Exploring the ability of CNNs to generalise to previously unseen scales over wide scale ranges

Ylva Jansson and Tony Lindeberg

Computational Brain Science Lab
Division of Computational Science and Technology
KTH Royal Institute of Technology
Stockholm, Sweden

ICPR 2020: International Conference on Pattern Recognition
January 10-15, 2021.

Topic: Handling scaling variations in deep network

Facts:

- Scaling variations common in image data — because of objects of different size in the world and at different distances to the camera
- Regular CNNs perform poorly when exposed to testing data at scales not spanned by the training data.

Goals:

- Equip deep networks with prior knowledge to handle scaling variations in image data,
- specifically the ability to *generalize to new scales* not spanned by the training data.

Topic: Handling scaling variations in deep network

We will show that:

- Scale generalization is possible with *scale channel networks* (without explicit use of data augmentation).

  “Train for training data at some scale(s), test at any other scale.”

- Scale channel networks lead to improvements in learning performance when training on data with variable scales in the *small sample regime*.
- Point out limitations of previous multi-scale approaches when exposed to testing data with scale variabilities over *wide scale ranges*.

Scale-channel networks

Classical computer vision — Scale-space theory:

- Multi-scale processing by convolving given image with Gaussian kernels and Gaussian derivatives over multiple scales.

Scale-channel networks:

- Apply same CNN to multiple rescaled copies of any image, with *shared weights* between the scale channels.

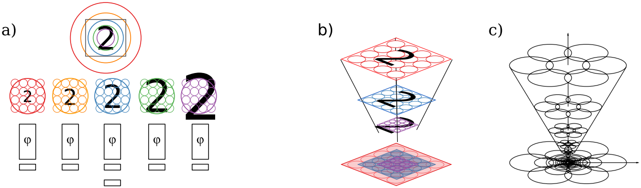
Mathematically:

- Convolution of given image with rescaled filters is *computationally equivalent* to applying fixed-size filter to multiple rescaled copies of the input image.

Scale-channel networks

Networks over multiple scale channels:

- FovMax - with max pooling over the multiple scale channels.
- FovAvg - with average pooling over the multiple scale channels.



(a) Architecture relative to the image frames as being processed.

(b) Architecture relative to the original image frame.

(c) Foveal scale-space model by Lindeberg and Florack (1994).

Continuous formalism for scale-channel networks

With scaling operator S_s defined by

$$(\mathcal{S}_s f)(x) = f(S_s^{-1} x) = f_s(x) = f(\tfrac{x}{s}).$$

and feature maps $\Gamma^{(i)}$, such scale channel networks are *scale covariant*:

$$(\Gamma^{(i)} \mathcal{S}_t f)(x, c, s) = (\Gamma^{(i)} f)(x, c, st).$$

> *"Resizing of the input image corresponds to a mere shift in the scale channels of the scale channel network."*

Proof in paper based on both (i) operator notation for the scaling group and (ii) an integral representation of the scale channel network.

Translational covariance: With shift operator $(\mathcal{D}_\delta f)(x) = f(x - \delta)$

$$(\Gamma^{(i)} \, \mathcal{D}_\delta \, f)(x, c, s) = (\Gamma^{(i)} f)(x - S_s \delta, c, s).$$

> *"Translational shift is rescaled depending upon the scale channel.'*

Scale-invariant scale-channel networks

Given an *infinite* number of scale channels, either continuous or discrete γ^i for $\gamma > 1$, the supremum over the scale channels

$$(\Lambda_{\sup} f)(x, c) = \sup_{s \in S} \left[(\phi_s f)(x, c, s) \right]$$

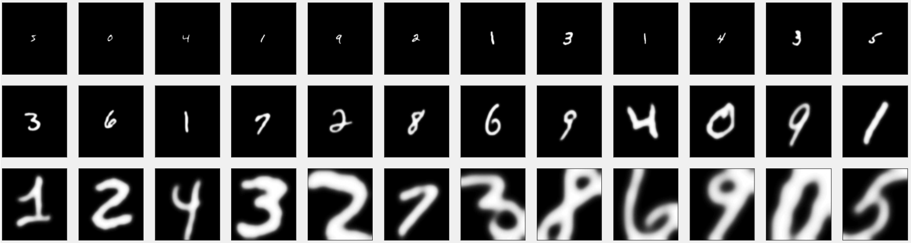
is *provably scale invariant* (proof in the paper).

Any other *permutation invariant pooling operation*, such as the average, is also provably scale invariant.

