IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 ©

number of nodes in the graph, degrading the performance of the model for online graph snapshots.

In Table II, we evaluate the performance of the student model Distill2Vec-$\mathcal{S}$ against the baseline approaches in the link prediction task. We observe that the student model Distill2Vec-$\mathcal{S}$ constantly outperforms the baseline approaches, in terms of AUC, for both datasets. This indicates that the proposed knowledge distillation strategy can efficiently transfer the knowledge of the pretrained model Distill2Vec-$\mathcal{T}$ to the student model Distill2Vec-$\mathcal{S}$. Therefore, Distill2Vec-$\mathcal{S}$ achieves high link prediction accuracy, while reducing the number of trainable parameters. Moreover, we observe that Distill2Vec-$\mathcal{L}$ exhibits similar behaviour as Distill2Vec-$\mathcal{S}$. However, the cross entropy function in Distill2Vec-$\mathcal{L}$ limits the prediction accuracy, when compared with the Kullback-Leibler divergence of the proposed the Distill2Vec-$\mathcal{S}$ model. Evaluated against TDGNN, which is the second best baseline approach in all datasets, Distill2Vec-$\mathcal{S}$ achieves relative gains 1.8 and 2.5% for the Yelp and ML-10M datasets, respectively. Note that as shown in Table I Distill2Vec-$\mathcal{S}$ achieves average compression ratio of 7:100 and 35:100, in terms of trainable parameters, when compared with TDGNN for the Yelp and ML-10M dataset, respectively. Thus, our model is able to capture the evolution of the graph in the learned node representations, while significantly reducing the model size. In addition, on inspection of Table II we observe that the student models Distill2Vec-$\mathcal{S}$ and DMTKG-$\mathcal{S}$ constantly outperform their respective teacher models Distill2Vec-$\mathcal{T}$ and DMTKG-$\mathcal{T}$. This demonstrates the capability of student models to overcome any bias introduced by the pretrained teacher models on the offline data. Thus, the student model Distill2Vec-$\mathcal{S}$ achieves relative gains of 5.5 and 4.2% against its teacher model for Yelp and ML-10M, respectively.

IV. Conclusion

In this paper, we presented a knowledge distillation strategy to reduce the size of a teacher model for dynamic graph representation learning. The proposed distillation strategy can efficiently generate a compact student model with low online inference latency, while achieving high link prediction accuracy. The experimental results demonstrate the compression efficiency of our distillation strategy. The proposed student model achieves a compression ratio up to 31:100 on two real-world datasets, when compared with the pretrained teacher model. Evaluated against several state-of-the-art approaches, the proposed student model achieves an average relative improvement of 2.2% on both datasets, by significantly reducing the number of required parameters. An interesting future direction is to explore the performance of data-free distillation strategies on dynamic graph representation learning approaches [27]. The main challenge is to design the student model so as to infer accurate embeddings on unobserved nodes by the teacher model.

References

[1] R. Trivedi, M. Farajtabar, P. Biswal, and H. Zha, "Dyrep: Learning representations over dynamic graphs," in *ICLR*, 2019.

[2] L. Zhou, Y. Yang, X. Ren, F. Wu, and Y. Zhuang, "Dynamic network embedding by modeling triadic closure process," in *AAAI*, 2018, pp. 571–578.

[3] A. Sankar, Y. Wu, L. Gou, W. Zhang, and H. Yang, "Dysat: Deep neural representation learning on dynamic graphs via self-attention networks," in *WSDM*, 2020, pp. 519–527.

[4] L. Zhu, D. Guo, J. Yin, G. V. Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks (extended abstract)," in *ICDE*, 2017, pp. 57–58.

[5] S. Mahdavi, S. Khoshraftar, and A. An, "Dynamic joint variational graph autoencoders," in *ECML*, 2019, pp. 385–401.

[6] E. Hajiramezanali, A. Hasanzadeh, K. R. Narayanan, N. Duffield, M. Zhou, and X. Qian, "Variational graph recurrent neural networks," in *NeurIPS*, 2019, pp. 10 700–10 710.

[7] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, 2006, pp. 535–541.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS*, 2015.

[9] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *CVPR*, 2019, pp. 7096–7104.

[10] M. Phuong and C. Lampert, "Towards understanding knowledge distillation," in *ICML*, 2019, pp. 5142–5151.

[11] J. Tang and K. Wang, "Ranking distillation: Learning compact ranking models with high performance for recommender system," in *KDD*, 2018, p. 2289–2298.

[12] H. Li, T. N. Chan, M. L. Yiu, and N. Mamoulis, "Fexipro: Fast and exact inner product retrieval in recommender systems," in *SIGMOD*, 2017, p. 835–850.

[13] Y. Cao, X. Wang, X. He, Z. Hu, and T.-S. Chua, "Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences," in *WWW*, 2019, p. 151–161.

[14] P. Goyal, N. Kamra, X. He, and Y. Liu, "Dyngem: Deep embedding method for dynamic graphs," vol. abs/1805.11273, 2018.

[15] R. Anil, G. Pereyra, A. Passos, R. Ormándi, G. E. Dahl, and G. E. Hinton, "Large scale distributed neural network training through online distillation," in *ICLR*, 2018.

[16] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *NIPS*, 2014, pp. 2654–2662.

[17] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *ICLR*, 2018.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *ICML*, 2017, pp. 1243–1252.

[20] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *ICLR*, 2020.

[21] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *KDD*, 2016, pp. 855–864.

[22] L. Qu, H. Zhu, Q. Duan, and Y. Shi, "Continuous-time link prediction via temporal dependent graph neural network," in *WWW*, 2020, p. 3026–3032.

[23] J. Ma and Q. Mei, "Graph representation learning via multi-task knowledge distillation," in *NeurIPS*, 2019.

[24] C. Li, X. Guo, and Q. Mei, "Deepgraph: Graph structure predicts network growth," 2016.

[25] S. Antaris, D. Rafailidis, and S. Girdzijauskas, "EGAD: Evolving graph representation learning with self-attention and knowledge distillation for live video streaming events," in *IEEE Big Data*, 2020.

[26] "Supplementary Material," https://github.com/stefanosantaris/Distill2Vec/blob/master/supplementary/supplementary.pdf, 2020, [Online; accessed 24-October-2020].

[27] P. Micaelli and A. J. Storkey, "Zero-shot knowledge transfer via adversarial belief matching," in *NeurIPS*, 2019, pp. 9547–9557.