Better Together: Text + Context

Kenneth W. Church,* John E. Ortega,*
Maria Antoniak,** Sergey Feldman** and Hui Guan***

*Northeastern University, **Allen Institute for Artificial Intelligence and ***Umass

Meta Issues:

(Important for the future of JSALT)

- Diversity:
  - Better Together: Text proposals + Speech proposals
- Deliverables:
  - Basic tasks + Stretch tasks
  - Evaluation
    - Better benchmarks
  - Resources: more embeddings
  - Tools: APIs for routing papers, recommendations
  - Theory: Unified Framework of Deep nets & SVD

work with conference organizers

New Stuff (since yesterday)

- New team member: Hui Guan
- Contribution from new team member:
  - GNNs with parallelism during training
- Evaluation (and related downstream evaluations)

Team

EAI The Institute for Experiential AI
Northeastern University

SEMANTIC SCHOLAR

University of Massachusetts Amherst

Semantic Scholar: Significant Effort (slide from Dan Weld)

50 person team

7 year project

207M+ scientific paper index

8M+ monthly active users

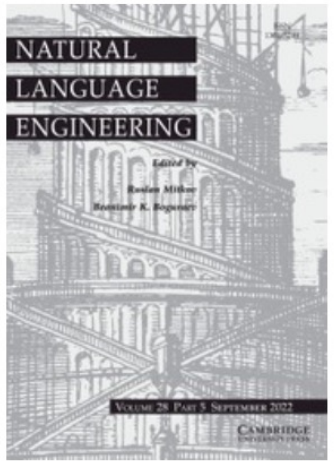
Hui Guan - Bio

- Assistant Professor, Computer Science, UMass Amherst
- Expertise:
  - Systems for Machine Learning
  - Multi-Task Learning
  - Graph Machine Learning
- Website: https://guanh01.github.io/
  - **[OSR'21] Scalable Graph Neural Network Training: The Case for Sampling.**
  - **[NeurIPS'2022] AutoMTL: A Programming Framework for Automating Efficient Multi-Task Learning.**
  - **[PLDI'19] Wootz: a Compiler-based Framework for Fast CNN Pruning via Composability**

University of Massachusetts Amherst

Teaser: Goals for Recommender System

- Recommender Systems
  - Relevance (on topic)
  - Importance (highly cited)
- Do not return papers that are
  - Buzz word compliant
  - But not credible



Emerging Trends: SOTA-Chasing

Published online by Cambridge University Press: **08 February 2022**

Kenneth Ward Church and Valia Kordoni

Article Figures Metrics

Save PDF | Share | Cite | Rights & Permissions

- Paper routing for conferences
  - Submissions → reviewers
- Conjectures:
  - Iffy software
    - Automatic assignments are worse than manual assignments
    - Reviewers are less qualified and less sympathetic to background
      - than target audience
- Better assignments → better conferences

ABSTRACT

This paper introduces ArtELingo, a new benchmark and dataset, designed to encourage work on diversity across languages and cultures. Following ArtEmis, a collection of 80k artworks from WikiArt with 0.45M emotion labels and English-only captions, ArtELingo adds another 0.79M annotations in Arabic and Chinese, plus 4.8K in Spanish to evaluate "cultural-transfer" performance. More than 51K artworks have 5 annotations or more in 3 languages. This diversity makes it possible to study similarities and differences across languages and cultures. Further, we investigate captioning tasks, and find diversity improves the performance of baseline models. ArtELingo is publicly available[1] with standard splits and baseline models. We hope our work will help ease future research on multilinguality and culturally-aware AI.

Semantic Scholar
Specter Embeddings

TEXT EMBEDDINGS

CITATIONS

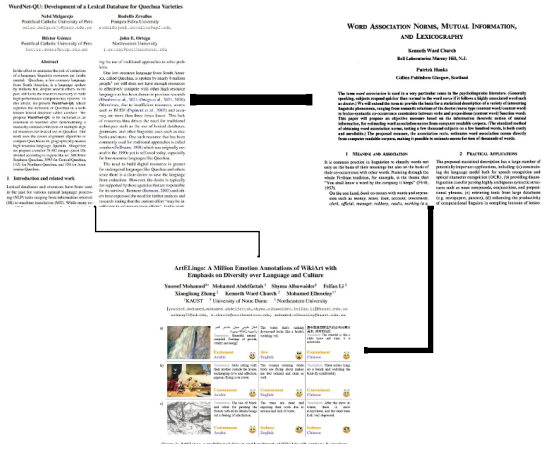
Malak Abdullah and Samira Shaikh. 2018. TeamUNCC at SemEval-2018 task 1: Emotion detection in English and Arabic tweets using deep learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 350–357, New Orleans, Louisiana. Association for Computational Linguistics.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Lila Abu-Lughod. 1990. The romance of resistance: Tracing transformations of power through bedouin women. *American ethnologist*, 17(1):41–55.

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. Artemis: Affective language for visual art. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.