Given a *finite* number of scale levels in an implementation, we propose to *ensure that there is a sufficient number of additional scale levels around the training scale(s),* with the intention that the learning scheme should learn that the image structures that occur at scales that are a bit off are less relevant, and thus lead to lower values of the feature maps, to reduce the risk that wrong classification results at wrong scales affect the performance.

MNISTLargeScale dataset

Images from the original MNIST dataset 28 × 28 rescaled by factors between 1/2 and 8 and embedded in images of size 112 × 112.



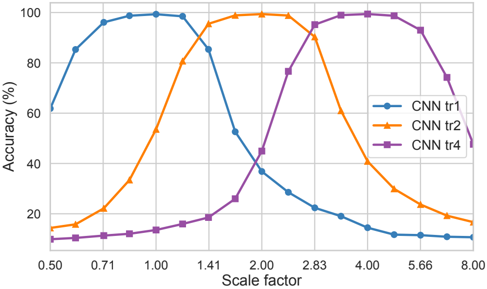
Training data for sizes 1, 2 and 4.

Testing data for sizes between 1/2 and 8 with relative size ratio $\sqrt[4]{2}$.

50 000 training images, 10 000 testing images, 10 000 validation images.

Performance of a vanilla CNN

8 conv-batchnorm-ReLU blocks + fully connected layer + final softmax
16-16-16-16-32-32-32-32-100-10 feature channels (≈ 90 000 parameters)
stride 2 in convolutional layers 2, 4, 6 and 8
Trained for each one of sizes 1, 2 and 4

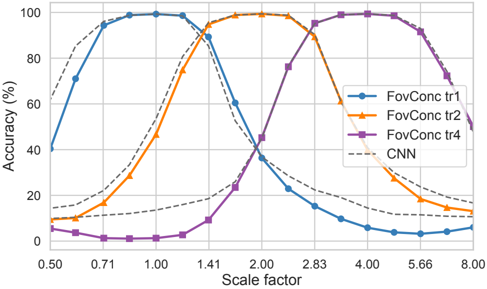


Poor generalization to previously unseen scales (no invariance mechanism)

Performance of a scale concatenated network

Proposed by Xu et al “Scale-invariant convolutional neural networks” (2014)

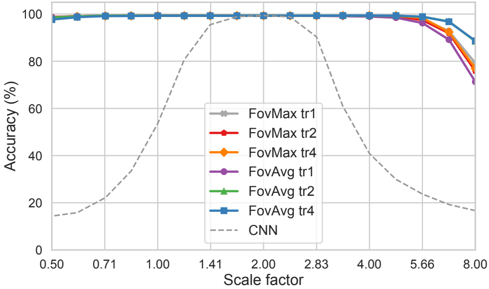
4 conv-batchnorm-ReLU blocks + fully connected layer + final softmax layer
16-16-32-32-100-10 feature channels ($\approx$ 70 000 parameters)
stride 2 in convolutional layers 2 and 4
Trained for each one of sizes 1, 2 and 4



Poor generalization to previously unseen scales (no invariance mechanism)

Performance of the FovMax and FovAvg networks

4 conv-batchnorm-ReLU blocks + fully connected layer + final softmax layer
16-16-32-32-100-10 feature channels (≈ 70 000 parameters)
stride 2 in convolutional layers 2 and 4
Trained for each one of sizes 1, 2 and 4

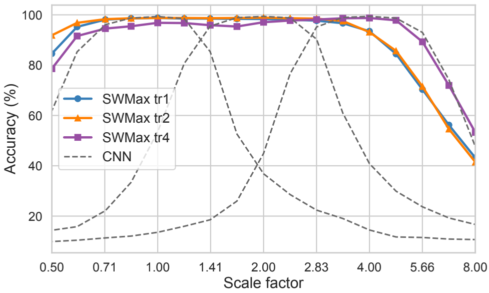


The invariance mechanism gives excellent generalization to previously unseen scales.

Performance of a sliding window approach

Sliding windows used in the OverFeat detector by Sermanet *et al.* (2013).

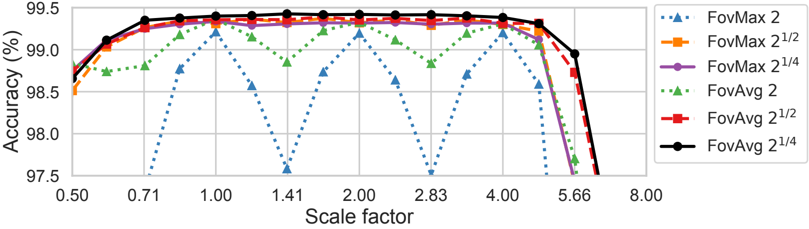
4 conv-batchnorm-ReLU blocks + fully connected layer + final softmax layer
16-16-32-32-100-10 feature channels ($\approx$ 70 000 parameters)
stride 2 in convolutional layers 2 and 4
Trained for each one of sizes 1, 2 and 4



Some scale generalization but not as good as for FovMax and FovAvg.

