Multiresolution Matrix Factorization and Wavelet Networks on Graphs

Truong Son Hy and Risi Kondor

Department of Computer Science, The University of Chicago

Abstract

Multiresolution Matrix Factorization (MMF) is unusual amongst fast matrix factorization algorithms in that it does not make a low rank assumption. This makes MMF especially well suited to modeling certain types of graphs with complex multiscale or hierarchical strucutre. While MMF promises to yield a useful wavelet basis, finding the factorization itself is hard, and existing greedy methods tend to be brittle. In this paper, we propose a "learnable" version of MMF that carfully optimizes the factorization with a combination of reinforcement learning and Stiefel manifold optimization through backpropagating errors. We show that the resulting wavelet basis far outperforms prior MMF algorithms and provides the first version of this type of factorization that can be robustly deployed on standard learning tasks. Furthermore, we construct the wavelet neural networks (WNNs) learning graphs on the spectral domain with the wavelet basis produced by our MMF learning algorithm. Our wavelet networks are competitive against other state-of-the-art methods in molecular graphs classification and node classification on citation graphs.

Source code: `https://github.com/risilab/Learnable_MMF`

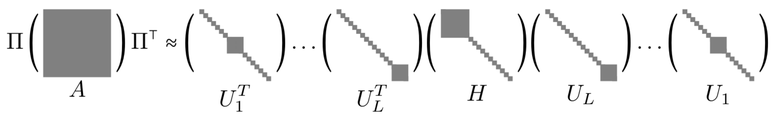
Multiresolution Matrix Factorization

MMF of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ (Kondor et al., 2014) is:

$$\boldsymbol{A} = \boldsymbol{U}_1^T \boldsymbol{U}_2^T \dots \boldsymbol{U}_L^T \boldsymbol{H} \boldsymbol{U}_L \dots \boldsymbol{U}_2 \boldsymbol{U}_1,$$

where:

- Each $\boldsymbol{U}_\ell$ is an orthogonal matrix that is a k-point rotation (small k),
- There is a nested sequence of sets $\mathbb{S}_L \subseteq \dots \subseteq \mathbb{S}_1 \subseteq \mathbb{S}_0 = [n]$ such that the coordinates rotated by $\boldsymbol{U}_\ell$ are a subset of $\mathbb{S}_\ell$,
- $\boldsymbol{H}$ is an $\mathbb{S}_L$-core-diagonal matrix meaning that is diagonal with a an additional small $\mathbb{S}_L \times \mathbb{S}_L$ dimensional "core".



Not based on the low-rank assumption

Stiefel Manifold Optimization

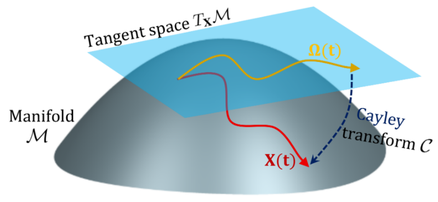
Finding the best MMF to a symmetric matrix $\boldsymbol{A}$ involves solving

$$\min_{\substack{\mathbb{S}_L \subseteq \dots \subseteq \mathbb{S}_1 \subseteq \mathbb{S}_0 = [n] \\ \boldsymbol{H} \in \mathbb{H}_n^{\mathbb{S}_L}; \boldsymbol{U}_1, \dots, \boldsymbol{U}_L \in \mathbb{O}}} \|\boldsymbol{A} - \boldsymbol{U}_1^T \dots \boldsymbol{U}_L^T \boldsymbol{H} \boldsymbol{U}_L \dots \boldsymbol{U}_1\|_{\mathcal{F}}.$$

It is equivalent to

$$\min_{\mathbb{S}_L \subseteq \dots \subseteq \mathbb{S}_1 \subseteq \mathbb{S}_0 = [n]} \left(\min_{\boldsymbol{U}_1, \dots, \boldsymbol{U}_L \in \mathbb{O}} \|\boldsymbol{U}_L \dots \boldsymbol{U}_1 \boldsymbol{A} \boldsymbol{U}_1^T \dots \boldsymbol{U}_L^T\|_{\text{resi}}^2 \right),$$

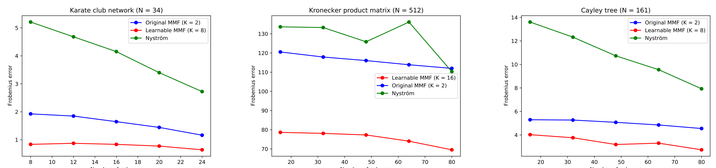
where $\|\cdot\|_{\text{resi}}^2$ is the squared residual norm $\|\boldsymbol{H}\|_{\text{resi}}^2 = \sum_{i \neq j; (i,j) \notin \mathbb{S}_L \times \mathbb{S}_L} |\boldsymbol{H}_{i,j}|^2$. The solution for $\boldsymbol{U}_\ell$ must satisfy the orthogonality constraint such that $\boldsymbol{U}_\ell^T \boldsymbol{U}_\ell = \boldsymbol{I}$. Given a fixed $\mathbb{S}_L \subseteq \dots \subseteq \mathbb{S}_1 \subseteq \mathbb{S}_0 = [n]$, we use gradient descent algorithm on the Stiefel manifold to optimize all rotations $\{\boldsymbol{U}_\ell\}_{\ell=1}^L$ simultaneously, to satisfy the orthogonality constraints.



