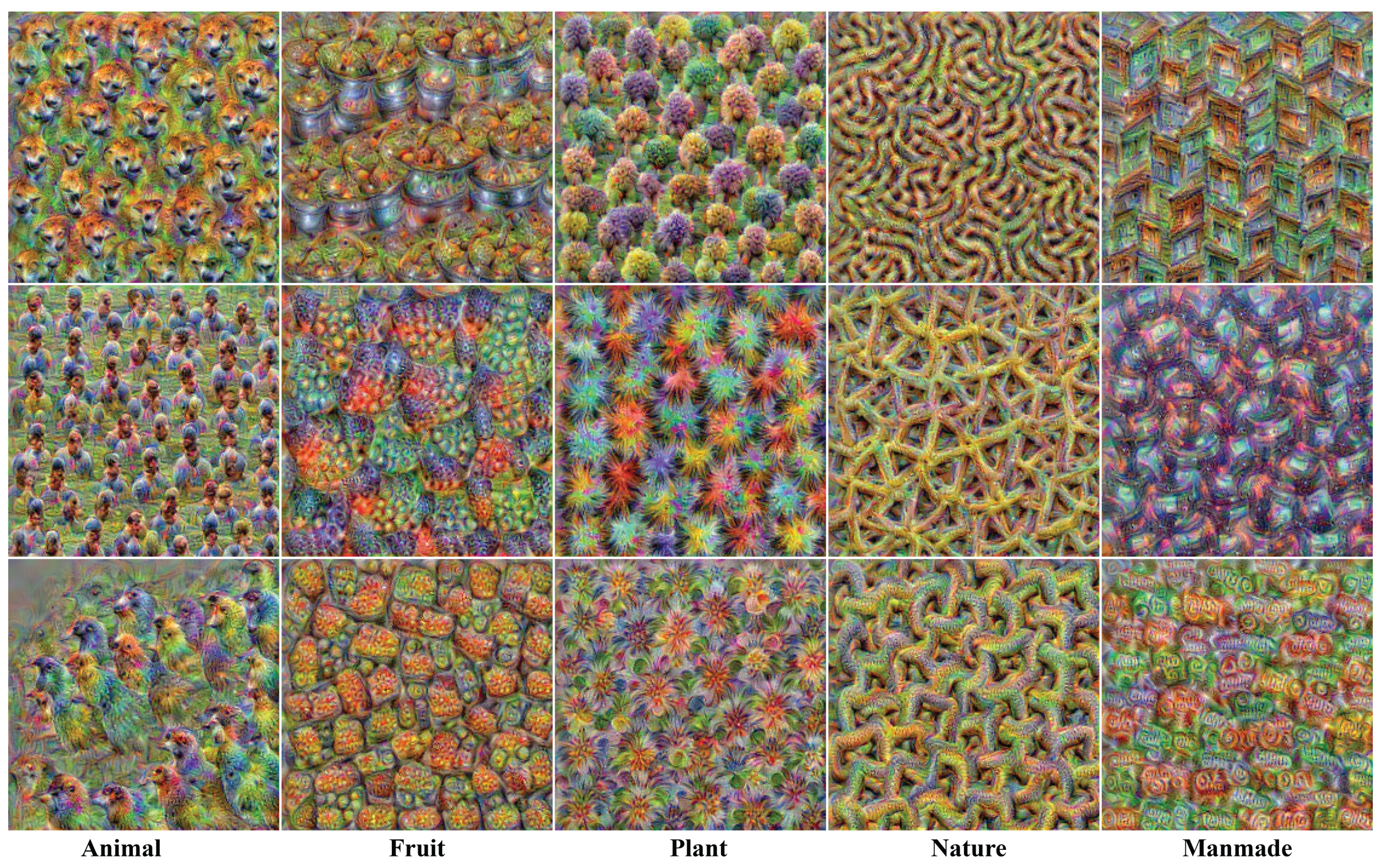
Deep CNN Sparse Coding Analysis

Michael Murray[1,2], **Jared Tanner**[1,2]
Contact Email: mmurray@turing.ac.uk
[1]University of Oxford, [2]Alan Turing Institute