Paper Citation
Graph

PRONE
Node2Vec

CONTEXT
EMBEDDINGS

BETTER
TOGETHER
PROPOSAL

Scale: Smaller than Web
(but challenging for academia)

	Source	Papers (millions)
1	CorpusId	207.80
2	MAG (Microsoft Academic Graph) 	182.18
3	DOI	113.54
4	PubMed	35.03
5	DBLP	6.06
6	PubMedCentral	4.86
7	ArXiv	2.15
8	ACL	0.08
	totals	551.71

Papers (millions)	8 Bits	Sources
79.18	11000000	CorpusId, MAG
68.37	11100000	CorpusId, MAG, DOI
19.69	11110000	CorpusId, MAG, DOI, PubMed
12.05	10100000	CorpusId, DOI
5.66	10000000	CorpusId
5.10	11010000	CorpusId, MAG, PubMed
4.11	11101000	CorpusId, MAG, DOI, DBLP
2.99	11110100	CorpusId, MAG,DOI, PubMed, PubMedCentral
2.96	10010000	CorpusId, PubMed
7.69	other	
207.80	**totals**	

Proposal: Build Multiple Embeddings to Capture Text and/or Context

- Text vs Context
  - Text (features within docs)
    - Examples: Titles, abstracts, body
  - Context (features in other docs)
    - Examples: Citation graph, citing sentences
- Embeddings: $M \in \mathbb{R}^{N \times K}$
  - N: number of documents (≈200M)
  - K: number of hidden dims (≈768)
  - Similarity of two docs: $\cos$
    - Similarity of all pairs of docs: MM^T
- Practical Apps of Embeddings (M):
  - Approximate nearest neighbors on M
    - Information Retrieval
    - Recommender Systems
    - Routing
    - Finding experts

- Text and/or Context Embeddings, M_t, M_c
  - Example of M_t: Specter
    - SciBERT + finetuning
    - Available from Semantic Scholar API
  - Example of M_c: (Proposed alternative)
    - Node2vec: $G \rightarrow M$ (Spectral Clustering)
      - G: citation graph (from API)
      - We use ProNE version of node2vec to compute M_c
- Interpretations:
  - large $\cos$ on M_t:
    - docs share similar text
  - large $\cos$ on M_c:
    - docs are close in commute distance on G
      - random walk from node i → j and back to i
      - commute dist: symmetric to match $\cos$

old

Latent Semantic Indexing

new

Currently, not in Semantic Scholar

Basic & Stretch Tasks

- **Basic Tasks**
  - Create multiple embeddings
  - Show some are better for text
    - and others are better for context
  - Show better together
    - Combination of text and context
    - is better than either by itself
  - Challenges: Scale
    - Semantic Scholar:
      - $N \approx 200$M docs
      - $G = (N, E)$ where E $\approx$ 2B citations
    - Currently, node2vec is expensive
      - time (≈week) and space (≈2TB of RAM),
      - but plenty of opportunities for more feasible approximations
  - GNN alternative: coming soon
  - Evaluation: next slide

Current status (K=280)

- **Stretch Tasks**
  - More interesting M_c embeddings
    - M_c: BERT encoding of citing sentences
    - Citing sentence(i):
      - sentence in another doc j that cites i
    - Example:
      - Turing (1936) has high impact (citations)
      - Why? Introduced:
        - *Turing Machine* & *Halting Problem*
      - Those terms are common in citing sents,
        - but not mentioned in Turing (1936)
  - Time Invariance:
    - M_c evolves with future citations unlike M_t
      - (M_t doesn't change after publication)
    - Can we model evolution of vectors in M_c over time like a flower blooming?
    - Should we think of literature as a conversation
      - (like social media)?
    - Impact combines contributions from
      - authors (text) +
      - audience appreciation (context)

Anchor Text

New Stuff (since yesterday)

✓New team member: Hui Guan

- Contribution from new team member:
  - GNNs with parallelism during training

➢**Evaluation (and related downstream evaluations)**

Evaluation

MAG240M

Dataset	# Nodes	# Edges	# Feat
ogbn-products (PR)	2.4M	62M	100
ogbn-papers100M (PA)	111M	1.6B	128
Amazon (AM)	1.56M	168M	200

- Established Benchmarks (⚠)
  - https://mimno.infosci.cornell.edu/data/nips_reviewer_data.tar.gz
  - https://github.com/allenai/scidocs
  - "Related" downstream tasks (⚠ ⚠ ⚠)
    - MAG240M https://arxiv.org/pdf/2103.09430.pdf
- Create a new Benchmark
  - Work with conference organizers (Bhuvana Ramabhadran)
- Predict held out citations
  - Remove some edges from citation graph
  - See how well you can predict the held out edges
  - (This method can also be used to train embeddings)

Most Relevant

Specter

“Related” Downstream Evaluations

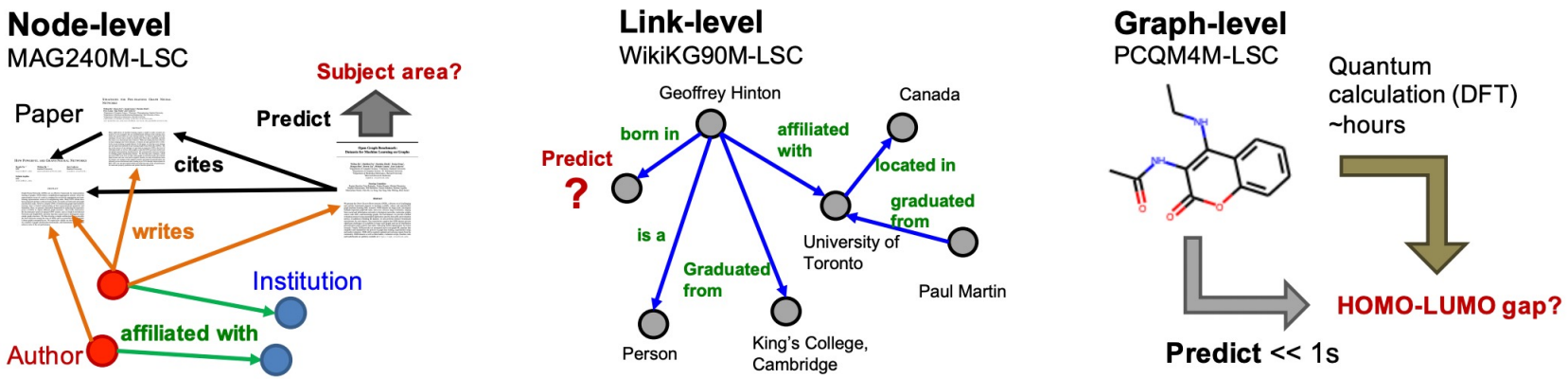
(from https://arxiv.org/pdf/2103.09430.pdf)

