Attentive Group Equivariant Convolutional Networks

David W. Romero [1]   **Erik J. Bekkers** [2]   **Jakub M. Tomczak** [1]   **Mark Hoogendoorn** [1]

Abstract

Although group convolutional networks are able to learn powerful representations based on symmetry patterns, they lack explicit means to learn meaningful relationships among them (e.g., relative positions and poses). In this paper, we present *attentive group equivariant convolutions*, a generalization of the group convolution, in which attention is applied during the course of convolution to accentuate meaningful symmetry combinations and suppress non-plausible, misleading ones. We indicate that prior work on visual attention can be described as special cases of our proposed framework and show empirically that our *attentive group equivariant convolutional networks* consistently outperform conventional group convolutional networks on benchmark image datasets. Simultaneously, we provide interpretability to the learned concepts through the visualization of equivariant attention maps.

1. Introduction

Convolutional Neural Networks (CNNs) have shown impressive performance in a wide variety of domains. The developments of CNNs, and many other machine learning approaches, have been fueled by intuitions and properties from visual systems of living organisms (LeCun et al., 1989; Delahunt & Kutz, 2019; Biederman, 1987; Blake & Lee, 2005; Zhaoping, 2014). While CNNs have resulted in a huge performance increase on many benchmark problems, their training efficiency and generalization capabilities are still open for improvement.

One concept that is being exploited for this purpose is *equivariance*, again drawing inspiration from human beings. Humans are able to identify familiar objects despite modifications in location, size, viewpoint, lighting conditions and background (Bruce & Humphreys, 1994). In addition, we
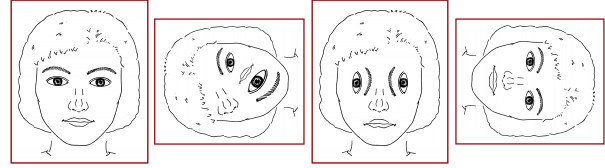
[1]Vrije Universiteit Amsterdam, [2]University of Amsterdam, The Netherlands. Correspondence to: David W. Romero <d.w.romeroguzman@vu.nl>.

Figure 1. Meaningful relationships among object symmetries. Though every figure is composed by the same elements, only the outermost examples resemble faces. The relative positions, orientations and scales of elements in the innermost examples do not match any meaningful face composition and hence, should not be labelled as such. Built upon Fig. 1 from (Schwarzer, 2000).

do not just recognize them but are able to describe in detail the type and amount of modification applied to them as well (von Helmholtz, 1868; Cassirer, 1944; Schmidt et al., 2016). Recently, several approaches have embraced these ideas as to preserve symmetries including translations (LeCun et al., 1989), planar rotations (Dieleman et al., 2016; Marcos et al., 2017; Worrall et al., 2017; Weiler et al., 2018; Li et al., 2018; Cheng et al., 2018; Bekkers et al., 2018; Lenssen et al., 2018), spherical rotations (Cohen et al., 2018; Worrall & Brostow, 2018; Cohen et al., 2019; Thomas et al., 2018; Kondor & Trivedi, 2018), scaling (Marcos et al., 2018; Worrall & Welling, 2019; Sosnovik et al., 2020) and general symmetry groups (Cohen & Welling, 2016a; Weiler & Cesa, 2019; Bekkers, 2020; Venkataraman et al., 2020) These approaches have been shown to improve the generalizability as well as training efficiency.

While group convolutional networks are able to learn powerful representations based on symmetry patterns, they lack any explicit means to learn meaningful relationships among them (e.g. relative positions, orientations and scales) as depicted in Fig. 1. In this paper, we draw inspiration from another promising development in the machine learning domain driven by neuroscience and psychology, namely *attention* (e.g., Pashler (2016)) to learn such relationships. The notion of attention is related to the idea that not all components of an input signal are *per se* equally relevant for a particular task. As a consequence, given a task and a particular input signal, task-relevant components of the input should be focused during its analysis while irrelevant, possibly misleading ones should be suppressed. Attention has been applied to different fields ranging from natural language processing (Bahdanau et al., 2014; Cheng et al.,

2016; Vaswani et al., 2017) to visual understanding (Ilse et al., 2018; Xu et al., 2015; Park et al., 2018; Woo et al., 2018; Ramachandran et al., 2019; Romero & Hoogendoorn, 2020) and graph analysis (Veličković et al., 2017; Zhang et al., 2020).

