Supplementary Material
Structured Receptive Fields in CNNs

Jörn-Henrik Jacobsen[1], Jan van Gemert[1,2], Zhongyou Lou[1], Arnold W. M. Smeulders[1]
[1]University of Amsterdam, The Netherlands
[2]TU Delft, The Netherlands
{j.jacobsen,z.lou,a.w.m.smeulders}@uva.nl, j.c.vangemert@tudelft.nl

1. Supplement Experiment 1 - Plot of Filter Weights for RFNiN-Scale 3^{rd}-order

To show that there is no substantial difference between 3^{rd} and 4^{th} order basis networks with and without scale, below we show a plot 1 of the weights from RFNiN-Scale 3^{rd}-order as listed in Table 1 in the submission. It shows, that filter weights are similarly distributed for a basis of order 3 as for order 4. Furthermore it can be observed that higher scales have highest weights for first and second order filters, indicating that higher orders are mainly important for the lowest layer and scale. Comparing figure 4 in the submission with this plot, we see that 3^{rd} and 4^{th} order basis weights after being trained on ILSVRC2012-100 don't show substantial differences, which was expected from their similar performance.

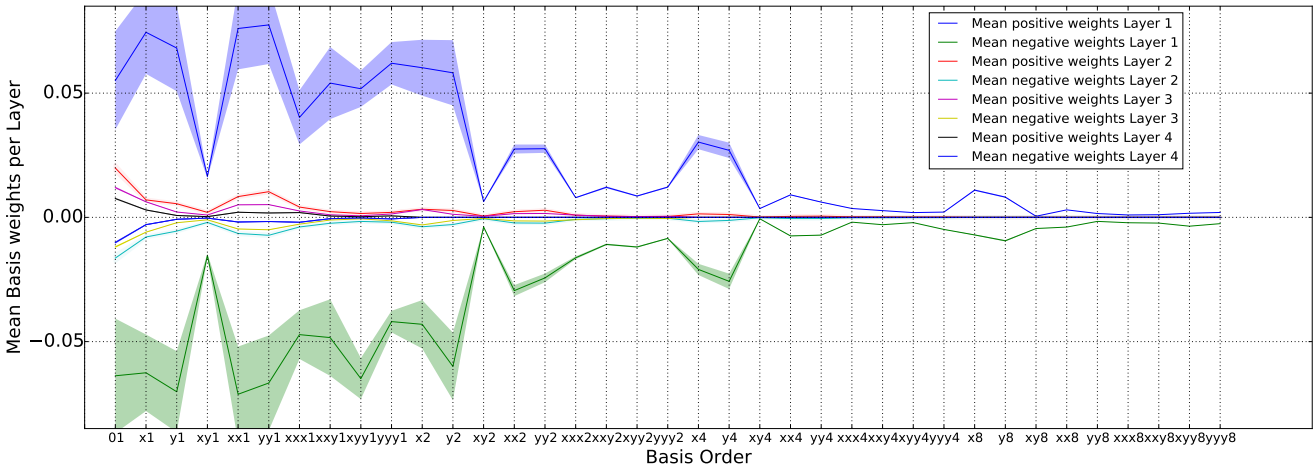

Figure 1: Mean of filter weights and variances per layer for 10 basis filters in 4 scales, as trained on the ILSVRC2012-100 subset and listed in Table 1 as RFNiN-Scale 3^{rd}-order. X and Y denote order of derivative in direction X and Y. The number denotes the sigma for the corresponding Gaussian derivative filter. Higher layers have the highest activation in zero order filters, indicating features being passed on from earlier layers. Whereas in the first layer most filters have high weights, in higher layers only first and second order have high weights. The higher scales have the highest energy in first order filters.

2. Supplement Experiment 2 - Full Result Table MNIST

(%) Accuracy on Subset	Scattering [1]	RFNN (ours)	CNN-A [5]	CNN-B [6]
60000	99.57	99.55±0.02	99.47	**99.58±0.02**
40000	99.47	**99.52±0.04**	99.35	99.50±0.06
20000	**99.42**	99.39±0.05	99.24	99.34±0.03
10000	99.12	**99.22±0.09**	99.15	99.17±0.09
5000	98.97	**99.06±0.03**	98.48	98.73±0.12
2000	**98.70**	98.35±0.09	97.47	97.84±0.20
1000	**97.70**	97.69±0.22	96.79	96.31±0.42
300	95.30	**96.32±0.43**	92.82	92.32±0.77

Table 1: **Results when training on various random subsets of MNIST:** The table shows performance of the Scattering network [1] and a published CNN [5]. We perform on par with Scattering for all subsets and thus outperform both CNNs when training set size is small. As expected the performance gap between CNNs and Scattering/RFNN increases with lower training set sizes.

