Bachelor's Programme in Science and Technology

Normalizing Flows for Graph Generation

Literature review

Hau Phan

Aalto University

BACHELOR'S THESIS

Aalto University
BACHELOR'S THESIS 2022

Normalizing Flows for Graph Generation

Literature review

Hau Phan

Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Technology.
Otaniemi, June 21, 2022

Supervisor: Anirudh Jain
Advisor: Anirudh Jain

Aalto University
School of Science
Bachelor's Programme in Science and Technology

Author
Hau Phan

Title
Normalizing Flows for Graph Generation

School School of Science

Degree programme Bachelor's Programme in Science and Technology

Major Data Science **Code** IL3011

Supervisor Anirudh Jain

Advisor Anirudh Jain

Level Bachelor's thesis **Date** June 21, 2022 **Pages** 12 **Language** English

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Keywords graph neural network, normalizing flows, continuous normalizing flows, graphs generation, machine learning, geometric machine learning

urn https://aaltodoc.aalto.fi

Contents

Introduction

0.1 Motivation

Deep learning has exploded in popularity in recent years due to its effectiveness in capturing hidden patterns of Euclidean data such as images, audio and natural languages. However, not all data can be represented in Euclidean space [Bronstein et al., 2017]. There is an increasing number of application for data that are represented as discrete graphs: sets of nodes and edges that can directly model abstract relations between objects. For example, molecules are essentially connected atoms, bounded to one another by chemical bonds; users-products graphs are bipartite graphs that imply preferences of a customer towards specific sets of products; citation networks represents citationships between academic papers and inter/cross connections between domains; road networks and traffics flows can also be considered as graphs with additional attributes [Wu et al., 2021]. Graphs are common since they model relationships, which are abstract and ubiquitous in nature. However, graphs are also complex mathematical object: they can be irregular, have arbitrary numbers of nodes and vertices, each node may have any number of neighbors; they are also invariant with respect to rotation and translation, since there are no markers of direction or origin with graphs [Bronstein et al., 2017]. This is in direct contrast to Euclidean data such as images, where the notion of directions and rotations are straightforward to define. The complexity that comes along the expressive power of graphs has imposed significant challenges to existing machine learning methods such as convolution, which are easy to compute in the image domain but difficult in graph domain [Wu et al., 2021]. Graph Neural Networks (GNNs) [Scarselli et al., 2009] are the first architecture to directly operate on graphs, bridging the gap between discrete space and Euclidean space, yet permeate the flexibility of composite neural network e.g. stacking and residual connections. GNN and its derivatives had achieved

significant success in various supervised and unsupervised learning tasks with graphs [Kipf and Welling, 2017, Veličković et al., 2018].

An interesting subdomain of graph machine learning are generative model for graphs. Generative modeling in particular has been popularized to various modalities since the introduction of Generative Adversarial Network (GANs) [**?**] and Variational Autoencoder (VAEs) [Kingma and Welling, 2014], where both had shown remarkable success in capturing the data distribution of images and even audio [Hershey et al., 2017]. For graphs generation in particular, the majority of research efforts are focused on solving the molecular graph generation problem, which have high practical value in drug discovery [Wu et al., 2021]. Specifically, there are two primary way for generating new molecular graphs that are currently explored: sequentially and globally. Sequential approaches generate graphs by proposing nodes and edges step by step, while global approaches output a graph all at once. Pioneers in sequential generation of graphs include graph generation using SMILE representation of graphs [Dai et al., 2018, Kusner et al., 2017, Gómez-Bombarelli et al., 2018], DeepGMG [Li et al., 2018] and GraphRNN [You et al., 2018]. For global approaches, existing neural network architecture such as GANs and VAEs are adapted to graph domain by replacing CNN layers with GNN layers, albeit, with certain differences in architectural choices. Notable works in this field include MolGAN [De Cao and Kipf, 2018], GraphVAE [Simonovsky and Komodakis, 2018] and NetGAN [Bojchevski et al., 2018].

An exciting path of research in generative modeling is to adapt *normalizing flows models* in discrete domain, which has experienced a surge in popularity due to the expressive power of discrete objects such as graphs and sets. Normalizing flows models aim to learn a continuous invertible and deterministic mapping between the data distribution and a latent distribution, which is usually a Gaussian that is easy to manipulate and sample from. Sampling can easily be done by sampling from the latent distribution and run through the reverse mapping to retrieve a sample from the data distribution. The mappings here are simply continuous invertible functions or a composition of them. Designing these functions is the main area of focus when improving the generalization power of these model [Kobyzev et al., 2021]. Notable works in apply normalizing flows to graph are GraphNVP [Madhawa et al., 2019], GraphAF [Shi et al., 2020] and MoFlow [Zang and Wang, 2020]. A recent development in normalizing flows model are *continuous flows*, where rather than a discrete sequence of functions that are composed together, the invertible mapping is modeled as a solution to a differential equation. Continuous time-dependent mappings allow data distribution to smoothly "flow" toward the latent distribution, albeit, require slightly more

computational power to approximate the solution. One of the most notable works in the domain of continuous flow is FFJORD [Grathwohl et al., 2018]. Application of continuous flows, however, is a largely unexplored domain.

0.2 Thesis outline

This thesis aim to provide a systematic review of normalizing flows models for graph generation, with a focus on graph neural networks as the primary computational modules. The thesis is structured with concept dependencies in mind, where later topics were written with the assumption that previous concepts had already been read. Therefore, it is recommended that the thesis should be read linearly, however, it is not a mandatory requirement for a pleasant read.

