Constraining neural networks to craft representations

"How can we structurally enforce desirable properties on neural networks?"

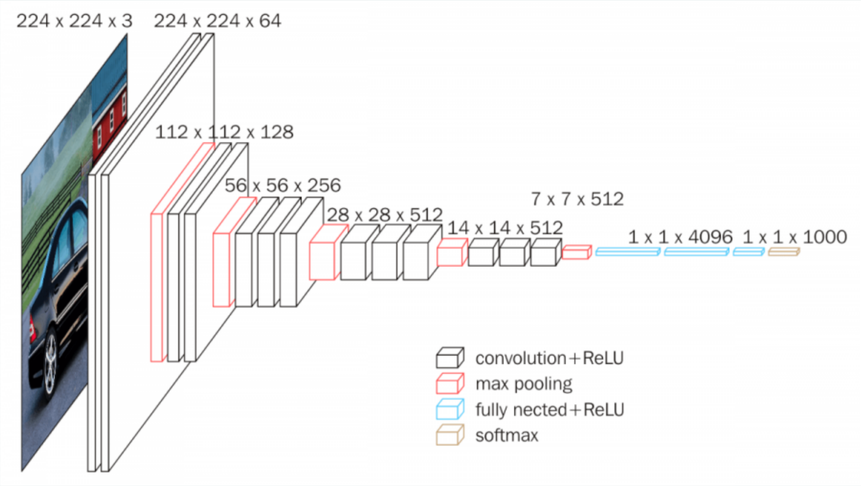
Friday, March 17, 2022

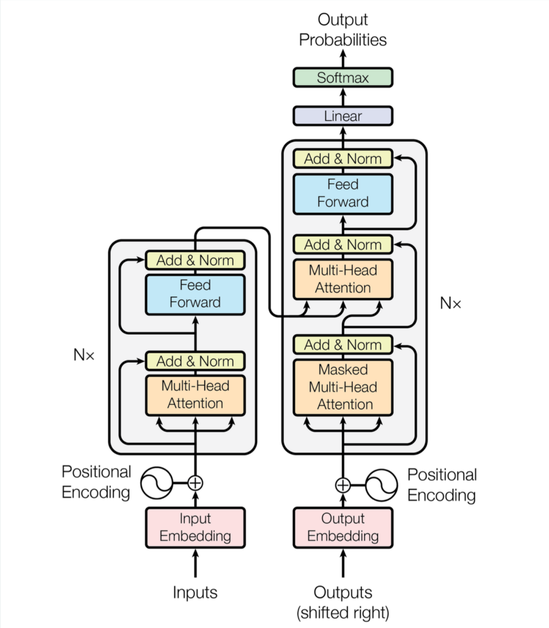
Emmanouil Theodosis
etheodosis@g.harvard.edu
manosth.github.io/
github.com/manosth/

Motivation

Interpretability

Deep learning is empirical and hard to reason





Crafting representations

How do we instill domain knowledge?

ReLU vs sigmoid vs tanh vs ….

Analysis is usually post-hoc.