Dependency on the scale sampling density

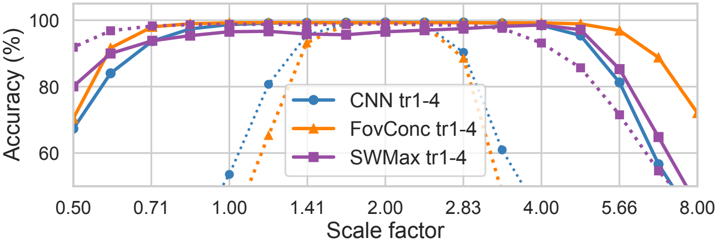
FovMax and FovAvg networks trained with scale sampling rates of 2, $\sqrt{2}$ and $\sqrt[4]{2}$



$\sqrt{2}$ is appropriate sampling rate, going to $\sqrt[4]{2}$ gives no major improvement.

Training over multiple scales

Training data with uniform distribution over sizes between 1 and 4.
Testing data with similar distribution over sizes between 1/2 and 8.

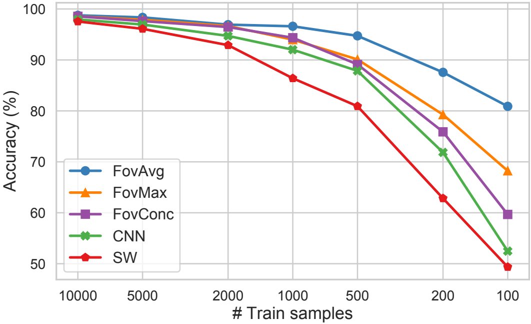


The vanilla CNN and the scale concatenation network are very much helped by training over multiple scales.

For the FovMax and FovAvg networks, training at a single scale is basically as good as training over multiple scales (see paper for details).

Small training sets with large scale variations

Training data with uniform distribution over sizes between 1 and 4.
Testing data with similar distribution over sizes between 1 and 4.



The FovMax and FovAvg networks make more efficient use of small amounts of training data that contain large scale variability.

Summary and conclusions

- Presented methodology to *handle scaling transformations in deep networks* by scale channel networks.
- Presented formalism to analyse scale channel networks and shown that they are *scale covariant* and translationally covariant.
- Combined with max pooling or average pooling over the scale channels, the foveated scale channel networks are also *provably scale invariant*.
- Shown that the FovMax and FovAvg are robust to scaling transformations and allow for *scale generalization*, with very good testing performance *at scales not spanned by the training data*.
- Investigated limited scale generalization performance of vanilla CNNs, scale concatenation networks and sliding window networks.
- Demonstrated that the FovMax and FovAvg networks lead to improvements for *multi-scale training data in the small sample regime*.

Further material on this topic

- Jansson and Lindeberg (2020) “Exploring the ability of CNNs to generalise to previously unseen scales over wide scale ranges”, `arXiv preprint arXiv:2004.01536`.

  Theoretical relationships to scale-space theory:

  - Show that *if the learning algorithm would learn receptive fields corresponding to Gaussian derivatives*, then the relationships between the such receptive fields over different scale channels are *computationally equivalent to the effect of computing scale-normalized derivatives over multiple scales*.
  - The net effect of max pooling or average pooling over the scale channels has *structural similarities to classical methods for scale selection*.

- Y. Jansson and T. Lindeberg (2020) “MNISTLargeScale dataset”, Available at: `https://www.zenodo.org/record/3820247`. DOI:10.5281/zenodo.3820247.

  The MNISTLargeScale dataset available for download.

Further material on this topic

- Lindeberg (2020) “Scale-covariant and scale-invariant Gaussian derivative networks”, `arXiv preprint arXiv:2011.14759`.

  Dual scale channel structure developed for Gaussian derivative networks, which instead preserve the same image resolution at all scale levels.

  Based on rescaling the filters as opposed to rescaling the input images.

- Finnveden, Jansson and Lindeberg (2021) “Understanding when spatial transformer networks do not support invariance, and what to do about it”, International Conference on Pattern Recognition (ICPR2020).

  Describes limitations of spatial transformer networks that are based on plain transformations of feature maps. They do, in general, not support true invariances, which also affects the classification performance. In the paper, we explore alternative approaches to remedy this problem.