We present *attentive group convolutions*, a generalization of the group convolution, in which attention is applied during convolution to accentuate meaningful symmetry combinations and suppress non-plausible, possibly misleading ones. We indicate that prior work on visual attention can be described as special cases of our proposed framework and show empirically that our *attentive group equivariant group convolutional networks* consistently outperform conventional group equivariant ones on rot-MNIST and CIFAR-10 for the $SE(2)$ and $E(2)$ groups. In addition, we provide means to interpret the learned concepts trough the visualization of the predicted equivariant attention maps.

Contributions:

- We propose a general group theoretical framework for equivariant visual attention, *the attentive group convolution*, and show that prior works on visual attention are special cases of our framework.

- We introduce a specific type of network referred to as *attentive group convolutional networks* as an instance of this theoretical framework.

- We show that our *attentive group convolutional networks* consistently outperform plain group equivariant ones.

- We provide means to interpret the learned concepts via visualization of predicted equivariant attention maps.

2. Preliminaries

Before presenting our approach, we start with defining group convolutions and attention mechanisms.

2.1. Spatial Convolution and Translation Equivariance

Let $f, \psi : \mathbb{R}^d \to \mathbb{R}^{N_{\tilde{c}}}$ be a vector valued signal and filter on $\mathbb{R}^d$, such that $f = \{f_{\tilde{c}}\}_{\tilde{c}=1}^{N_{\tilde{c}}}$ and $\psi = \{\psi_{\tilde{c}}\}_{\tilde{c}=1}^{N_{\tilde{c}}}$. The spatial convolution ($\star_{\mathbb{R}^d}$) is defined as:

$$[f \star_{\mathbb{R}^d} \psi](y) = \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_{\mathbb{R}^d} f_{\tilde{c}}(x)\psi_{\tilde{c}}(x - y) \, dx. \quad (1)$$

Intuitively, Eq. 1 resembles a collection of $\mathbb{R}^d$ inner products between the input signal f and y-translated versions of ψ. Since the continuous integration in Eq. 1 is usually performed on signals and filters captured in a discrete grid $\mathbb{Z}^d$, the integral on $\mathbb{R}^d$ is reduced to a sum on $\mathbb{Z}^d$. In our derivations, however, we stick to the continuous case as to guarantee the validity of our theory for techniques defined on continuous spaces, e.g., steerable and Lie group convolutions

(Cohen & Welling, 2016b; Worrall et al., 2017; Bekkers et al., 2018; Weiler et al., 2018; Kondor & Trivedi, 2018; Thomas et al., 2018; Bekkers, 2020; Sosnovik et al., 2020).

To study (and generalize) the properties of the convolution, we rewrite Eq. 1 using the translation operator $\mathcal{L}_y$:

$$[f \star_{\mathbb{R}^d} \psi](y) = \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_{\mathbb{R}^d} f_{\tilde{c}}(x)\mathcal{L}_y[\psi_{\tilde{c}}](x) \, dx, \quad (2)$$

where $\mathcal{L}_y[\psi_{\tilde{c}}](x) = \psi_{\tilde{c}}(x - y)$. Note that the translation operator $\mathcal{L}_y$ is indexed by an amount of translation y. Resultantly, we actually consider a set of operators $\{\mathcal{L}_y\}_{y \in \mathbb{R}^d}$ that indexes the set of all possible translations $y \in \mathbb{R}^d$. A fundamental property of the convolution is that it commutes with translations:

$$\mathcal{L}_y[f \star_{\mathbb{R}^d} \psi] = \mathcal{L}_y[f] \star_{\mathbb{R}^d} \psi. \quad (3)$$

In other words, convolving a y-translated signal $\mathcal{L}_y[f]$ with a filter is equivalent to first convolving the original signal f with the filter ψ, and y-translating the obtained response next. This property is referred to as *translation equivariance* and, in fact, convolution (and reparametrizations thereof) is the *only* linear *translation equivariant* mapping (Cohen & Welling, 2016a; Kondor & Trivedi, 2018; Bekkers, 2020).

