DeepMind

Large-Scale Representation Learning on Graphs via Bootstrapping

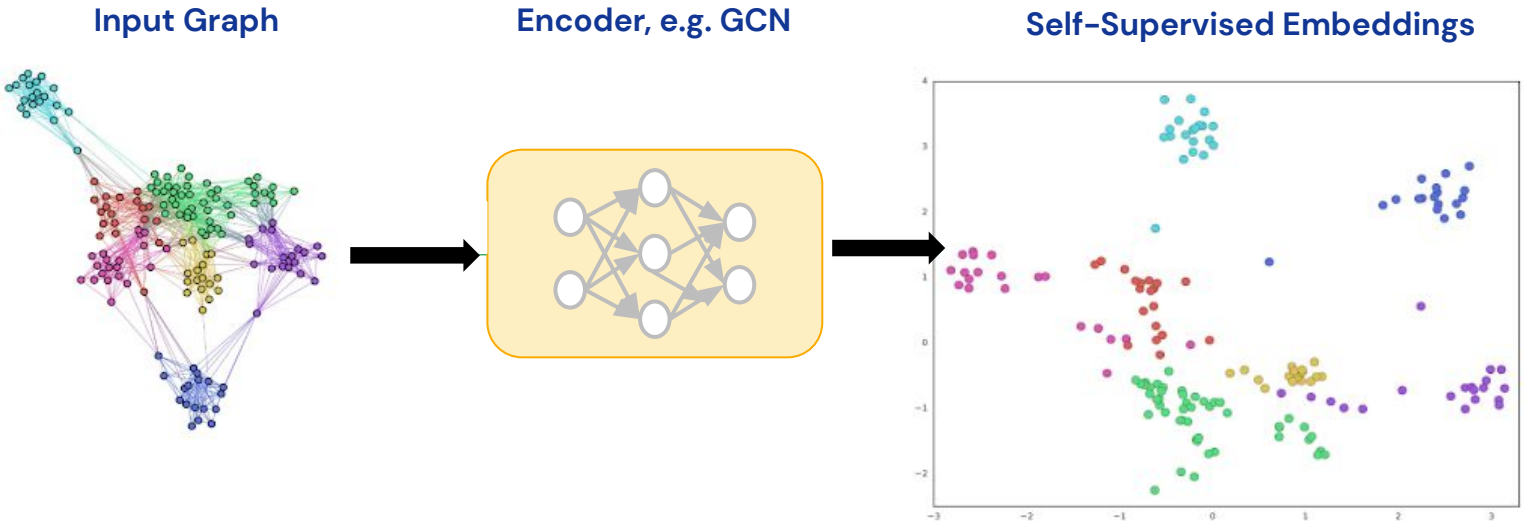
Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L Dyer, Remi Munos, Petar Veličković, Michal Valko

ICLR 2022

Graph Representation Learning

- Goal: Learn meaningful node representations *without supervision*

- Why?
  - Unlabeled data cheaper than obtaining labels
  - Pre-training for downstream tasks
  - Auxiliary signal for semi-supervised training

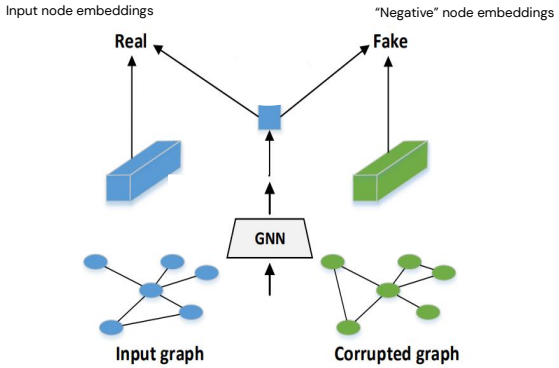


https://ai.googleblog.com/2019/06/innovations-in-graph-representation.html

Drawbacks of Existing Methods

- Contrastive methods
  - Push together similar objects (*positive examples*)
  - Pull apart dissimilar objects (*negative examples*)

