In: Proc. International Conference on Pattern Recognition (ICPR 2020), Milan, Italy, Jan 10-15, 2021, pages 1181-1188.

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

***Abstract*—The ability to handle large scale variations is crucial for many real world visual tasks. A straightforward approach for handling scale in a deep network is to process an image at several scales simultaneously in a set of *scale channels*. Scale invariance can then, in principle, be achieved by using weight sharing between the scale channels together with max or average pooling over the outputs from the scale channels. The ability of such *scale channel networks* to generalise to scales not present in the training set over significant scale ranges has, however, not previously been explored. We, therefore, present a theoretical analysis of invariance and covariance properties of scale channel networks and perform an experimental evaluation of the ability of different types of scale channel networks to generalise to previously unseen scales. We identify limitations of previous approaches and propose a new type of *foveated scale channel architecture*, where the scale channels process increasingly larger parts of the image with decreasing resolution. Our proposed FovMax and FovAvg networks perform almost identically over a scale range of 8, also when training on *single scale training data*, and do also give improvements in the small sample regime.**

I. Introduction

Scaling transformations are as pervasive in natural image data as translations. In any natural scene, the size of the projection of an object on the retina or a digital sensor varies continuously with the distance between the object and the observer. Convolutional neural networks (CNNs) already encode structural assumptions about translation invariance and locality. A vanilla CNN is, however, not designed for multi-scale processing, since the fixed size of the filters together with the depth and max-pooling strategy applied implies a preferred scale. Encoding structural priors about visual transformations, including scale or affine invariance, is an integrated part of a range of successful classical computer vision approaches. There is also a growing body of work on invariant CNNs, especially concerning invariance to 2D/3D rotations and flips (see e.g. [1]–[3]). The possibilities for CNNs to generalise to previously unseen scales have, however, not been well explored. We propose that structural assumptions about scale could, similarly to translation covariance, be a useful prior in convolutional neural networks. *Scale-invariant CNNs* could enable both multi-scale processing and predictable behaviour when encountering objects at novel scales, without the need to fully span all possible scales in the training set.

One of the simplest CNN architectures used for covariant and invariant image processing is a channel network (also referred to as siamese network) [3]. In such an architecture, transformed copies of the input image are processed in parallel by different "channels" (subnetworks) corresponding to a set of image transformations. If combined with weight sharing and max or average pooling over the output from the channels, this approach can enable invariant recognition for finite transformation groups.

An *invariant scale channel network* is a natural extension of invariant channel networks for rotations [3]. It can equivalently be seen as a way of extending ideas underlying the classical scale-space methodology [4]–[6] to deep learning. It should be noted that a channel architecture for scale-invariant recognition poses additional challenges compared to recognition over finite groups. First, scaling transformations are, as opposed to 2D or 3D rotations, not a compact group (intuitively, there is no smallest or largest scale). Second, scaling implies a change in image size and resolution for discrete image data. The subject of this paper is to investigate the possibility for CNNs to generalise to previously unseen scales by means of a scale channel architecture.

A. Contribution and novelty

The key contributions of our work are as follows:

- We perform a theoretical analysis of invariance and covariance properties of scale channel networks.
- We present a new family of invariant foveated scale channel networks.
- We evaluate different types of scale channel networks and a vanilla CNN on the task of *scale generalisation over wide scale ranges*, using a new variation of the MNIST dataset with large scale variations.
- We demonstrate inherent limitations of previous scale channel approaches and show that our proposed foveated networks can enable very good generalisation to unseen scales and improvements in the small sample regime.

This is, to our knowledge, the first study to evaluate and demonstrate means for CNNs to *generalise to unseen scales over significant scale ranges*.

B. Related work

In classical scale-space theory [4]–[6], a multi-scale representation of an input image is created by convolving the image with a set of rescaled Gaussian kernels and Gaussian derivative filters, which are then often combined in non-linear ways. The scale channel networks described in this paper can

The support from the Swedish Research Council (contract 2018-03586) is gratefully acknowledged.

be seen as an extension of this philosophy of processing an image *at all scales simultaneously*, but using deep non-linear feature extractors learned from data.

CNNs can give impressive performance but they are sensitive to scale variations. Performance degrades for scales not present in the training set [7]–[9], different network structure is optimal for small vs large images [9] and it is possible to construct adversarial examples by means of small translations rotations and scalings [7], [8]. State-of-the-art CNN based object detection approaches all employ different mechanisms to deal with scale variability, e.g. branching off classifiers at different depths [10], learning to transform the input or the filters [11], [12], or using different types of image pyramids [13]–[15]. The goal of these approaches has, however, not been to generalise to *previously unseen scales* and they lack the structure necessary for true scale invariance.

There exist some previous work aimed explicitly at scale invariant recognition in CNNs [16]–[19]. These approaches have, however, either not been evaluated for the task of generalisation to scales *not present in the training set* [17]–[19] or only across a very limited scale range [16]. Previous scale channel networks exist, but are explicitly designed for multi-scale processing [20], [21] rather than scale invariance or have not been evaluated with regard to their ability to generalise to unseen scales over any significant scale range [13], [16].

II. Theory

In this section, we will introduce a mathematical framework for scale channel networks based on a continuous model of the image space. This model enables straightforward analysis of the covariance and invariance properties of channel networks and generalise previous analysis of invariance properties of channel networks [3] to scale channel networks. We further analyse covariance properties and additional options for aggregating information across the transformation channels.