2.2. Group Convolution and Group Equivariance

The convolution operation can be extended to general transformations by utilizing a larger set of transformations $\{\mathcal{L}_g\}_{g \in G}$, s.t. $\{\mathcal{L}_y\}_{y \in \mathbb{R}^d} \subseteq \{\mathcal{L}_g\}_{g \in G}$. However, in order to preserve equivariance, we must restrict the class of transformations allowed in $\{\mathcal{L}_g\}_{g \in G}$. To formalize this intuition, we first present some important concepts from *group theory*.

2.2.1. PRELIMINARIES FROM GROUP THEORY

Groups. A *group* is a tuple $(G, \cdot)$ consisting of a set G, $g \in G$, and a binary operation $\cdot : G \times G \to G$, referred to as the *group product*, that satisfies the following axioms:

- *Closure:* For all $h, g \in G$, $h \cdot g \in G$.
- *Identity:* There exists an $e \in G$, such that $e \cdot g = g \cdot e = g$.
- *Inverse:* For all $g \in G$, there exists an element $g^{-1} \in G$, such that $g \cdot g^{-1} = g^{-1} \cdot g = e$.
- *Associativity:* For all $g, h, k \in G$, $(g \cdot h) \cdot k = g \cdot (h \cdot k)$.

Group actions. Let G and X be a group and a set, respectively. The (left) *group action* of G on X is a function $\odot : G \times X \to X$ that satisfies the following axioms:

- *Identity:* If e is the identity of G, then, for any $x \in X$, $e \odot x = x$.
- *Compatibility:* For all $g, h \in G$, $x \in X$, $g \odot (h \odot x) = (g \cdot h) \odot x$.

In other words, the action of G on X describes how the elements $x \in X$ are transformed by $g \in G$. For brevity, we omit the operations $\cdot$ and $\odot$ and refer to the set G as a group, to elements $g \cdot h$ as gh and to actions $(g \odot x)$ as gx.

Semi-direct product and affine groups. In practice, the analysis of data (and hence convolutions) defined on $\mathbb{R}^d$ is of main interest. Consequently, one is mainly interested in groups of the form $G = \mathbb{R}^d \rtimes H$ resulting from the *semi-direct product* ($\rtimes$) between the translation group $\mathbb{R}^d$, and an arbitrary (Lie) group H that acts on $\mathbb{R}^d$ (e.g., rotation, scaling, mirroring). This family of groups is referred to as *affine groups* and their group product is defined as

$$g_1 g_2 = (x_1, h_1)(x_2, h_2) = (x_1 + h_1 x_2, h_1 h_2), \quad (4)$$

where $g_1 = (x_1, h_1)$, $g_2 = (x_2, h_2) \in G$, $x_1, x_2 \in \mathbb{R}^d$ and $h_1, h_2 \in H$. Some important affine groups are the roto-translation ($SE(d) = \mathbb{R}^d \rtimes SO(d)$), the scale-translation ($\mathbb{R}^d \rtimes \mathbb{R}^+$) and the euclidean ($E(d) = \mathbb{R}^d \rtimes O(d)$) groups.

Group representations. Let G be a group and $\mathbb{L}_2(X)$ be a space of functions defined on some vector space X. The (left) regular *group representation* of G on functions $f \in \mathbb{L}_2(X)$ is a transformation $\mathcal{L} : G \times \mathbb{L}_2(X) \to \mathbb{L}_2(X)$, $(g, f) \mapsto \mathcal{L}_g[f]$, such that it shares the group structure via:

$$\mathcal{L}_g \mathcal{L}_h[f](x) = \mathcal{L}_{gh}[f](x), \quad (5)$$

$$\mathcal{L}_g[f](x) := f(g^{-1}x) \quad (6)$$

for any $g, h \in G$, $f \in \mathbb{L}_2(X)$, $x \in X$. That is, concatenating two such transformations, parametrized by g and h, is equivalent to one transformation parametrized by $gh \in G$. Intuitively, the representation of G on a function $f \in \mathbb{L}_2(X)$ describes how the function as a whole, i.e. $f(x), \forall x \in X$, is transformed by the effect of group elements $g \in G$.

