Research notes on metrics for GNNs applied to biological problems

Philip Hartout

Monday January 24, 2022

1 Graph Neural Networks

1.1 Reviews

1.1.1 Graph neural networks:

A review of methods and applications Zhou et al mention several GNN approaches in use today Generative models popular today:

Sequential graph generation process

- GraphRNN - generates the adjacency matrix of a graph by generating the adjacency vector of each node step by step, with graph outputs with different number of nodes.
- Li 2018 - also generates nodes and edges sequentially uses the hidden state to decide what to do at the next step
- GraphAF - also a sequential process, Conducts a validity check of each molecule generated at each step to see if it's valid.

Non-sequential graph generation process

- MolGAN - to generate small molecules. Uses a permutation-invariant to solve the node adjacency matrix at once. Also implements an RL-based optimization toward desired chemical properties
- Ma et al 2018 - constrained VAE for semantic validity of generated graph
- GCPN similar to MolGAN, uses RL based methods to ensure validity of domain-specific rules Example work showing EMD kernel:
- Graph Normalizing Flows

This one has a fairly comprehensive website: `https://sites.google.com/view/graph-normalizing-f`
Full architecture

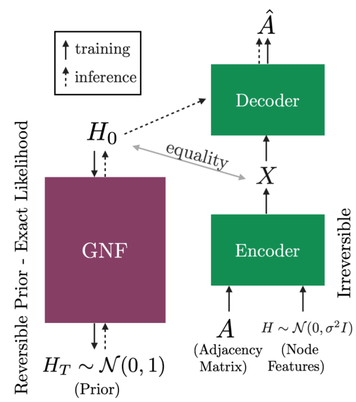

Figure 1: figure name

- Graphite isotropic gaussian for VAE + iterative refinement for decoding

1.2 Three most popular according to O'Bray 2021:

- GraphRNN, GRAN, Graph Score Matching.
- Graph Recurrent Attention Networks

> In previous work, You et al. [37] computed degree distributions, clustering coefficient distributions, and the number of occurrence of all orbits with 4 nodes, and then used the maximum mean discrepancy (MMD) over these graph statistics, relying on Gaussian kernels with the first Wasserstein distance, i.e., earth mover's distance (EMD), in the MMD.In practice, we found computing this MMD with the Gaussian EMD kernel to be very slow for moderately large graphs. Therefore, we use the total variation (TV) distance, which greatly speeds up the evaluation and is still consistent with EMD. In addition to the node degree, clustering coefficient and orbit counts (used by [36]), we also compare the spectra of the graphs by computing the eigenvalues of the normalized graph Laplacian (quantized to approximate a probability density). This spectral comparison provides a view of the global graph properties, whereas the previous metrics focus on local graph statistics.

- Graph Score Matching On MMD, they say the following:

2 Generative modelling metrics

a common practice of applying Langevin dynamics. We chose the value of the hyper-parameters based on the MMD metrics on the validation set, which contains 32 samples from the training set.

$$\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \epsilon_s\sqrt{\alpha_i}\ \mathbf{z}_t$$

Figure 2: MMD optimization strategy

2.1 Objective:

2.1.1 Generative graph dist close to the input graph dist

2.1.2 (pseudo)-metric to assess dissimilarity between G (generated graphs) and G* (input graphs)

2.2 On images

2.2.1 Frechet Inception Distance

The idea here is to use deeper representational layers of an ANN and used the squared Wasserstein metric to compare two multinomial Gaussians. Introduced 2017

2.2.2 LPIPS Project page

Introduced 2017

2.2.3 Why comparing graphs is hard:

- Metrics need to deal with spatial invariances such as cycles.
- Graph edit distance is NP-hard (Zeng 2009) and therefore does not satisfy efficiency criterion.
- Other publications:

2.3 Desiderata for good metrics:

1. Robust to noise
2. Expressive, if they don't arise from the same dist, then metric should detect this.
3. Computationally efficient.