A. Images and image transformations

We consider images $f : \mathbb{R}^N \to \mathbb{R}$ that are measurable functions in $L_\infty(\mathbb{R}^N)$ and denote this space of images as V. A *group of image transformations* corresponding to a group G is a family of image transformations $\mathcal{T}_g$ ($g \in G$) with a group structure. We denote the combination of two group elements $g, h \in G$ by gh and the cardinality of G as $|G|$. Formally, a group G induces an *action on functions* by acting on the underlying space on which the function is defined (here the image domain). We are here interested in the group of *uniform scalings* around x_0 with the group action

$$(\mathcal{S}_{s,x_0} f)(x') = f(x), \quad x' = S_s(x - x_0) + x_0, \tag{1}$$

where $S_s = \mathrm{diag}(s)$. For simplicity, we often assume $x_0 = 0$ and denote $\mathcal{S}_{s,0}$ as $\mathcal{S}_s$ corresponding to

$$(\mathcal{S}_s f)(x) = f(S_s^{-1} x) = f_s(x). \tag{2}$$

We will also consider the translation group with the action (where $\delta \in \mathbb{R}^N$)

$$(\mathcal{D}_\delta f)(x') = f(x), \quad x' = x + \delta. \tag{3}$$

B. Invariance and covariance

Consider a general feature extractor $\Lambda : V \to \mathbb{K}$ that maps an image $f \in V$ to a feature representation $y \in \mathbb{K}$. In our continuous model, $\mathbb{K}$ will typically correspond to a set of M feature maps (functions) so that $\Lambda f \in V^M$. This is a continuous analogue of a discrete convolutional feature map with M features.

A feature extractor Λ is *covariant* to a transformation group G (formally to the group action) if there exists an *input independent* transformation $\tilde{\mathcal{T}}_g$ that can align the feature maps of a transformed image with those of the original image

$$\Lambda(\mathcal{T}_g f) = \tilde{\mathcal{T}}_g(\Lambda f) \tag{4}$$

for all $g \in G$ and $f \in V$. Thus, for a covariant feature extractor it is possible to predict the feature maps of a transformed image from the feature maps of the original image.

A feature extractor Λ is *invariant* to a transformation group G if the feature representation of a transformed image is *equal to* the feature representation of the original image

$$\Lambda(\mathcal{T}_g f) = \Lambda(f) \tag{5}$$

for all $g \in G$ and $f \in V$. Invariance is thus a special case of covariance where $\tilde{\mathcal{T}}_g$ is the identity transformation.

C. Continuous model of a CNN

Let $\phi : V \to V^{M_k}$ denote a continuous CNN with k layers and M_i feature channels in layer i. Let $\theta^{(i)}$ represent the transformation between layers $i-1$ and i such that

$$(\phi^{(i)} f)(x, c) = (\theta^{(i)} \theta^{(i-1)} \cdots \theta^{(2)} \theta^{(1)} f)(x, c), \tag{6}$$

where $c \in \{1, 2, \ldots M_k\}$ denotes the feature channel and $\phi = \phi^{(k)}$. We model the transformation $\theta^{(i)}$ between two adjacent layers $\phi^{(i-1)} f$ and $\phi^{(i)} f$ as a convolution followed by the addition of a bias term $b_{i,c} \in \mathbb{R}$ and the application of a pointwise non-linearity $\sigma_i : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} &(\phi^{(i)} f)(x, c) \\ &= \sigma_i \left(\sum_{m=1}^{M_{i-1}} \int_{\xi \in \mathbb{R}^N} (\phi^{(i-1)} f)(x - \xi, m)\, g_{m,c}^{(i)}(\xi)\, d\xi + b_{i,c} \right) \end{aligned} \tag{7}$$

where $g_{m,c}^{(i)} \in L_1(\mathbb{R}^N)$ denotes the convolution kernel that propagates information from feature channel m in layer $i-1$ to output feature channel c in layer i. A final fully connected classification layer with compact support can also be modelled as a convolution combined with a non-linearity σ_k that represents *a softmax operation* over the feature channels.

D. Scale channel networks

The key idea underlying *channel networks* is to process transformed copies of an input image in parallel, in a set of network "channels" (subnetworks) with shared weights. For finite transformation groups, such as discrete rotations, using one channel corresponding to each group element and applying

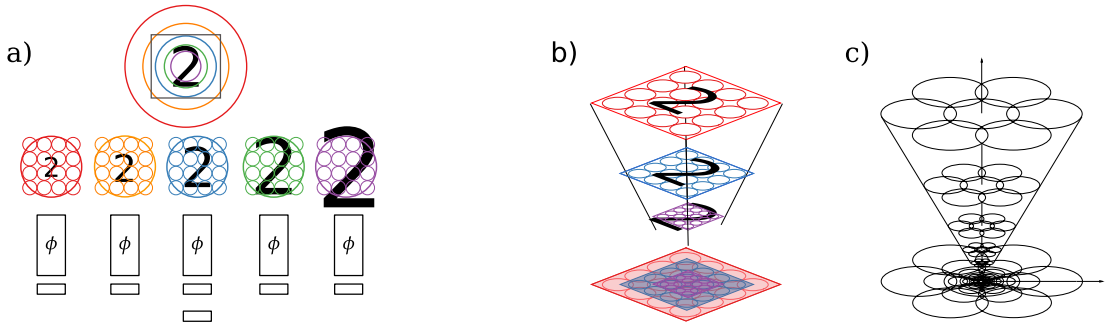


Fig. 1: *Foveated scale channel networks.* a) Foveated scale channel network that process an image of the digit 2. Since each scale channel has a fixed size receptive field/support region in the scale channels, they will together process input regions corresponding to varying sizes in the original image (circles of corresponding colors). b) This corresponds to a type of foveated processing, where the center of the image is processed with high resolution, which works well to detect small objects, while larger regions are processed using gradually reduced resolution, which enables detection of larger objects. c) There is a close similarity between this model and the foveal scale space model [22], which was motivated by a combination of regular scale space axioms with a complementary assumption of a uniform limited processing capacity at all scales.

max pooling over the channel dimension can give an invariant output code. For continuous but compact groups, invariance can instead be achieved for a discrete subgroup.