Table 1: **Basic statistics of the OGB-LSC datasets used in KDD Cup 2021.** Datasets marked by † has been updated to v2 after the KDD Cup (*cf.* Section 3).

Task type	Dataset	Statistics	
Node-level	**MAG240M**	#nodes: #edges:	244,160,499 1,728,364,232
Link-level	**WikiKG90M**†	#nodes: #edges:	87,143,637 504,220,369
Graph-level	**PCQM4M**†	#graphs: #edges (total):	3,803,453 55,399,880

	Source	Papers (millions)
1	CorpusId	207.80
2	MAG (Microsoft Academic Graph)	182.18
3	DOI	113.54
4	PubMed	35.03
5	DBLP	6.06
6	PubMedCentral	4.86
7	ArXiv	2.15
8	ACL	0.08

“Related” but not quite the same as our problem



(a) MAG240M (b) WikiKG90M (c) PCQM4M

Figure 1: **Overview of the three OGB-LSC datasets, covering node-, link-, and graph-level prediction tasks, respectively.** **(a)** MAG240M is a heterogeneous academic graph, and the task is to predict the subject areas of papers situated in the heterogeneous graph (*cf.* Section 2.1). **(b)** WikiKG90M is a knowledge graph, and the task is to impute missing triplets (*cf.* Section 2.2). **(c)** PCQM4M is a quantum chemistry dataset, and the task is to predict an important molecular property—the HOMO-LUMO gap—of a given molecule (*cf.* Section 2.3).

Related Task: predict subject areas
Our Task: est similarity of two docs

	Source	Papers (millions)
1	CorpusId	207.80
2	MAG (Microsoft Academic Graph)	182.18
3	DOI	113.54

Text Embedding

Graph. We extract 121M academic papers in English from MAG (version: 2020-11-23) to construct a heterogeneous academic graph. The resultant paper set is written by 122M author entities, who are affiliated with 26K institutes. Among these papers, there are 1.3 billion citation links captured by MAG. Each paper is associated with its natural language title and most papers' abstracts are also available. We concatenate the title and abstract by period and pass it to a RoBERTa sentence encoder (Liu *et al.*, 2019; Reimers and Gurevych, 2019), generating a 768-dimensional vector for each paper node. Among the 121M paper nodes, approximately 1.4M nodes are arXiv papers annotated with 153 arXiv subject areas, *e.g.*, `cs.LG (Machine Learning)`. On the paper nodes, we attach the publication years as meta information.

Prediction task and evaluation metric. The task is to predict the primary subject areas of the given arXiv papers, which is cast as an ordinary multi-class classification problem. The metric is the classification accuracy.

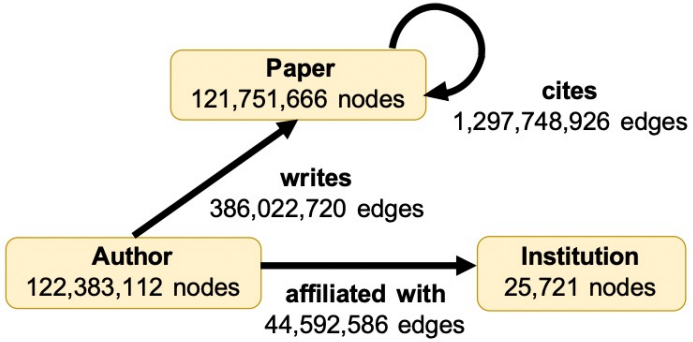


Figure 2: **A schema diagram of MAG240M.**

Table 2: **Results of MAG240M measured by the accuracy (%).**

Model	#Params	Validation	Test
MLP	0.5M	52.67	52.73
LABELPROP	0	58.44	56.29
SGC	0.7M	65.82	65.29
SIGN	3.8M	66.64	66.09
MLP+C&S	0.5M	66.98	66.18
GRAPHSAGE (NS)	4.9M	66.79	66.28
GAT (NS)	4.9M	67.15	66.80
R-GRAPHSAGE (NS)	12.2M	69.86	68.94
R-GAT (NS)	12.3M	**70.02**	**69.42**
KDD 1ST: BD-PGL			**75.49**
KDD 2ND: ACADEMIC			**75.19**
KDD 3RD: SYNERISE AI			**74.60**

Table 3: **Analysis of graph homophily for different meta-paths connecting 1,251,341 arXiv papers (only train+validation).** Connection strength indicates the number of different possible paths along the template meta-path, *e.g.*, meta-path "Paper-Author-Paper (P-A-P)" with connection strength 3 means that at least 3 authors are shared for the two papers of interest. Homophily ratio is the ratio of two nodes having the same target labels.

