EXPLOITING REDUNDANCY: SEPARABLE GROUP CONVOLUTIONAL NETWORKS ON LIE GROUPS

Anonymous authors
Paper under double-blind review

ABSTRACT

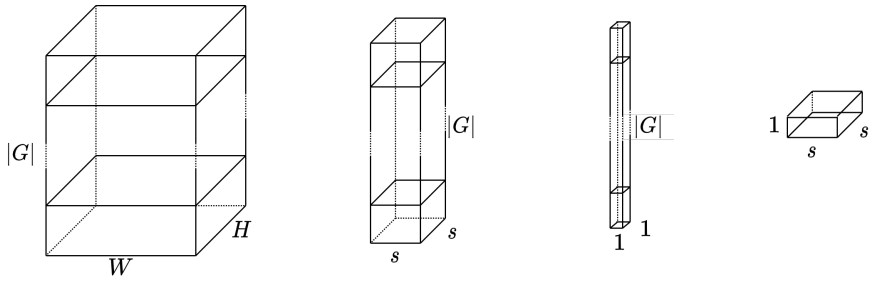
Group convolutional neural networks (G-CNNs) have been shown to increase parameter efficiency and model accuracy by incorporating geometric inductive biases. In this work, we investigate the properties of representations learned by regular G-CNNs, and show considerable parameter redundancy in group convolution kernels. This finding motivates further weight-tying by sharing convolution kernels over subgroups. To this end, we introduce convolution kernels that are separable over the subgroup and channel dimensions. In order to obtain equivariance to arbitrary affine Lie groups we provide a continuous parameterisation of separable convolution kernels. We evaluate our approach across several vision datasets, and show that our weight sharing leads to improved performance and computational efficiency. In many settings, separable G-CNNs outperform their non-separable counterpart, while only using a fraction of their training time. In addition, thanks to the increase in computational efficiency, we are able to implement G-CNNs equivariant to the $\mathrm{Sim}(2)$ group; the group of dilations, rotations and translations of the plane. $\mathrm{Sim}(2)$-equivariance further improves performance on all tasks considered.

1 INTRODUCTION

Minsky & Papert (1988) suggest that the power of the perceptron comes from its ability to *learn to discard* information irrelevant to the problem it is tasked with solving. In other words; information that is not relevant to the task does not influence the representation built by the network. According to Minsky & Papert (1988), this leads to a definition of perceptrons in terms of the *symmetry groups* their learned representations are invariant to. Progress in geometric deep learning has shown the power of pro-actively equipping models with such geometric structure as inductive bias, reducing model complexity and improving generalisation and performance (Bronstein et al., 2017). An early example of such geometric inductive bias at work can be seen in the convolutional layer in a CNN (LeCun et al., 1998). CNNs have been instrumental in conquering computer vision tasks, and much of their success has been attributed to their use of the convolution operator, which commutes with the action of the translation group. This property, known as *equivariance* to translation, comes about as a result of the application of the same convolution kernel throughout an input signal, enabling the CNN to learn to detect the same features at any location in the input signal, directly exploiting translational symmetries that naturally occur in many tasks.

Although invariance to object-identity preserving transformations has long been recognised as a desirable model characteristic in machine learning literature (Kondor, 2008; Cohen, 2013; Sifre & Mallat, 2014), only recently Cohen & Welling (2016a) introduced the Group Equivariant CNN (G-CNN) as a natural extension of the CNN (LeCun et al., 1998), generalising its equivariance properties to group actions beyond translation. It has since been shown that equivariant deep learning approaches may serve as a solution in fields that as of yet remain inaccessible to machine learning due to scarce availability of labelled data (Winkels & Cohen, 2018; Linmans et al., 2018; Bekkers, 2019), or when compact model design due to limited computational power is required. The layers of a G-CNN are explicitly designed to be equivariant to these transformations, hence the model is no longer burdened with *learning* invariance to transformations that leave object identity intact.

Complexity and redundancy issues impeding regular group convolutions A growing body of work shows applications of G-CNNs consistently and decisively outperforming classical CNNs (Worrall et al., 2017; Weiler et al., 2018a; Bekkers et al., 2018; Esteves et al., 2018; Bekkers,



a) $f : G \to \mathbb{R}$ **b)** $k : G \to \mathbb{R}$ **c)** $k_H : H \to \mathbb{R}$ **d)** $k_{\mathbb{R}^2} : \mathbb{R}^2 \to \mathbb{R}$

Figure 1: In group convolutions on affine Lie groups, a feature map f defined over the group, depicted in figure **a)** is convolved with a filter k shown in figure **b)** of size $|G| \times s \times s$, where s is the kernel size, and $|G|$ the size of the group. We propose factorising this convolution into a convolution over the group with a kernel k_H shown in figure **c)** with size $|G| \times 1 \times 1$, followed by a convolution over the spatial dimensions with a kernel $k_{\mathbb{R}^2}$ shown in figure **d)** with size $1 \times s \times s$.

2019; Worrall & Welling, 2019; Sosnovik et al., 2021b). However, a practical challenge impeding application to larger groups is the computational complexity of regular group convolutions, which scales exponentially with the dimensionality of the group. Furthermore, Lengyel & van Gemert (2021) show that group convolution filters in the original formulation of the G-CNN by Cohen & Welling (2016a) exhibit considerable redundancies along the group axis for the $p4m$ and $\mathbb{Z}^2$ groups. Similar observations motivated depthwise separable convolutions (Chollet, 2017), which not only increased parameter efficiency but also model performance; observed correlations between weights are explicitly enforced with further parameter sharing through the use of kernels separable along spatial and channel dimensions. We address the observations of redundancy along with the scalability issues of regular G-CNNs in their current form. Our paper contains the following contributions:

- We introduce separable group convolutions for affine Lie groups $\mathbb{R}^n \rtimes H$, sharing the kernels for translation elements $x \in \mathbb{R}^n$ along subgroup elements $h \in H$. See figure 1 for an overview.
- We propose the use of a SIREN (Sitzmann et al., 2020) as kernel parameterisation in the Lie algebra - imposing a fixed number of parameters per convolution kernel, regardless of the resolution at which this kernel is sampled, and ensuring smoothness over the Lie group.
- Separable group convolutions allow us to build $\mathrm{Sim}(2)$-equivariant G-CNNs, which we thoroughly experiment with. We show equivariance to this group increases model performance over a range of vision benchmarks.
- To achieve equivariance to continuous affine Lie groups, we propose a random sampling method over H for approximating the group convolution operation.

First, we position this work within the area of equivariant deep learning by giving an overview of related works, and explaining which current issues we are addressing with this work. We derive separable group convolutions, and show how they may be applied to continuous groups. Lastly, we apply these ideas in practice by experimenting with implementations for roto-translations in 2D ($\mathrm{SE}(2)$), dilation and translation in 2D ($\mathbb{R}^2 \rtimes \mathbb{R}^+$) and dilation, rotation and translation in 2D ($\mathrm{Sim}(2)$).

2 Related Work

Group Equivariant Convolutional Neural Networks Broadly speaking, research on G-CNNs can be divided into two approaches. The first, the *regular* G-CNN, uses the left-regular representation of the group of interest to learn representations of scalar functions over the group manifold, or a quotient space of the group. We let the left-regular representation act on the convolution kernels, yielding an orbit of the kernel under the group action. Convolving the input using these transformed filters, we obtain a feature map defined over the group at each layer. This approach most naturally extends the conventional CNN, in which each convolution kernel is transformed under elements of the translation group. Regular G-CNNs have been implemented for both discrete groups (Cohen & Welling, 2016a; Winkels & Cohen, 2018; Worrall & Brostow, 2018), compact continuous groups (Marcos et al., 2017;

Bekkers et al., 2018) and arbitrary non-compact continuous Lie groups (Bekkers, 2019; Finzi et al., 2020; Romero et al., 2020). However, practical implementations for continuous groups often require some form of discretisation of the group, possibly introducing discretisation artefacts, and requiring a choice of *resolution* over the group. In the second approach, that of *steerable* G-CNNs, representation theory is used to compute a basis of equivariant functions for a given group, which are subsequently used to parameterise convolution kernels in (Cohen & Welling, 2016b; Weiler et al., 2018a;b; Sosnovik et al., 2019; 2021a). Although steerable G-CNNs decouple the cardinality of the group from the dimensionality of the feature maps, this approach is only compatible with compact groups.