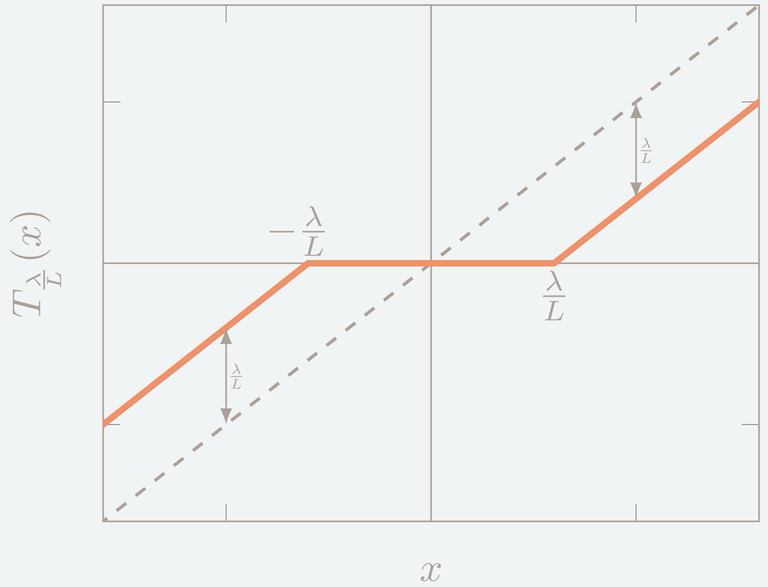
Optimization-informed deep learning

Linear generative model

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

Inverse problem

$$\boldsymbol{x}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$



In deep learning $\boldsymbol{A}$ is overcomplete

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{x} \text{ sparse}$$

Iterative soft thresholding

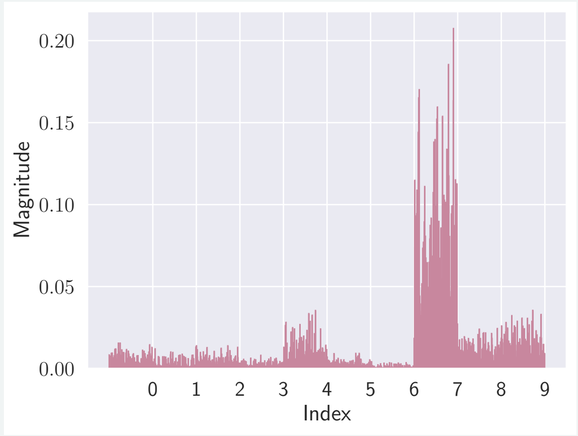
$$\boldsymbol{x}^{(l+1)} \leftarrow T_{\frac{\lambda}{L}}(\boldsymbol{x}^{(l)} + \frac{1}{L}\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(l)}))$$

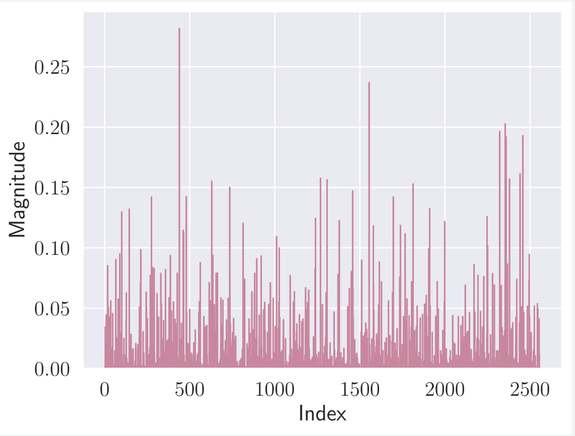
Group sparse priors

Model

Assume a group sparse prior

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \sum_{g \in S} \|\boldsymbol{x}_g\|_2$$

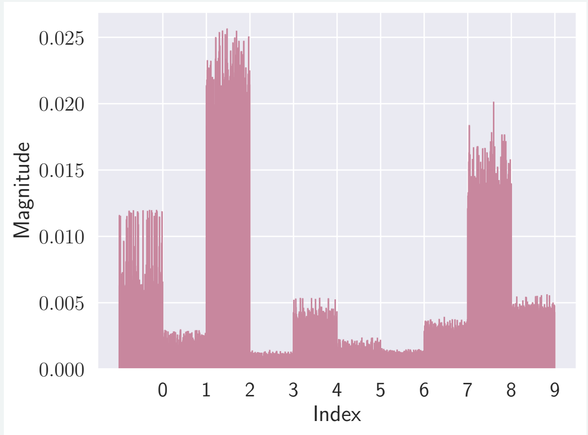


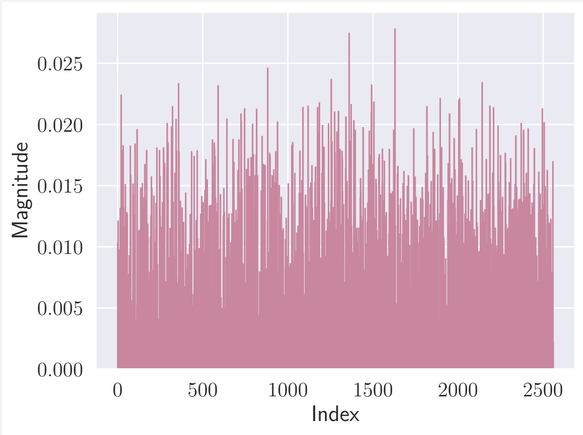


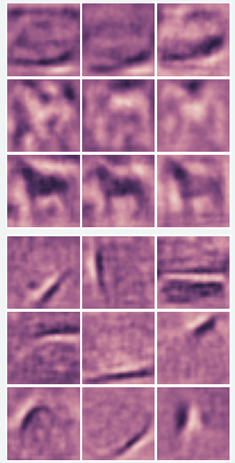
Proximal operator

Projection to the solution set via

$$\sigma_\lambda(\boldsymbol{x}_g) = \left(1 - \frac{\lambda}{\|\boldsymbol{x}_g\|_2}\right)_+ \boldsymbol{x}_g$$