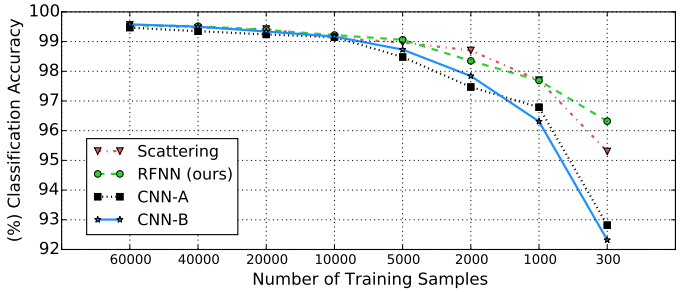

Figure 2: Corresponding figure to table 1 for side by side comparison.

Table 1 shows the full result table underlying figure 5 in experiment 2. The results for the Scattering network are taken from [1] and the CNN-A results from [5]. These are compared to our RFNN based on the network in [6] and the identical CNN-B architecture without structured RFs to illustrate, that our performance is not due to the architecture, but due to structured receptive fields. The results for RFNN and CNN-B are averaged over 3 runs and include standard deviations. Scattering is the state of the art method for small subsets of MNIST, while CNN-A is the best published CNN on this problem. The Scattering network and RFNN perform on par and show clear superiority over the CNNs for small dataset sizes.

3. Supplement Experiment 4 - Alzheimer's Classification Data Details

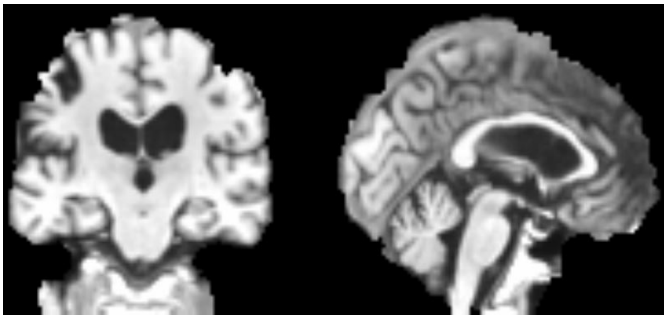

Figure 3: As an example, two 2D slices of a pre-processed 3D brain scan from the ADNI cohort as used in experiment 4. The steps performed were: 1) Re-orientation to standard MNI orientation. 2) Automatic cropping of empty border. 3) Bias-field correction. 4) Linear registration to standard MNI space. 5) Brain extraction.

The data used in experiment 4, classification of Alzheimer's diseased vs. cognititve normal subjects, were obtained from the cohort of the Alzheimer's disease Neuroimaging Intitiative (ADNI) [4]. The dataset is a standard benchmark for neuroimaging classification methods [2]. We follow the experimental procedure outlined in the extensive review paper of Cuingnet et al. [2] and use the exact same criteria for including subjects into the dataset, which results in the identical subset of 150 training (81 cognitive normal, 69 Alzheimer's diseased) and 149 testing images (81 coginitive normal, 68 Alzheimer's diseased). The criteria for including subjects into the train and test set are according to the ADNI protocol, the exact list of subjects can be found in the supplementary material of [2]. We selected the raw scans and performed standard pre-processing with the FSL library in default settings [3]. Specifically we applied the FSL anatomical pre-processing script to all brain scans with no non-linear registration and no segmentation applied. Subsequently we applied the brain extraction tool of the FSL library in default settings. An example of a pre-processed T1 MRI scan as used in the experiment can be seen in figure 3.

4. Deriving Independence between Parameters and Basis Layer

Here we show, that from RFNN backpropagation derivation directly follows separability of basis and weight layer. As mentioned in the paper, a 2D filter kernel function $F(x, y)$ in all layers, is a linear combination of m unique (non-symmetric) Gaussian derivative basis functions ϕ

$$F(x, y) = \alpha_1\phi_1 + \cdots + \alpha_n\phi_m, \tag{1}$$

where $\alpha_1, ..., \alpha_m$ are the parameters being learned.