The scaling group does, however, imply additional challenges, since it is neither finite nor compact. The key question that we address here, is whether a scale channel network can still support invariant recognition.

We will define a multi-column *scale channel network* $\Lambda : V \to V^{M_k}$ for the group of scaling transformations S by using a single base network $\phi : V \to V^{M_k}$ to define a set of *scale channels* $\{\phi_s\}_{s \in S}$

$$(\phi_s f)(x, c) = (\phi \, \mathcal{S}_s f)(x, c) = (\phi f_s)(x, c), \tag{8}$$

where each channel thus applies exactly the same operation to a scaled copy of the input image (see Figure 1a). We will denote the mapping from the input image to the scale channel feature maps at depth i as $\Gamma^{(i)} : V \to V^{M_i|S|}$

$$(\Gamma^{(i)} f)(x, c, s) = (\phi_s^{(i)} f)(x, c) = (\phi^{(i)} \mathcal{S}_s f)(x, c). \tag{9}$$

A scale channel network invariant to the continuous group of uniform scaling transformations $S = \{s \in \mathbb{R}_+\}$ can be constructed using an *infinite* set of scale channels $\{\phi_s\}_{s \in S}$. The following analysis also holds for a set of scale channels corresponding to a discrete subgroup of the group of uniform scaling transformations such that $S = \{\gamma^i | i \in \mathbb{Z}\}$ for $\gamma > 1$.

The final output Λf from the scale channel network is an aggregation across the scale dimension of the last layer scale channel feature maps. In our theoretical treatment, we combine the output of the scale channels by the supremum

$$(\Lambda_{\sup} f)(x, c) = \sup_{s \in S} \left[(\phi_s f)(x, c, s) \right]. \tag{10}$$

Other permutation invariant operators such as averaging operations, could also be used. For this construction, the network output will be invariant to *rescalings around* $x_0 = 0$ *for all* x such that $(\Lambda_{\sup} f)(x, c) = (\Lambda_{\sup} S_s f)(x, c)$ (global scale invariance). This architecture is appropriate when characterising a single centered object that might vary in scale and it is the main architecture we explore in this paper. Alternatively, one may instead pool over *corresponding image points* in the original image by operations of the form

$$(\Lambda_{\sup}^{\text{local}} f)(x, c) = \sup_{s \in S} \{(\phi_s f)(S_s x, c)\} \tag{11}$$

This descriptor instead has the invariance property $(\Lambda_{\sup}^{\text{local}} f)(x_0, c) = (\Lambda_{\sup}^{\text{local}} S_{s,x_0} f)(x_0, c)$ *for all* x_0, i.e. when scaling around an arbitrary image point, the output *at that specific point* does not change (local scale invariance). This property makes it more suitable to describe scenes with multiple objects that might vary in size.

1) Scale covariance: Consider a scale channel network Λ (10) that expands the input over the group of uniform scaling transformations S. We can relate the feature map representation $\Gamma^{(i)}$ for a scaled image copy $\mathcal{S}_t f$ for $t \in S$ and its original f in terms of operator notation as

$$\begin{aligned}(\Gamma^{(i)} \mathcal{S}_t f)(x, c, s) &= (\phi_s^{(i)} \, \mathcal{S}_t f)(x, c) \\ &= (\phi^{(i)} \, \mathcal{S}_s \, \mathcal{S}_t f)(x, c) = (\phi^{(i)} \, \mathcal{S}_{st} f)(x, c) \\ &= (\phi_{st}^{(i)} f)(x, c) = (\Gamma^{(i)} f)(x, c, st),\end{aligned} \tag{12}$$

where we have used the definitions (8) and (9) together with the fact that S is a group. A scaling of an image thus only results in a multiplicative shift in the scale dimension of the feature maps. A more general and more rigorous proof using an integral representation of a scale channel network is given in Section II-E.

2) Scale invariance: Consider the scale channel network $\Lambda_{\sup}$ (10) that selects the supremum over scales. We will show that $\Lambda_{\sup}$ is scale invariant i.e. that

$$(\Lambda_{\sup} \, \mathcal{S}_t f)(x, c) = (\Lambda_{\sup} f)(x, c). \tag{13}$$

First, (12) gives $\{\phi_s^{(i)}(\mathcal{S}_t f)\}_{s \in S} = \{\phi_{st}^{(i)}(f)\}_{s \in S}$. Then, we note that $\{st\}_{s \in S} = St = S$. This holds both in the case

when $S = \mathbb{R}_+$ and in the case when $S = \{\gamma^i | i \in \mathbb{Z}\}$. Thus, we have

$$\{(\phi_s^{(i)} \mathcal{S}_t f)(x, c)\}_{s \in S} = \{(\phi_{st}^{(i)} f)(x, c)\}_{s \in S} = \{(\phi_s^{(i)} f)(x, c)\}_{s \in S}, \quad (14)$$

i.e. *the set* of outputs from the scale channels for a transformed image is equal to the set of outputs from the scale channels for its original image. For any permutation invariant aggregation operator, such as the supremum, we have that

$$(\Lambda_{\sup} \mathcal{S}_s f)(x, c) = \sup_{s \in S} \{(\phi_{st}^{(k)} f)(x, c)\} = \sup_{s \in S} \{(\phi_s^{(k)} f)(x, c)\} = (\Lambda_{\sup} f)(x, c), \quad (15)$$

and, thus, Λ is invariant to uniform rescalings.