Meta-path	Connect. strength	Homophily ratio (%)	#Edges
P-P	1	57.80	2,017,844
P-A-P	1	46.12	88,099,071
	2	57.02	12,557,765
	4	64.03	1,970,761
	8	66.65	476,792
	16	70.46	189,493
P-A-I-A-P	1	3.83	159,884,165,669
	2	4.61	81,949,449,717
	4	5.69	33,764,809,381
	8	6.85	12,390,929,118
	16	7.70	4,471,932,097
All pairs	0	1.99	782,926,523,470

GNNs (Addanki *et al.*, 2021). As real-world large-scale graphs are almost always dynamic, exploiting the temporal information is a promising direction of future research.

New Stuff (since yesterday)

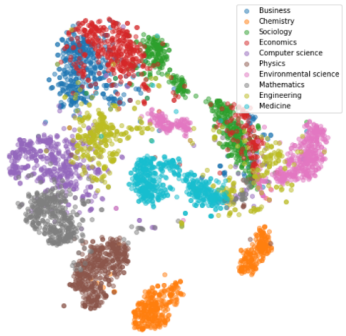
✓New team member: Hui Guan

➢**Contribution from new team member:**

- **GNNs with parallelism during training**

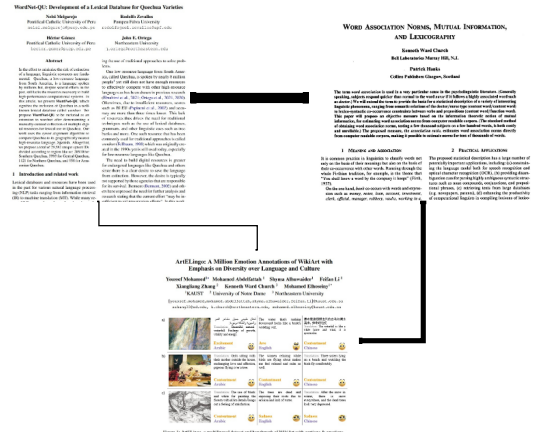
✓Evaluation (and related downstream evaluations)