Reinforcement Learning

We formulate the problem of finding the optimal nested sequence $\mathbb{S}_L \subseteq \dots \subseteq \mathbb{S}_1 \subseteq \mathbb{S}_0 = [n]$ as learning a Markov Decision Process (MDP) that can be subsequently solved by the gradient policy method of Reinforcement Learning (RL), in which the RL agent (or stochastic policy) is modeled by graph neural networks (GNNs).

Learnable MMF outperforms the original MMF and the Nyström method in matrix approximation:



Karate club network | Kronecker matrix | Cayley tree

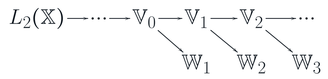
Multiresolution Analysis

The functional analytic view of wavelets is provided by Multiresolution Analysis (Mallat, 1989) is a way of filtering a function space into a sequence of subspaces

$$\dots \subset \mathbb{V}_{-1} \subset \mathbb{V}_0 \subset \mathbb{V}_1 \subset \mathbb{V}_2 \subset \dots$$

Iteratively, each $\mathbb{V}_\ell$ is splitted into the orthogonal sum $\mathbb{V}_\ell = \mathbb{V}_{\ell+1} \oplus \mathbb{W}_{\ell+1}$:

- **Approximation space:** The a smoother part $\mathbb{V}_{\ell+1}$. Each space $\mathbb{V}_\ell$ has an orthonormal basis $\Phi_\ell \triangleq \{\phi_m^\ell\}_m$ in which each ϕ is called a **father** wavelet.
- **Detail space:** The rougher part $\mathbb{W}_{\ell+1}$. Each space $\mathbb{W}_\ell$ is also spanned by an orthonormal basis $\Psi_\ell \triangleq \{\psi_m^\ell\}_m$ in which each ψ is called a **mother** wavelet.



Instead of diagonalizing $\boldsymbol{A}$ in a single step as in PCA, multiresolution analysis will involve a sequence of basis transforms $\boldsymbol{U}_1, \boldsymbol{U}_2, \dots, \boldsymbol{U}_L$, transforming $\boldsymbol{A}$ step by step as:

$$\boldsymbol{A} \to \boldsymbol{U}_1 \boldsymbol{A} \boldsymbol{U}_1^T \to \boldsymbol{U}_2 \boldsymbol{U}_1 \boldsymbol{A} \boldsymbol{U}_1^T \boldsymbol{U}_2^T \to \dots \to \boldsymbol{U}_L \dots \boldsymbol{U}_2 \boldsymbol{U}_1 \boldsymbol{A} \boldsymbol{U}_1^T \boldsymbol{U}_2^T \dots \boldsymbol{U}_L^T,$$

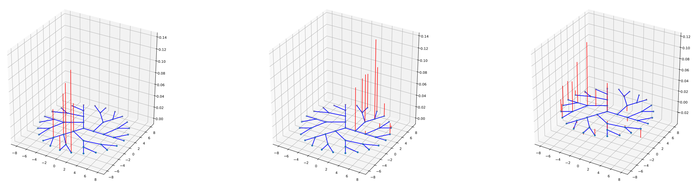
Each individual rotation $\boldsymbol{U}_\ell : \mathbb{V}_{\ell-1} \to \mathbb{V}_\ell \oplus \mathbb{W}_\ell$ is a sparse basis transform that expresses $\Phi_\ell \cup \Psi_\ell$ in the previous basis $\Phi_{\ell-1}$ such that:

$$\phi_m^\ell = \sum_{i=1}^{\dim(\mathbb{V}_{\ell-1})} [\boldsymbol{U}_\ell]_{m,i} \phi_i^{\ell-1}, \qquad \psi_m^\ell = \sum_{i=1}^{\dim(\mathbb{V}_{\ell-1})} [\boldsymbol{U}_\ell]_{m+\dim(\mathbb{V}_{\ell-1}),i} \phi_i^{\ell-1}.$$

In the case $\boldsymbol{A}$ is the normalized graph Laplacian of a graph $\mathcal{G} = (V, E)$, the wavelet transform (up to level L) expresses a graph signal (function over the vertex domain) $f : V \to \mathbb{R}$, without loss of generality $f \in \mathbb{V}_0$, as:

$$f(v) = \sum_{\ell=1}^{L} \sum_m \alpha_m^\ell \psi_m^\ell(v) + \sum_m \beta_m \phi_m^L(v), \qquad \text{for each } \; v \in V,$$

where $\alpha_m^\ell = \langle f, \psi_m^\ell \rangle$ and $\beta_m = \langle f, \phi_m^L \rangle$ are the wavelet coefficients.



