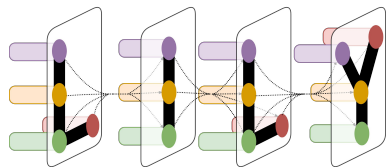


Unsupervised Cross-Domain Prerequisite Chain Learning using Variational Graph Autoencoders

Irene Li, Vanessa Yan, Tianxiao Li, Rihao Qu and Dragomir Radev

Yale University, USA

ACL 2021

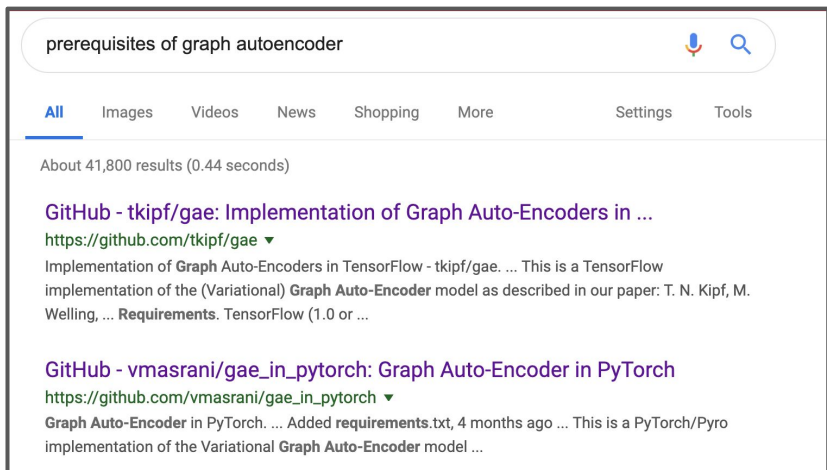


Yale

Motivation

Concept prerequisite chain learning: automatically determining the existence of prerequisite relationships among concept pairs.

Cross-domain prerequisite chains: borrow existing knowledge from the known domain.



prerequisites of graph autoencoder

All Images Videos News Shopping More Settings Tools

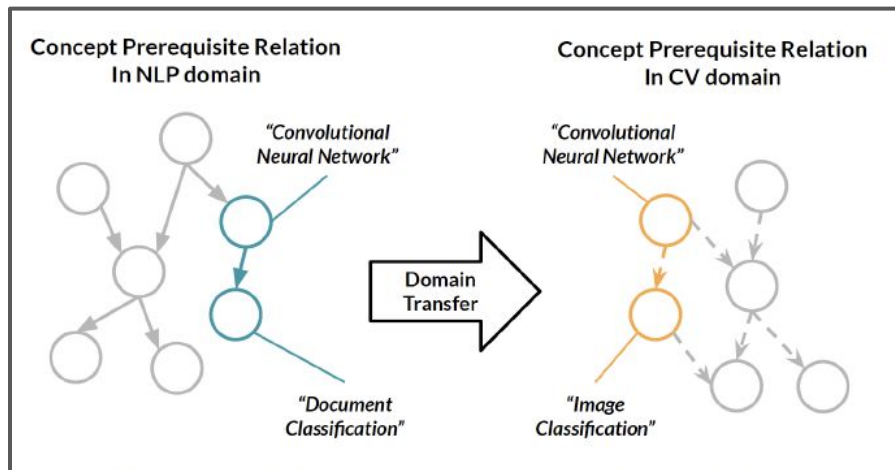
About 41,800 results (0.44 seconds)

GitHub - tkipf/gae: Implementation of Graph Auto-Encoders in ...
<https://github.com/tkipf/gae>

Implementation of **Graph Auto-Encoders** in TensorFlow - tkipf/gae. ... This is a TensorFlow implementation of the (Variational) **Graph Auto-Encoder** model as described in our paper: T. N. Kipf, M. Welling, ... **Requirements**. TensorFlow (1.0 or ...

GitHub - vmasrani/gae_in_pytorch: Graph Auto-Encoder in PyTorch
https://github.com/vmasrani/gae_in_pytorch

Graph Auto-Encoder in PyTorch. ... Added requirements.txt, 4 months ago ... This is a PyTorch/Pyro implementation of the Variational **Graph Auto-Encoder** model ...