E. Proof of scale and translation covariance using an integral representation of a scale channel network

We, here, prove the transformation property

$$(\Gamma^{(i)} h)(x, s, c) = (\Gamma^{(i)} f)(x + S_s S_t x_1 - S_t x_2, st, c) \quad (16)$$

of the scale channel feature maps under a more general combined scaling transformation and translation of the form

$$h(x') = f(x) \quad \text{for} \quad x' = S_t(x - x_1) + x_2 \quad (17)$$

corresponding to

$$h(x) = f(S_t^{-1}(x - x_2) + x_1) \quad (18)$$

using an integral representation of the deep network. In the special case when $x_1 = x_2 = x_0$, this corresponds to a uniform scaling transformation around x_0 (i.e. $S_{x_0, s}$). With $x_1 = x_0$ and $x_2 = x_0 + \delta$, this corresponds to a scaling transformation around x_0 followed by a translation $\mathcal{D}_\delta$.

Consider a deep network $\phi^{(i)}$ (6) and assume the integral representation (7), where we for simplicity of notation incorporate the offsets $b_{i,c}$ into the non-linearities $\sigma_{i,c}$. By expanding the integral representation of the rescaled image h (18), we have that that the feature representation in the scale channel network is given by (with $M_0 = 1$ for a scalar input image):

$$\begin{aligned}
&(\Gamma^{(i)} h)(x, s, c) = \{\text{definition (9)}\} = (\phi_s^{(i)} h)(x, c) \\
&= \{\text{definition (8)}\} = (\phi^{(i)} h_s)(x, c) = \{\text{equation (6)}\} \\
&= (\theta^{(i)} \theta^{(i-1)} \dots \theta^{(2)} \theta^{(1)} h_s)(x, c) = \{\text{equation (7)}\} \\
&= \sigma_{i,c} \left(\sum_{m_i=1}^{M_{i-1}} \int_{\xi_i \in \mathbb{R}^N} \sigma_{i-1,m_i} \left(\sum_{m_{i-1}=1}^{M_{i-2}} \int_{\xi_{i-1} \in \mathbb{R}^N} \dots \right. \right. \\
&\quad \sigma_{1,m_2} \left(\sum_{m_1=1}^{M_0} \int_{\xi_1 \in \mathbb{R}^N} h_s(x - \xi_i - \xi_{i-1} - \dots - \xi_1) \times \right. \\
&\quad \left. g_{m_1,m_2}^{(1)}(\xi_1)\, d\xi_1 \right) \dots g_{m_{i-1},m_i}^{(i-1)}(\xi_{i-1})\, d\xi_{i-1} \Bigg) \\
&\quad \left. g_{m_i,c}^{(i)}(\xi_i)\, d\xi_i \right). \quad (19)
\end{aligned}$$

Under the scaling transformation (17), the part of the integrand $h_s(x - \xi_i - \xi_{i-1} - \dots - \xi_1)$ transforms as follows:

$$\begin{aligned}
&h_s(x - \xi_i - \xi_{i-1} - \dots - \xi_1) \\
&= \{h_s(x) = h(S_s^{-1} x) \text{ according to definition (2)}\} \\
&= h(S_s^{-1}(x - \xi_i - \xi_{i-1} - \dots - \xi_1)) \\
&= \{h(x) = f(S_t^{-1}(x - x_2) + x_1) \text{ according to (18) }\} \\
&= f(S_t^{-1} S_s^{-1}((x - \xi_i - \xi_{i-1} - \dots - \xi_1) - S_s x_2 + S_s S_t x_1) \\
&= \{S_s S_t = S_{st} \text{ for scaling transformations}\} \\
&= f(S_{st}^{-1}((x + S_s S_t x_1 - S_s x_2 - \xi_i - \xi_{i-1} - \dots - \xi_1)) \\
&= \{f_{st}(x) = f(S_{st}^{-1} x) \text{ according to definition (2)}\} \\
&= f_{st}(x + S_s S_t x_1 - S_s x_2 - \xi_i - \xi_{i-1} - \dots - \xi_1). \quad (20)
\end{aligned}$$

Inserting this transformed integrand into the integral representation (19) gives

$$\begin{aligned}
&(\Gamma^{(i)} h)(x, s, c) = \\
&= \sigma_{i,c} \left(\sum_{m_i=1}^{M_{i-1}} \int_{\xi_i \in \mathbb{R}^N} \sigma_{i-1,m_i} \left(\sum_{m_{i-1}=1}^{M_{i-2}} \int_{\xi_{i-1} \in \mathbb{R}^N} \dots \right. \right. \\
&\quad \sigma_{1,m_2} \left(\sum_{m_1=1}^{M_0} \int_{\xi_1 \in \mathbb{R}^N} f_{st}(x + S_s S_t x_1 - S_s x_2 - \right. \\
&\quad \left. \xi_i - \xi_{i-1} - \dots - \xi_1) \times \right. \\
&\quad \left. g_{m_1,m_2}^{(1)}(\xi_1)\, d\xi_1 \right) \dots g_{m_{i-1},m_i}^{(i-1)}(\xi_{i-1})\, d\xi_{i-1} \Bigg) \\
&\quad \left. g_{m_i,c}^{(i)}(\xi_i)\, d\xi_i \right), \quad (21)
\end{aligned}$$

which we recognise as

$$\begin{aligned}
&(\Gamma^{(i)} h)(x, s, c) \\
&= (\theta^{(i)} \theta^{(i-1)} \dots \theta^{(2)} \theta^{(1)} f_{st})(x + S_s S_t x_1 - S_s x_2, c) \\
&= (\phi^{(i)} f_{st})(x + S_s S_t x_1 - S_s x_2, c) \\
&= (\phi_{st}^{(i)} f)(x + S_s S_t x_1 - S_s x_2, c) \\
&= (\Gamma^{(i)} f)(x + S_s S_t x_1 - S_s x_2, st, c) \quad (22)
\end{aligned}$$

and which proves the result. Note that for a pure translation ($S_t = I$, $x_1 = x_0$ and $x_2 = x_0 + \delta$) this gives

$$(\Gamma^{(i)} \mathcal{D}_\delta f)(x, c, s) = (\Gamma^{(i)} f)(x - S_s \delta, s, c). \quad (23)$$

Thus, translation covariance is preserved in the scale channel network but the magnitude of the spatial shift in the feature maps will depend on the scale channel.