The current paper may, in approach, be compared to recent works by Bekkers (2019) and Finzi et al. (2020), who define convolution kernels on the algebra corresponding to a Lie group of interest, to enable convolutions on the manifold of continuous, non-compact groups. We similarly make use of the Lie algebra and exponential map to obtain convolution kernels on the group, but separate the kernels by subgroup.

Bekkers (2019) defines a set of basis vectors in the Lie algebra, which, when combined with the exponential map, allow for the identification of group elements by a vector in $\mathbb{R}^n$. Subsequently, a set of B-splines is defined on the algebra, which form a basis to expand convolution kernels in. A linear combination of these bases creates a locally continuous function on the Lie algebra defining a convolution kernel and its behaviour under transformations of the group. Although this method allows for direct control over kernel smoothness, the learned convolution filters are limited in their expressivity by their basis functions. Finzi et al. (2020) instead use an MLP to learn this same continuous function of convolution kernels on the Lie algebra, while extending the group convolution operator to handle point cloud data, whereas the convolutional layer is generally defined for data lying on a regular grid. The MLP is constructed to learn kernel values at (arbitrary) relative offsets in the Lie algebra. We propose using SIRENs (Sitzmann et al., 2020) to parameterise convolution kernels, as they offer more explicit control over kernel smoothness, which is a desirable property for addressing discretisation artefacts that occur in modelling features on continuous groups.

Separable filters in machine learning In image processing, spatially separable filters have long been known to increase parameter- and computational efficiency, and learning such constrained filters may even increase model performance (Rigamonti et al., 2013). In Sifre & Mallat (2014), authors investigate SE(2)-invariant feature learning through scattering convolutions, and propose separating the group convolution operation for affine groups into a cascade of two filterings, the first along the spatial dimensions $\mathbb{R}^n$, and the second along subgroup dimension H. From this, authors derive a separable approach to the convolution operation with learnable filters as used in CNNs. This formulation has since been named the *depthwise-separable* convolution (Chollet, 2017), a special case of the Network-In-Network principle (Lin et al., 2013) which forms the basis for the success of the Inception architectures (Szegedy et al., 2015). In depthwise separable convolutions, each input channel first gets convolved using a (set of) kernel(s) with limited spatial support. Afterwards, a 1x1 convolution is used to project the feature set detected in the input channels to the output space. Chollet (2017) speculates that the Inception architectures are successful due to the explicit separation of spatial and channel mapping functions, whereas in conventional CNNs, kernels are tasked with simultaneously mapping inter-channel and spatial correlations.

Haase & Amthor (2020) argue that the original formulation of the depthwise-separable convolution reinforces inter-kernel correlations, but does not in fact leverage intra-kernel correlations. Subsequently, they propose an inverse ordering of the operations given in depthwise-separable convolutions, sharing the same spatial kernel along the input channels, and show convincing results. Extending this investigation of learned convolution filters to the original G-CNN proposed by Cohen & Welling (2016a), Lengyel & van Gemert (2021) remark on the high degree of correlation found among filters along the rotation axis, and propose to share the same spatial kernel for every rotation feature map. We attempt to generalise this approach, proposing separable convolutions on arbitrary affine Lie groups.

3 Background

In the following section we give the theoretical background for implementing group convolutions for arbitrary affine Lie groups (Bekkers, 2019). We assume familiarity with the basics of group theory. For those not familiar, see appendix A.1 for a concise introduction of the relevant concepts. For simplicity of notation, we initially assume that our input signal/feature map f is defined over a single channel.

Lifting Convolutions To preserve information on the pose of features in the input, an equivariant convolution operation is achieved by *lifting* a function from the input space to (a homogeneous space of) the group. As we are interested in signal data, and specifically image data living on $\mathbb{R}^2$, we assume the Lie group of interest H is taken in semidirect product with the domain of our data; $G = \mathbb{R}^2 \rtimes H$. In group convolutions, a given kernel is left-translated by all transformations in G, thereby generating a signal on the higher dimensional space G instead of $\mathbb{R}^2$. The output feature maps thus disentangle poses (e.g. positions plus rotations or scales) through a domain expansion. For a given group element $h \in H$, kernel k, and location $\vec{x}$ in the input domain $\mathbb{R}^2$, the lifting convolution is given by:

$$(f *_{\text{lifting}} k)(g) = \int_{\mathbb{R}^2} f(\tilde{\boldsymbol{x}}) k_h(\tilde{\boldsymbol{x}} - \boldsymbol{x}) \mathrm{d}\tilde{\boldsymbol{x}}, \tag{1}$$

where $g = (\boldsymbol{x}, h)$ and $k_h = \frac{1}{|\det h|}\mathcal{L}_h(k)$ is the kernel $k : \mathbb{R}^2 \to \mathbb{R}$ transformed via the action of group element h via $[\mathcal{L}_h k](\boldsymbol{x}) := k(h^{-1}\boldsymbol{x})$, and with $\det h$ the determinant of the matrix representation of h that acts on $\mathbb{R}^d$. The output of lifting convolutions yields a G-feature map with the original two spatial input dimensions ($\mathbb{R}^2$), and an additional group dimension (H).

Group convolutions Now that our data is lifted to the domain of the group, we continue with group convolutions in subsequent layers. Given a kernel k (now a function on G), Haar measures $\mathrm{d}\tilde{g}$ and $\mathrm{d}\tilde{h}$ on the group G and sub-group H respectively, group convolutions are given by

$$\begin{aligned}(f *_{\text{group}} k)(g) &= \int_G f(\tilde{g}) k(g^{-1} \cdot \tilde{g}) \mathrm{d}\tilde{g} = \int_G f(\tilde{g}) \mathcal{L}_g k(\tilde{g}) \mathrm{d}\tilde{g} \\ &= \int_{\mathbb{R}^2} \int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) \mathcal{L}_x \mathcal{L}_h k(\tilde{\boldsymbol{x}}, \tilde{h}) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}} \mathrm{d}\tilde{h} \\ &= \int_{\mathbb{R}^2} \int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) k(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x}), h^{-1} \cdot \tilde{h}) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}} \mathrm{d}\tilde{h}.\end{aligned} \tag{2}$$

If we evaluate this convolution for every group element G, we again obtain a function defined on G. As we know $G = \mathbb{R}^2 \rtimes H$, we can factor this operation into a transformation of a kernel k by a group element $h \in H$, $k_h = \mathcal{L}_h(k)$, followed by a convolution at every spatial location in our feature map.

Achieving invariance Using the lifting and group convolution operations, we are able to construct convolutional layers that co-vary with the action of the group, explicitly preserving pose information throughout the representations created in the network. In most cases, we would ultimately like to create representations that are invariant to transformations in the input, as we want our model to be invariant to these identity-preserving transformations. This may be achieved by aggregating the information at all group elements in a feature map using an operation invariant to the group action, such as max- or mean-projection or simply integrating over the group. In practice, this is often done after the last group convolution operation, and followed by one or more fully connected layers.

