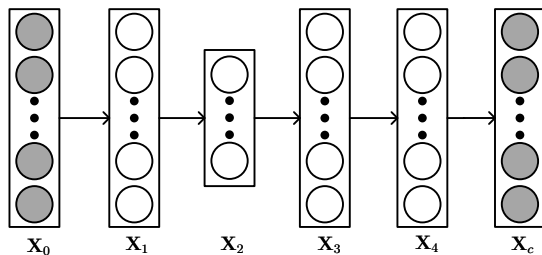- **LDA**: sparse, relatively high dimension
- **MF**: low rank, low dimension

Problems to Explore

1. Can SDAE learn effective representation for recommendation?
2. How to incorporate relational information into SDAE?
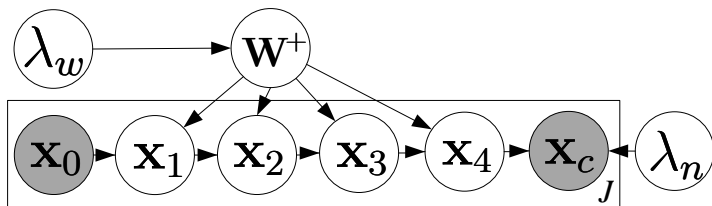3. How is the performance?

Outline

Stacked Denoising Autoencoder
(Vincent et al. JMLR 2010)

$$\min_{\{\mathbf{W}_l\},\{\mathbf{b}_l\}} \|\mathbf{X}_c - \mathbf{X}_L\|_F^2 + \lambda \sum_l \|\mathbf{W}_l\|_F^2,$$

where λ is a regularization parameter and $\|\cdot\|_F$ denotes the Frobenius norm.

Generalized Probabilistic SDAE

1. For each layer l of the SDAE network,
   1. For each column n of the weight matrix $\mathbf{W}_l$, draw
      $\mathbf{W}_{l,*n} \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   2. Draw the bias vector $\mathbf{b}_l \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   3. For each row j of $\mathbf{X}_l$, draw
      $$\mathbf{X}_{l,j*} \sim \mathcal{N}(\sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l), \lambda_s^{-1}\mathbf{I}_{K_l}).$$
2. For each item j, draw a clean input
   $$\mathbf{X}_{c,j*} \sim \mathcal{N}(\mathbf{X}_{L,j*}, \lambda_n^{-1}\mathbf{I}_B).$$

Outline

Relational SDAE: Generative Process

1. Draw the relational latent matrix $\mathbf{S}$ from a *matrix variate normal distribution*:

$$\mathbf{S} \sim \mathcal{N}_{K,J}(0, \mathbf{I}_K \otimes (\lambda_l \mathscr{L}_a)^{-1}).$$

2. For layer l of the SDAE where $l = 1, 2, \ldots, \frac{L}{2} - 1$,
   1. For each column n of the weight matrix $\mathbf{W}_l$, draw $\mathbf{W}_{l,*n} \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   2. Draw the bias vector $\mathbf{b}_l \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   3. For each row j of $\mathbf{X}_l$, draw

   $$\mathbf{X}_{l,j*} \sim \mathcal{N}(\sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l), \lambda_s^{-1}\mathbf{I}_{K_l}).$$

3. For layer $\frac{L}{2}$ of the SDAE network, draw the representation vector for item j from the product of two Gaussians (PoG):

$$\mathbf{X}_{\frac{L}{2},j*} \sim \mathsf{PoG}(\sigma(\mathbf{X}_{\frac{L}{2}-1,j*}\mathbf{W}_l + \mathbf{b}_l), \mathbf{s}_j^T, \lambda_s^{-1}\mathbf{I}_K, \lambda_r^{-1}\mathbf{I}_K).$$
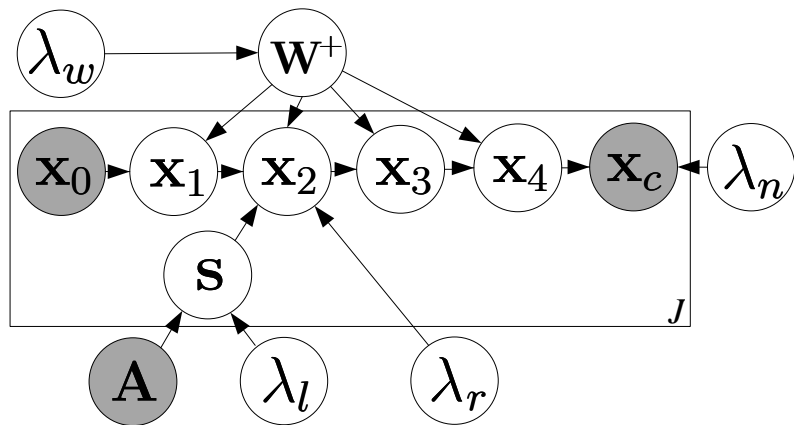
Relational SDAE: Generative Process