III. Discrete scale channel networks

Discrete scale channel networks are implemented by using a standard discrete CNN as the base network ϕ. For practical applications, it is also necessary to restrict the network to include a finite number of scale channels $\hat{S} = \{\gamma^i\}_{-K_{min} \leq i \leq K_{max}}$. The input image $f : \mathbb{Z}^2 \to \mathbb{R}$ is assumed to be of finite support.

The outputs from the scale channels are, here, aggregated using e.g. max pooling

$$(\Lambda_{\max} f)(x, c) = \max_{s \in \hat{S}} \{(\phi_s f)(x, c, s)\} \tag{24}$$

or average pooling

$$(\Lambda_{\text{avg}} f)(x, c) = \text{avg}_{s \in \hat{S}} \{(\phi_s f)(x, c, s)\}. \tag{25}$$

We will also implement discrete scale channel networks that concatenate the outputs from the scale channels followed by an additional transformation $\varphi : \mathbb{R}^{M_i |\hat{S}|} \rightarrow \mathbb{R}^{M_i}$ that mixes the information from the different channels

$$\begin{aligned} &(\Lambda_{\text{conc}} f)(x, c) \\ &= \varphi \left([(\phi_{s_1} f)(x, c), (\phi_{s_2} f)(x, c) \cdots (\phi_{s_{|\hat{S}|}} f)(x, c)] \right). \end{aligned} \tag{26}$$

Λ_{conc} does not have any theoretical guarantees of invariance, but since scale concatenation of outputs from the scale channels has been previously used with the explicit aim of scale invariant recognition [16], we will evaluate it also here.

A. Foveated processing

A standard convolutional neural network ϕ has a finite support region Ω in the input. When rescaling an input image of fixed size/finite support in the scale channels, it is necessary to decide how to process the resulting images of varying size using a feature extractor with fixed support. One option is to process regions of *constant size* in the scale channels corresponding to regions of *different sizes* in the input image. This results in *foveated image operations*, where a smaller region around the center of the input image is processed with high resolution, while gradually larger regions of the input image are processed with gradually reduced resolution (see Figure 1b-c). We will refer to the foveated network architectures $\Lambda_{\max}$, Λ_{avg} and Λ_{conc} as the FovMax network, the FovAvg network and the FovConc network respectively.

B. Approximation of scale invariance

Foveated processing combined with max or average pooling will give an approximation of the scale invariance in the continuous model (Section II-D2) over *a limited scale range*. The numerical scale warpings of the input images in the scale channels approximate continuous scaling transformations. A discrete set of scale channels will approximate the representation for a continuous scale parameter. A possible issue is problems at the scale boundaries. Boundary effects can, however, be mitigated if the network learns to suppress responses for both very zoomed in and very zoomed out objects. If including a large enough number of scale channels and training the network from scratch, this is, in fact, a likely scenario, since the network will otherwise classify based on object views that hardly provide useful information.

C. Sliding window processing in the scale channels

An alternative option for dealing with varying image sizes is to, in each scale channel, process the entire rescaled image by applying the base network in *a sliding window manner*. The output from the scale channels can then be combined by max (or avarage) pooling over space followed by max (or average) pooling over scales

$$(\Lambda_{sw,\max} f)(c) = \max_{s \in S} \max_{x \in \Omega_s} \{(\phi_s f)(x, c, s)\}, \tag{27}$$

where $\Omega_s = \{sx | x \in \Omega\}$. We will here only evaluate the architecture using max pooling, which is structurally similar to the popular multi-scale OverFeat detector [13]. We refer to this network as the SWMax network. For this scale channel network to support invariance, it is not enough that boundary effects resulting from using a finite number of scale channels are mitigated. When processing regions in the scale channels corresponding to only a single region in the input image, new structures can appear (or disappear) in this region when objects appear at a new scale. For a linear approach, one would not expect e.g. partial views of zoomed in objects to cause problems. For a non-linear method, however, this could result in strong erroneous responses. We are, here, interested in how this affects scale generalisation in deep neural networks.

IV. Experiments

A. The MNISTLargeScale dataset

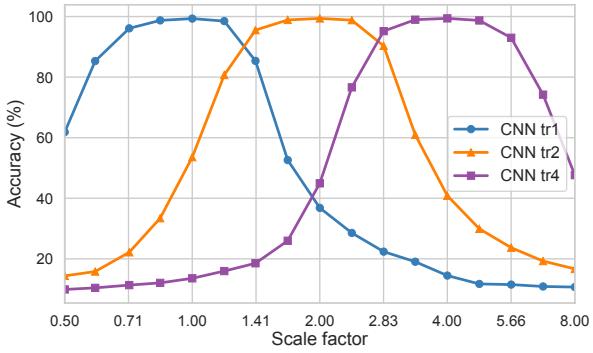
To evaluate the ability of vanilla CNNs and scale channel networks to generalise to unseen scales over *a wide scale range*, we have created a new version of the standard MNIST dataset [23]. This new dataset *MNISTLargeScale*, which is available online [24], is composed of images of size 112×112 with scale variations of a factor 16 for scale factors $s \in [0.5, 8]$ relative to the original images. The train and test sets for the different scale factors are created by resampling the original MNIST training and test sets using bilinear interpolation followed by smoothing and soft thresholding to reduce discretization effects. Note that for scale factors > 4, the full digit might not be visible in the image. These scale values are nonetheless included to study the limits of generalisation. More details on the dataset creation and numerical performance scores for the experiments are given in [25].