1. Towards an understanding of DCNNs

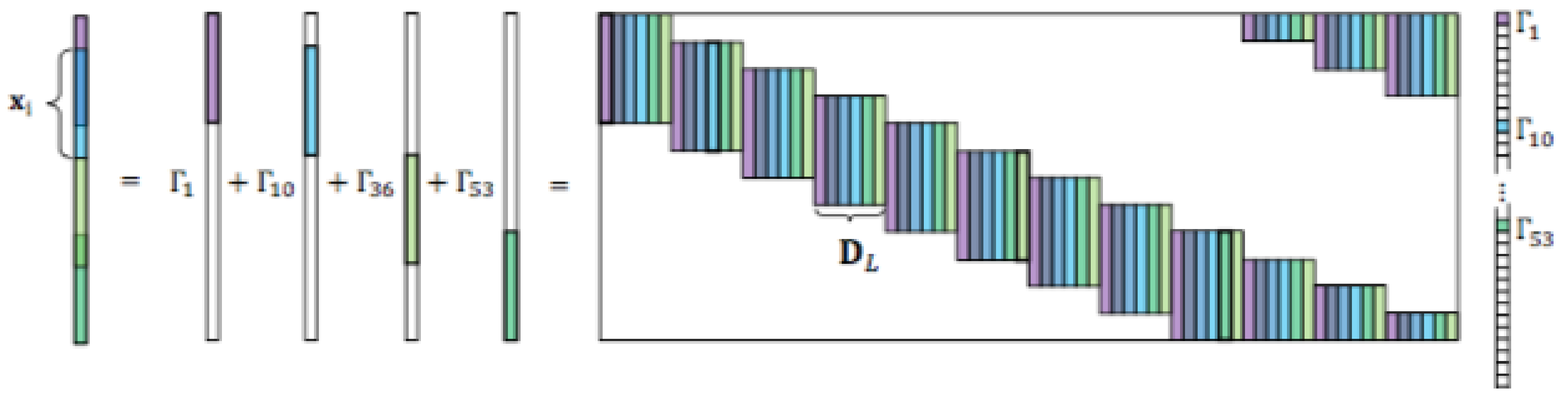
- Since AlexNet (Krizhevsky et al, [1]) in 2012, Deep Convolutional Neural Networks (DCNNs) have been the state of the art for many tasks in computer vision.
- The activation pathway of data through a DCNN is the pattern of nonzero node outputs at every layer. It indicates how a DCNN matches data to its own internal representations and hence how it makes decisions, e.g., for tasks like classification.
- Much experimental work has been done to understand and visualise these activations (e.g., [5] and [7]), but a thorough theoretical understanding is lacking.



Visualisation dislaying the image patterns that activate given neurons - Wei et al [6]