Related Work

Prerequisite chain learning

Feature engineering and classifiers that determine if there's a prerequisite relation between any given concept pair: (Liu et al., 2016; Liang et al., 2017);

Materials including university course descriptions and materials as well as online educational data: (Liu et al., 2016; Liang et al., 2017) .

Our previous work

Our previous work Li et al. (2019) and Li et al. (2020): applied GCN (Kipf and Welling, 2017), GAE and VGAE (Kipf and Welling, 2017) to predict relations in a concept graph.

In this work we introduce the new task cross-domain prerequisite chain learning.

LectureBank 2.0: A collection of English slides in NLP-related courses

Li et al., R-VGAE: Relational-variational Graph Autoencoder for Unsupervised Prerequisite Chain Learning, COLING 2020

Domain	#courses	#files	#tokens	#pages	#tokens/page
NLP	45	953	1,521,505	37,213	40.886
ML	15	312	722,438	12,556	57.537
DL	7	259	450,879	7,420	60.765
AI	5	98	139,778	3,732	37.454
IR	5	95	205,359	4,107	50.002
Overall	77	1,717	3,039,959	65,028	46.748

Table 1: Dataset Statistics. In each category, we have a given number of courses (#courses); each course consists of lecture files (#files); each lecture file has a number of individual slides (#pages). We also show the number of total tokens (#tokens) and average token number per slide (#tokens/page).

This work: LectureBank CD

Cross-domain version of LectureBank.

Covers 3 domains: NLP, CV and BIO.

Same annotation strategy.

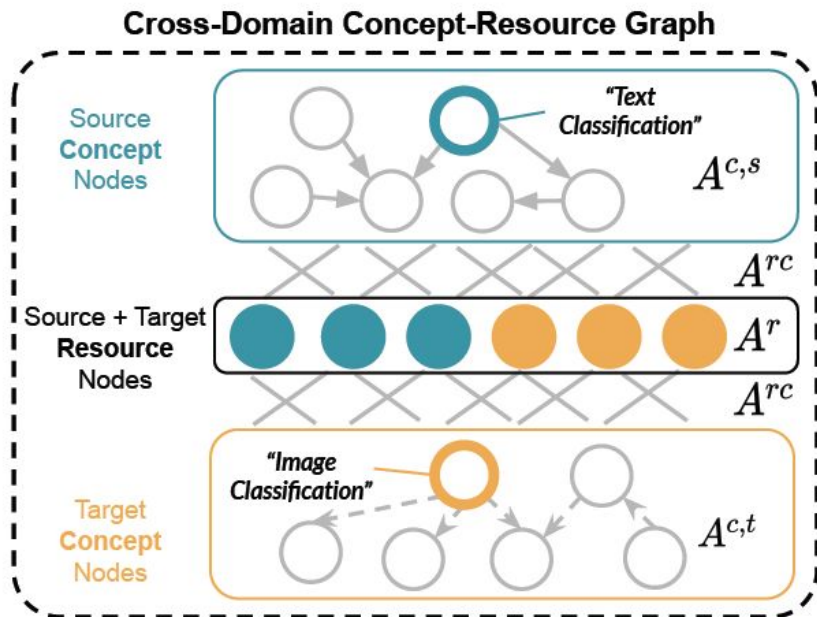
Collected for cross-domain prerequisite chain learning research. [NLP-> CV and NLP -> BIO]

Domain	Files	Pages	Tks/pg	Con.	PosRel
NLP	1,717	65,028	47	322	1,551
CV	1,041	58,32	43	201	871
BIO	148	7,13	135	100	234

Comparison of statistics from NLP, CV and BIO datasets: Tks/pg (Tokens per slide page), Con. (Number of concepts), PosRel (Positive Relations).

<https://github.com/Yale-LILY/LectureBank/tree/master/LectureBankCD>

Proposed Method: Cross-domain Variational Graph Autoencoders (CD-VGAE)



CD-VGAE Model Illustration

We model the resource nodes (solid nodes) and concept nodes (hollow nodes) from two domains (in blue and orange) in a heterogeneous graph.

Node Feature (X)

Phrase2vec (Artetxe et al., 2018)

BERT (Devlin et al., 2018)

$A^{c,s}$ Annotated;

$A^{r,c}$ A^r Cosine similarity;

$A^{c,t}$ To be predicted.