B. Network and training details

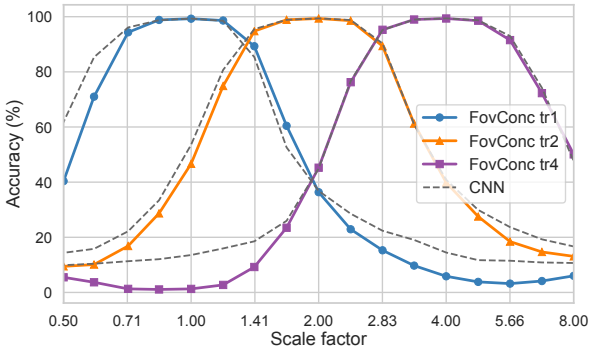
The baseline CNN is composed of 8 conv-batchnorm-ReLU blocks followed by a fully connected layer and a final softmax layer. The number of features/filters in each layer is 16-16-16-16-32-32-32-32-100-10. A stride of 2 is used in convolutional layers 2, 4, 6 and 8. The reason for using a quite deep network is to avoid a network structure that is heavily biased towards recognising either small or large digits.

The FovMax, FovAvg, FovConc and SWMax[1] *scale channel networks* are constructed using scale channels with 4 conv-

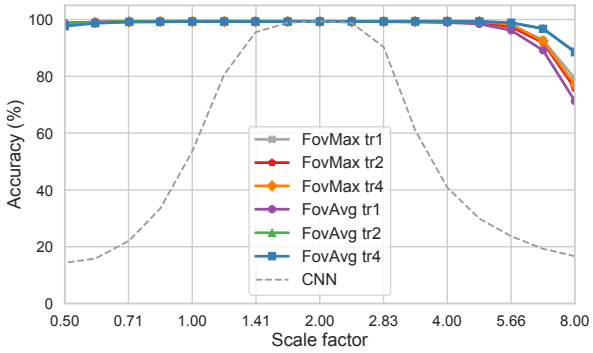
[1]We noted that batchnorm impairs performance when training the SWMax network from scratch. We believe this is because the sliding window approach implies in a change in the feature distribution when processing data of different scales. We, therefore, train the SWMax network without batchnorm.



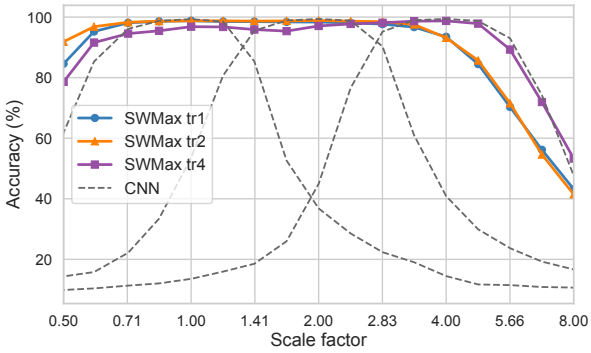
(a) The baseline CNN



(b) The FovConc network



(c) The FovMax and FovAvg networks



(d) The SWMax network

Fig. 2: *Generalisation ability to unseen scales for the baseline CNN and the different scale channel network architectures*. The networks are trained on digits of scale 1 (tr1), scale 2 (tr2) or scale 4 (tr4) and evaluated for varying rescalings of the test set. We note that the CNN (a) and the FovConc network (b) have poor generalisation ability to unseen scales, while the FovMax and FovAvg networks (c) generalise extremely well. The SWMax network (d) generalises considerably better than the CNN, but there is some drop in performance for scales not seen during training.

batchnorm-ReLU blocks followed by a fully connected layer and a final softmax layer. Batchnorm layers are *shared* across the scale channels. The number of features/filters in each layer is 16-16-32-32-100-10. A stride of 2 is used in convolutional layers 2 and 4. All scale channel architectures have ~70 000 parameters, while the CNN has ~90 000 parameters.

All networks are trained with 50 000 training samples from the MNISTLargeScale dataset for 20 epochs using the Adam optimiser. During training, 15 % dropout is applied to the first fully connected layer. The learning rate starts at $3e^{-3}$ and decays with a factor $1/e$ every second epoch towards a minimum learning rate of $5e^{-5}$. Results are reported for the MNISTLargeScale test set (10 000 samples) as the average of training each network using three different random seeds. The remaining 10 000 samples constitute a validation set.

C. Generalisation to unseen scales

We, first, evaluate the ability of the baseline CNN and the different scale channel networks to generalise to previously unseen scales. We train each network on each of the scales 1, 2, and 4 from the MNISTLargeScale training sets and evaluate the performance on the test sets with scale factors between $1/2$ and 8. The FovMax, FovAvg and SWMax networks have 17 scale channels spanning the scale range $[\frac{1}{2}, 8]$. The FovConc network has 3 scale channels spanning the scale range $[1, 4]$.[2] The results are presented in Figure 2. We note that all networks achieve similar top performance for the scales seen during training. There are, however, large differences in the abilities of the networks to generalise to unseen scales:

1) The baseline CNN: The baseline CNN shows limited generalisation ability to unseen scales with a large drop in accuracy for scale variations larger than a factor $\sqrt{2}$. This illustrates that, while the network can recognise digits of all sizes, a vanilla CNN includes no structural prior to promote scale invariance.