GNN Alternative



Semantic Scholar
Specter Embeddings

Text



Paper Citation
Graph

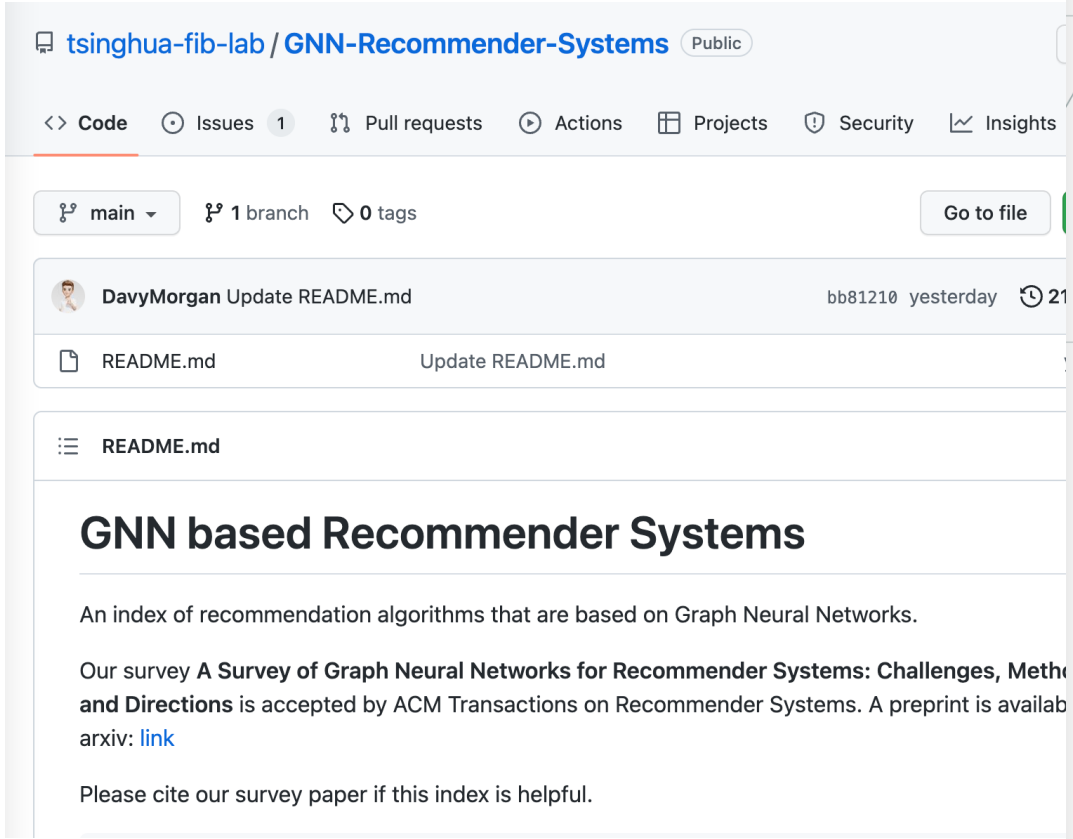
Context

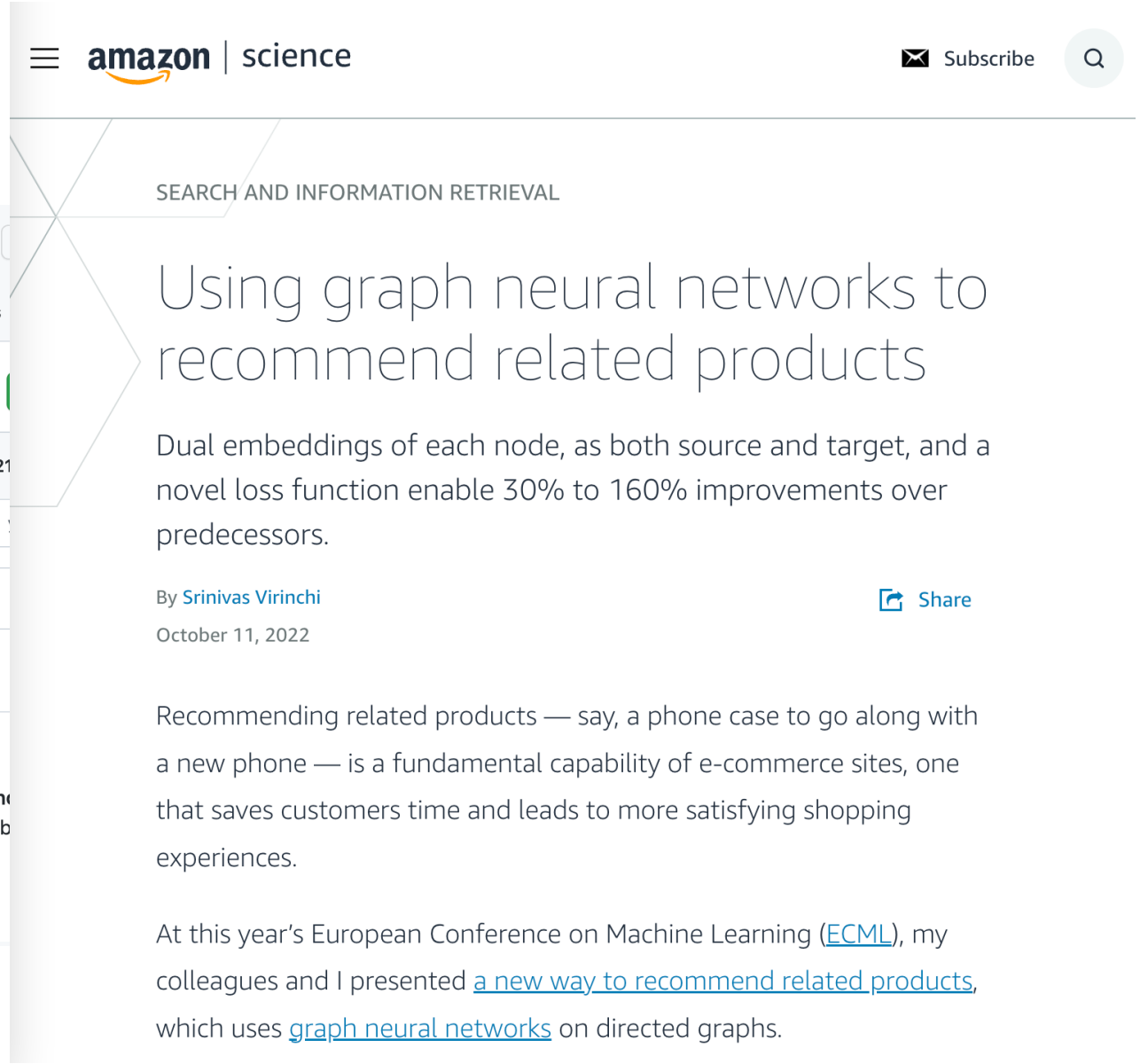
Black Box

GNN

Embeddings
$M \in \mathbb{R}^{N \times K}$

Why GNN?


tsinghua-fib-lab / GNN-Recommender-Systems
Public
Code
Issues 1
Pull requests
Actions
Projects
Security
Insights
main
1 branch
0 tags
Go to file
DavyMorgan Update README.md
bb81210 yesterday
README.md
Update README.md
README.md
GNN based Recommender Systems
An index of recommendation algorithms that are based on Graph Neural Networks.
Our survey A Survey of Graph Neural Networks for Recommender Systems: Challenges, Meth
and Directions is accepted by ACM Transactions on Recommender Systems. A preprint is availab
arxiv: link
Please cite our survey paper if this index is helpful.

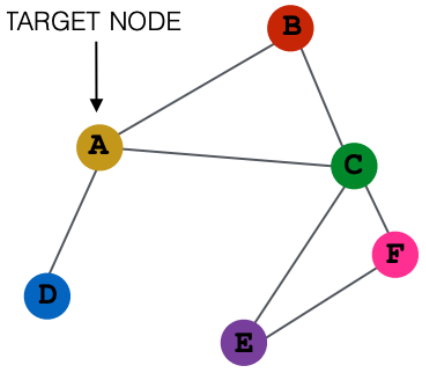

amazon | science
Subscribe
SEARCH AND INFORMATION RETRIEVAL
Using graph neural networks to recommend related products
Dual embeddings of each node, as both source and target, and a novel loss function enable 30% to 160% improvements over predecessors.
By Srinivas Virinchi
Share
October 11, 2022
Recommending related products — say, a phone case to go along with a new phone — is a fundamental capability of e-commerce sites, one that saves customers time and leads to more satisfying shopping experiences.
At this year's European Conference on Machine Learning (ECML), my colleagues and I presented a new way to recommend related products, which uses graph neural networks on directed graphs.

What is GNN?

