MSc. Thesis - Deep Restoration
Paper-ish Summary

Frederik Harder

June 30, 2017

1 Motivation

- Inverting Convnets is a good way of looking at learned representations
- Inversion is not trivial and is only approximately possible
- Limitations of inverse networks
- Limitations of optimization based approaches

1.1 Theoretical Considerations

1.1.1 Limitations of inverse networks

Claim: Convolutions are not sufficient for inverting convolutions.

Argument sketch: Given a feature map as f_{j+1}, where $\boldsymbol{W_j}$ is toeplitz and f_j is a variable with log prior given as $\frac{1}{2} f_j^T \boldsymbol{\Sigma_{f_j}^{-1}} f_j + const$, the optimization over f_j is as follows:

$$\min_{f_j} \frac{1}{2}\lambda||f_{j+1} - \boldsymbol{W_j} f_j||^2 + \frac{1}{2} f_j^T \boldsymbol{\Sigma_{f_j}^{-1}} f_j + const \tag{1}$$

which resolves to

$$\hat{f}_j = \left(\frac{1}{\lambda}\boldsymbol{\Sigma_{f_j}^{-1}} + \boldsymbol{W_j^T W_j}\right)^{-1} \boldsymbol{W_j^T} f_{j+1} \tag{2}$$

It can be shown, that $\left(\frac{1}{\lambda}\boldsymbol{\Sigma_{f_j}^{-1}} + \boldsymbol{W_j^T W_j}\right)^{-1} \boldsymbol{W_j^T}$ is not toeplitz (under what conditions?, why?), and therefore the optimal inverse operation is not a convolution.

Assuming no prior and going for the maximum likelihood estimate instead yields $\hat{f}_j = \boldsymbol{W_j^{-1}} f_{j+1}$ and $\boldsymbol{W_j^{-1}}$ is also not toeplitz (under what conditions?). Therefore using Convolutions to approximate $\boldsymbol{W_j^{-1}}$ limits reconstruction performance by design.

1.1.2 Limitations of optimization based approaches

There was a thought here. Any hints?

2 Related Work

- in detail:

  Dosovitskiy & Brox

  Mahendran & Vedaldi

  Zeiler & Fergus ?

  Yosinsky et al.

- also others (will make list later)

3 Methods

- Modular inversion
- Optimization + Natural Image priors
- Sparse coding priors

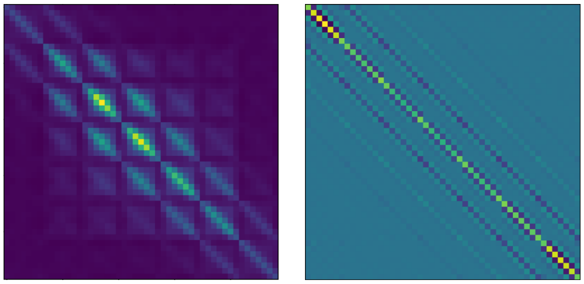
3.1 Priors

3.1.1 ICA Sparse coding priors

3.1.2 field of experts priors

Given equation (1) in Roth and Black 2005, changing notation of $\mathbf{J}$ to $\mathbf{w}$, we get the following prior for patches $\mathbf{x}$

$$\begin{aligned} p(\mathbf{x}) =& \frac{1}{Z(\mathbf{\Theta})} \prod_{i=1}^{N} \phi_i(\mathbf{w}_i^T \mathbf{x}; \alpha_i) \\ =& \frac{1}{Z(\mathbf{\Theta})} \prod_{i=1}^{N} \left(1 + \frac{1}{2}(\mathbf{w}_i^T \mathbf{x})^2\right)^{-\alpha_i} \\ \log p(\mathbf{x}) =& \sum_{i=1}^{N} -\alpha_i \log \left(1 + \frac{1}{2}(\mathbf{w}_i^T \mathbf{x})^2\right) - \log Z(\mathbf{\Theta}) \qquad (3) \\ \propto& \sum_{i=1}^{N} -\alpha_i \log \left(1 + \frac{1}{2}(\mathbf{w}_i^T \mathbf{x})^2\right) \\ \propto& \sum_{i=1}^{N} \alpha_i G(\mathbf{w}_i^T \mathbf{x}) \end{aligned}$$



(a) Average covariance

(b) Inverse average covariance

Figure 1: both figures are based on flattened feature maps in the fifth 7 by 7 pooling layer of VGG16 over 200 images

In order to use score matching, the derivatives of G need to be computed

$$\begin{aligned} G(s) &= -\log(1 + \frac{1}{2}s^2) \\ g(s) &= -\frac{2s}{s^2+2} \\ g'(s) &= \frac{2(s^2-2)}{(s^2+2)^2} \end{aligned} \tag{4}$$

Directly following Hyvarinen 2003, equations (11) and (12), this results in the sampled loss function $\tilde{J}$:

$$\tilde{J} = \sum_{k=1}^{m} \alpha_k \frac{1}{T} \sum_{t=1}^{T} g'(\mathbf{w}_k^T \mathbf{x}(t)) + \frac{1}{2} \sum_{j,k=1}^{m} \alpha_j \alpha_k \mathbf{w}_j^T \mathbf{w}_k \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{w}_k^T \mathbf{x}(t)) g(\mathbf{w}_j^T \mathbf{x}(t)) \tag{5}$$