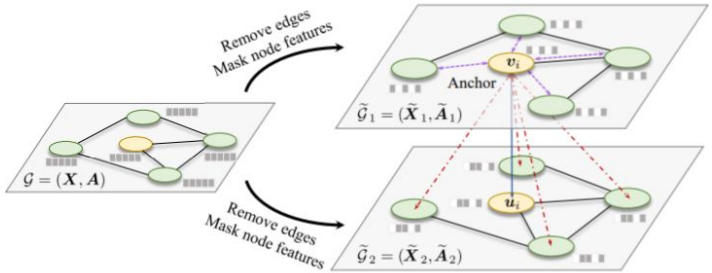
Deep Graph Infomax (DGI) [Veličković et al, 2017]
Contrasts against negative graph



Downside: Hard to define negative examples at graph level

https://dl.acm.org/doi/fullHtml/10.1145/3366423.3380112

Deep Graph Contrastive Representation Learning (GRACE) [Zhu et al, 2020]
Contrasts every node against each other

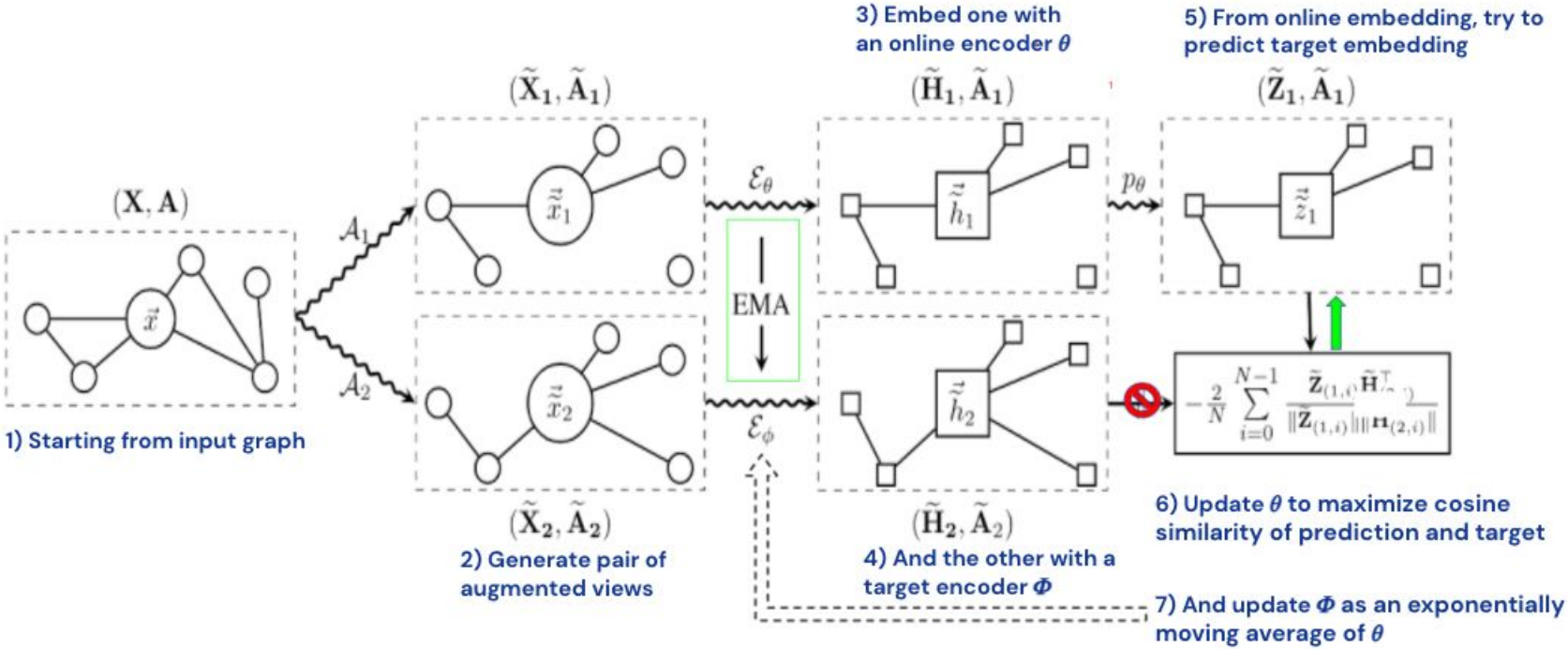


Downside: All-vs-all contrastive scales quadratically

https://medium.com/syncedreview/deep-graph-contrastive-representation-learning-39ecaeee45f5