4 Separable Continuous Kernel Group Convolutions

Redundancies in group convolution filters Similar to Haase & Amthor (2020); Lengyel & van Gemert (2021) we perform an analysis of the parameter efficiency of learned convolution kernels to motivate separable filters. We train an $\mathrm{SE}(2)-$equivariant CNN on the Galaxy10 dataset (see 5 for experimental details) and analyse the resulting group convolution kernels. We apply PCA by treating the values of the group convolution kernel for each subgroup element $h \in H$ as distinct spatial kernels with $k * k$ features. The ratio of variance explained by the first principle component gives an indication of the variability of the group convolution kernel along the subgroup axis. If the ratio of explained variance is high, the distinct spatial convolution kernels along the subgroup axis are well-characterised by a single shared kernel. In figure 2, results are shown for the group convolution layers contained in our $\mathrm{SE}(2)$-equivariant CNN before and after training. We find that during the training process, redundancy along the subgroup axis increases considerably. This motivates sharing a single spatial kernel along the subgroup elements, which can be achieved by separating the group convolution operation into a convolution over the subgroup H, followed by a convolution over the spatial domain $\mathbb{R}^2$.

Separable group convolutions To separate the group convolution operation over G given in equation 2 into separate convolutions over $\mathbb{R}^2$ and H, let us assume that convolution kernel $k : G \to \mathbb{R}$ is separable $k(g){=}k_{\mathbb{R}^2}(\boldsymbol{x})k_H(h)$. We can derive the factorised separable group

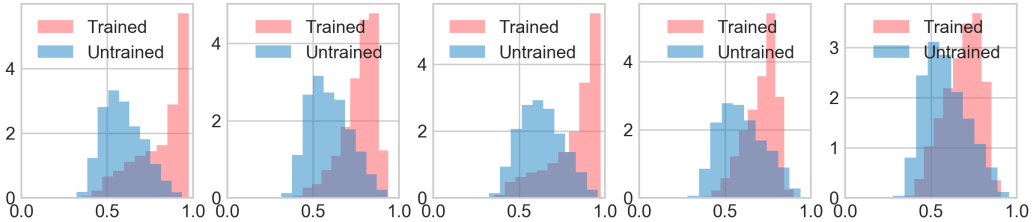


Figure 2: A set of histograms showing redundancy in learned group convolution kernels. On the x-axis is the ratio of variance explained by the first principal component when applying PCA on the set of spatial kernels along the group axis of a group convolution kernel. Y-axis shows the proportion of kernels with this explained variance ratio, where all bins sum to 1. Throughout the training process, redundancy in the group convolution kernels increases. Left to right: subsequent layers in the network.

convolution as (see appendix A.2 for the full derivation):

$$
\begin{aligned}
(f *_{group} k)(g) &= \int_{\mathbb{R}^2} \int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) k(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x}), h^{-1} \cdot \tilde{h}) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}} \mathrm{d}\tilde{h} \\
&= \int_{\mathbb{R}^2} \left[\int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) k_H(h^{-1} \cdot \tilde{h}) \mathrm{d}\tilde{h} \right] k_{\mathbb{R}^2}(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x})) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}}. \qquad (3)
\end{aligned}
$$

Here k_H is a convolution kernel over the group H, and $k_{\mathbb{R}^2}$ is a convolution kernel over the spatial domain $\mathbb{R}^2$. We set $k^\theta = k_H^\theta k_{\mathbb{R}^2}^\theta$, with k_H^θ constant along $\mathbb{R}^2$ and $k_{\mathbb{R}^2}^\theta$ constant along H. This can be thought of as parameterising a convolution kernel k over the group G by sharing same spatial kernel $k_{\mathbb{R}^2}^\theta$ weighed by a value $k_H^\theta(h)$ at every input group element $h \in H$. Importantly, this factorisation greatly increases the efficiency of the group convolution operation. Since we can precompute the inner integral in equation 3, convolving over a single channel of a feature map f of size $|H| \times x \times y$ goes from $O(|H|^2 \times x \times y \times k^2)$ to $O(|H| \times x \times y \times (|H| + k^2))$.

Defining convolution kernels on Lie algebras In order to perform lifting- and group convolutions given in equations 1 and 2, we are required to evaluate our convolution kernels k at arbitrary relative offsets g' on the group. As in Bekkers et al. (2017); Weiler et al. (2018b); Bekkers (2019); Finzi et al. (2020), we express our group convolution kernel k in an analytical form, as a function of relative group elements yielding kernel values. Motivated by findings in Romero et al. (2021), we use a Sinusoidal Representation Network (SIREN) as parameterisation, which offers explicit control of kernel smoothness through a frequency multiplier parameter ω_0. Since the affine Lie group of interest can be non-euclidean, and neural networks are functions generally defined over euclidean spaces, we resort to defining the kernel function on the Lie algebra of our group of interest. Thus, our kernel function k_θ maps from points in the Lie algebra (which may be associated with the relative offsets on the group we are convolving over by the exponential map) to kernel values, $k^\theta : \mathfrak{g} \to \mathbb{R}$. For a given group element g, we have: $k^\theta(g) = \mathrm{SIREN}^\theta(\log(\mathrm{g}))$.

Approximating equivariance for compact continuous and non-compact continuous groups In the case of small discrete subgroups H, such as the $p4m$ group of rotations by 90 degrees, we are able to perform lifting- and group convolution operations which are exactly equivariant to the action of the group, as the integral given in equation 2 is tractable and computationally feasible. In the current paper we consider image data in particular, which is defined over a discrete grid $\mathbb{Z}^2$, a subgroup of $\mathbb{R}^2$. This makes both integrals over the group discrete sums:

$$
(f *_{\text{group}} k)(g) = \sum_{\tilde{\boldsymbol{x}} \in \mathbb{Z}^2} \sum_{\tilde{h} \in H} f(\tilde{\boldsymbol{x}}, \tilde{h}) k(\mathcal{T}_{h^{-1}}(\tilde{\boldsymbol{x}} - \boldsymbol{x}), h^{-1} \cdot \tilde{h}) \frac{1}{|h|} \Delta\tilde{\boldsymbol{x}} \Delta\tilde{h}, \qquad (4)
$$

with $\Delta\tilde{\boldsymbol{x}}$ and $\Delta\tilde{h}$ denoting the volume elements corresponding to the grid points. In the case of continuous groups, it is possible to either make a discretisation of the subgroup H, or approximate the group convolution integral by means of random sampling. In order to obtain a volumetrically uniform sampling grid over the group, we sample a set of n equidistant points in the Lie algebra as in (Bekkers, 2019), and map those to the group using the exponential map to obtain a grid $\mathcal{H} := [h_e, ..., h_n]$. For noncompact H such as the dilation group $\mathbb{R}^+$, we localize the support in the Lie algebra. In

the case of compact continuous groups such as $\mathrm{SO}(2)$ we approximate the integral by convolving over a uniformly spaced grid $\mathcal{H}$ which is perturbed by left-multiplication with a uniformly randomly sampled group element $h_\epsilon \sim \mathrm{d}\tilde{h}$, i.e., $h_\epsilon \mapsto \mathcal{H}_\epsilon := h_\epsilon \mathcal{H}$. Note that by uniform sampling of H we obtain left-invariant $\Delta\tilde{h}$ which only scales the overall convolution result, allowing us to omit it from 4. We further let $\Delta\tilde{\boldsymbol{x}} = 1$ and obtain the discrete separable group convolution for continuous groups:

$$
\begin{aligned}
(f *_{group} k)(g) &\approx \sum_{\tilde{\boldsymbol{x}} \in \mathbb{Z}^2} \sum_{\tilde{h} \in \mathcal{H}_\epsilon} f(\tilde{\boldsymbol{x}}, \tilde{h}) k(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x}), h^{-1} \cdot \tilde{h}) \frac{1}{|h|} \\
&= \sum_{\tilde{\boldsymbol{x}} \in \mathbb{Z}^2} \left[\sum_{\tilde{h} \in \mathcal{H}_\epsilon} f(\tilde{\boldsymbol{x}}, \tilde{h}) k_H(h^{-1} \cdot \tilde{h}) \right] \frac{1}{|h|} k_{\mathbb{R}^2}(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x})). \qquad (5)
\end{aligned}
$$

During training and inference, randomly sampling over the rotation group makes the network approximately equivariant to the continuous group $\mathrm{SO}(2)$, whereas discretisation makes the network exactly equivariant only to a discretised subgroup of $\mathrm{SO}(2)$. In our experiments we show that randomly sampling outperforms discretisation when modelling equivariance to $\mathrm{SO}(2)$.