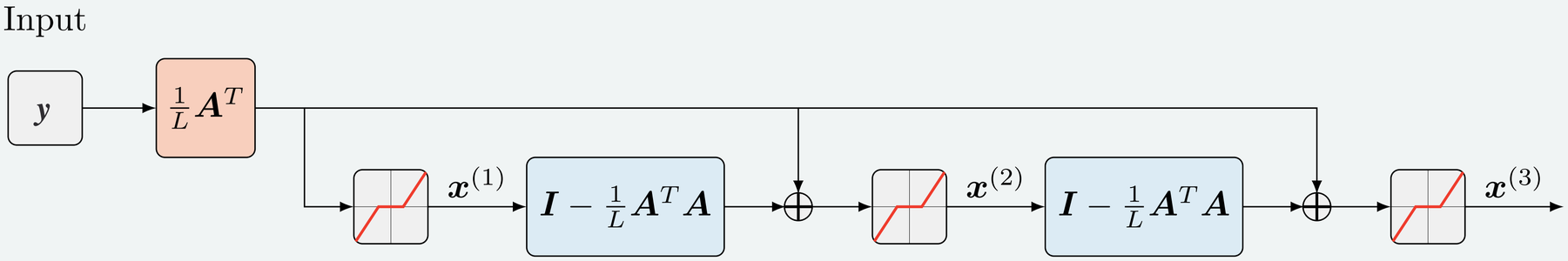






Architecture

Residual connections to the input

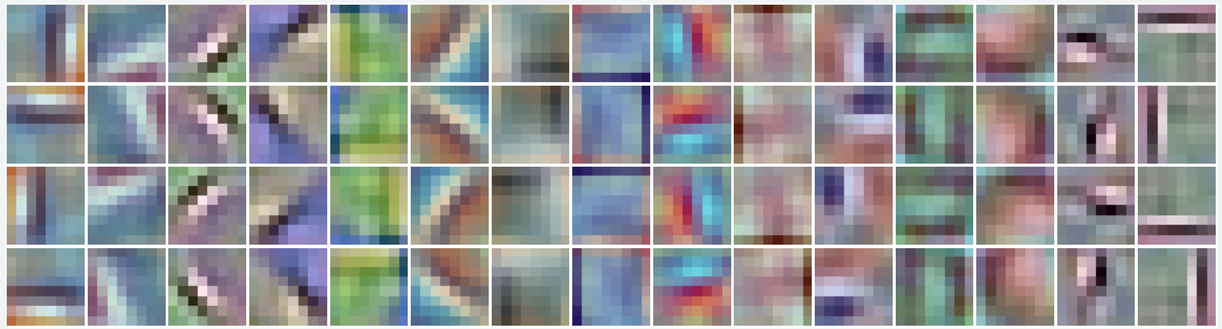


Unrolling equivariances

Equivariance

"Consistent" behavior w.r.t. an operator

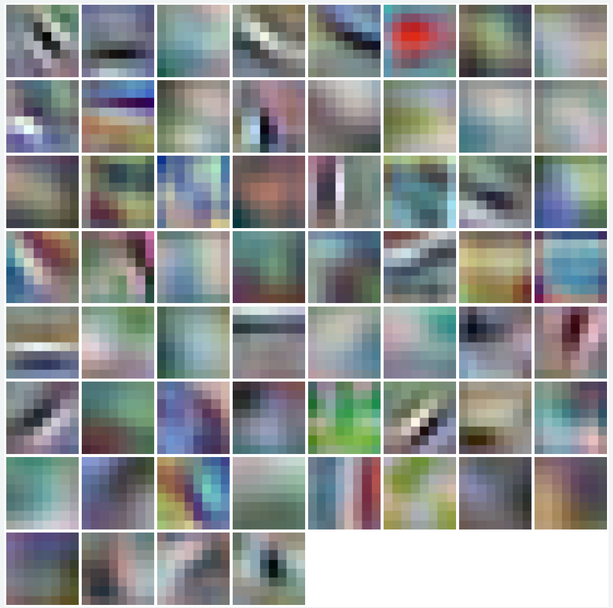
$$f(T(\boldsymbol{x})) = T'(f(\boldsymbol{x}))$$



Symmetry

Model fixed rotations

$$G = \{e, R_\theta, R_\theta^2, \ldots, R_\theta^{k-1}\}$$



Unrolled network

Layer weights

$$\boldsymbol{W}_l = [\boldsymbol{w}_l \quad R_\theta(\boldsymbol{w}_l) \quad \ldots R_\theta^{k-1}(\boldsymbol{w}_l)]$$