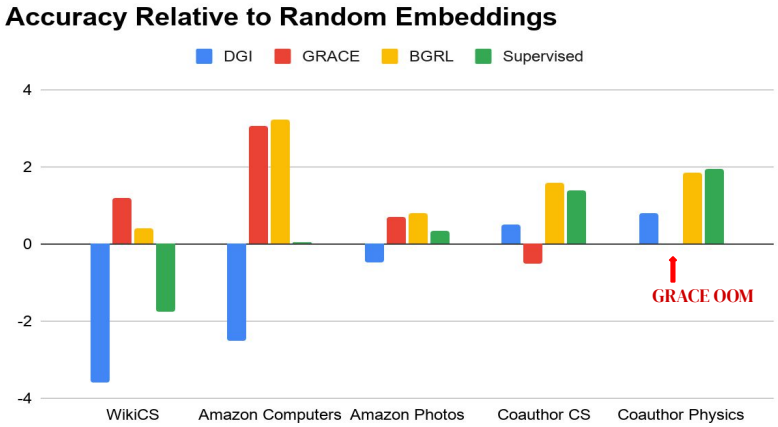
Our Solution: Bootstrapped Graph Latents (BGRL)

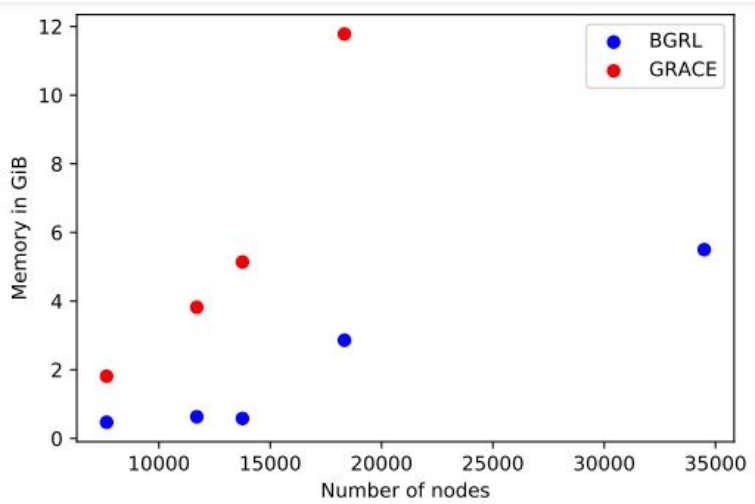
- Key idea: avoid negative examples by bootstrapping embeddings from each node
  Follows framework of Bootstrap Your Own Latent (BYOL) [Grill et al, 2020]
  Loss takes linear time to compute



Improved Performance with Lower Computational Cost

- Linear evaluation protocol on standard benchmarks
  - BGRL competitive with both GRACE and supervised learning
  - BGRL uses **5–10x** less memory and scales to largest dataset





Application to OGB-Large Scale Challenge

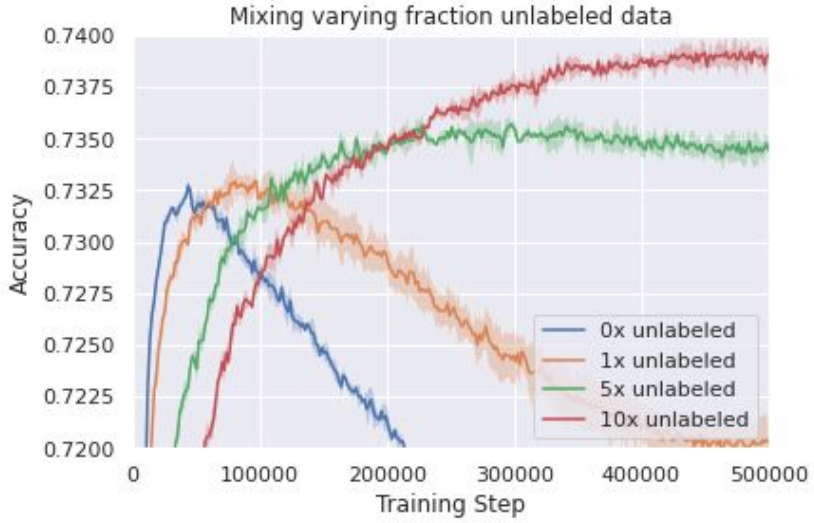
- Microsoft Academic Graph - single connected graph, 240M nodes, 1.7B edges, 350GB data
  - Task is to classify arXiv papers by category
  - Only 1% of nodes are classifiable arXiv papers - how to effectively use other data?