1. For layer l of the SDAE network where $l = \frac{L}{2} + 1, \frac{L}{2} + 2, \ldots, L$,
   1. For each column n of the weight matrix $\mathbf{W}_l$, draw $\mathbf{W}_{l,*n} \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   2. Draw the bias vector $\mathbf{b}_l \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   3. For each row j of $\mathbf{X}_l$, draw

      $$\mathbf{X}_{l,j*} \sim \mathcal{N}(\sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l), \lambda_s^{-1}\mathbf{I}_{K_l}).$$
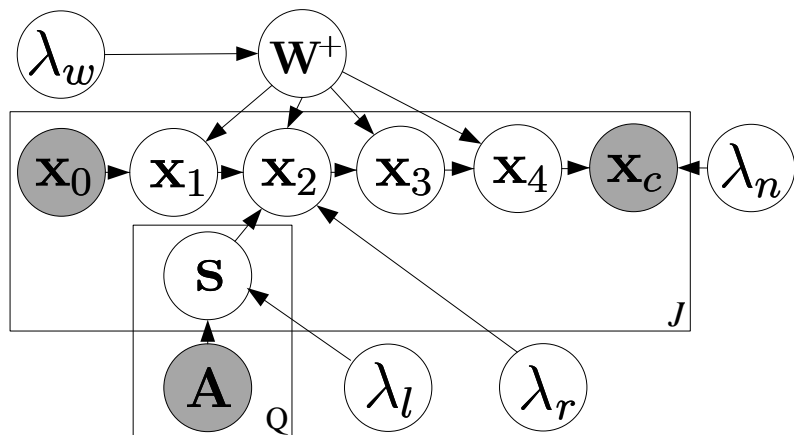
2. For each item j, draw a clean input

   $$\mathbf{X}_{c,j*} \sim \mathcal{N}(\mathbf{X}_{L,j*}, \lambda_n^{-1}\mathbf{I}_B).$$

Relational SDAE: Objective function

The log-likelihood:

$$
\begin{aligned}
\mathscr{L} = &- \frac{\lambda_l}{2} \mathrm{tr}(\mathbf{S}\mathscr{L}_a\mathbf{S}^T) - \frac{\lambda_r}{2} \sum_j \|(\mathbf{s}_j^T - \mathbf{X}_{\frac{L}{2},j*})\|_2^2 \\
&- \frac{\lambda_w}{2} \sum_l (\|\mathbf{W}_l\|_F^2 + \|\mathbf{b}_l\|_2^2) \\
&- \frac{\lambda_n}{2} \sum_j \|\mathbf{X}_{L,j*} - \mathbf{X}_{c,j*}\|_2^2 \\
&- \frac{\lambda_s}{2} \sum_l \sum_j \|\sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l) - \mathbf{X}_{l,j*}\|_2^2,
\end{aligned}
$$

where $\mathbf{X}_{l,j*} = \sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l)$. Similar to the generalized SDAE, taking λ_s to infinity, the last term of the joint log-likelihood will vanish.

Updating Rules

For $\mathbf{S}$:

$$\mathbf{S}_{k*}(t+1) \leftarrow \mathbf{S}_{k*}(t) + \delta(t)r(t)$$

$$r(t) \leftarrow \lambda_r \mathbf{X}^T_{\frac{L}{2},*k} - (\lambda_l \mathscr{L}_a + \lambda_r \mathbf{I}_J)\mathbf{S}_{k*}(t)$$

$$\delta(t) \leftarrow \frac{r(t)^T r(t)}{r(t)^T (\lambda_l \mathscr{L}_a + \lambda_r \mathbf{I}_J) r(t)}.$$

For $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{b}$: Use Back Propagation.

From Representation to Tag Recommendation

Objective function:

$$\mathscr{L} = - \frac{\lambda_u}{2} \sum_i \|\mathbf{u}_i\|_2^2 - \frac{\lambda_v}{2} \sum_j \|\mathbf{v}_j - \mathbf{X}_{\frac{L}{2},j*}^T\|_2^2$$
$$- \sum_{i,j} \frac{c_{ij}}{2} (\mathbf{R}_{ij} - \mathbf{u}_i^T \mathbf{v}_j)^2,$$

where λ_u and λ_v are hyperparameters. c_{ij} is set to 1 for the existing ratings and 0.01 for the missing entries.

