Geometric Scattering for Graph Data Analysis

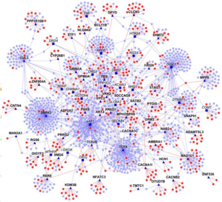
Feng Gao[1], Guy Wolf[2], Matthew Hirn[1]
[1] Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA
[2] Department of Mathematics and Statistics, Université de Montréal, Montreal, QC, Canada

ICML, Long Beach, June 13, 2019

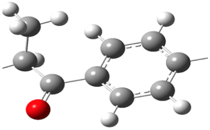
Graphs

- Many data can be modelled as graphs, e.g. social networks, protein-protein interaction networks and molecules.

Social Network

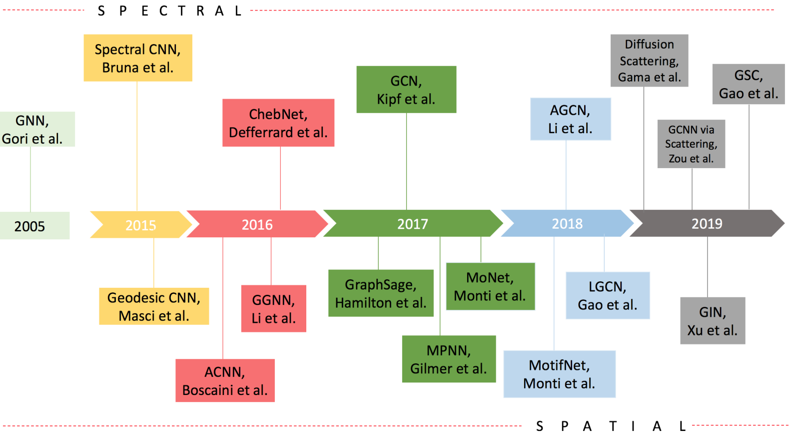


Protein-protein Interaction Network



Molecular Graph

Brief Review of Graph Convolutional Networks



Can we build GCN in an unsupervised way?

Euclidean Scattering Transform

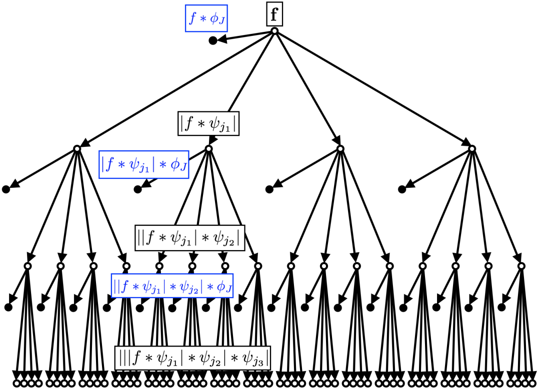


Figure: Illustration of scattering transform for feature extraction

Graph Wavelets

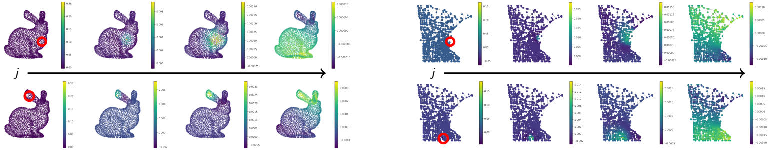
- Graph Wavelet: defined as the difference between lazy random walks at different time scales:

$$\Psi_j = \mathbf{P}^{2^{j-1}} - \mathbf{P}^{2^j} = \mathbf{P}^{2^{j-1}}(\mathbf{I} - \mathbf{P}^{2^{j-1}}) \, .$$

- Graph wavelet transform up to the scale 2^J:

$$W_J f = \{P^{2^J} f \, , \, \Psi_j f : j \leq J\} = \{f * \phi_J \, , \, f * \psi_j : j \leq J\} \, .$$

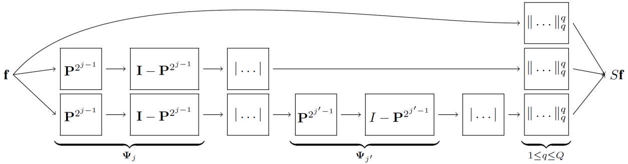
Graph Wavelet Transform



(a) Sample graph of bunny manifold (b) Minnesota road network graph

Figure: Wavelets Ψ_j for increasing scale 2^j left to right, applied to Diracs centered at two different locations (marked by red circles) in two graphs.