- Node (a paper): specter embedding
- Edge: citation

Matrix Computation

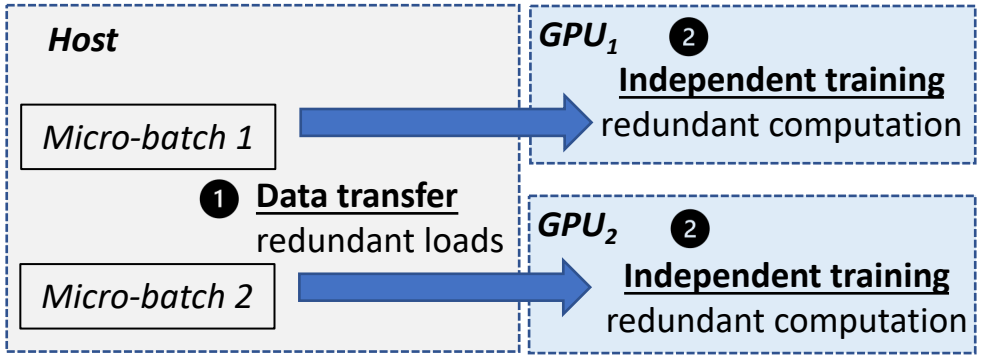
Input specter embedding
$x_i \in R^K$, i= {1,…N}



Input Citation Graph
$G \in R^{N \times N}$

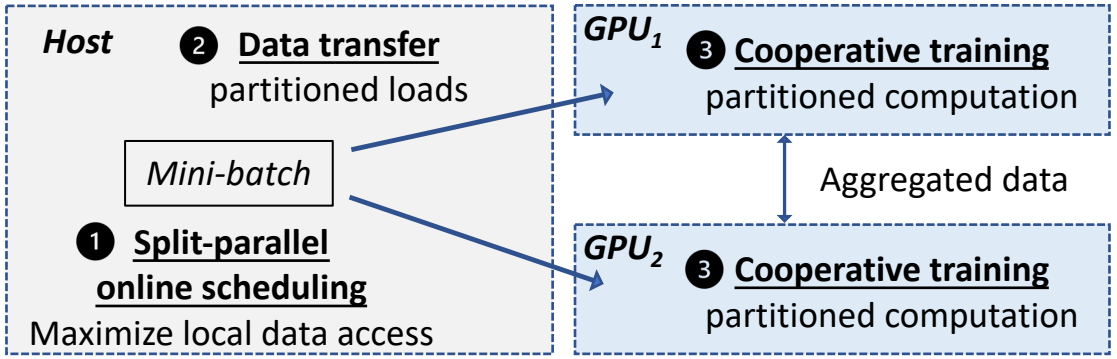
Ref: https://snap-stanford.github.io/cs224w-notes/machine-learning-with-networks/graph-neural-networks

Challenge Scale
Data Parallelism v.s. Gsplit Parallelism



Today's data parallelism (e.g., DGL, PaGraph)

Standard



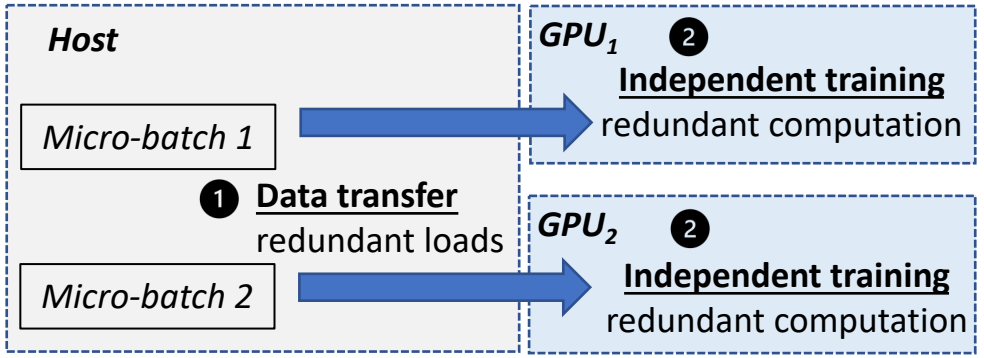
Proposed split parallelism (Gsplit)