Depthwise separability So far we have only considered unidimensional input features, whereas feature maps in deep convolutional models generally consist of a number of input channels. That is, the domain of our input function f is $\mathbb{R}^{C^l}$, where C^l is the number of channels at the current layer l. In a regular convolutional layer, a kernel has support over the entire input channel dimension. This could be interpreted as having a distinct spatial kernel for every channel dimension. Analogously, the regular group convolutional kernel is defined over the entire input channel dimension. For every input channel, each group element within the support of the convolution kernel is weighted separately. Factorising the group convolution operation into distinct subgroup and spatial dimensions presents us with a choice: we can either (1) define both k^θ_H and $k^\theta_{\mathbb{R}^2}$ over the input channels, or choose to (2) reduce the support of our spatial kernel to a single channel, and share a reweighing of it along the input channels in a depth-wise separable manner, similar to Haase & Amthor (2020). An ablation study detailed in appendix A.5 shows that on a fixed parameter budget, additional depthwise separation (2) consistently outperforms (1) defining both kernels over the input space. As such, we will keep to the additional depthwise separation and refer to this implementation as the separable group convolution.

Intuition behind separable group convolutions Separable convolution kernels are strictly less expressive than their non-separable counterpart, but to what extent would this limit the expressivity of their learned representations? In separable group convolutions, we are sharing a weighted version of a single spatial kernel along the input group axis. In contrast, we could view non-separable group convolutions as having the ability to learn distinct spatial configurations of features along the group axis. For example in the case of the roto-translation group $\mathrm{SE}(2)$, this would enable non-separable group convolution kernels to learn to represent features that are built up of different spatial configurations at different orientations. As an oversimplified example, a single non-separable $\mathrm{SE}(2)$ group convolution kernel could detect line segments at *one* and corners at a *different* orientation, whereas a single separable group convolution kernel can only detect the presence of *either* line segments *or* corners at different orientations. For a visual example, see appendix A.3. Although this reduction in expressivity could theoretically prove limiting in the application of G-CNNs on vision problems, in the next section we experimentally show that separable group convolutions in fact often outperform their non-separable counterpart.

5 Experiments

We empirically motivate the use of seperable group convolutions by studying the performance of three G-CNNs that incorporate equivariance to three distinct groups. As our goal is to isolate the effect of separating the group convolution, we use the same shallow ResNet architecture throughout all experiments, only varying the sampling resolution over the group, see appendix A.6 for more details. After the last convolution block, we apply max-projection over the remaining group dimensions to achieve an invariant representation. We experiment with three different groups: the roto-translation group in 2 dimensions ($\mathrm{SE}(2)$), the translation-dilation group in two dimensions ($\mathbb{R}^2 \rtimes \mathbb{R}^+$) and the group of rotations, dilations and translations in two dimensions ($\mathrm{Sim}(2)$). For $\mathrm{SE}(2)$, we first evaluate the influence of random sampling versus discretisation of $\mathrm{SO}(2)$. Next, we assess the difference in performance of separable and non-separable group convolutions on transformed MNIST

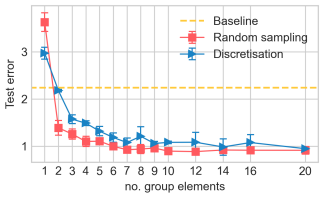


Figure 3: Test error versus $\mathrm{SO}(2)$ resolution for separable $\mathrm{SE}(2)$-CNNs on MNIST-rot, discretisation and random sampling.

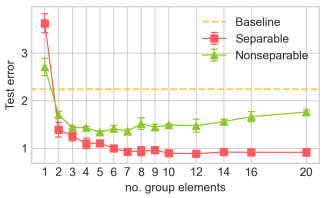


Figure 4: Test error versus sampling resolution of $\mathrm{SO}(2)$ on MNIST-rot for separable and non-separable $\mathrm{SE}(2)$-CNNs.

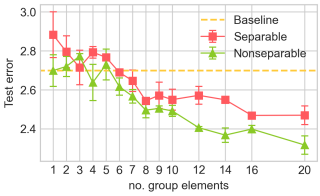


Figure 5: Test error versus sampled extent of the dilation group for separable and non-separable $\mathbb{R}^2 \rtimes \mathrm{R}^+$-CNNs.

variants; MNIST-rot, MNIST-scale for $\mathrm{SE}(2)$ and $\mathbb{R}^2 \rtimes \mathbb{R}^+$ respectively. Due to the reduction in computational complexity separable group convolutions bring, we are able to model equivariance to higher-dimensional groups: we implement two variations of a $\mathrm{Sim}(2)$-CNN and evaluate it on MNIST-rot-scale. Lastly, to investigate the advantages equivariance brings in more complex problem settings, we experiment with all setups on three vision benchmark datasets; CIFAR10, CIFAR100 Krizhevsky et al. (2009) and Galaxy10 Leung & Bovy (2019). See appendix A.7 for details on the training regimes, and A.4 for an additional experiment empirically validating the equivariance properties of our models.

Discretisation versus random sampling over subgroup We investigate the impact of approximating the group convolution integral through random sampling, instead of making a discretisation. We trained the architecture listed in appendix A.6 on rotated MNIST for different rotation group resolutions, for both sampling methods. Results are shown in figure 3. We can clearly see the advantage of approximating the group convolution integral through random sampling, which likely aids the model in overcoming aliasing effects that may occur in between grid points on the group for a discretisation of the rotation groups.

Separable and non-separable G-CNN performance on MNIST variants Following Cohen & Welling (2016a); Weiler et al. (2018b); Sosnovik et al. (2019); Finzi et al. (2020) we conduct experiments on MNIST variants with separable and non-separable implementations of equivariant models for each of the three groups. We discuss results per dataset in-depth.

Rotated MNIST The $\mathrm{SE}(2)-$CNNs are evaluated on rotated MNIST, the standard benchmark for rotation equivariant models. The 62.000 MNIST images are split into a training, validation and test set of 10.000, 2.000 and 50.000 images respectively, and randomly rotated to orientations between $[0, 2\pi)$. To assess the influence of sampling resolution over the group, we vary the number of sampled group elements from 1 to 20. The group integral is approximated through random sampling. Results are shown in figure 4. We clearly see the influence of increasing the resolution over the rotation group from 1 to 4 elements, after which the non-separable model saturates, and performance starts to drop when further increasing the number of sampled rotations. This is in line with findings by Bekkers et al. (2018), who note that as the group resolution increases, so does the possibility for overfitting on specific spatial configurations. The separable implementation saturates around 7 elements, but performance does not drop significantly beyond this point. This may imply that the reduction in kernel expressivity also has a regularising effect that benefits generalisation. In this experiment, separable group convolutions decisively outperform the non-separable variant.