2) The FovConc network: The generalisation ability of the FovConc network is quite similar to that of the baseline CNN, sometimes slightly worse. The reason for limited gen-

[2]The FovConc network performs considerably worse when including too many scale channels or spanning a too large scale range. Since we are more interested in the best case rather than the worst case scenario, we, here, picked the best network out of a large range of configurations.

eralisation is that although the scale channels share weights, when simply concatenating the outputs from the scale channels there is no structural constraint to support invariance. This is consistent with our observation that spanning a too large scale range or using too many channels degrades generalisation for the FovConc network. For scales *not present during training*, there is, simply, no useful training signal to learn the correct weights in the fully connected layers that combine the scale channel outputs. Note that our results are not contradictory to those previously reported for a similar network structure [16], since they train on data that contain natural scale variations and test over a quite narrow scale range. What we do show, however, is that this network structure is *not scale invariant*.

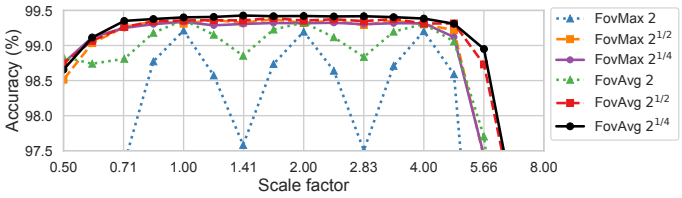


Fig. 3: *Varying the sampling density of the scale channels.* FovMax and FovAvg networks spanning the scale range $[\frac{1}{4}, 8]$ are trained with varying spacing between the scale channels (2, $2^{1/2}$ and $2^{1/4}$). All networks are trained on scale 2.

3) The FovAvg and FovMax networks: We note that the FovMax and FovAvg networks generalise very well, independently of which scale the network is trained on. The maximum difference in performance in the scale range $[1, 4]$ between training on scale 1, scale 2 or scale 4 is less than 0.2 percentage points for these network architectures. Importantly, this shows that, if including a large enough number of scale channels and training the networks from scratch, boundary effects at the scale boundaries do not prohibit invariant recognition. For the FovAvg and FovMax networks, we also investigate how densely it is necessary to sample the scale channels for good performance. The result is presented in Figure 3. Accuracy is considerably improved when decreasing the distance between consecutive channels from a factor 2 (5 channels) to a factor of $2^{1/2}$ (9 channels), while a further reduction to $2^{1/4}$ (17 channels) provide very small additional benefits.

4) The SWMax network: We note that the SWMax network generalises considerably better than the baseline CNN, but there is some drop in performance for scales not seen during training. We believe that the reason for this is that since all scale channels are processing a fixed sized region in the input (as opposed to for foveated processing), new structures can leave or enter this region when an object is rescaled. This can give erroneous high responses for unfamiliar views (Section III-C). We also noted that the SWMax networks are harder to train (more sensitive to learning rate etc) compared to the foveated network architectures as well as more computationally expensive. Thus, the SWMax network seems to work best for spanning a more limited scale range where fewer scale channels are needed (as was indeed the use case in [13]).

D. Multiscale vs. single scale training

All the scale channel architectures support multiscale processing although they might not support scale invariance. We, here, test the performance of the different scale channel networks when training on multiscale training data spanning the scale range [1,4]. The results are presented in Figure 4. The difference between training on multiscale and single scale data is striking for the baseline CNN and the FovConc network. It can, however, be noted that the FovConc network does generalise slightly better to unseen scales than the baseline CNN. For the SWMax network, including multiscale data improves generalisation somewhat for larger scales but impairs generalisation somewhat for smaller scales. The difference in generalisation ability between training on a single scale or multiscale image data is almost indiscernible for the FovMax and FovAvg networks.

E. Generalisation from fewer training samples

Another scenario of interest is when the training data does span a relevant range of scales, but there are few training samples. Theory would predict a correlation between the performance in this scenario and the ability to generalise to unseen scales. To test this prediction, we trained the baseline CNN and the different scale channel networks on multi scale training data spanning the scale range $[1, 4]$, while gradually reducing the number of samples in the training set. Here, the same scale channel setup with 17 channels spanning the scale range $[\frac{1}{2}, 8]$ is used for all the architectures. The results are presented in Figure 5. We note that the FovConc network shows some improvement over the baseline CNN. The SWMax network, on the other hand, does not, and we hypothesise that when using fewer samples, the problem with partial views of objects (see Section III-C) might be more severe. Note that the way the OverFeat detector is used is the original study [13], is more similar to our single scale training scenario, since they use base networks pretrained on ImageNet. The FovAvg and FovMax networks show the highest robustness also in this scenario. This illustrates that these networks can give improvements when multiscale training data is available but there are few training samples.