Algorithm

1. **Learning representation:**
   repeat
      Update $\mathbf{S}$ using the updating rules
      Update $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{b}$
   until convergence
   Get resulting representation $\mathbf{X}_{\frac{L}{2},j*}$

2. **Learning $\mathbf{u}_i$ and $\mathbf{v}_j$:**
   Optimize the objective function $\mathscr{L}$

3. **Recommend tags to items according to the predicted $\mathbf{R}_{ij}$:**
   $\mathbf{R}_{ij} = \mathbf{u}_i^T \mathbf{v}_j$
   Rank $\mathbf{R}_{1j}, \mathbf{R}_{2j}, \ldots, \mathbf{R}_{Ij}$
   Recommend tags with largest $\mathbf{R}_{ij}$ to item j

1. Can SDAE learn effective representation for recommendation?
2. How to incorporate relational information into SDAE?
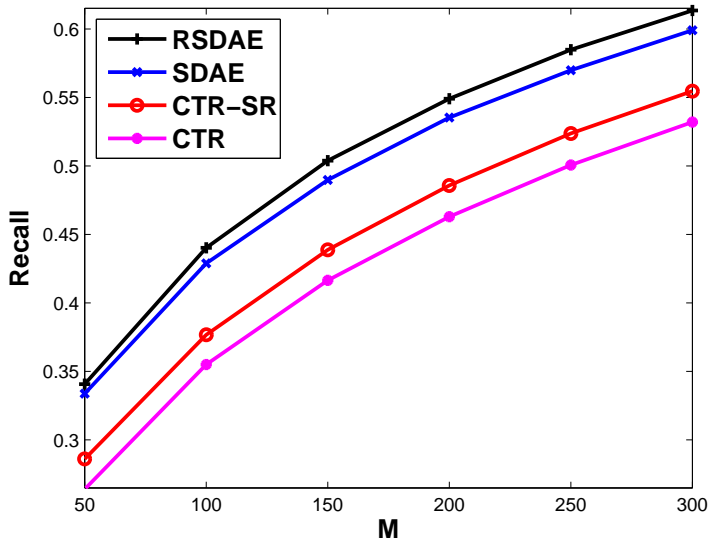3. How is the performance?

Outline

Datasets

Description of datasets

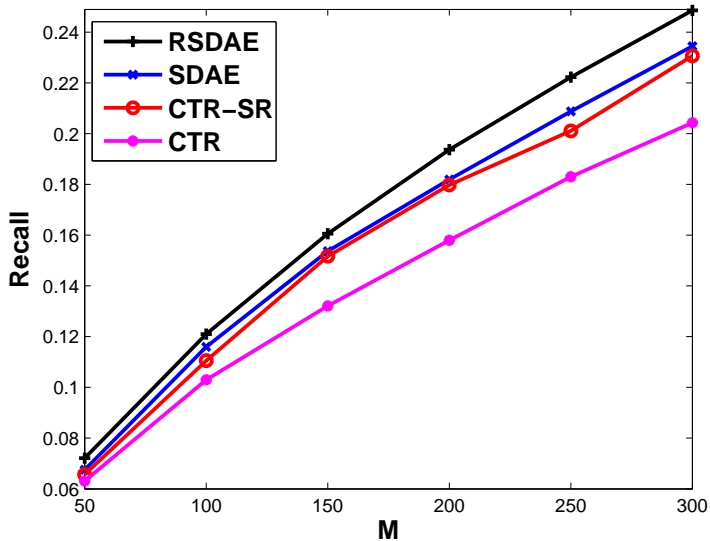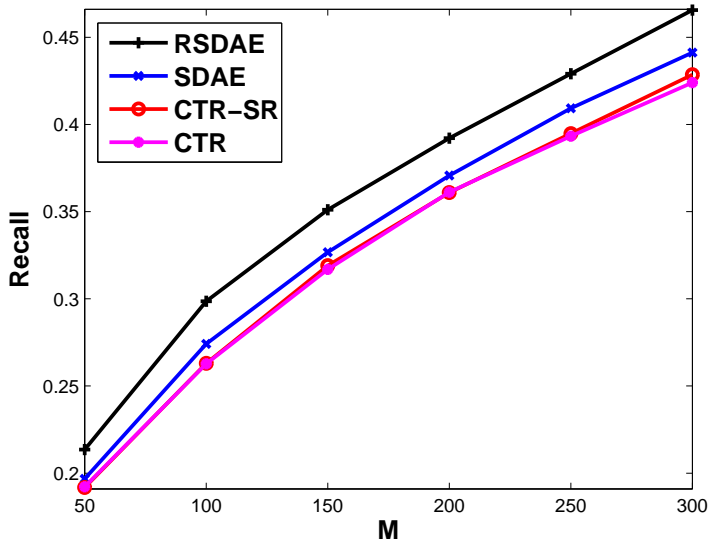|	citeulike-a	citeulike-t	movielens-plot
#items	16980	25975	7261
#tags	7386	8311	2988
#tag-item paris	204987	134860	51301
#relations	44709	32665	543621
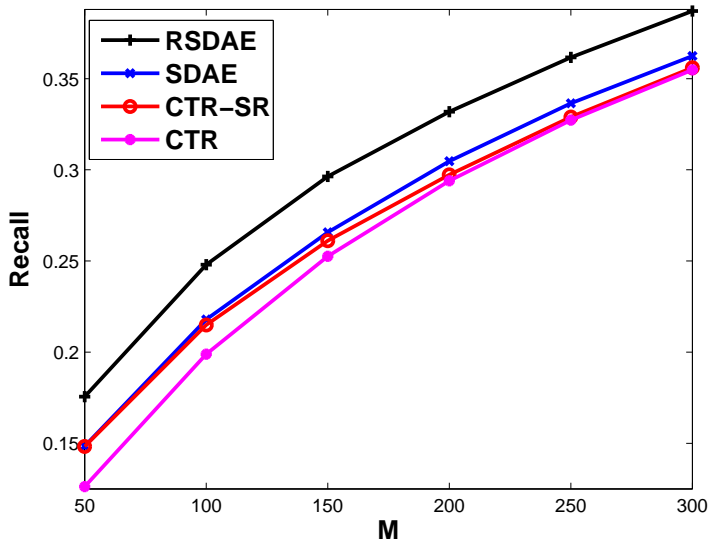
citeulike-a, Dense Settting