Scaled MNIST The $\mathbb{R}^2 \rtimes \mathrm{R}^+$-CNNs are evaluated on MNIST scale, where the same training and test splits are made from the original MNIST dataset, but now randomly scaled by a factor [0.3, 1] and padded with zeros to retain the original resolution. The group convolution integral is approximated through discretisation. In this experiment, we retain the same resolution over the dilation group in all experiments, but instead vary the value at which the group is truncated, ranging from 1.0 at a grid of 1 element in the Lie algebra to 3.0 at a grid of 20 group elements. As in Sosnovik et al. (2019), we found that the inter-scale interactions in the group convolution operation reduced performance, likely because of increase in equivariance error due to truncation of the group. Therefore, we limited support of the group convolution kernel on R^+ to two neighbouring elements for all experiments in this paper. Results are shown in figure 5. Here, non-separable group convolutions outperform the

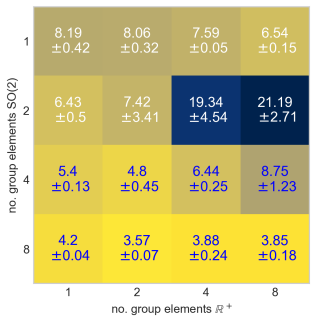


Figure 6: Test error versus resolution over rotation and scale group for separable $\mathrm{Sim}(2)-$CNN.

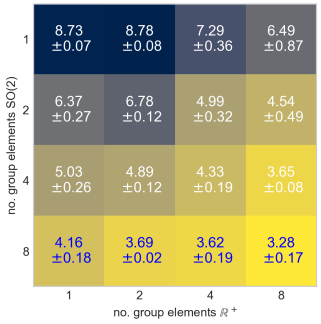


Figure 7: Test error versus resolution over rotation and scale group for H-separable $\mathrm{Sim}(2)-$CNN.

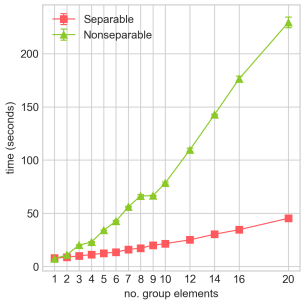


Figure 8: Process time per epoch (one pass over training and test sets) in seconds for different numbers of sampled group elements.

separable variant, suggesting that the ability of non-separable group convolutions to model spatial feature patterns over different scales is beneficial for the scaled MNIST dataset.

Scaled rotated MNIST Lastly, the $\mathrm{Sim}(2)-$CNN is benchmarked on MNIST-rot-scale, in which the same aforementioned training and test splits are randomly transformed by a scaling between [0.3, 1] and a rotation between $[0, 2\pi)$. We limit our experiments to 2,4,6 and 8 elements for each subgroup. We test two implementations: (1) separable, where the group convolution is factorised into two convolutions, one over $\mathbb{R}^+ \rtimes \mathrm{SO}(2)$ and one over $\mathbb{R}^2$, and (2) H-separable, which factorises the convolution into three steps: a convolution over $\mathbb{R}^+$, one over $\mathrm{SO}(2)$, and one over $\mathbb{R}^2$. As can be seen in figure 6 the separable implementation turned out to be unstable in the training process for low $\mathrm{SO}(2)$ resolutions. We suspect that an approximation of $\mathrm{SO}(2)$ of only 2 rotation elements, paired with possible aliasing effects and equivariance error occurring over the dilation group, impedes the model in learning robust representations. For higher resolution over $\mathrm{SO}(2)$, this effect seems to decrease. The results for the H-separable implementation, shown in figure 7, highlight a much clearer increase in performance as the resolution over both groups increases, which suggests decoupling these information mappings stabilises the learning process.

Application to computer vision benchmarks To evaluate the value of equivariance to different transformations groups in natural image classifcation, we apply the $\mathrm{SE}(2)-, \mathbb{R}^2 \rtimes \mathbb{R}^+-$ and $\mathrm{Sim}(2)-$CNNs to three vision datasets. For $\mathrm{SE}(2)$ and $\mathbb{R}^2 \rtimes \mathbb{R}^+$ we use resolutions of 4 and 8 group elements, and truncate the scale group at a value of $\sqrt{3}$, as we found this to work well for our experiments. For $\mathrm{Sim}(2)$, we use a resolution of 4 elements over the $\mathrm{SO}(2)$ group and 4 elements over the $\mathbb{R}^2 \rtimes \mathbb{R}^+$ subgroups and again truncate $\mathbb{R}^+$ at $\sqrt{3}$. On CIFAR10, as GPU memory allows, we also experiment with using resolutions of 8 group elements for both subgroups. As baseline, we implement the same architecture of our non-separable G-CNN, but restrict it to equivariance to translations. Results are summarized in table 1 and exhibit the following general pattern: For most groups, separable group convolutions outperform non-separable ones, and in all experiments best performance is achieved by a $\mathrm{Sim}(2)$-CNN. Next, we discuss results per dataset in-depth.

CIFAR10 First, we evaluate our models on the CIFAR10 dataset Krizhevsky et al. (2009). Results in table 1 show that on CIFAR10, the non-separable variant of our $\mathrm{SE}(2)$-CNN is outperformed by the translation equivariant baseline. With separable group convolutions and a group resolution of 8 elements, $\mathrm{SE}(2)$-CNN show improved performance. Interestingly, $\mathbb{R}^2 \rtimes \mathbb{R}^+$-CNNs hugely outperform both baseline and $\mathrm{SE}(2)$-CNNs, so it seems scale invariance is a better inductive bias than rotation invariance in CIFAR10. Combining scale and rotation invariance ultimately leads to the best performance, with separable $\mathrm{Sim}(2)-$CNNs and a resolution of 8 scale and 8 rotation group elements achieving 89.38% test accuracy. The H-separable $\mathrm{Sim}(2)$-CNN, which models no interactions between scale and rotation, performs slightly worse.

CIFAR100 To assess the impact of separating spatial and group information in more complex problem settings, we experiment with CIFAR100. Notably, the non-separable $\mathbb{R}^2 \rtimes \mathrm{R}^+$-CNNs perform well under baseline. We discuss this particular finding in section 6. For $\mathrm{SE}(2)$-CNNs, non-separable

Table 1: Test accuracy (%) on different vision benchmark datasets. Best performance per group convolution implementation is underlined in blue. Best overall performance is boldfaced. † Separable along rotation and scale dimensions.

GROUP	NO. ELEM	SEPARABLE	CIFAR10	CIFAR100	GALAXY10
SE(2)	4	✗	74.50(±0.14)	48.39 (±1.28)	85.58(±0.45)
		✓	78.30(±0.75)	48.11 (±0.25)	86.96(±0.07)
	8	✗	76.81(±0.57)	51.67 (±0.65)	84.97(±0.07)
		✓	80.89(±0.85)	50.19 (±0.40)	86.61(±0.38)
$\mathbb{R} \rtimes \mathbb{R}^+$	4	✗	81.44(±0.62)	44.80 (±0.56)	85.13(±0.07)
		✓	83.08(±0.43)	52.50 (±0.60)	84.82(±0.25)
	8	✗	80.56(±0.82)	44.18(±0.92)	84.83(±0.22)
		✓	83.60(±0.23)	53.26 (±0.10)	85.30(±0.06)
Sim(2)	4×4	✓	86.63(±0.60)	49.93 (±0.86)	84.90(±0.19)
		✓†	85.65(±0.66)	**54.62** (±1.41)	**87.45(±0.31)**
	8×8	✓	**89.38(±0.25)**	-	-
		✓†	87.64(±0.16)	-	-
Baseline	-	-	77.79(±0.27)	47.91(±1.50)	85.10(±0.42)

convolutions outperform separable ones, suggesting the model benefits from modelling distinct spatial configurations at different poses. Again, best performance is achieved by the $\mathrm{Sim}(2)$-CNN.

Galaxy10 Lastly, we apply our models in the domain of astro-photography, on the Galaxy10 dataset Leung & Bovy (2019). This dataset contains 21785 69x69 color images of galaxies divided into 10 unbalanced classes. Rotation and scaling symmetries exist naturally in this dataset, making it an interesting application for our invariant models. Here, results show little difference between performance of non-separable and separable approaches. Rotation invariance seems to be the stronger inductive bias in this dataset, but again, combining rotation and scale invariance leads to the best performance, with the H-separable $\mathrm{Sim}(2)$-CNN achieving 87.45% test accuracy. The higher separable $\mathrm{Sim}(2)$ test error seems to indicate that interactions between scale and rotation features for this dataset impede model performance.