V. Summary and Conclusions

We have presented a theoretical analysis of covariance and invariance properties of continuous scale channel networks. Moreover, we have performed an experimental evaluation of different types of discrete scale channel networks on the task of generalising to unseen scales over wide scale ranges. The tested networks include a new family of scale channel networks that combine foveated processing with max or average pooling over the scale channels (the FovMax and FovAvg networks). The experimental evaluation illustrates the strong invariance properties of these networks in practice and limitations of previous approaches and vanilla CNNs. We believe that our proposed foveated scale channel networks will prove useful in situations where a simple approach that can generalise to unseen scales or learn from small datasets

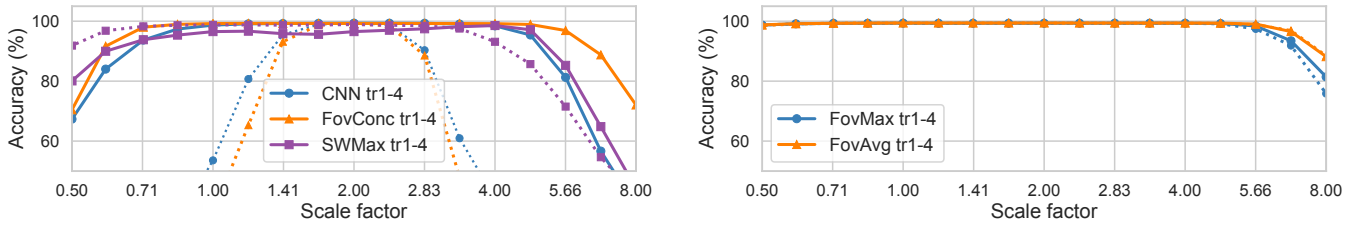


Fig. 4: *Multiscale image data.* All networks are trained on digits in the scale range $[1, 4]$ (tr1-4) and evaluated for varying scale factors in the test set. The difference in generalisation ability between training on multiscale and single scale data (dotted lines) is very large for both the CNN and the FovConc network. For the FovMax and FovAvg networks, the difference is negligible between multiscale and single scale training, which illustrates the strong invariance properties of these networks.

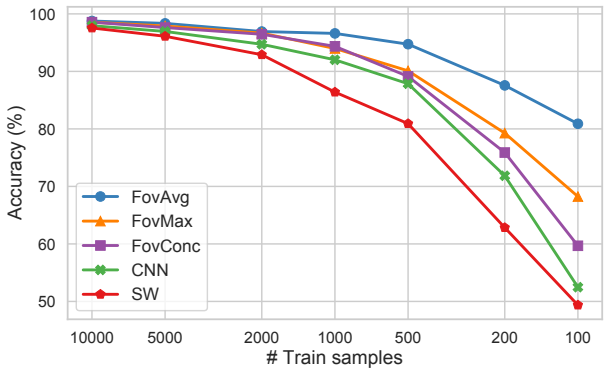


Fig. 5: *Training with smaller training sets with large scale variations*. All network architectures are evaluated on their ability to classify data with large scale variations while reducing the number of training samples. Both the training and test set here span the scale range $[1, 4]$. The FovAvg network shows the highest robustness when decreasing the number of training samples followed by the FovMax network.

with large scale variations is needed. This type of foveated scale invariant processing could also be included as subparts in more complex frameworks dealing with large scale variations. A more overarching aim of this study have been to test the limits of CNNs to generalise to unseen scales over a wide scale range. The key take home message is a proof of concept that such generalisation is possible if including structural assumptions about scale in the network design.

References

[1] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[2] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International Conference on Machine Learning*, 2016, pp. 2990–2999.

[3] D. Laptev, N. Savinov, J. M. Buhmann, and M. Pollefeys, "TI-pooling: transformation-invariant pooling for feature learning in convolutional neural networks," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 289–297.

[4] A. P. Witkin, "Scale-space filtering," in *Proc. 8th Int. Joint Conf. on Artificial Intelligence.*, Karlsruhe, Germany, Aug. 1983, pp. 1019–1022.

[5] J. J. Koenderink, "The structure of images," *Biological Cybernetics*, vol. 50, pp. 363–370, 1984.

[6] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *Journal of Applied Statistics*, vol. 21, no. 1-2, pp. 225–270, 1994.

[7] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International Conference on Machine Learning (ICML)*, 2019, pp. 1802–1811.

[8] A. Fawzi and P. Frossard, "Manitest: Are classifiers really invariant?" *British Machine Vision Conference (BMVC)*, 2015.

[9] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection SNIP," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3578–3587.

[10] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 354–370.

[11] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 2017–2025.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Int. Conf. on Learning Representations (ICLR)*, 2016.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[14] R. Girshick, "Fast R-CNN," in *Proc. International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[16] Y. Xu, T. Xiao, J. Zhang, K. Yang, and Z. Zhang, "Scale-invariant convolutional neural networks," *arXiv preprint arXiv:1411.6369*, 2014.

[17] A. Kanazawa, A. Sharma, and D. W. Jacobs, "Locally scale-invariant convolutional neural networks," *arXiv preprint arXiv:1412.5104*, 2014.

[18] D. Marcos, B. Kellenberger, S. Lobry, and D. Tuia, "Scale equivariance in CNNs with vector fields," *arXiv preprint arXiv:1807.11783*, 2018.

[19] D. Worrall and M. Welling, "Deep scale-spaces: Equivariance over scale," in *Advances in Neural Information Processing Systems (NIPS)*, 2019, pp. 7364–7376.

[20] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[21] N. Van Noord and E. Postma, "Learning scale-variant and scale-invariant features for deep image classification," *Pattern Recognition*, vol. 61, pp. 583–592, 2017.

[22] T. Lindeberg and L. Florack, "Foveal scale-space and linear increase of receptive field size as a function of eccentricity," Dept. of Numerical Analysis and Computer Science, KTH, report ISRN KTH/NA/P--94/27--SE, Aug. 1994, available from https://people.kth.se/~tony/papers/cvap166.pdf.

[23] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] Y. Jansson and T. Lindeberg. MNISTLargeScale dataset. [Online]. Available at: https://www.zenodo.org/record/3820247. DOI:10.5281/zenodo.3820247.

[25] ——, "Exploring the ability of CNNs to generalise to previously unseen scales over wide scale ranges," *arXiv preprint arXiv:2004.01536*, 2020.