If the group G is affine, i.e., $G = \mathbb{R}^d \rtimes H$, the (left) group representation $\mathcal{L}_g$ can be split as:

$$\mathcal{L}_g[f](x) = \mathcal{L}_y \mathcal{L}_h[f](x) \quad (7)$$

with $g = (y, h) \in G$, $y \in \mathbb{R}^d$ and $h \in H$. This property is key for the efficient implementation of functions on groups.

2.2.2. THE GROUP CONVOLUTION

Let $f, \psi : G \to \mathbb{R}^{N_{\tilde{c}}}$ be a vector valued signal and kernel on G. The group convolution ($\star_G$) is defined as:

$$[f \star_G \psi](g) = \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_G f_{\tilde{c}}(\tilde{g}) \psi_{\tilde{c}}(g^{-1}\tilde{g}) \, d\tilde{g} \quad (8)$$

$$= \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_G f_{\tilde{c}}(\tilde{g}) \mathcal{L}_g[\psi_{\tilde{c}}](\tilde{g}) \, d\tilde{g}. \quad (9)$$

Differently to Eq. 2, the domain of the signal f, the filter ψ and the group convolution itself $[f \star_G \psi]$ are now defined on
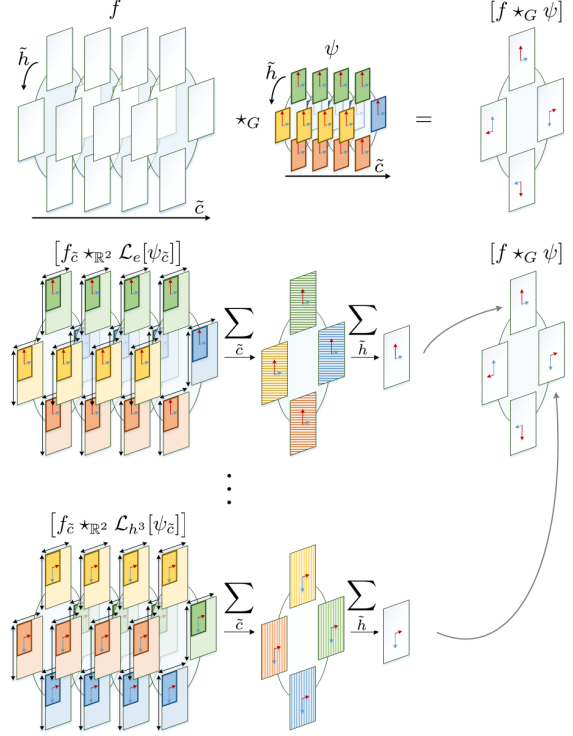

Figure 2. Group convolution on the roto-translation group $SE(2)$ for discrete rotations by 90 degrees (also called the $p4$ group). The $p4$ group is defined as $H = \{e, h, h^2, h^3\}$, with h depicting a 90° rotation. The group convolution corresponds to $|H| = 4$ convolutions between the input f and h-transformations of the filter ψ, $\mathcal{L}_h[\psi], \forall h \in H$. Each of these convolutions is equal to the sum over group elements $\tilde{h} \in H$ and channels $\tilde{c} \in [N_{\tilde{c}}]$ of the spatial channel-wise convolutions $[f_{\tilde{c}} \star_{\mathbb{R}^2} \mathcal{L}_h[\psi_{\tilde{c}}]]$ among f and $\mathcal{L}_h[\psi]$.

the group G. [1] Intuitively, the group convolution resembles a collection of inner products between the input signal f and g-transformed versions of ψ. A key property of the group convolution is that it generalizes equivariance (Eq. 3) to arbitrary groups, i.e., it commutes with g-transformations:

$$\mathcal{L}_g[f \star_G \psi] = \mathcal{L}_g[f] \star_G \psi. \quad (10)$$

In other words, group convolving a g-transformed signal $\mathcal{L}_g[f]$ with a filter ψ is equivalent to first convolving the original signal f with the filter ψ, and g-transforming the obtained response next. This property is referred to as *group equivariance* and, just as for spatial convolutions, the group convolution (or reparametrizations thereof) is the *only* linear G-equivariant map (see e.g. (Bekkers, 2020)).