4 Results (and methods... distinction later)

4.1 Feature map statistics

In an initial investigation, three feature map characteristics were looked at. The average covariance and its inverse, shown in figure 1, substantiate an assumption of spatial stationarity (ground in math) in the feature map activation. Border effects are visible, especially in the inverse matrix, but leaving these aside, the inverse is close to toeplitz, which is expected to become more accurate for larger sets of images. (and this matters because math). This finding serves to motivate convolutional and patch-based priors. Gram matrices and sparsity statistics of the feature maps were also evaluated, and may be used at a later stage in evaluating the trained feature map priors.

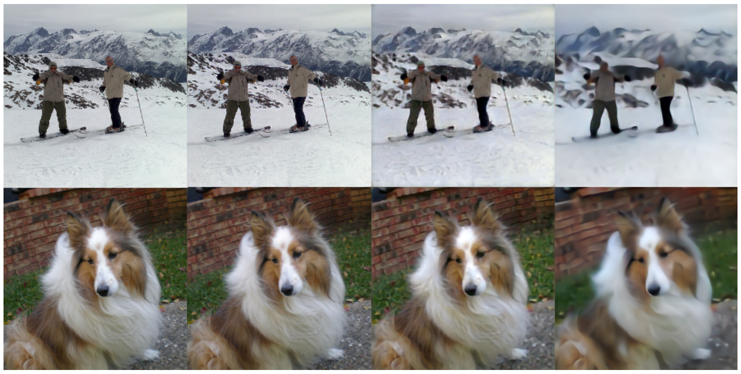


Figure 2: on the left: original image, then: inverse networks reconstructions based on VGG16 feature maps from pooling layers 1, 2, and 3 (will need more detail & deeper layers)

4.2 modular inverting networks

Following the work by Dosovitskiy & Brox, a first baseline has been set using inverting networks. Here, a modular approach is used, where each module is trained to in invert a small part of the network, i.e. either a convolution operation followed by a nonlinearity, a pooling operation, or, in case of AlexNet, a local response normalization operation. These modules can then be combined and trained synchronously, where each module output computes a reconstruction loss and is fed as input to the next module. Different modules have been tested. A transpose convolution, followed by a ReLU and then a convolution has performed best. Future results will follow Odena, et al. and use upscaling and convolution instead of a regular transpose convolution in order to prevent artifacts. Some preliminary results are shown in 2.

Besides setting a baseline in terms of reconstruction quality, this approach is also intended to explore two questions. The first minor insight concerns the performance of stacked modules that were trained individually based on feature map pairs as compared to models, which are trained together. As figure 3 shows, simultaneous training yields better results. (possibly trivial and to be dropped)

The second and major incentive is to motivate the use of constraints in the intermediate feature map layers. For this purpose, the modules have been trained end to end, leaving out all intermediate losses. A comparison with stacked modules is shown in figure 4. This result shows, that constraining the intermediate representations in the model to match the feature maps of the classifier to be inverted can lead to overall better reconstructions. It is therefore reasonable to assume that the same will hold in the following approach using optimization based methods.

Because differences between image and reconstruction can be quite small, two difference measures have been added. the first is a color/channel accurate measure, showing the difference between image and reconstruction, renormalized

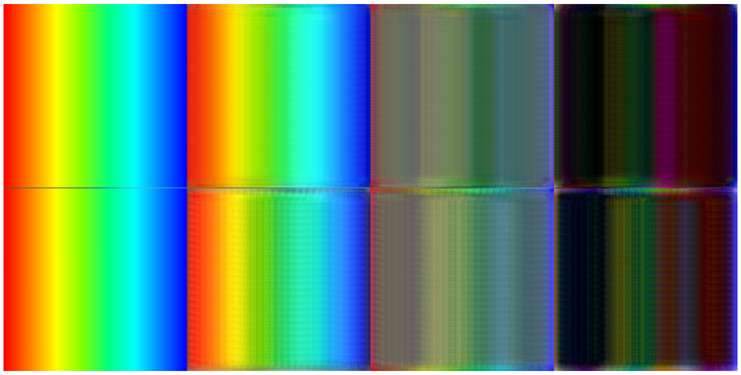


Figure 3: top row: modules trained in parallel, bottom row: stacked modules. left to right: original image, reconstruction from AlexNet, two pixel error measures for visualization

to the interval [0,1]. So if, for example, the reconstructed image has less blue in it, the measure will be blue. Areas without error (or equal errors across channels) are gray. The second difference measure is the renormalized absolute difference, which highlights areas with errors more clearly. Areas without error are black. Given renormalization, pixel intensity only informs about relative error within the image, not absolute error compared to other images. Without normalization, the errors were hardly visible. The figures are ordered left to right as: image, reconstruction, color measure, absolute measure.