2. Convolutional Sparse Coding

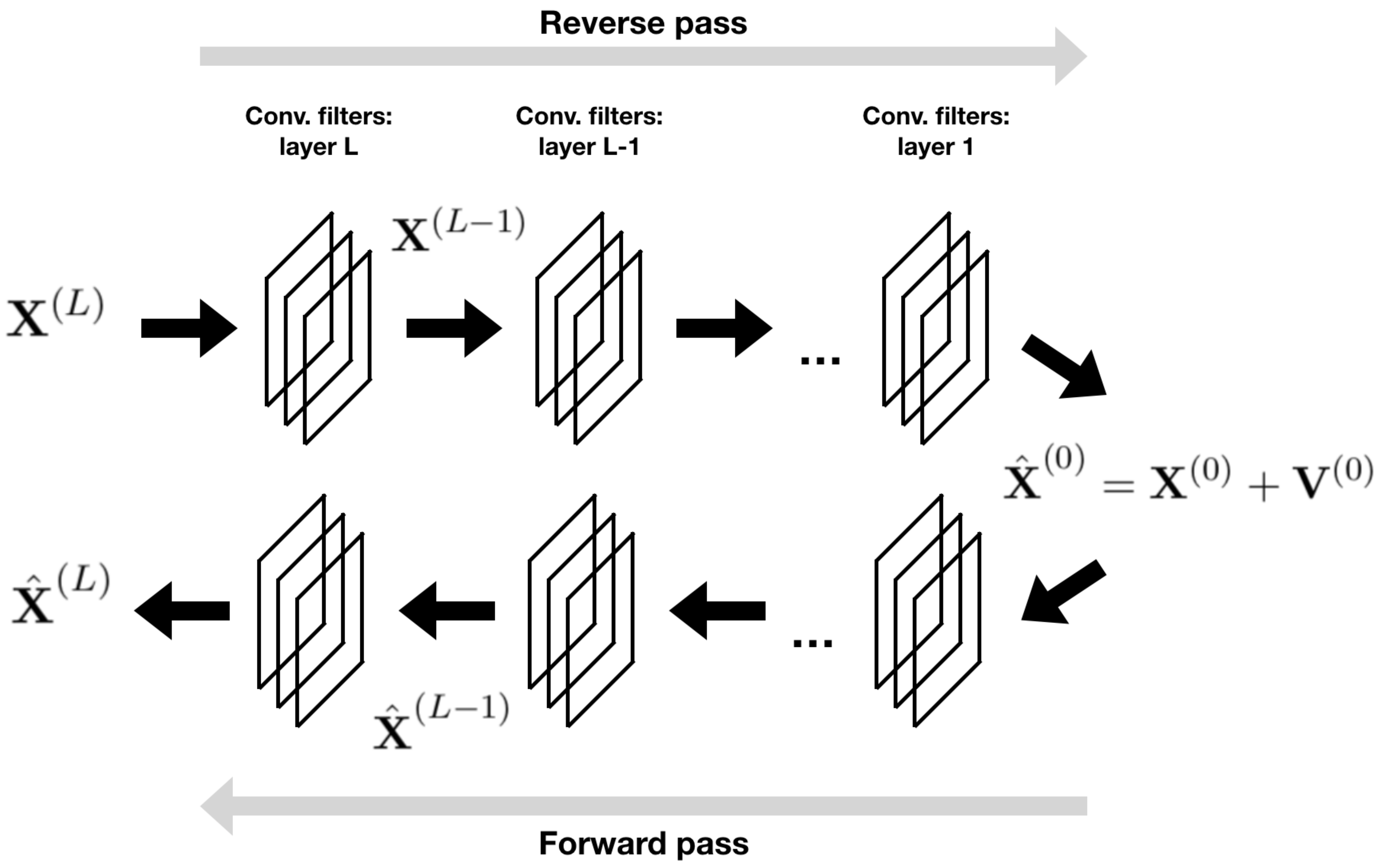
- The Convolutional Sparse Coding model (CSC), proposed by Papyan and Sulam [3] allows us to connect CNNs with sparse coding. In the CSC model the global dictionary is created by shifting a local dictionary across different spatial locations.
- The mutual coherence, μ, of the global CSC dictionary is high due to a) the small support of the local atoms and b) the large inner product between any atom and its shifted versions. This makes recovery of the support challenging.
- Papyan and Sulam where able to partially alleviate this issue by introducing a local sparsity measure, proving guarantees for recovery based upon local "stripe sparsity" instead of global sparsity.
- This model has interesting connections with DCNNs - indeed the forward pass across a single layer of a DCNN can be viewed as solving a CSC problem.



CSC model - Papyan et al [3]

3. Interpreting the forward pass as approximatly solving a sequence of sparse coding problems

Here we build on the work of Papyan et al, investigating the role of the forward pass algorithm and its ability to recover the **reverse** activation pathway for data belonging to the Deep Convolutional Sparse Coding model (D-CSC) [2]. The D-CSC model interprets the forward pass of a ReLU activated DCNN as approximatly solving a sequence of Convolutional Sparse Coding (CSC) problems.



D-CSC Model as proposed by Papyan et al in [2]

- The **reverse pass** creates a set of representations of $\mathbf{X}^{(L)}$ by applying a sequence of filter convolutions. These can be expressed as convolutional matrices of the form $\mathbf{A}^{(l)}\mathbf{D}^{(l)}$.
- In the **forward pass** we seek to recover these representations from the noisy measurement $\hat{\mathbf{X}}^{(0)}$
- The activations in the presence of noise are estimated recursively from the data matrix $\hat{\mathbf{X}}^{(0)}$ according to

$$\hat{\mathbf{X}}^{(l)} = Proj_{\|\cdot\|_{0,\infty}^{Q^{(l)}} \leq S_l} \left((\mathbf{D}^{(l)}\mathbf{A}^{(l)})^T \hat{\mathbf{X}}^{(l-1)} \right). \quad (1)$$

- In a standard DCNN the projection operator is typically a **ReLU** - however to allow us to conduct our analysis, which relies on random filter signs at each layer, we deploy a **Hard Thresholding (HT)** function.