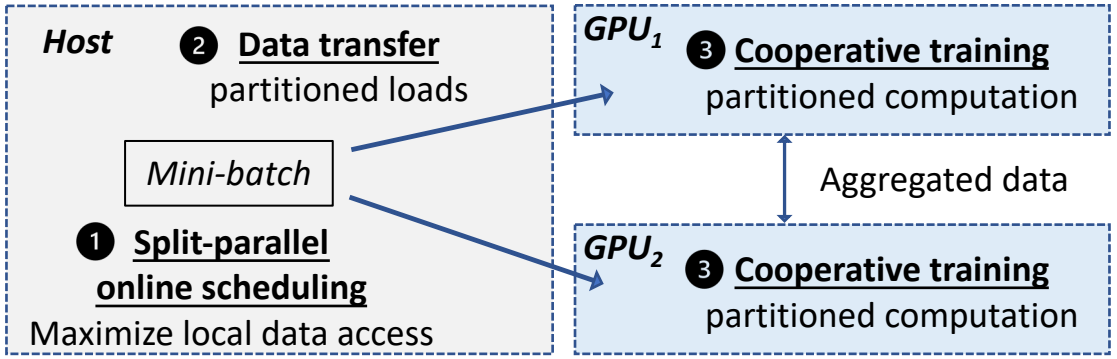
Proposed

- Data parallelism – too much communication
- **Split parallelism (Proposed)**

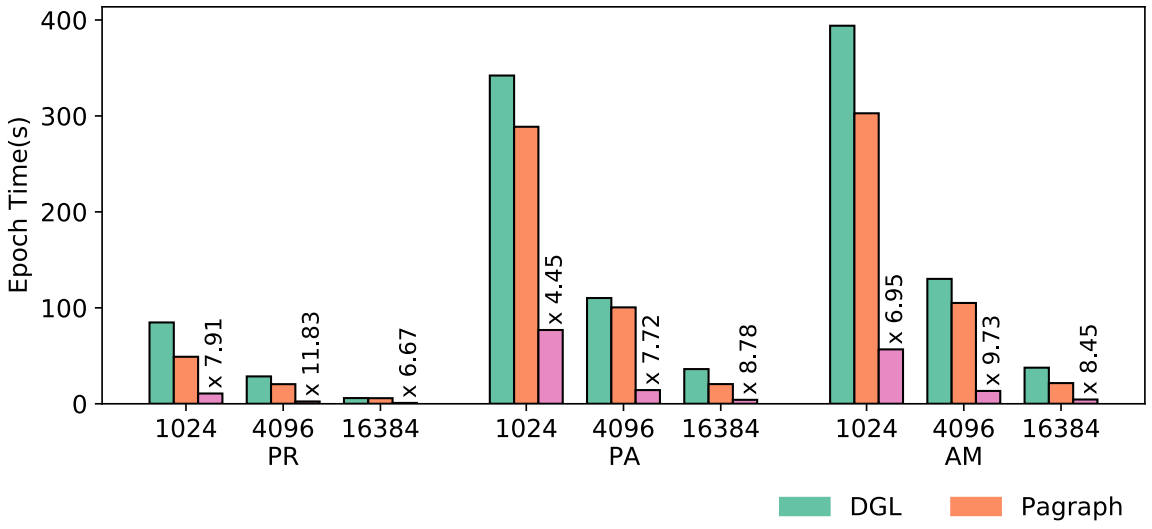
Gsplit is Faster

Dataset	# Nodes	# Edges	# Feat
ogbn-products (PR)	2.4M	62M	100
ogbn-papers100M (PA)	111M	1.6B	128
Amazon (AM)	1.56M	168M	200

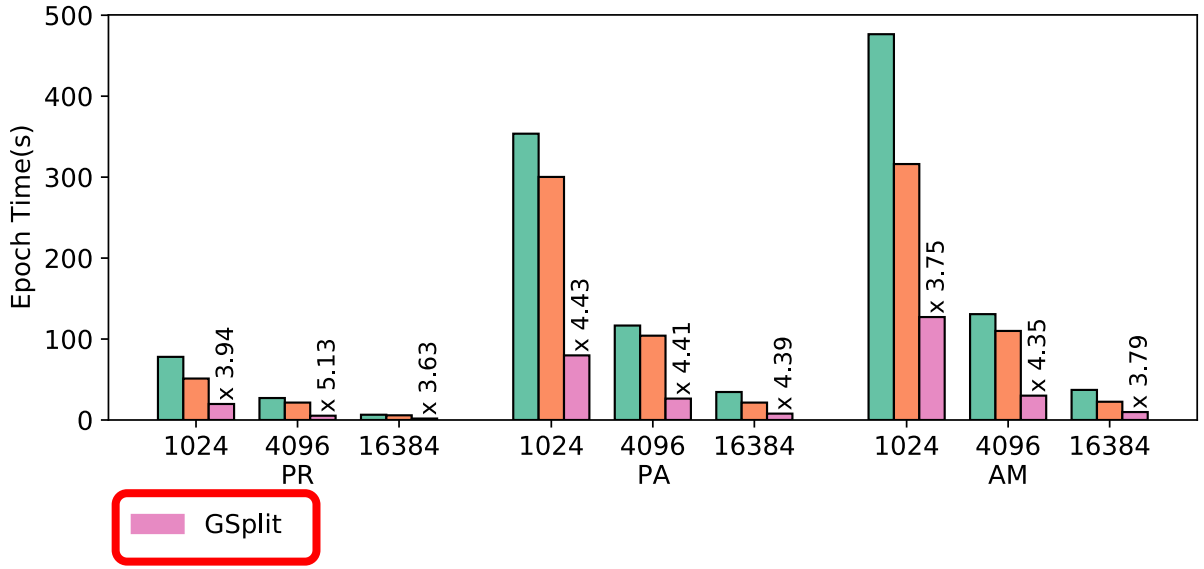




Today's data parallelism (e.g., DGL, PaGraph)



Proposed split parallelism (Gsplit)



Bigger Stretch: Theoretical combinations of deep nets & SVD

- Conjecture: More reps → more understanding
  - Unified theory of deep nets & SVD
- Compare & contrast: deep nets & SVD
  - BERT: example of deep nets
  - ProNE (node2vec): based on SVD
  - Both produce (similar) embeddings
- In speech, a spectrogram is just a different representation of wave file
  - So too, node2vec emb (M) is just a diff rep of G
  - Some representations are more convenient than others (depends on what you want to do)
    - Embeddings make it easy to estimate $\cos$
- If M is computed from G,
  - then M may have more parameters than G
  - but no more information: $H(M) \leq H(G)$
- In particular, SVD on sparse G increases params
    - but SVD does not create information
  - M_c has NK params, more than G (E params)
    - Since info is not created,
    - Most of the NK params must be redundant
  - Redundancy is easier to see with SVD than deep nets
- With traditional regression,
    - too many params → overfitting
  - But deep nets thrive on scale:
    - better results with more N, more K, etc. Why?
- Suggestions:
  - Larger K improves estimates of $\cos$
    - Easier to see with node2vec (SVD) than deep nets
    - More K → Less dimension reduction
  - Network effects (Metcalfe's law)
    - Larger N makes search easier (not harder)
    - Web search $\gg$ Enterprise search
    - Easier to see with G than other representations