$\ell = 1$ | $\ell = 20$ | $\ell = 39$

The low index wavelets (low ℓ) are highly localized, whereas the high index ones are smoother and spread out over large parts of the Cayley tree/graph of 46 nodes.

Wavelet Neural Networks

Analogous to the convolution based on Graph Fourier Transform (GFT) (Bruna et al., 2014), each convolution layer $k = 1, .., K$ of our wavelet network transforms an input vector $\boldsymbol{f}^{(k-1)}$ of size $|V| \times F_{k-1}$ into an output $\boldsymbol{f}^{(k)}$ of size $|V| \times F_k$ as

$$\boldsymbol{f}_{:,j}^{(k)} = \sigma\left(\boldsymbol{W} \sum_{i=1}^{F_{k-1}} \boldsymbol{g}_{i,j}^{(k)} \boldsymbol{W}^T \boldsymbol{f}_{:,i}^{(k-1)} \right) \quad \text{for} \quad j = 1, \dots, F_k,$$

where $\boldsymbol{W} = [\overline{\phi}, \overline{\psi}]$ is our wavelet basis matrix of a total of N wavelets returned from our learnable MMF:

- L mother wavelets $\overline{\psi} = \{\psi^1, .., \psi^L\}$,
- $N - L$ father wavelets $\overline{\phi} = \{\phi_m^L = \boldsymbol{H}_{m,:}\}_{m \in \mathbb{S}_L}$;

and $\boldsymbol{g}_{i,j}^{(k)}$ is a parameter/filter in the form of a diagonal matrix, and σ is an element-wise linearity. Since the wavelet basis is **sparse**, the wavelet transform can be implemented efficiently by sparse matrix multiplication.

Experiments

Our WNNs outperform 7/8, 7/8, 8/8, and 2/8 competing methods on molecular graph classification datasets, respectively. Average percentages of non-zero elements of the wavelet basis: 19.23% (MUTAG), 18.18% (PTC), 2.26% (PROTEINS), 11.43% (NCI1). In constrast, the graph Fourier basis is completely dense.

Method	MUTAG	PTC	PROTEINS	NCI1
DGCNN (Zhang et al., 2018)	85.83 ± 1.7	58.59 ± 2.5	75.54 ± 0.9	74.44 ± 0.5
PSCN (Niepert et al., 2016)	88.95 ± 4.4	62.29 ± 5.7	75 ± 2.5	76.34 ± 1.7
DCNN (Atwood and Towsley, 2016)	N/A	N/A	61.29 ± 1.6	56.61 ± 1.0
CCN (Hy et al., 2018)	**91.64 ± 7.2**	**70.62 ± 7.0**	N/A	76.27 ± 4.1
GK (Shervashidze et al., 2009)	81.39 ± 1.7	55.65 ± 0.5	71.39 ± 0.3	62.49 ± 0.3
RW (Vishwanathan et al., 2010)	79.17 ± 2.1	55.91 ± 0.3	59.57 ± 0.1	N/A
PK (Neumann et al., 2016)	76 ± 2.7	59.5 ± 2.4	73.68 ± 0.7	82.54 ± 0.5
WL (Shervashidze et al., 2011)	84.11 ± 1.9	57.97 ± 2.5	74.68 ± 0.5	**84.46 ± 0.5**
IEGN (Maron et al., 2019)	84.61 ± 10	59.47 ± 7.3	75.19 ± 4.3	73.71 ± 2.6
MMF	86.31 ± 9.47	67.99 ± 8.55	**78.72 ± 2.53**	71.04 ± 1.53

Our WNNs outperforms other SOTAs in the task of node classification on citation networks (i.e. Cora & Citeseer). The splits are (1) 20%/20%/60%, (2) 40%/20%/40% and (3) 60%/20%/20%. The sparsity is 4.69% on Cora and 15.25% on Citeseer.

Method	Cora	Citeseer
MLP	55.1%	46.5%
ManiReg (Belkin et al., 2006)	59.5%	60.1%
SemiEmb (Weston et al., 2008)	59.0%	59.6%
LP (Zhu et al., 2003)	68.0%	45.3%
DeepWalk (Perozzi et al., 2014)	67.2%	43.2%
ICA (Getoor, 2005)	75.1%	69.1%
Planetoid (Yang et al., 2016)	75.7%	64.7%
Spectral CNN (Bruna et al., 2014)	73.3%	58.9%
ChebyNet (Defferrard et al., 2016)	81.2%	69.8%
GCN (Kipf & Welling, 2017)	81.5%	70.3%
MoNet (Monti et al., 2017)	81.7%	N/A
GWNN (Xu et al., 2019)	82.8%	71.7%
MMF$_1$	**84.35%**	68.07%
MMF$_2$	**84.55%**	**72.76%**
MMF$_3$	**87.59%**	**72.90%**

Reference

Kondor, R., Teneva, N., and Garg, V. Multiresolution matrix factorization. Proceedings of the 31st International Conference on Machine Learning, volume 32, pp. 1620–1628.

Mallat, S. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7):674–693, 1989.