4. Pior Art - Uniform bounds

For signals consistent with the D-CSC model with $\mathbf{D}^{(l)} = \mathbf{I}\ \forall l$ then under worst case assumptions, if $\|\mathbf{X}^{(l)}\|_{0,\infty}^{Q^{(l)}} \leq S_l$ and $\|\mathbf{V}^{(l)}\|_{2,\infty}^{P^{(l)}} \leq \zeta_l$ for some $\{S_l\}_{l=1}^{L}$ and $\{\zeta_l\}_{l=1}^{L}$, then so long as

$$S_l < \frac{\mu^{(l)})^{-1}}{|X_{max}^{(l)}|} \left(\frac{1}{2}|X_{min}^{(l)}| - \zeta_l \right) + \frac{1}{2} \quad (2)$$

the activation of $\hat{\mathbf{X}}^{(l)}$ is exactly the same as the activation of $\mathbf{X}^{(l)}$. Notable in the sparsity bound (2) is the proportionality to μ_l^{-1} [2]. This will typically be **small** for convolutional matrices thereby limiting the complexity of signals guaranteed to be recovered.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[2] V. Papyan, Y. Romano, and M. Elad. Convolutional Neural Networks Analyzed via Convolutional Sparse Coding. *ArXiv e-prints*, July 2016.

[3] V. Papyan, J. Sulam, and M. Elad. Working Locally Thinking Globally: Theoretical Guarantees for Convolutional Sparse Coding. *IEEE Transactions on Signal Processing*, 65:5687–5701, Nov. 2017.

[4] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14(11):828–831, Nov 2007.

[5] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv e-prints*, Dec. 2013.

[6] D. Wei, B. Zhou, A. Torrabla, and W. Freeman. Understanding Intra-Class Knowledge Inside CNN. *ArXiv e-prints*, July 2015.

[7] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. *ArXiv e-prints*, Feb. 2017.

5. Contribution - probabilistic bounds

Our extension follows from incorporating the prior work on one step thresholding by Schnass and Vandergheynst [4] into an appropriately modified D-CSC model. We introduce $\mathbf{D}^{(l)}$, a diagonal matrix whose diagonal entries are independent Rademacher random variables, at each layer which applies a random sign pattern to the columns of $\mathbf{A}^{(l)}$.

Theorem: Let $\hat{\mathbf{X}}^{(l-1)}$ be consistent with the D-CSC model, with $\|\mathbf{V}^{(l)}\|_{2,\infty}^{P^{(l)}} \leq \zeta_l$ and $\|\mathbf{X}^{(l)}\|_{0,\infty}^{Q^{(l)}} \leq S_l$ for all $l = 0, \dots, L-1$. Furthermore assume that $\mathbf{D}^{(l)}$ is a random diagonal matrix with independent Rademacher random variables on the diagonal entries, drawn independent of the dictionaries $\mathbf{A}^{(l)}$. Finally suppose the estimate at each layer is as in (1) and denote as Z_L the event that the activation path is successfully recovered. Then

$$P(\bar{Z}_L) \leq 2dM \sum_{l=1}^{L} n_l \exp\left(- \frac{|X_{min}^{(l)}|^2}{8\left(|X_{max}^{(l)}|^2 \mu_l^2 S_l + \zeta_{l-1}^2\right)} \right). \quad (3)$$

A key implication is that the derived probability bound scales proportional to μ_l^{-2} across a given layer, rather than μ_l^{-1}. To be precise, for a given representation $\hat{\mathbf{x}}^{(l-1)}$ and an arbitrary $\delta \in [0,1]$, then assuming the support is recovered at layer $l-1$, and denote W_l as the event that the support is recovered at layer l. $P(\bar{W}_l) \leq \delta$ if

$$S_l \leq \left(\frac{|x_{min}^{(l)}|^2}{8|x_{max}^{(l)}|^2 \ln\left(\frac{2Mn_l}{\delta}\right)} - \frac{\zeta_{l-1}^2}{|x_{max}^{(l)}|^2} \right) \mu_l^{-2}. \quad (4)$$