Group convolution on affine groups. For affine groups, the group convolution (Eq. 9) can be decomposed, without modifying its properties, by taking advantage of the group

---

[1]Note that Eq. 2 matches Eq. 9 with the substitution $G = \mathbb{R}^d$. It follows that $\mathcal{L}_g[f](x) = f(g^{-1}x) = f(x - y)$, where $g^{-1} = -y$ is the inverse of g in the translation group $(\mathbb{R}^d, +)$ for $g = y$.
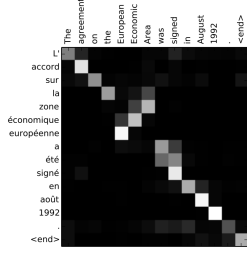
Figure 3. English to French translation. Brighter depicts stronger influence. Note how relevant parts of the input sentence are highlighted as a function of the current output word during translation. Taken from Bahdanau et al. (2014).

structure and the representation decomposition (Eq. 7) as:

$$[f \star_G \psi](g)$$

$$= \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_H \int_{\mathbb{R}^2} f_{\tilde{c}}(\tilde{x}, \tilde{h}) \mathcal{L}_g[\psi_{\tilde{c}}](\tilde{x}, \tilde{h}) \, d\tilde{x} \, d\tilde{h} \quad (11)$$

$$= \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_H \int_{\mathbb{R}^2} f_{\tilde{c}}(\tilde{x}, \tilde{h}) \mathcal{L}_x \mathcal{L}_h[\psi_{\tilde{c}}](\tilde{x}, \tilde{h}) \, d\tilde{x} \, d\tilde{h} \quad (12)$$

where $g = (x, h)$, $\tilde{g} = (\tilde{x}, \tilde{h}) \in G$, $x, \tilde{x} \in \mathbb{R}^d$ and $h, \tilde{h} \in H$. By doing so, the group convolution can be separated into $|H|$ spatial convolutions of the input signal f for each h-transformed filter $\mathcal{L}_h[\psi]$ (Fig. 2):

$$[f \star_G \psi](x, h) = \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_H \left[f_{\tilde{c}} \star_{\mathbb{R}^2} \mathcal{L}_h[\psi_{\tilde{c}}] \right](x, \tilde{h}) \, d\tilde{h}. \quad (13)$$

Resultantly, the computational cost of a group convolution is roughly equivalent to that of a spatial convolution with a filter bank of size $N_{\tilde{c}} \times |H|$ (Cohen & Welling, 2016a; Worrall & Welling, 2019; Cohen et al., 2019).

2.3. Attention, Self-Attention and Visual Attention

Attention mechanisms find their roots in recurrent neural network (RNN) based machine translation. Let $\varphi(\cdot)$ be an arbitrary non-linear mapping (e.g., a neural network), $\underline{y} = \{y_j\}_{j=1}^m$ be a sequence of target vectors y_i, and $\underline{x} = \{x_i\}_{i=1}^n$ be a source sequence, whose elements influence the prediction of each value $y_j \in \underline{y}$. In early models (e.g. Kalchbrenner & Blunsom (2013); Cho et al. (2014)), features in the input sequence are aggregated into a context vector $c = \sum_i \varphi(x_i)$ which is used to augment the hidden state in RNN layers. These models assume that source elements x_i contribute *equally* to every target element y_j and hence, that the same context vector c can be utilized for all target positions y_j, which does not generally hold (Fig. 3).

