What Do Neural Networks Learn? A Mathematical Comparison of Convolution Kernels and Image Processing Features

Jonas A. Actor, [1] Béatrice Rivière, [1] and David Fuentes [2]

[1]Computational and Applied Mathematics, Rice University
[2]Imaging Physics, University of Texas MD Anderson Cancer Center

Goal : Compare CNN kernels vs. imaging features

In the last decade, the traditional image segmentation methods have been replaced by techniques using deep convolutional neural networks (CNN). These CNNs are treated as 'black boxes', which limits clinical interpretation of what image features a CNN uses in its decision-making. However, convolution-based image processing features are already used by clinicians for manual image segmentation.

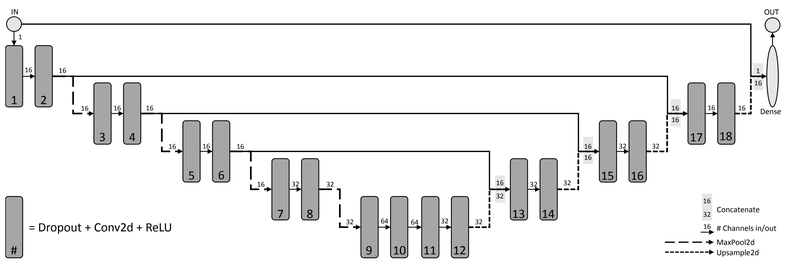
How similar are CNN convolution kernels to these known clinical image processing features?

Clincal image processing kernels

- Laplacian
- Edge detection
- Sharpen
- Identity
- Gaussian blur
- Local mean

Learned convolution kernel

$$K = \begin{bmatrix} k_{-1,-1} & k_{-1,0} & k_{-1,1} \\ k_{0,-1} & k_{0,0} & k_{0,1} \\ k_{1,-1} & k_{1,0} & k_{1,1} \end{bmatrix}$$



Structure of our sample UNet.

Conclusions

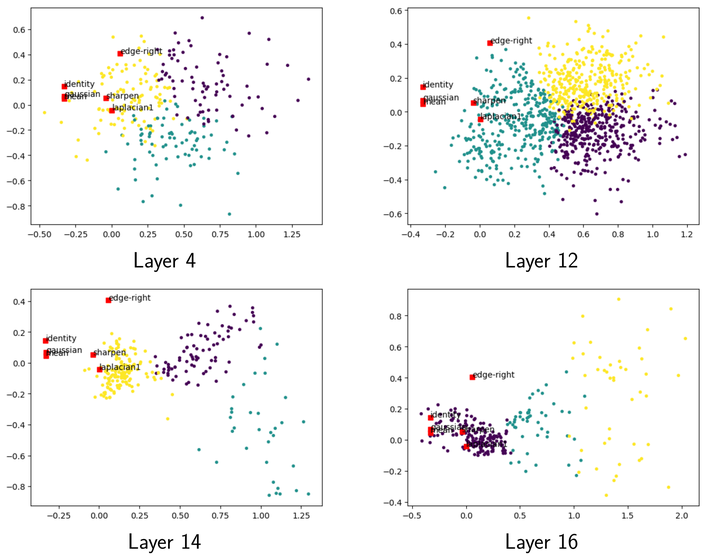
- Kernels roughly cluster around known clinical imaging features
- Clustering patterns emerge in decoder half of UNet
- Distribution of kernels closest to each imaging feature remains consistent across encoder half of UNet

JAA is supported by the NLM Training Program in Biomedical Informatics & Data Science (T15LM007093), supplemented by the Ken Kennedy Institute Computer Science & Engineering Enhancement Fellowship, funded by the Rice Oil & Gas HPC Conference.

1. Find image processing kernel with closest values

1. Flatten each 3×3 kernel into a vector in $\mathbb{R}^9$
2. Assemble data matrix from flattened kernels
3. Cluster 9-dimensional kernel data by k-means
4. Visualize via PCA
5. Superimpose clinical image processing kernels

$$\begin{bmatrix} k_{-1,-1} & k_{-1,0} & k_{-1,1} \\ k_{0,-1} & k_{0,0} & k_{0,1} \\ k_{1,-1} & k_{1,0} & k_{1,1} \end{bmatrix} \longrightarrow \begin{bmatrix} k_{-1,-1} \\ k_{-1,0} \\ k_{-1,1} \\ k_{0,-1} \\ k_{0,0} \\ k_{0,1} \\ k_{1,-1} \\ k_{1,0} \\ k_{1,1} \end{bmatrix} = \text{data for clustering}$$



Clustering visualizations at different layers of our UNet. Encoder layers (top) show a roughly normal distribution of convolution features, while distinct patterns emerge in the decoder layers (bottom).

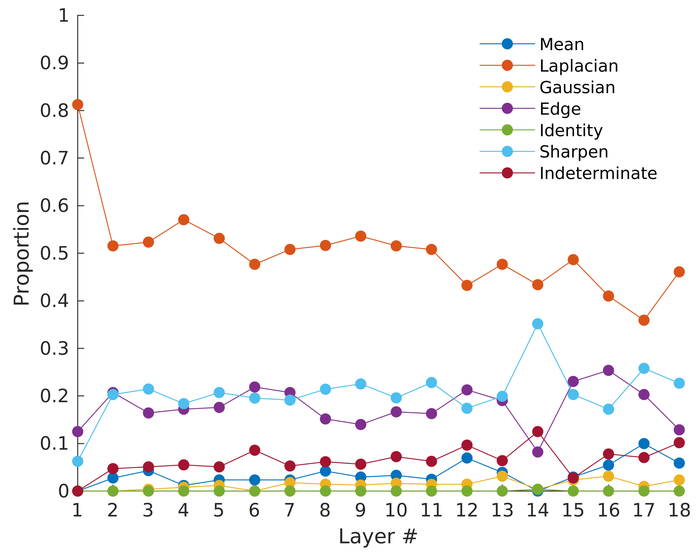
2. Find image processing kernel with closest actions

1. Construct matrix $A_{[K]} \in \mathbb{R}^{n_x n_y \times n_x n_y}$ describing convolution with K
2. Compute singular values of linear operator
3. Compute singular values of clinical image processing kernels
4. Assign closest clinical feature F that has smallest spectral distance to K

$$A_{[K]} = \begin{bmatrix} T_0 & T_1 & & \\ T_{-1} & \ddots & \ddots & \\ & \ddots & \ddots & T_1 \\ & & T_{-1} & T_0 \end{bmatrix} \quad \text{with} \quad T_i = \begin{bmatrix} k_{i,0} & k_{i,1} & & \\ k_{i,-1} & \ddots & \ddots & \\ & \ddots & \ddots & k_{i,1} \\ & & k_{i,-1} & k_{i,0} \end{bmatrix}$$

$$A_{[K]} = U_K \Sigma_K V_K^T \qquad \forall K \in \{\text{layers}\}$$
$$A_{[F]} = U_F \Sigma_F V_F^T \qquad \forall F \in \{\text{features}\}$$

$$\text{find} \quad \underset{F \in \{\text{features}\}}{\arg\min} \; \|\Sigma_K - \Sigma_F\|_1$$



Assignment of convolution kernels to their closest image processing feature counterparts. A label of 'indeterminate' was given if the L_1 distance was not within 10% of the largest singular value.

jonasactor@rice.edu https://caam.rice.edu/~jaa5 https://github.com/jonasactor/livermask