Geometric Scattering Transform



- Zero order feature:

$$S\mathbf{f}(q)=\sum_{\ell=1}^{n}\mathbf{f}(v_\ell)^q,\quad 1\leq q\leq Q$$

- First order feature:

$$S\mathbf{f}(j,q)=\sum_{\ell=1}^{n}|\Psi_j\mathbf{f}(v_\ell)|^q,\ 1\leq j\leq J,\ 1\leq q\leq Q$$

- Second order feature:

$$S\mathbf{f}(j,j',q)=\sum_{\ell=1}^{n}|\Psi_{j'}|\Psi_j\mathbf{f}(v_\ell)||^q,\ \begin{matrix}1\leq j<j'\leq J\\ 1\leq q\leq Q\end{matrix}$$

Graph Classification on Social Networks

	COLLAB	IMDB-B	IMDB-M	REDDIT-B	REDDIT-5K	REDDIT-12K
WL	77.82 ± 1.45	71.60 ± 5.16	N/A	78.52 ± 2.01	50.77 ± 2.02	34.57 ± 1.32
Graphlet	73.42 ± 2.43	65.40 ± 5.95	N/A	77.26 ± 2.34	39.75 ± 1.36	25.98 ± 1.29
WL-OA	80.70 ± 0.10	N/A	N/A	89.30 ± 0.30	N/A	N/A
DGK	73.00 ± 0.20	66.90 ± 0.50	44.50 ± 0.50	78.00 ± 0.30	41.20 ± 0.10	32.20 ± 0.10
DGCNN	73.76 ± 0.49	70.03 ± 0.86	47.83 ± 0.85	N/A	48.70 ± 4.54	N/A
2D CNN	71.33 ± 1.96	70.40 ± 3.85	N/A	89.12 ± 1.70	52.21 ± 2.44	48.13 ± 1.47
PSCN	72.60 ± 2.15	71.00 ± 2.29	45.23 ± 2.84	86.30 ± 1.58	49.10 ± 0.70	41.32 ± 0.42
GCAPS-CNN	77.71 ± 2.51	71.69 ± 3.40	48.50 ± 4.10	87.61 ± 2.51	50.10 ± 1.72	N/A
S2S-P2P-NN	81.75 ± 0.80	73.80 ± 0.70	51.19 ± 0.50	86.50 ± 0.80	52.28 ± 0.50	42.47 ± 0.10
GIN-0 (MLP-SUM)	80.20 ± 1.90	75.10 ± 5.10	52.30 ± 2.80	92.40 ± 2.50	57.50 ± 1.50	N/A
GS-SVM	79.94 ± 1.61	71.20 ± 3.25	48.73 ± 2.32	89.65 ± 1.94	53.33 ± 1.37	45.23 ± 1.25

Table: Comparison of the proposed GS-SVM classifier with leading deep learning methods on social graph datasets.

Classification with Low Training-data Availability

Graph classification with four training/validation/test splits:

- 80%/10%/10%
- 70%/10%/20%
- 40%/10%/50%
- 20%/10%/70%

Training data reduced from 80% to 20% only results in a decrease of 3% in classification accuracy on social network datasets

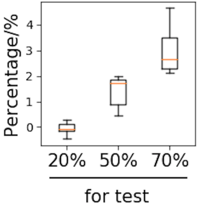


Figure: Drop in SVM classification accuracy over social graph datasets when reducing training set size

Dimensionality Reduction

ENZYME dataset: on average 124.2 edges, 29.8 vertices, and 3 features per vertex per graph

Geometric scattering combined with PCA enables significant dimensionality reduction with only a small impact on classification accuracy

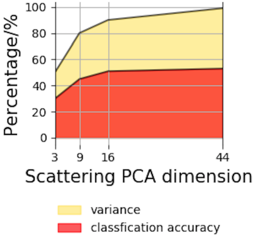
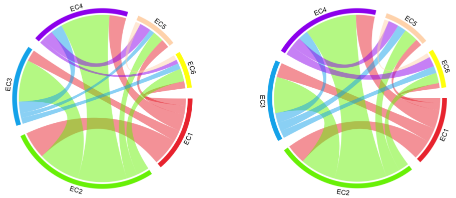


Figure: Relation between explained variance, SVM classification accuracy, and PCA dimensions over scattering features in ENZYMES dataset.

Data Exploration: Enzyme Class Exchange Preferences

- ENZYME dataset contains enzymes from six top level enzyme classes and are labelled by their Enzyme Commission (EC) numbers.
- Geometric scattering features are considered as signature vectors for individual enzymes, and can be used to infer EC exchange preferences during enzyme evolution.

Scattering features are sufficiently rich to capture relations between enzyme classes



(a) observed (b) inferred

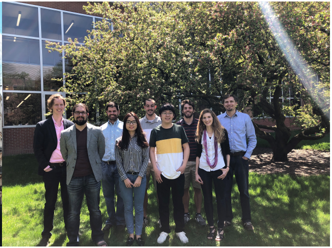
Figure: Comparison of EC exchange preferences in enzyme evolution: (a) observed in Cuesta et al. (2015), and (b) inferred from scattering features

Conclusion

- A generalization of Euclidean scattering transform to graph.
- Scattering features can serve as universal representations of graphs.
- Geometric scattering transform provides a new way for computing and considering global graph representations, independent of specific learning tasks.

Acknowledgement

- NIEHS grant P42 ES004911
- Alfred P. Sloan Fellowship (grant FG-2016-6607)
- DARPA YFA (grant D16AP00117)
- NSF grant 1620216



Guy Wolf CEDAR Team

Thank you!