3 MMD - current accepted method to evaluate generative GNNs

- The MMD formula goes as follows:

$\text{MMD}(X,Y) := \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i,x_j) + \frac{1}{m^2}\sum_{i,j=1}^{n} k(y_i,y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(y_i,y_j)$

- use it for hypothesis/two-sample testing.
- In practice, we evaluate $d_{MMD}(\mathcal{G},\mathcal{G}*) := MMD(f(\mathcal{G}), f(\mathcal{G}*))$ for a distribution $\mathcal{G}$. Given multiple distributions $G_1, G_2, \ldots$, the values of d_{MMD} can be used to rank models, where smaller values are assumed to indicate a larger agreement with the original distribution $\mathcal{G}*$.
- Commonly used kernels: first Wasserstein distance, total variation distance, radial basis function.
- Commonly used descriptor functions: degree distribution histogram, clustering coefficient, Laplacian spectrum histogram.
- Recommended kernels: RBF, Laplacian kernel, linear kernel (expressivity & robustness need to be analyzed)

3.1 Potential pitfalls of descriptors

- Degree distributions are ok seemingly
- Clustering does not distinguish fully connected vs disconnected cliques
- Spectral methods are not clearly expressive. Does not seem to be for certain classes of graphs.
- Parameters and descriptors are set a priori in the best case
- Model performance is highly dependent on parameters and descriptor functions.

4 Research objectives

There are multiple objectives here:

1. Find optimal kernel/hyperparameter combination based on controlled experiments on a given dataset to evaluate a good MMD configuration.
    - For this we will need `https://www.alphafold.ebi.ac.uk/download`, because it's clean. Also filter single chain proteins to extract graphs in the first place.
    - This can be built as a first step to get the pipeline going.
2. Show which parameters influence evaluation and how?
    - Conduct perturbation experiments on graphs
3. Find novel domain-agnostic evaluation & domain-specific evaluation metrics

(a) Domain-agnostic evaluation measures

- Correlation with graph-edit distance
- Correlation with perturbation
- Topology/persistence based approaches could be useful for modelling features like binding pockets, etc?

(b) Domain-specific evaluation measures

- Alignment
- Energy?

4.1 From Tim: gather literature sources. Intro structure

4.1.1 Evaluation of generative models (different domains)

4.1.2 Evaluation of generative models for graphs

Check how it was done before, why combo of parameters/kernels were used.

4.1.3 Evaluation of proteins (.../molecules/drugs) (What makes a valid protein?)

4.1.4 Evaluation of generative models for proteins

5 Module-wise breakdown of the plan

- Graph extraction
- Descriptor functions
- kernels, MMD
- Domain agnostic
- Domain specific
- Other metrics
- TDA descriptors
- Labeled edge graph
- NSPDK
- Other metrics
- Extract graph from real datasets

6 Annotations

Online approach to k-NN graph construction epsilon nearest neighbor graphs dissertation with history and construction methods. Giotto-TDA library Heat kernel on persistence diagrams original mmd papers Specific MMD for biological data: Introduction several metrics based on the features extracted by an untrained random GNN, showing more expressive metrics of GNN performance Neighborhood Subgraph Pairwise Distance graph kernel (NSPDK) Examples using the Neighborhood Subgraph Pairwise Distance graph kernel (NSPDK) Inverse protein folding problem, maybe good to look at to see what makes a good protein. Papers discussing Delaunay graphs (dual graph of Voronoi diagram), obtained through Delaunay triangulation. Useful for extracting hierarchical structures. Graph edges can also be added on the basis of the Delaunay triangulation. Delaunay triangles correspond to joining points that share a face in the 3D Voronoi diagram of the protein structures. For distance-based edges, a Long Interaction Network (LIN) parameter controls the minimum required separation in the amino acid sequence for edge creation. This can be useful in reducing the number of noisy edges under distance-based edge creation schemes. Edge featurisation for atom-level graphs is provided by annotations of bond type and ring status. Not a great review, but still useful source of references and gives overview of the field.

References

References

[1] David C Anastasiu. "Algorithms for Constructing Exact Nearest Neighbor Graphs". PhD thesis. University of Minnesota, 2016.

[2] Karsten M Borgwardt et al. "Integrating structured biological data by kernel maximum mean discrepancy". In: *Bioinformatics* 22.14 (2006), e49–e57.