Comparison in training time between separable and non-separable group convolutions As one of the main motivators for the separable group convolution is its theoretical increase in computational efficiency, a comparison is made in training time between separable and non-separable G-CNNs. We tracked processing times per epoch, which consists of a train- and inference procedure, for the rotated MNIST experiments, and show them in figure 8. Because of the overhead of two sequential convolution operations offset by the fact that convolution operations on GPUs are heavily optimised, the theoretical decrease in computational complexity, described in section 4, is not seen in practice. Nonetheless, separable group convolutions present a remarkable decrease in inference time.

6 DISCUSSION & FUTURE WORK

On implicit kernel representations Some of the non-separable configurations generated results below baseline, which is unexpected given that G-CNNs usually improve upon regular CNNs. We conjecture that these results are caused by the SIREN parameterisation, which take as input a coordinate vector of different units (mixing both spatial and sub-group h coordinates). This could limit kernel expressivity, as sharing of a frequency multiplier ω_0 sets kernel smoothness to be identical along spatial and subgroup dimensions. The fact that for the separable variant we use two distinct SIRENs would lift this restriction. Despite useful advantages of MLP parametrisations for the kernels, such as their flexibility and ease of implementation for arbitrary Lie groups, it remains an open problem to have full control over their smoothness; ideally one band-limits the kernel MLPs to the discretised resolution on each axis, which would be an important direction for future work.

Conclusion Motivated by observed redundancies in learned group convolution filters, we introduced separable group convolutions, a computationally efficient implementation of regular group convolutions. We showed that separable group convolutions not only drastically increase computational efficiency, but in many settings also outperform their non-separable counterpart. Furthermore, we demonstrated the value of separable group convolutions as a solution for modelling equivariance to larger groups; $\mathrm{Sim}(2)$-CNNs decivisely outperform models only equivariant to subgroups of $\mathrm{Sim}(2)$, clearly reinforcing equivariance as means of model generalisation.

REFERENCES

Erik J Bekkers. B-spline cnns on lie groups. In *International Conference on Learning Representations*, 2019.

Erik J Bekkers, Marco Loog, Bart M ter Haar Romeny, and Remco Duits. Template matching via densities on the roto-translation group. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):452–466, 2017.

Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pp. 440–448. Springer, 2018.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Taco Cohen. Learning transformation groups and their invariants. Master's thesis, University of Amsterdam, 2013.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016a.

Taco Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 52–68, 2018.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pp. 3165–3176. PMLR, 2020.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Daniel Haase and Manuel Amthor. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14600–14609, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Imre Risi Kondor. *Group theoretical methods in machine learning*. PhD thesis, Columbia University, 2008.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Attila Lengyel and Jan C van Gemert. Exploiting learned symmetries in group equivariant convolutions. *arXiv preprint arXiv:2106.04914*, 2021.

Henry W Leung and Jo Bovy. Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astronomical Society*, 483(3):3255–3277, 2019.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Jasper Linmans, Jim Winkens, Bastiaan S Veeling, Taco Cohen, and Max Welling. Sample efficient semantic segmentation using rotation equivariant convolutional networks. *arXiv preprint arXiv:1807.00583*, 2018.

Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5048–5057, 2017.

Marvin L Minsky and Seymour A Papert. Perceptrons: expanded edition, 1988.

Roberto Rigamonti, Amos Sironi, Vincent Lepetit, and Pascal Fua. Learning separable filters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2754–2761, 2013.

David W Romero, Erik J Bekkers, Jakub M Tomczak, and Mark Hoogendoorn. Wavelet networks: Scale equivariant learning from raw waveforms. *arXiv preprint arXiv:2006.05259*, 2020.

David W Romero, Anna Kuzina, Erik J Bekkers, Jakub M Tomczak, and Mark Hoogendoorn. Ckconv: Continuous kernel convolution for sequential data. *arXiv preprint arXiv:2102.02611*, 2021.

Laurent Sifre and Stéphane Mallat. Rigid-motion scattering for texture classification. *arXiv preprint arXiv:1403.1687*, 2014.

Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020.

Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. *arXiv preprint arXiv:1910.11093*, 2019.

Ivan Sosnovik, Artem Moskalev, and Arnold Smeulders. Disco: accurate discrete scale convolutions. *arXiv preprint arXiv:2106.02733*, 2021a.

Ivan Sosnovik, Artem Moskalev, and Arnold WM Smeulders. Scale equivariance improves siamese tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2765–2774, 2021b.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *arXiv preprint arXiv:1807.02547*, 2018a.

Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 849–858, 2018b.

Marysia Winkels and Taco S Cohen. 3d g-cnns for pulmonary nodule detection. *arXiv preprint arXiv:1804.04656*, 2018.

Daniel E Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 567–584, 2018.

Daniel E Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. *arXiv preprint arXiv:1905.11697*, 2019.

Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5028–5037, 2017.

A APPENDIX

A.1 GROUP THEORETIC PRELIMINARIES

In this section, group theoretical prerequisites used throughout the paper are briefly refreshed. This is by no means intended as an exhaustive exposition of the fundamentals of group theory, we only introduce those concepts relevant to the current work.

Group. A group is defined by a set G of *group elements*, along with a binary operator $\cdot : G \times G \to G$, called the *group product*. The group product defines a way to combine each pair of elements $g_1, g_2 \in G$. For the binary operator $\cdot$ to be considered a group product, it needs to satisfy four constraints:

1. Closure. G is closed under $\cdot$; for all $g_1, g_2 \in G$ we have $g_1 \cdot g_2 \in G$.
2. Identity. There exists an identity element e s.t. for each $g \in G$, we have $e \cdot g = g \cdot e = g$.
3. Inverse. For every element $g \in G$ we have an element $g^{-1} \in G$, s.t. $g \cdot g^{-1} = e$.
4. Associativity. For every set of elements $g_1, g_2, g_3 \in G$, we have $(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3)$.

Lie groups. A Lie group is a group of which the elements form a smooth manifold. Since the group itself is not necessarily a vector space, combination of elements through addition or subtraction is not defined. However, to each Lie group G, an algebra $\mathfrak{g}$ may be associated, given by the tangent space of the Lie group at the identity $T_e(G)$. The Lie algebra may be interpreted as a vector space of infinitesimal generators of the group, a set of elements from which we can obtain the group G, by repeated application.

Exponential and logarithmic map. The exponential map $\exp : \mathfrak{g} \to G$ is a function mapping elements from the Lie algebra to the group. For many transformation Lie groups of interest this map is surjective, and it is possible to define an inverse mapping; the logarithmic map which maps from the group to the algebra.

Semi-direct product groups. In practice, we are often only interested in data defined over $\mathbb{R}^d$, and hence in this paper only consider affine Lie groups of the form $G = \mathbb{R}^d \rtimes H$, where $\mathbb{R}^d$ is the translation group in d dimensions and H is a transformation Lie group of interest.