Dense and sparse decomposition

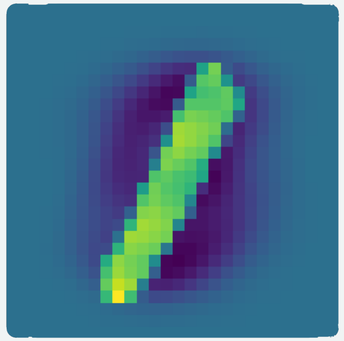
Model

Decompose observations

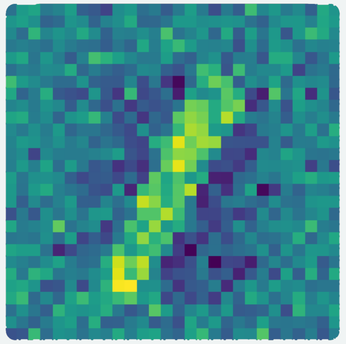
$$y = Ax + Bu$$

where

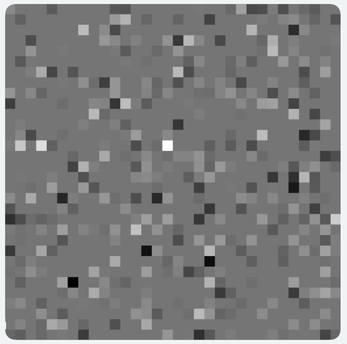
- u is sparse
- Ax is dense



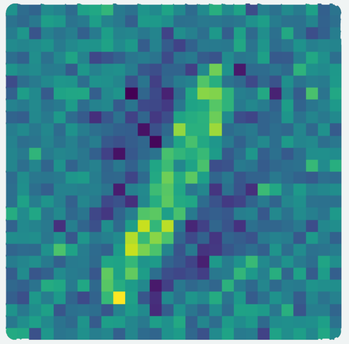
y



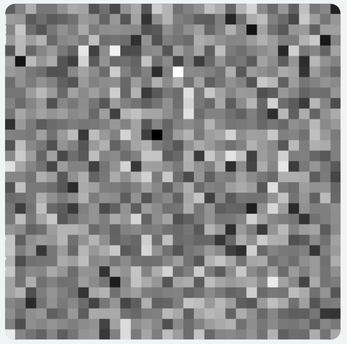
Bu



u



Ax



x

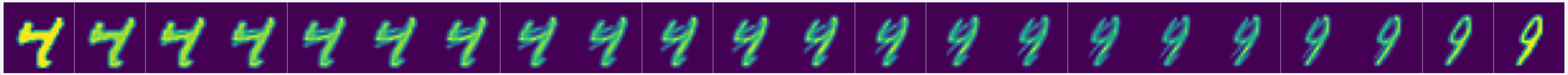
Optimization

Unrolled network for

$$\min_{x,u} \|y - Ax - Bu\|_2^2 + \mu\|Ax\|_2^2 + \lambda\|u\|_1$$

Current and future ideas

Robustness



- Group sparse supports are harder to change
- Robustness to noise (natural or adversarial)

Generalization

$$\textbf{sample complexity}_{\text{GS}} < \textbf{sample complexity}_{\text{S}}$$

- Able to reach similar performance with fewer samples
- Generalization error is smaller

(Partial) Symmetry learning

$$\boldsymbol{W}_l = [\boldsymbol{w}_l \quad \boldsymbol{A}\boldsymbol{w}_l \quad \ldots \boldsymbol{A}^{K-1}\boldsymbol{w}_l]$$

- Learn residual symmetries in convolutional settings
- General framework for implicit group learning

Collaborators and publications

Publications

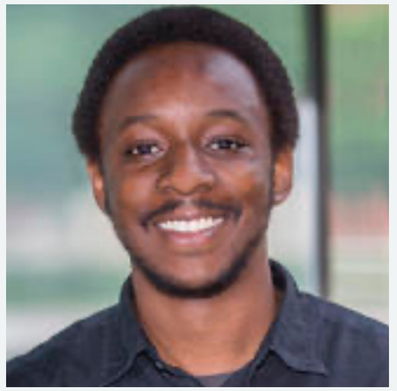
E. Theodosis and D. Ba, "Learning silhouettes with group sparse autoencoders", in *International Conference on Acoustics, Speech, and Signal Processing*, 2023

A. Tasissa, **E. Theodosis**, B. Tolooshams, and D. Ba, "Discriminative reconstruction via simultaneous dense and sparse coding", *Under review*, 2022

E. Theodosis and D. Ba, "Learning unfolded networks with a cyclic group structure", in *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, 2022

E. Theodosis, B. Tolooshams, P. Tankala, A. Tasissa, and D. Ba, "On the convergence of group sparse autoencoders", in *arXiv*, 2020

Bahareh Tolooshams
btolooshams.github.io/

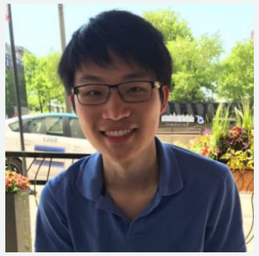


Demba Ba
demba-ba.org/

Abiy Tasissa
sites.tufts.edu/atasissa/

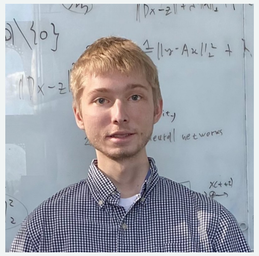
Acknowledgements

Members of CRISP



Alex

Shubham



Jon

THANK YOU

Questions?