[3] Fabrizio Costa and Kurt De Grave. "Fast neighborhood subgraph pairwise distance kernel". In: *ICML*. 2010.

[4] Nicola De Cao and Thomas Kipf. "MolGAN: An implicit generative model for small molecular graphs". In: *arXiv preprint arXiv:1805.11973* (2018).

[5] Romanos Fasoulis, Georgios Paliouras, and Lydia E Kavraki. "Graph representation learning for structural proteomics". In: *Emerging Topics in Life Sciences* 5.6 (2021), pp. 789–802.

[6] Nikhil Goyal, Harsh Vardhan Jain, and Sayan Ranu. "GraphGen: a scalable approach to domain-agnostic labeled graph generation". In: *Proceedings of The Web Conference 2020*. 2020, pp. 1253–1263.

[7] Arthur Gretton et al. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[8] Aditya Grover, Aaron Zweig, and Stefano Ermon. "Graphite: Iterative generative modeling of graphs". In: *International conference on machine learning*. PMLR. 2019, pp. 2434–2444.
[9] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017).
[10] John Ingraham et al. "Generative models for graph-based protein design". In: (2019).
[11] Wataru Kawai, Yusuke Mukuta, and Tatsuya Harada. "Scalable Generative Models for Graphs with Graph Attention Mechanism". In: *arXiv preprint arXiv:1906.01861* (2019).
[12] Yujia Li et al. "Learning deep generative models of graphs". In: *arXiv preprint arXiv:1803.03324* (2018).
[13] Renjie Liao et al. "Efficient graph generation with graph recurrent attention networks". In: *arXiv preprint arXiv:1910.00760* (2019).
[14] Jenny Liu et al. "Graph normalizing flows". In: *arXiv preprint arXiv:1905.13177* (2019).
[15] Tengfei Ma, Jie Chen, and Cao Xiao. "Constrained generation of semantically valid graphs via regularizing variational autoencoders". In: *arXiv preprint arXiv:1809.02630* (2018).
[16] Chenhao Niu et al. "Permutation invariant graph generation via score-based generative modeling". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 4474–4484.
[17] Leslie O'Bray et al. "Evaluation Metrics for Graph Generative Models: Problems, Pitfalls, and Practical Solutions". In: *arXiv preprint arXiv:2106.01098* (2021).
[18] Marco Podda and Davide Bacciu. "GraphGen-Redux: a Fast and Lightweight Recurrent Model for labeled Graph Generation". In: *arXiv preprint arXiv:2107.08396* (2021).
[19] Jan Reininghaus et al. "A stable multi-scale kernel for topological machine learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4741–4748.
[20] Chence Shi et al. "GraphAF: a flow-based autoregressive model for molecular graph generation". In: *arXiv preprint arXiv:2001.09382* (2020).
[21] Guillaume Tauzin et al. "giotto-tda:: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration." In: *J. Mach. Learn. Res.* 22 (2021), pp. 39–1.
[22] Todd J Taylor and Iosif I Vaisman. "Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures". In: *Physical Review E* 73.4 (2006), p. 041925.
[23] Lucas Theis, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models". In: *arXiv preprint arXiv:1511.01844* (2015).
[24] Rylee Thompson et al. "On Evaluation Metrics for Graph Generative Models". In: *arXiv preprint arXiv:2201.09871* (2022).

[25] Jiaxuan You et al. "Graph convolutional policy network for goal-directed molecular graph generation". In: *arXiv preprint arXiv:1806.02473* (2018).

[26] Jiaxuan You et al. "GraphRNN: Generating realistic graphs with deep auto-regressive models". In: *International conference on machine learning*. PMLR. 2018, pp. 5708–5717.

[27] Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[28] Wan-Lei Zhao, Hui Wang, and Chong-Wah Ngo. "Approximate k-NN graph construction: a generic online approach". In: *IEEE Transactions on Multimedia* (2021).

[29] Jie Zhou et al. "Graph neural networks: A review of methods and applications". In: *AI Open* 1 (2020), pp. 57–81.