Interpretable Neuron Structuring with Graph Spectral Regularization

Alexander Tong, David van Dijk, Jay S. Stanley, Matthew Amodio, Kristina Yim, Rebecca Muhle, James Noonan, Guy Wolf, and Smita Krishnaswamy
Convolutioanal NN filter interpretability

- Filter maps
- Activation maps
- Gradient based methods [Olah et al. 2017]
- Up-convolutional net [Dosovitskiy and Brox 2016]
Can we make interpretable activation maps for fully-connected NNs?
Analogy to real neural networks

• Often preprocessed into “functional regions”
• X condition has activation / suppression in Y region
• We can gain a high-level understanding of real brains by summarizing $10^{11}$ neurons into localized groups
Organizing layers with graph structure

Enforcing graph structure
  • Take a predefined graph and force activations to be smooth on that graph

Learning graph structure
  • Simultaneously optimize the graph structure and activation smoothness
Enforcing a Grid Structure on MNIST

- MNIST classification with dense encoder
- 64 width layer enforcing an 8x8 grid structure
- Two methods
  - Convolutional classifier
  - Graph smoothing

\[ \text{Loss}(z, L) = z^T L z \]
\[ L = D - A \]
### Activation Maps for MNIST

| No Convolution | Convolution + Classification |
|----------------|-----------------------------|
|                |                             |
| **No Smoothing** |                             |
| 0              | 0                           |
| 1              | 1                           |
| 2              | 2                           |
| 3              | 3                           |
| 4              | 4                           |
| 5              | 5                           |
| 6              | 6                           |
| 7              | 7                           |
| 8              | 8                           |
| 9              | 9                           |
| **Laplacian Smoothing** |                             |
| 0              | 0                           |
| 1              | 1                           |
| 2              | 2                           |
| 3              | 3                           |
| 4              | 4                           |
| 5              | 5                           |
| 6              | 6                           |
| 7              | 7                           |
| 8              | 8                           |
| 9              | 9                           |
Convolution + Graph regularization

Segmentation

| Label, Prediction | (9,9) | (9,9) | (9,7) | (3,3) | (3,3) | (3,7) |
|-------------------|-------|-------|-------|-------|-------|-------|
| **Input**         | ![Image](input_1.png) | ![Image](input_2.png) | ![Image](input_3.png) | ![Image](input_4.png) | ![Image](input_5.png) | ![Image](input_6.png) |
| **Embedding**     | ![Image](embedding_1.png) | ![Image](embedding_2.png) | ![Image](embedding_3.png) | ![Image](embedding_4.png) | ![Image](embedding_5.png) | ![Image](embedding_6.png) |
Learning a Graph Structure

Repeatedly do:
- Create graph from gaussian kernel on activations
- Train for M steps with GSR loss

\[ K(z_i, z_j) = \exp(-\frac{||z_i - z_j||^2}{2\sigma_{ij}}) \]

\[ L = D - A \]

\[ \text{Loss}(z, L) = z^T L z \]
Learning the graph in a single-cell (cell X gene) dataset

a) Training Time

b) Extracted Graph Structure of Genes

c) Developing T-cells

setty et al. 2016

d) Visualization of cells

Setty et al. 2016
Summary

- Fully connected layers have no natural coherent structure.
- Imposing a graph structure can create locality like a brain.
- Graph structure can be learned from the data.
Acknowledgements

Colleagues
• Krishnaswamy Lab
• Noonan Lab

Funding
• IVADO
• Chan-Zuckerberg Initiative
• NIH

Lab Website: www.krishnaswamylab.org
Code: https://github.com/KrishnaswamyLab/GraphSpectralRegularization