Bahdanau et al. (2014) proposed the inclusion of *attention coefficients* $\alpha_i = \{\alpha_{i,j}\}$, $[n] = \{1, ..., n\}$, $i \in [n]$, $j \in [m]$, $\sum_i \alpha_{i,j} = 1$, to modulate the contributions of the source elements x_i as a function of the current target element y_j by means of an adaptive context vector $c_j = \sum_i \alpha_{i,j} \varphi(x_i)$. Thereby, they obtained large improvements both in performance and interpretability. Recently, attention has been extended to several other machine learning tasks (e.g. Vaswani

et al. (2017); Park et al. (2018); Veličković et al. (2017)). The main development behind these extensions was *self-attention* (Cheng et al., 2016), where, in contrast to conventional attention, the target and source sequences are equal, i.e., $\underline{x} = \underline{y}$. Hence, the attention coefficients $\alpha_{i,j}$ encode correlations among input element pairs (x_i, x_j). For vision tasks, self-attention has been proposed to encode visual co-occurrences in data (Hu et al., 2018; Wang et al., 2018; Park et al., 2018; Woo et al., 2018; Cao et al., 2019; Bello et al., 2019; Ramachandran et al., 2019; Romero & Hoogendoorn, 2020). Unfortunately, its application on visual and, in general, on high-dimensional data is non-trivial.

2.3.1. VISUAL ATTENTION

In the context of visual attention, consider a feature map $f : X \to \mathbb{R}^{N_c}$ to be the source "sequence" [2]. Self-attention then imposes the learning of a total $n^2 = |X|^2$ attention vectors $\alpha_{i,j} \in \mathbb{R}^{N_c}$, which rapidly becomes unfeasible with increasing feature map size. Interestingly, Cao et al. (2019) and Zhu et al. (2019) empirically demonstrated that, for visual data, the attention coefficients $\{\alpha_{i,j}\}$ are approximately invariant to changes in the target position x_j. Consequently, Cao et al. (2019) and Zhu et al. (2019) proposed to approximate the attention coefficients $\{\alpha_{i,j}\} \in \mathbb{R}^{|X|^2 \times N_c}$ by a single vector $\{\alpha_i\} \in \mathbb{R}^{|X| \times N_c}$ which is independent of target position x_j. Despite this significant reduction in complexity, the dimensionality of $\{\alpha_i\}$ is still very large and further simplifications are mandatory. To this end, existing works (e.g. Hu et al. (2018); Woo et al. (2018)) replace the input f with a much smaller vector of input statistics s that summarizes relevant information from f.

For instance, the SE-Net (Hu et al., 2018) utilizes global average pooling to produce a vector of channel statistics of f, $s^{\mathcal{C}} \in \mathbb{R}^{N_{\tilde{c}}}$, $s^{\mathcal{C}} = \frac{1}{|\mathbb{R}^d|} \int_{\mathbb{R}^d} f_{\tilde{c}}(x) \, dx$, which is subsequently passed to a small fully-connected network $\varphi^{\mathcal{C}}(\cdot)$ to compute channel attention coefficients $\alpha^{\mathcal{C}} = \{\alpha_{\tilde{c}}^{\mathcal{C}}\}_{\tilde{c}=1}^{N_{\tilde{c}}} = \varphi_{\mathcal{C}}(s^{\mathcal{C}})$. These attention coefficients are then utilized to modulate the corresponding input channels $f_{\tilde{c}}$.

Complementary to channel attention akin to that of the SE-Net, Park et al. (2018) utilize a similar strategy for spatial attention. Specifically, they utilize channel average pooling to generate a vector of spatial statistics of f, $s^{\mathcal{X}} \in \mathbb{R}^d$, $s^{\mathcal{X}} = \frac{1}{N_{\tilde{c}}} \sum_{\tilde{c}=1}^{N_{\tilde{c}}} f_{\tilde{c}}(x)$, which is subsequently passed to a small convolutional network $\varphi^{\mathcal{X}}(\cdot)$ to compute spatial attention coefficients $\alpha^{\mathcal{X}} = \{\alpha^{\mathcal{X}}(x)\}_{x \in \mathbb{R}^2} = \varphi^{\mathcal{X}}(s^{\mathcal{X}})$. These attention coefficients are then utilized to modulate the corresponding spatial input positions $f(x)$. Recent works include extra statistical information, e.g., max responses (Woo et al., 2018), or replace pooling by convolutions (Cao et al., 2019).