- Main question: *Does BGRL help at super-large scale, for semi-supervised training?*
  - First step - simply add BGRL loss on top of supervised loss on 1% labeled data only

Application to OGB-Large Scale Challenge

- Microsoft Academic Graph - single connected graph, 240M nodes, 1.7B edges
  - Task is to classify arXiv papers by category
  - Only 1% of nodes are classifiable arXiv papers - how to effectively use other data?

- Main question: *Does BGRL help at super-large scale, for semi-supervised training?*
  - Second step - try mixing varying amounts of unlabeled nodes for BGRL loss



Application to OGB-Large Scale Challenge

- Microsoft Academic Graph - single connected graph, 240M nodes, 1.7B edges
  - Task is to classify arXiv papers by category
  - Only 1% of nodes are classifiable arXiv papers - how to effectively use other data?

- Main question: *Does BGRL help at super-large scale, for semi-supervised training?*
  - Final step - mixing 10x unlabeled data delays overfitting, thus can run 10x longer
  - BGRL key to our submission achieving 2nd place in KDD Cup OGB-LSC

Final leaderboard for MAG240M-LSC

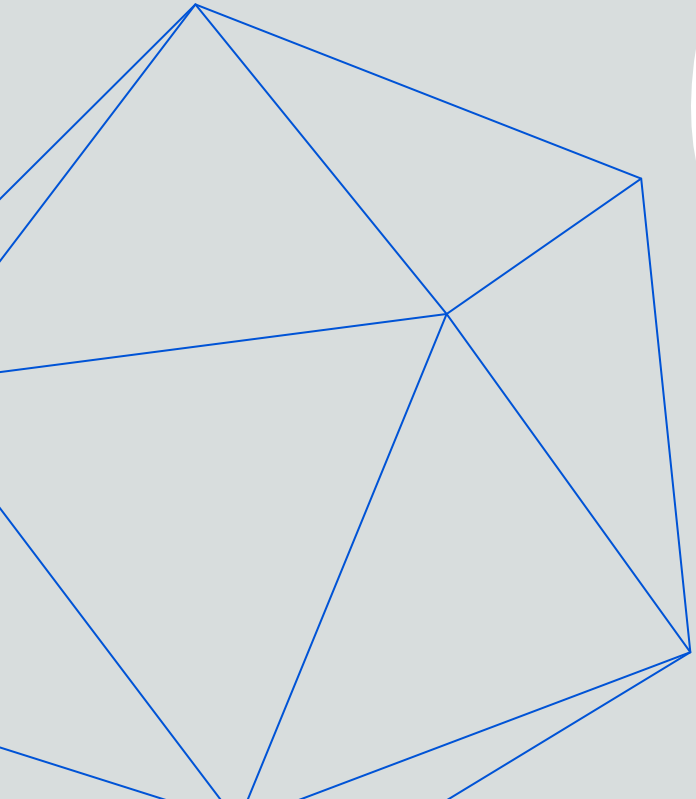
Classification accuracy. The higher, the better.

Rank	Team	Test Accuracy	
1	BD-PGL	0.7549	
2	Academic	0.7519	5.3M Parameters
3	Synerise AI	0.7460	

Conclusions

- BGRL: new method for self-supervised representation learning
- Competitive with contrastive methods, without using negative examples
- Works in wide range of settings: frozen/semi-supervised training, full-graph/subsampled neighborhood training, with simple augmentations
- Update step takes time and memory linear in graph size, highly scalable
- Effective method to leverage unlabeled data on huge graphs, key to our entry on OGB-Large Scale Challenge

DeepMind



Thank You!

For any questions please contact:
thakoor@google.com

Experiment code: https://github.com/nerdslab/bgrl

OGB-LSC Contest submission:
https://github.com/deepmind/deepmind-research/tree/master/ogb_lsc/mag