Group action. A group G may have an action on a given space $\mathcal{X}$. Given a group element $g \in G$ and a set $\mathcal{X}$, the group action $\mathcal{T}_g$ defines what happens to any element of $x \in \mathcal{X}$ when we apply the transformation given by element g to it. This action is given by:

$$\mathcal{T} : G \times \mathcal{X} \to \mathcal{X} \text{ and } \mathcal{T}_g : \mathcal{X} \to \mathcal{X} \tag{6}$$

such that for any two elements $g, h \in G$, we can combine their actions into a single action; $\mathcal{T}_{g,h} = \mathcal{T}_g \circ \mathcal{T}_h$. To avoid clutter, we write the action $\mathcal{T}_g(x)$ as $g \cdot x$. Note that the action of a group G on domain $\mathcal{X}$ also extends to functions defined on this domain. The *induced group action* $\mathbb{T}_g$ of an element $g \in G$ maps a square integrable function f to a transformed function f' under the action of group G. For a given element of the transformed domain $x' \in \mathcal{X}'$, the value of the transformed function f' is obtained by applying the *inverse* transformation of g to the domain of the function f:

$$\mathbb{T}_g : f \to f' \text{ and for } x' \in \mathcal{T}_g(\mathcal{X})\text{: } f'(x') = f(g^{-1}x') \tag{7}$$

Representations In practice, the object on which the group operates is often defined as a vector space $\mathbb{R}^2$. We operate on 2d images, which can be seen as functions on $\mathbb{R}^2$. This makes the object our group operates on a vector space, meaning the group actions (or elements) need to be constructed as linear transformations from this vector space onto itself. Simply put, every group element $g \in G$ needs to be represented in this vector space by a matrix. We say that linear transformations $\mathcal{L}_g^{G \to \mathbb{L}_2(X)} : \mathbb{L}_2 \to \mathbb{L}_2$ on some space X are a representation of a group G when they satisfy the group structure, meaning for $g_1, g_2 \in G, x \in X$:

$$(\mathcal{L}_{g_1}^{G \to \mathbb{L}_2(X)} \circ \mathcal{L}_{g_2}^{G \to \mathbb{L}_2(X)})(x) = \mathcal{L}_{g_1 \cdot g_2}^{G \to \mathbb{L}_2(X)}(x) \tag{8}$$

A representation may be seen as the action of a group on a space in matrix form. For simplicity of notation, we write $\mathcal{L}_{g_1}^{G \to \mathbb{L}_2(X)}$ as $\mathcal{L}_g$.

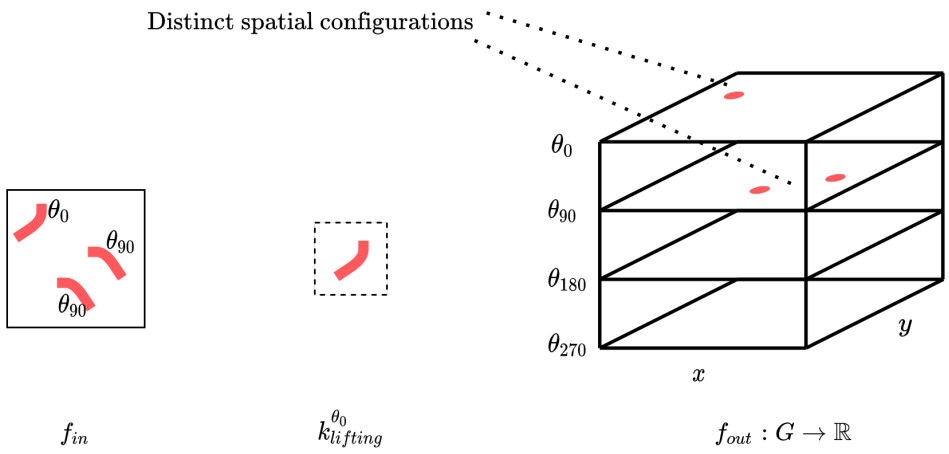


Figure 9: An example of a configuration of features a single separable group convolution kernel is unable to recognise.

Equivariance. An operator is equivariant with respect to a group, if it commutes with the action of the group group. For an operator $\Phi : \mathbb{L}_2(X) \to \mathbb{L}_2(Y)$:

$$\forall g \in G : \mathcal{L}_g \circ \Phi = \Phi \circ \mathcal{L}_g \tag{9}$$

A.2 Deriving separable group convolutions

If we assume

$$k(g) = k(\boldsymbol{x}, h) \tag{10}$$
$$= k_{\mathbb{R}^2}(\boldsymbol{x}) k_H(h) \tag{11}$$

We can derive separable group convolutions as:

$$(f *_{\text{group}} k)(g) = \int_G f(\tilde{g}) k(g^{-1} \cdot \tilde{g}) \mathrm{d}\mu(\tilde{g}) \tag{12}$$
$$= \int_{\mathbb{R}^2} \int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) \mathcal{L}_{x^{-1}} \mathcal{L}_{h^{-1}} k(\tilde{\boldsymbol{x}}, \tilde{h}) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}} \mathrm{d}\tilde{h} \tag{13}$$
$$= \int_{\mathbb{R}^2} \int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) k(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x}), h^{-1} \cdot \tilde{h}) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}} \mathrm{d}\tilde{h} \tag{14}$$
$$\to \int_{\mathbb{R}^2} \int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) k_{\mathbb{R}^2}(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x})) k_H(h^{-1} \cdot \tilde{h}) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}} \mathrm{d}\tilde{h} \tag{15}$$
$$= \int_{\mathbb{R}^2} \int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) k_H(h^{-1} \cdot \tilde{h}) k_{\mathbb{R}^2}(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x})) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}} \mathrm{d}\tilde{h} \tag{16}$$
$$= \int_{\mathbb{R}^2} \left[\int_H f(\tilde{\boldsymbol{x}}, \tilde{h}) k_H(h^{-1} \cdot \tilde{h}) \mathrm{d}\tilde{h} \right] k_{\mathbb{R}^2}(h^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{x})) \frac{1}{|h|} \mathrm{d}\tilde{\boldsymbol{x}} \tag{17}$$

A.3 What features separable group kernels cannot learn

Separable group convolution kernels share a reweighting of the same spatial kernel along the group axis. This makes them less expressive than non-separable group convolution kernels. To give an intuition for the restrictions this imposes on the types of features the separable group convolution can learn, we show a simplified visual example, seen in figure 9. In this example, we are modelling equivariance for translations and rotations of 90°. Assume we have an input image f_{in}, which contains three translated instances of the same archetypal feature r, one occuring under the canonical pose denoted with θ_0, and two under a rotation of 90°, denoted with θ_{90}. We are given a lifting

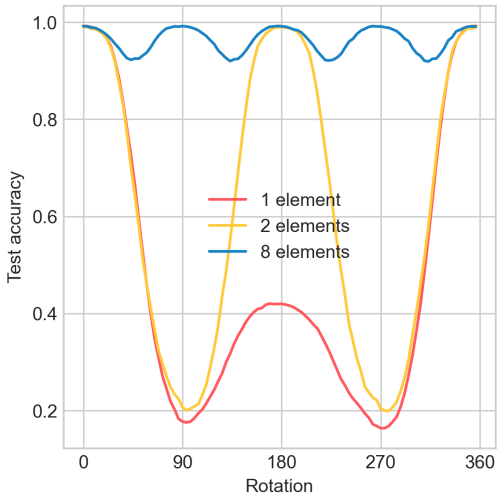


Figure 10: Test error vs rotation angle of MNIST test set, when trained on upright MNIST, for separable $\mathrm{SE}(2)$-CNN.

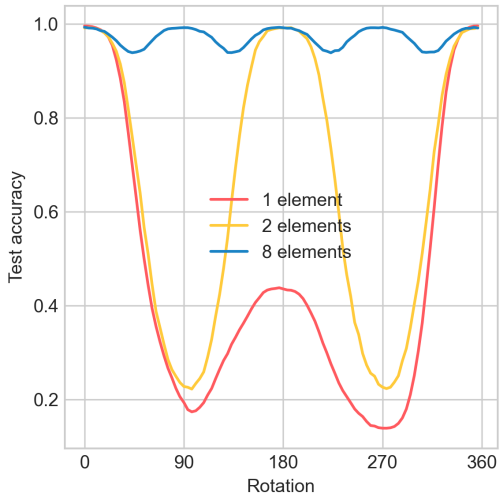


Figure 11: Test error vs rotation angle of MNIST test set, when trained on upright MNIST, for non-separable $\mathrm{SE}(2)$-CNN.

kernel $k_{lifting}$ that is able to match the feature r. If we apply the lifting convolution, as defined in equation 1, we obtain a feature map f_{out} defined over the group; for every rotations of $90°$ of our lifting kernel $k_{lifting}$, we obtain a spatial feature map containing the response for every location in the input. In this case, f_{in} contained three archetypal features at two distinct orientations. Note that the spatial configurations for the two orientations of r differ, and as such, the spatial configuration of responses along the rotation axis in f_{out} differs. A non-separable group convolution kernel in a following layer would be able to learn this set of distinct spatial configurations along the rotation axis, whereas separable group convolutions would not. Although theoretically this restricts the separable group convolution, our experiments show that in practice this rarely seems a problem, and may even help prevent overfitting.