We learn the filter weights α by mini-batch stochastic gradient descent. We compute the derivatives of the loss function E with respect to the parameters α by applying the chain rule

$$\frac{\partial E}{\partial \alpha_{ij}^l} = \frac{\partial E}{\partial o_{jn}^l}\frac{\partial o_{jn}^l}{\partial t_{jn}^l}\frac{\partial t_{jn}^l}{\partial \alpha_{ij}^l}, \tag{2}$$

where α_{ij}^l are the parameters at layer l between the input feature map i and the output feature map j indexed by neuron n. And t_{jn}^l is the weighted sum of outputs of previous neurons. o_{jn}^l is the n^{th} neural value of output feature map j in the layer l by applying the activation function to t_{jn}^l(i.e. $o_{jn}^l = \psi(t_{jn}^l)$). ψ is the activation function (i.e. sigmod function in this derivation). $E = Loss(y, y^*)$ is the loss function, y is ground truth label and y^* is the prediction.

For clarity, we split 2 into two parts, δ_{jn}^l and the derivative of the convolutional function D_{ij}^l.

The first part contain the first two term of 2 which is

$$\delta_{jn}^l = \frac{\partial E}{\partial o_{jn}^l}\frac{\partial o_{jn}^l}{\partial t_{jn}^l}, \tag{3}$$

It is trivial to solve 3 if l is the last layer since $o_{jn}^{l} = y^{*}$ in this case and the second term is the derivative of the activation function (ψ'). For the inner layers, by applying the chain rule, $\delta_{jn}^{l} =$

$$(\sum_{k}\sum_{q}\delta_{kq}^{l+1}(\alpha_{ij1}\cdot\phi_1 + ... + \alpha_{ijM}\cdot\phi_M))\psi'(t_{jn}^{l}) \tag{4}$$

Here, k is the feature map index of the layer $l+1$ and q is the neural index of feature map k on the layer $l+1$. $\psi'(t_{jn}^{l})$ is the derivative of the activation function.

The second part of equation 2 is only dependent on the parameters α_{ij}. Let o_i^{l-1} denote the output feature map of layer $l-1$ (which is also the output feature of layer l), the second part of the equation can be calculated as:

$$\begin{aligned} D_c = \frac{\partial t_{jn}^{l}}{\partial \alpha_{ij}} &= \frac{\partial[o_i^{l-1}\cdot(\alpha_{ij1}\cdot\phi_1 + ... + \alpha_{ijM}\cdot\phi_M)]}{\partial\alpha_{ij}} \\ &= \begin{bmatrix} o_i^{l-1}\cdot\phi_1 \\ o_i^{l-1}\cdot ... \\ o_i^{l-1}\cdot\phi_M \end{bmatrix} \end{aligned} \tag{5}$$

where $\phi_m \in \{1, 2, 3, ..., M\}$ denotes the basis functions up to the order M. By substituting the two terms in 2 with 4, we have the derivative of the error with respect to all parameters in the network. The result is:

$$\begin{aligned} \frac{\partial E}{\partial \alpha_{ij}^{l}} &= \sum_{n}\delta_{jn}^{l}\cdot\begin{bmatrix} o_i^{l-1}\cdot\phi_1 \\ o_i^{l-1}\cdot ... \\ o_i^{l-1}\cdot\phi_M \end{bmatrix} \\ wrt.\quad \delta_{jn}^{l} &= \begin{cases} a & \text{if } l \text{ is the last layer} \\ b & \text{if } l \text{ is an inner layer} \end{cases} \end{aligned} \tag{6}$$

where

$$\begin{aligned} a &= (y-t)\phi'(t_{jn}^{l}) \\ b &= \sum_{k}\sum_{q}\delta_{kq}^{l+1}(\alpha_{ij1}\cdot\phi_1 + ... + \alpha_{ijM}\cdot\phi_M))\phi'(t_{jn}^{l}) \end{aligned} \tag{7}$$

Thus we can separate the basis and basis weights into two distinct layers. One is a fixed basis layer and the other is a 1x1 convolution layer, linearly recombining the basis outputs and learning the basis function weights.

Acknowledgement. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] J. Bruna and S. Mallat. Invariant scattering convolution networks. *TPAMI*, 35(8):1872–1886, 2013.

[2] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative, et al. Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *neuroimage*, 56(2):766–781, 2011.

[3] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[4] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in alzheimers disease: the alzheimers disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

[5] M. A. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*. IEEE, 2007.

[6] M. D. Zeiler and R. Fergus. Stochastic pooling for regularization of deep convolutional neural networks. *arXiv preprint arXiv:1301.3557*, 2013.