movielens-plot, Dense Settting

Outline

Tagging Scientific Articles

An example article with recommended tags

Example Article	Title: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews			
	Top topic 1: language, text, mining, representation, semantic, concepts, words, relations, processing, categories			
	SDAE	True?	RSDAE	True?
	1. instance	no	1. sentiment_analysis	no
	2. consumer	**yes**	2. instance	no
	3. sentiment_analysis	no	**3. consumer**	**yes**
	4. summary	no	4. summary	no
Top 10 tags	5. 31july09	no	**5. sentiment**	**yes**
	6. medline	no	**6. product_review_mining**	**yes**
	7. eit2	no	**7. sentiment_classification**	**yes**
	8. l2r	no	8. 31july09	no
	9. exploration	no	**9. opinion_mining**	**yes**
	10. biomedical	no	**10. product**	**yes**

Tagging Movies

An example movie with recommended tags

Example Movie	Title: E.T. the Extra-Terrestrial	
	Top topic 1: crew, must, on, earth, human, save, ship, rescue, by, find, scientist, planet	
Top 10 recommended tags	SDAE	True tag?
	1. Saturn Award (Best Special Effects)	**yes**
	2. Want	no
	3. Saturn Award (Best Fantasy Film)	no
	4. Saturn Award (Best Writing)	**yes**
	5. Cool but freaky	no
	6. Saturn Award (Best Director)	no
	7. Oscar (Best Editing)	no
	8. almost favorite	no
	9. Steven Spielberg	**yes**
	10. sequel better than original	no

Tagging Movies

An example movie with recommended tags

Example Movie	Title: E.T. the Extra-Terrestrial	
	Top topic 1: crew, must, on, earth, human, save, ship, rescue, by, find, scientist, planet	
Top 10 recommended tags	RSDAE	True tag?
	1. Steven Spielberg	**yes**
	2. Saturn Award (Best Special Effects)	**yes**
	3. Saturn Award (Best Writing)	**yes**
	4. Oscar (Best Editing)	no
	5. Want	no
	6. Liam Neeson	no
	7. AFI 100 (Cheers)	**yes**
	8. Oscar (Best Sound)	**yes**
	9. Saturn Award (Best Director)	no
	10. Oscar (Best Music - Original Score)	**yes**

Outline

Conclusion

Contribution:

1. Adapt SDAE for tag recommendation
2. A probabilistic relational model for relational deep learning
3. State-of-the-art performance

Take-home Message:

1. Deep models significantly boost recommendation accuracy
2. Probabilistic formulation facilitates relational deep learning
3. Incorporating relational information further boosts accuracy

1. Applications other than tag recommendation
2. Adaption for other deep learning models
3. Integrated model instead of separate ones
4. Fully Bayesian methods

Other Work

1. Collaborative CNN for recommender systems
2. Relational deep learning for link prediction