Independent Study Project Proposal

Ashutosh Tiwari (ashutiwa@iu.edu)

August 17, 2022

1 Project

1.1 Debiased Walk: Learning Representations using Debiased Embeddings

2 Number of work hours

12 hours / week (equivalent to 3 credit hours)

3 Required meetings

2-3 meetings / week

4 Required Readings and Assignments with due dates

4.1 Readings

- [Khajehnejad et al., 2021] (1^{st} week)
- [Rahman et al., 2019] (2^{nd} week)
- [Laclau et al., 2021] (3^{rd} week)
- [Gonen and Goldberg, 2019](4^{th} week)
- [Bolukbasi et al., 2016](5^{th} week)
- [Perozzi et al., 2014](6^{th} week)
- [Ravfogel et al., 2020](7^{th} week)
- [Garg et al., 2018](8^{th} week)
- [Kojaku et al., 2021](9^{th} week)
- [Brunet et al., 2019](10^{th} week)
- [Kenna, 2021](11^{th} week)

4.2 Assignments

- Investigate and possibly publish work on comparison of different methods to debias graph embeddings with residual2vec. Possibly write utilities which facilitate these experiments by allowing plug and play of models, techniques, datasets and configurations. (first half)
- Investigate and possibly publish work on a framework to analyze the bias manifold structure of different kinds of biases in different datasets using graph embeddings generated by different models.(second half)

5 Assessment

Student is going to be assessed on the amount of understanding he gains and the development work he undertakes during this study. This study banks upon the ability of student to be curious about a rather fundamental and at the same time under investigated topic in machine learning.

6 Work Plan

Primary purpose of this study is to understand and contribute to the understanding the manifold of existing classes of biases. Historically it is assumed that most of those are linear in nature and hence most of metrics used to quantify those are linear in nature. But there is sufficient evidence to doubt that size of embeddings, type of dataset do play a significant role in determining same.

Second part of this study explores different methods of tackling these biases. There already exist different methods (some included in reading list 4.1) some of which work on embeddings while some on data itself. Study also focuses on comparing all these methods. In particular we are interested use of residual2vec as such a method.

References

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Brunet et al., 2019] Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., and Zemel, R. (2019). Understanding the Origins of Bias in Word Embeddings. Technical Report arXiv:1810.03611, arXiv. arXiv:1810.03611 [cs, stat] type: article.

[Garg et al., 2018] Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16). arXiv:1711.08412 [cs].

[Gonen and Goldberg, 2019] Gonen, H. and Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. arXiv:1903.03862 [cs].

[Kenna, 2021] Kenna, D. (2021). Using Adversarial Debiasing to Remove Bias from Word Embeddings. arXiv:2107.10251 [cs].

[Khajehnejad et al., 2021] Khajehnejad, A., Khajehnejad, M., Babaei, M., Gummadi, K. P., Weller, A., and Mirzasoleiman, B. (2021). CrossWalk: Fairness-enhanced Node Representation Learning.

[Kojaku et al., 2021] Kojaku, S., Yoon, J., Constantino, I., and Ahn, Y.-Y. (2021). Residual2Vec: Debiasing graph embedding with random graphs. *undefined*.

[Laclau et al., 2021] Laclau, C., Redko, I., Choudhary, M., and Largeron, C. (2021). All of the Fairness for Edge Prediction with Optimal Transport. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1774–1782. PMLR. ISSN: 2640-3498.

[Perozzi et al., 2014] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. *CoRR*, abs/1403.6652.

[Rahman et al., 2019] Rahman, T., Surma, B., Backes, M., and Zhang, Y. (2019). Fairwalk: Towards Fair Graph Embedding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3289–3295, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

[Ravfogel et al., 2020] Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. arXiv:2004.07667 [cs].