CD-VGAE

Considering “domain neighbors” when doing Graph Convolution:

$$h_i^{(l+1)} = \sigma \left(\underbrace{\sum_{j \in N_i} W^{(l)} h_j^{(l)}}_{\text{Information from its direct neighbors}} + \underbrace{W^{(l)} h_i^{(l)}}_{\text{Node itself}} + \underbrace{\sum_{k \in N_i^D} W_D^{(l)} h_k^{(l)}}_{\text{Information from its domain neighbors}} \right)$$

Then we apply Cross-domain GCN as the encoder of a VGAE (Kipf and Welling, 2016) model to recover the adjacency matrix, evaluated on concept edges only.

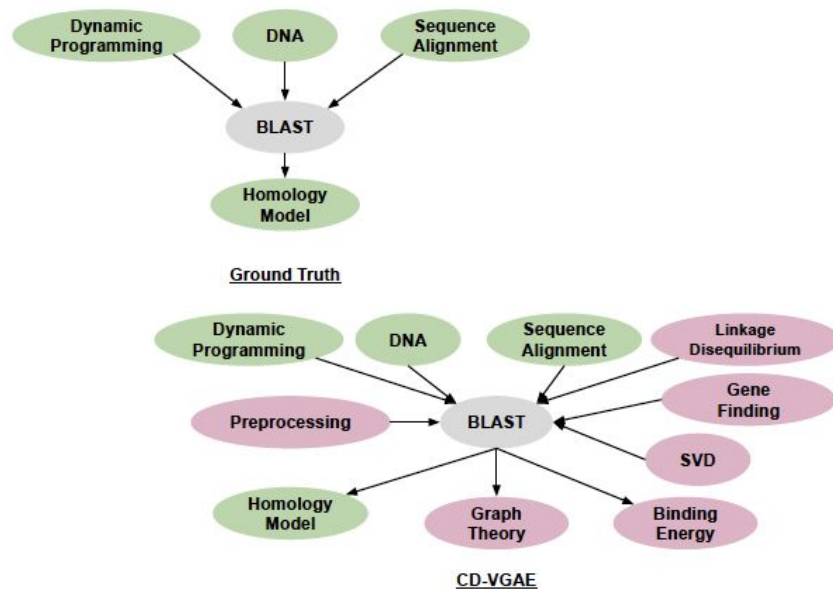
Benchmark Evaluation

Method	NLP→CV				NLP→BIO			
	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec
Baseline Models								
CLS + BERT	0.4277	0.5480	0.5743	0.3419	0.3930	0.6000	0.7481	0.2727
CLS + P2V	0.4881	<u>0.5757</u>	0.6106	0.4070	0.2222	0.5333	0.6000	0.1364
VGAE + BERT (Li et al., 2019)	0.5885	<u>0.5477</u>	0.5398	0.6488	0.6011	0.6091	0.6185	0.5909
VGAE + P2V (Li et al., 2019)	<u>0.6202</u>	0.5500	0.5368	0.7349	<u>0.6177</u>	<u>0.6273</u>	0.6521	0.6091
Proposed Method								
CD-VGAE + BERT	0.6391	0.5593	0.5441	0.7884	0.6289	0.6273	0.6425	0.6364
CD-VGAE + P2V	0.6754	0.5759	0.5468	0.8837	0.6512	0.6591	0.6667	0.6364
Supervised Performance – Upper Bound								
CLS + Node2vec (Grover and Leskovec, 2016)	0.8172	0.8197	0.8223	0.8140	0.8060	0.7956	0.7547	0.8727

Case Studies

Base	VGAE
Image Representation OCR	Image Representation Computer graphics Eye Tracking
CD-VGAE	Ground Truth
Video/Image augmentation Image Representation Face Detection Emotion Recognition Feature Extraction Feature Learning OCR Computer Graphics Eye Tracking	Video/Image augmentation Image Representation Face detection Emotion Recognition Feature Extraction Feature Learning OCR Computer Graphics Eye Tracking

CV: Successors of the concept **Image Processing**



Bio: Case study of direct neighbors of BLAST, including successors and prerequisites.

Conclusion

- We propose cross-domain variational graph autoencoders to perform unsupervised prerequisite chain learning in a heterogeneous graph.
- We are the first to perform domain transfer within a single graph, to the best of our knowledge.
- We introduce the LectureBankCD dataset by collecting and annotating resources and concepts in two new target domains, promoting research on cross-domain prerequisite chain learning.

Thanks
Q&A