backup

Basic & Stretch Tasks

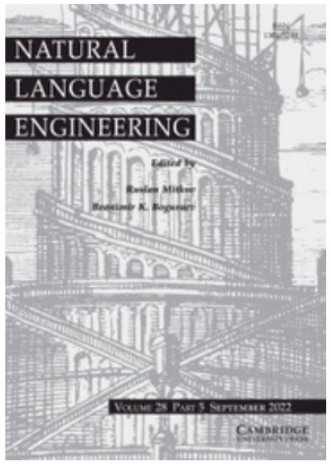
- **Basic Tasks**
  - Create multiple embeddings
  - Show some are better for text
    - and others are better for context
  - Show better together
    - Combination of text and context
    - is better than either by itself
  - Challenges: Scale
    - Semantic Scholar:
      - $N \approx 200$M docs
      - $G = (N, E)$ where E $\approx$ 2B citations
    - Currently, node2vec is expensive
      - time (≈week) and space (≈2TB of RAM),
      - but plenty of opportunities for more feasible approximations

Current status (K=280)

- **Stretch Tasks**
  - More interesting M_c embeddings
    - M_c: BERT encoding of citing sentences
    - Citing sentence(i):
      - sentence in another doc j that cites i
    - Example:
      - Turing (1936) has high impact (citations)
      - Why? Introduced:
        - *Turing Machine* & *Halting Problem*
      - Those terms are common in citing sents,
        - but not mentioned in Turing (1936)
  - Time Invariance:
    - M_c evolves with future citations unlike M_t
      - (M_t doesn't change after publication)
    - Can we model evolution of vectors in M_c over time like a flower blooming?
      - Can we invert the process
      - going back before ``big bang'' (publication)?
    - Should we think of literature as a conversation (like social media)?
    - Impact combines contributions from
      - authors (text) +
      - audience appreciation (context)

Anchor Text

SOTA-Chasing



Emerging Trends: SOTA-Chasing

Published online by Cambridge University Press: **08 February 2022**

Kenneth Ward Church and Valia Kordoni

- Root causes for SOTA-chasing
  1. Iffy reviewing
  2. Lack of leadership
- Iffy reviewing
  - Root cause: Bad assignments of submissions to reviewers
    - We used to do this by hand (better than we do these days)
    - Current status: Conference software (open review, softconf, easy chair) use iffy programs
    - Evaluation: Safe and effective (???)
  - Reviewers should be
    - More qualified (and sympathetic to background assumptions)
    - than target audience
  - If not, reviewers will ``abstain'' (average grade or worse) → kill paper
- Bad assignments
  - → Teach authors to write papers that can be reviewed by unqualified reviewers
  - → SOTA-chasing
  - → Lack of meaningful progress

ABSTRACT

This paper introduces ArtELingo, a new benchmark and dataset, designed to encourage work on diversity across languages and cultures. Following ArtEmis, a collection of 80k artworks from WikiArt with 0.45M emotion labels and English-only captions, ArtELingo adds another 0.79M annotations in Arabic and Chinese, plus 4.8K in Spanish to evaluate "cultural-transfer" performance. More than 51K artworks have 5 annotations or more in 3 languages. This diversity makes it possible to study similarities and differences across languages and cultures. Further, we investigate captioning tasks, and find diversity improves the performance of baseline models. ArtELingo is publicly available[1] with standard splits and baseline models. We hope our work will help ease future research on multilingual-ity and culturally-aware AI.

CITATIONS

Malak Abdullah and Samira Shaikh. 2018. TeamUNCC at SemEval-2018 task 1: Emotion detection in English and Arabic tweets using deep learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 350–357, New Orleans, Louisiana. Association for Computational Linguistics.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Lila Abu-Lughod. 1990. The romance of resistance: Tracing transformations of power through bedouin women. *American ethnologist*, 17(1):41–55.

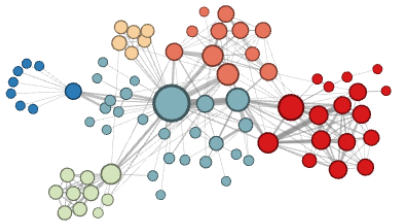
Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. Artemis: Affective language for visual art. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

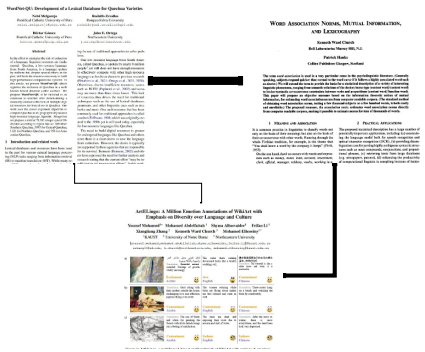
Semantic Scholar
Specter Embeddings

$$\begin{bmatrix} 8 & 2 & 1 & 9 \\ 6 & 5 & 4 & 0 \\ 7 & 1 & 6 & 2 \\ 1 & 3 & 5 & 8 \\ 0 & 4 & 9 & 1 \end{bmatrix}$$

TEXT EMBEDDINGS



PRONE
Node2Vec



Paper Citation
Graph

CONTEXT
EMBEDDINGS

BETTER
TOGETHER
PROPOSAL