A.4 Additional experiment: validating model invariance to transformation groups

Following Weiler et al. (2018b), we empirically assess the level of model-invariance to rotation transformations, by training on upright MNIST, and subsequently assessing performance on a test set of MNIST images that have been transformed by a subgroup element $h \in H$. Each model is trained on the MNIST training set containing 60,000 images, and tested on 10,000 transformed images. Performance is tested for rotations over a range between $[0, 2\pi)$, in 100 steps. The group convolution integral is approximated through random sampling. Results are shown in figures 10 and 11 for 1, 2 and 8 group elements for the separable and non-separable variants. We can see that equivariance error is very similar for separable and non-separable variants, and greatly reduces with increased resolution over the group. We do note that for 8 elements, the non-separable group convolution achieves slightly better performance.

A.5 Additional experiment: Support of the separable convolution kernels over the channel dimensions

As explained in section 4, we can choose to either define both k_H and $k_{\mathbb{R}^2}$ over the input channel dimension, or incorporate depthwise separability as well; additionally sharing the same spatial kernel $k_{\mathbb{R}^2}$ over the input channels as well as the group input elements. In an additional ablation study we find that at any resolution above $|H| = 1$, additional depthwise separability outperforms defining both kernels over the input channel dimensions on a fixed parameter budget. For this experiment on rotated MNIST, we use the same depthwise group separable $\mathrm{SE}(2)-$CNN as in section 5, with the architecture described in appendix A.6. For the non-depthwise group separable $\mathrm{SE}(2)-$CNN,

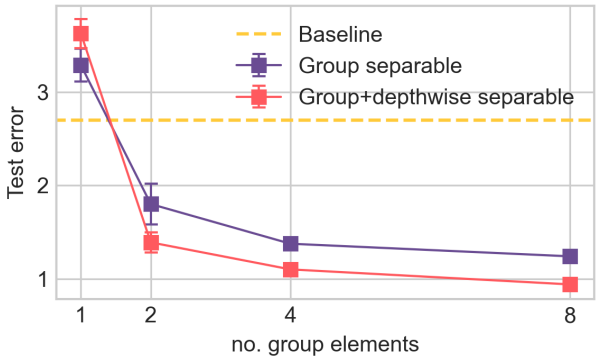


Figure 12: Results for depthwise group separable and group separable $\mathrm{SE}(2)-$CNN on rotated MNIST for different resolutions on the group.

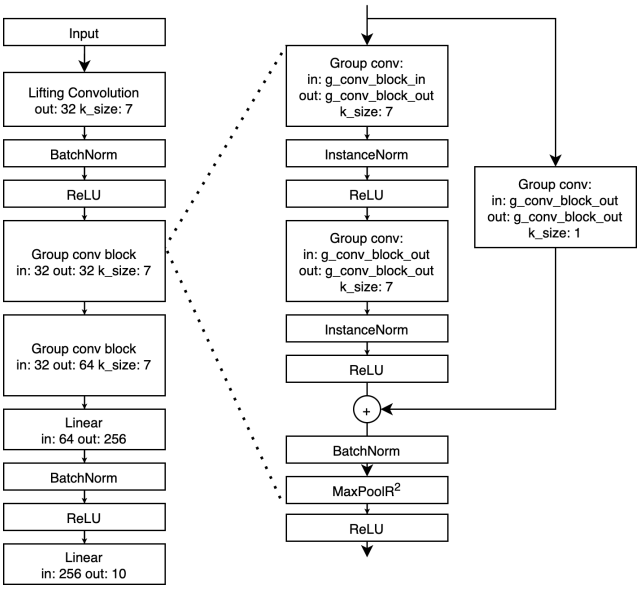


Figure 13: Architecture used over all experiments $\mathrm{Sim}(2)-$CNN.

we limit the number of channels to 48 instead of 64, and limit the hidden sizes of the SIRENs used to 32 instead of 64 to arrive at approximately the same number of parameters. Results, shown in figure 12 highlight that additional depthwise separability outperforms non-depthwise group separable convolutions, which is why we stick with additional depthwise separation of the spatial kernel, and name this approach the separable group convolution.

A.6 Architectural details

As architecture for our experiments we use a simple ResNet model He et al. (2016). We use a single lifting convolution with 32 output channels, followed by two residual blocks of 2 group convolutions each. After the first residual block we apply max-pooling over the spatial dimensions of the feature map. After the last residual block, we apply max pooling over the group and remaining spatial and subgroup dimensions, followed by two linear layers with batchnorm and ReLU in between. An overview is given in figure 13.

Our kernels are parameterised by a SIREN network Sitzmann et al. (2020), where output y^l for a layer l and input x^{l-1} is defined by:

$$\boldsymbol{y}^l = \sin(\omega_0 \boldsymbol{W}^l \boldsymbol{x}^{\boldsymbol{l-1}} + \boldsymbol{b}^l) \tag{18}$$

In this equation, ω_0 acts as a multiplier for low dimensional frequencies found in the input domain (the grid of relative offsets on the group), which explicitly introduces higher frequencies, allowing the neural net to learn high-frequency functions (such as kernels). We found a value for ω_0 of 10 to work well in all our experiments. For the SIREN, we used an architecture of two hidden layers of 64 units each. In non-separable G-CNNs we have a single SIREN with a final layer mapping to a vector $\mathbb{R}^{c_{in} \times c_{out}}$. In separable G-CNNs, we use two SIRENs, the first mapping to a function the Lie algebra of the subgroup H; $\mathbb{R}^{c_{in} \times c_{out}}$, and the second mapping to $\mathbb{R}^{c_{out}}$.

A.7 Experimental details

Here, we list training regimes for all experiments. We keep these as consistent as possible, as we are only interested in the effect of separating the group convolution operation, and in isolating the effect of incorporating equivariance into our models.

All architectures are trained for 200 epochs, batch sizes and learning rates specific to the dataset as detailed below, using Adam optimisation Kingma & Ba (2014), and $1e^{-5}$ weight decay. All models are trained on a single Titan V.

For MNIST-rot, MNIST-scale, MNIST-rot-scale, CIFAR10 and CIFAR100, we use a batch size of 128, and a learning rate of $1 \cdot 10^{-4}$. For Galaxy10, we limit batch size to 32 images, and scale learning rate accordingly to $2.5 \cdot 10^{-5}$, as suggested by Goyal et al. (2017).

A.8 Group convolution blocks and random sampling on the group

In our residual blocks, we subsequently convolve the input to the block $\boldsymbol{x}_{\text{in}}$ by two group convolutions, `gconv_1` and `gconv_2`, yielding $\boldsymbol{x}_{\text{out}}$ and apply elementwise addition with the input $\boldsymbol{x}_{\text{in}}$, followed by ReLU activation. When approximating the group convolution integral through random sampling, we must take care to define the input and output of the two group convolution layers on the same grid over the group as the skip-connection to ensure a well-defined equivariant group convolution block. This may be done by adding a group shortcut layer which maps from the set of input elements on the group to the set of output elements. We implement this as a group convolution with a 1×1 spatial extent. The group shortcut layer simultaneously serves as a channelwise projection from the input space of `gconv_1` to the output space of `gconv_2`.