4.3 Optimization based reconstruction

The central approach to network visualization is based on work by Mahendran & Vedaldi 2015/16. Their method optimizes a 'pre-image' variable to match one of its feature map representations with that of an image, while constraining the 'pre-image' additionally by a hand-crafted natural image prior. These results have been reproduced, with the goal of improving upon them by using learned priors instead of handcrafted ones. In addition to a natural image prior, priors trained on the feature maps will be used to constrain the problem further and hopefully lead to better reconstructions as a result. A very initial result using a prior based on over-complete ICA is displayed in figure 5, but should not be representative of the potential of this model.

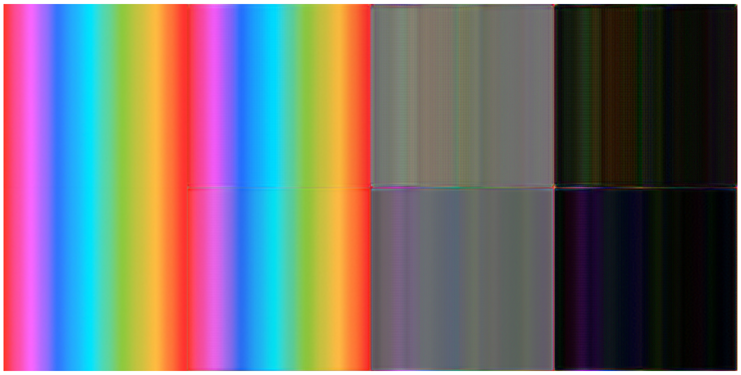


Figure 4: Top row: end to end trained model, bottom row: stacked modules. left to right: original image, reconstruction from the first VGG pooling layer, two pixel error measures for visualization

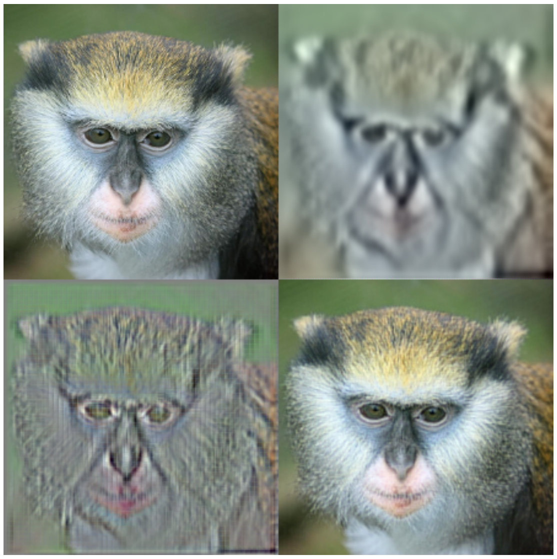


Figure 5: Comparison of reconstructions from the first layer of VGG16. top left: original image, top right: ICA image prior + MSE optimization, bottom left: pure MSE optimization, bottom right: M&V handcrafted prior + MSE

5 Next steps and research goals

5.1 Immediate next steps:

The training of image and feature map priors based on over-complete ICA is set up and will be completed shortly. There is one potential open question with regards to the patch sizes on feature maps. Given that the number of channels grows quickly, the patches should be chosen as small as possible. The size of the receptive field in the previous layer seems like a good first guess. Should the number of patch features grow too large regardless, separate priors for subsets of channels may have to be considered. This, however, would make independence assumptions that are almost certainly wrong and can hopefully be avoided.

Given the priors, a first concern will be the adequate relative weighting in optimization based reconstruction, as each one of them represents an unnormalized negative log probability and as a result can vary greatly in scale. If a principled approach is not found, parameter exploration schemes will have to be designed.

Experiments will then investigate the quality of reconstructions following Mahendran & Vedaldi 2016 with learned priors either augmenting or replacing the handcrafted ones. If these efforts prove fruitful, related applications like activation maximization and caricaturization Mahendran & Vedaldi 2016 can be explored in a quick follow-up. If, instead, the results point to limitations of the prior model, alternative models will need to be explored immediately. Here I propose two options: The field of experts model is related to the model using ICA based filters, but uses Student-t experts instead of a logistically distributed components (category mismatch, work out details) and is well established as a model. The second option comes from Nguyen et al. 2016, which uses denoising auto-encoders to approximate the gradient of natural image priors. This should also work on feature maps and is distinct, in that it is not explicitly a patch-based approach. On the other hand, training times may become more of a concern in that case.

5.2 Goals:

The minimal empirical project goal at this point is the reproduction of the experiments conducted by Mahendran & Vedaldi 2016, with the use of learned image and feature map priors.

A variety of new experiments could be enabled with a good sampling technique. As far as I am aware, unfortunately, neither the patch-based methods, nor the DAE prior lend themselves very well to this task. The highest goal, of sampling images jointly from weighted priors at different feature maps poses an additional problem, which I can currently offer no good solution to. Of course the modes of these distributions can be explored through optimizing pre-images. Generative models, for instance Variational Auto-Encoders, can be trained on individual feature maps and serve as targets for optimization based approaches or as inputs to inverting networks. This could show, how the distribution of feature map activations corresponds to natural images.

A wider review of related work may also provide some additional setups that can be improved with priors. (A day or two of reading may be necessary though)