The thesis contains seven chapters, starting with Chapter 1, where a general overview of notations and related works on graph generation was given. Chapter 2 primary focus on graphs neural networks, explored under a similar categorization framework as introduced by [Zhou et al., 2020], however are compressed and filtered in terms of relevance to normalizing flows. Chapter 3 considers a brief walkthrough of the normalizing flows framework, including motivation, formulation, categorization of different flows and a comparison with others generative modeling architectures such as GANs [**?**] and VAEs [Kingma and Welling, 2014]. Chapter 4 consists of different application of normalizing flows on graphs data, with a particular focus on molecules generation due to its high practical values and conveniently being extensively studied in the literature. As for closing remarks, Chapter 5 give an open discussion on graph generative models along with personal thoughts on future works and research direction. Chapter 6 provide a final conclusion and closing remarks on the thesis.

1. Background

1.1 Notations

1.2 Related works

2. Graph Neural Network

Graph neural networks (GNNs) in their most general definition can be considered as deep learning based methods that directly operates on graphs. There have been multiple surveys published in effort of standardizing and categorizing graph neural networks: [Zhang et al., 2020] [Zhang et al., 2020] [Chami et al., 2022] are notable mentions. However, GNN is a field of active research and its taxonomy can change drastically in the future. For convenience and consistency, this thesis adopts the categorization framework by [Zhang et al., 2020] where graph neural networks are perceived through the lens of deep learning architectural design. Under this view, graph neural networks pipeline are composed of primarily *computational modules* falling into three main categories: propagation modules, sampling modules and pooling modules. Propagation modules can be considered as the central information aggregators that through the process of propagating information between nodes, are capable of capturing both feature and topological information. There are two main operators used in propagation: convolution operators and recurrent operators, which can be generalized to an encompassing superset of convolutional graph neural network (ConvGNNs) and recurrent graph neural networks (RecGNNs) from the taxonomy proposed by [Wu et al., 2021]. The central idea of convolution operators is to generalize convolutions from the other domain to graph domain. Convolutions abstractly can be represented as layers with different parameters or weights that "convolve" with the information passed in as inputs and then output a latent representation. The main distinction between recurrent approaches and convolution approaches is the reusing of weights during information propagation. While convolution use separate weights for each operating layers, recurrent approaches recycle the layer and pass the output as the new input and hence reuse the weights. As the result, the notion of layer depth are mostly applied to convolution methods and not recurrent ones.

In an effort of scoping the field of GNN for application in normalizing flows models, only convolutional graph neural networks are explored in detail in this

chapter. There are two main methods for convolution operators: spectral methods and spatial methods.

2.1 Spectral methods

2.2 Spatial methods

2.2.1 Basic approaches

2.2.2 Attention mechanism

2.2.3 Framework

3. Normalizing Flows

3.1 Background

3.2 Linear Flows

3.3 Planar and Radial Flows

3.4 Coupling Flows

3.5 Autoregressive Flows

3.6 Residual Flows

3.7 Continuous Flows

4. Normalizing Flows for Graph Generation

4.1 GraphNVP

4.2 GraphAF

4.3 MoFlow

4.4 GraphDF

5. Discussion

5.1 Future work

6. Conclusion

Bibliography

[Bojchevski et al., 2018] Bojchevski, A., Shchur, O., Zügner, D., and Günnemann, S. (2018). NetGAN: Generating Graphs via Random Walks.

[Bronstein et al., 2017] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

[Chami et al., 2022] Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C., and Murphy, K. (2022). Machine Learning on Graphs: A Model and Comprehensive Taxonomy.

[Dai et al., 2018] Dai, H., Tian, Y., Dai, B., Skiena, S., and Song, L. (2018). Syntax-directed variational autoencoder for molecule generation. In *Proceedings of the International Conference on Learning Representations*.

[De Cao and Kipf, 2018] De Cao, N. and Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs.

[Gómez-Bombarelli et al., 2018] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276.

[Grathwohl et al., 2018] Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. (2018). FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models.

[Hershey et al., 2017] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

[Kingma and Welling, 2014] Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes.

[Kipf and Welling, 2017] Kipf, T. N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks.

[Kobyzev et al., 2021] Kobyzev, I., Prince, S. J. D., and Brubaker, M. A. (2021). Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979.

[Kusner et al., 2017] Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. (2017). Grammar Variational Autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1945–1954. PMLR.

[Li et al., 2018] Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. (2018). Learning Deep Generative Models of Graphs.

[Madhawa et al., 2019] Madhawa, K., Ishiguro, K., Nakago, K., and Abe, M. (2019). GraphNVP: An Invertible Flow Model for Generating Molecular Graphs.

[Scarselli et al., 2009] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

[Shi et al., 2020] Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., and Tang, J. (2020). GraphAF: A Flow-based Autoregressive Model for Molecular Graph Generation.

[Simonovsky and Komodakis, 2018] Simonovsky, M. and Komodakis, N. (2018). GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. In Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., and Maglogiannis, I., editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, Lecture Notes in Computer Science, pages 412–422, Cham. Springer International Publishing.

[Veličković et al., 2018] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph Attention Networks.

[Wu et al., 2021] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

[You et al., 2018] You, J., Ying, R., Ren, X., Hamilton, W., and Leskovec, J. (2018). GraphRNN: Generating Realistic Graphs with Deep Auto-regressive Models. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5708–5717. PMLR.

[Zang and Wang, 2020] Zang, C. and Wang, F. (2020). MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 617–626.

[Zhang et al., 2020] Zhang, Z., Cui, P., and Zhu, W. (2020). Deep Learning on Graphs: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270.

[Zhou et al., 2020] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.