---

[2]In the machine translation context we can think of f as a sequence $\underline{x} = \{f(x_i)\}_{i=1}^n$, with $n = |X|$ number of elements.
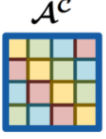
Figure 4. Same colors depict equal weights. The first column of $\mathcal{A}^{\mathcal{C}}$ corresponds to ψ and the following ones to $\mathcal{L}_h[\psi]$, obtained via cyclic permutations. See how $\{\mathcal{L}_h[\psi]\}_{h \in H}$ resembles a circulant matrix. Taken from Romero & Hoogendoorn (2020).

3. Attentive Group Equivariant Convolution

In this section, we propose our generalization of visual self-attention, discuss its properties and relation to prior work on visual attention.

Let $f, \psi : G \to \mathbb{R}^{N_{\tilde{c}}}$, and let $\alpha : G \times G \to [0, 1]^{N_{\tilde{c}}}$ be an *attention map* that takes, respectively, target and source elements $g, \tilde{g} \in G$ as input. We define *attentive group convolution* $(\star_G^\alpha)$ as:

$$[f \star_G^\alpha \psi](g) = \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_G \alpha_{\tilde{c}}(g, \tilde{g}) f_{\tilde{c}}(\tilde{g}) \mathcal{L}_g[\psi_{\tilde{c}}](\tilde{g}) \, d\tilde{g}, \quad (14)$$

in which $\alpha = \mathcal{A}[f]$ is computed by some *attention operator* $\mathcal{A}$. As such, attentive group convolutions modulate contributions of group elements at different channels $\tilde{c} \in [N_{\tilde{c}}]$ during pooling[3]. Properties and conditions on $\mathcal{A}$ are summarized in Thm. 1 and proven in the supplementary materials.

Theorem 1. *The attentive group convolution is an equivariant operator if and only if the attention operator $\mathcal{A}$ satisfies*

$$\forall_{\bar{g}, g, \tilde{g} \in G} : \quad \mathcal{A}[\mathcal{L}_{\bar{g}} f](g, \tilde{g}) = \mathcal{A}[f](\bar{g}^{-1}g, \bar{g}^{-1}\tilde{g}). \quad (15)$$

If, moreover, the maps generated by $\mathcal{A}$ are invariant to either one of the arguments, and, thus, exclusively attends either the input or output domain, then $\mathcal{A}$ satisfies Eq. 15 iff it is equivariant, and, thus, based on group convolutions.

3.1. Tying Together Equivariance and Visual Attention

Interestingly, and, perhaps in some cases unaware of it, *all* of the visual attention approaches outlined in Section 2.3.1, as well as *all of those we are aware of* (Xu et al., 2015; Hu et al., 2018; Park et al., 2018; Woo et al., 2018; Wang et al., 2018; Ilse et al., 2018; Ramachandran et al., 2019; Cao et al., 2019; Chen et al., 2019; Bello et al., 2019; Lin et al., 2019; Romero & Hoogendoorn, 2020) *exclusively utilize translation equivariance preserving maps for the generation of the attention coefficients and, hence, constitute altogether group equivariant networks by which they satisfy Thm. 1.*

As will be explained in the following sections, all these works resemble special cases of Eq. 14 by substituting G with the corresponding group and modifying the specifications of how α is calculated (Sec. 3.2, 3.4).

---

[3]Eq. 14 corresponds to Eq. 9 up to a multiplicative factor $\alpha_{\tilde{c}}(g, \tilde{g})^{-1}$, if $\alpha_{\tilde{c}}(g, \tilde{g})$ is constant for every $g, \tilde{g} \in G, \tilde{c} \in [N_{\tilde{c}}]$.

3.1.1. TRANSLATION EQUIVARIANT VISUAL ATTENTION

Since convolutions as well as popular pooling operations[4] are translation equivariant, the visual attention approaches outlined in Sec. 2.3.1 are translation equivariant as well. One particular case worth emphasising is that of SE-Nets. Here, a fully-connected network $\varphi^{\mathcal{C}}$, a non-translation equivariant map, is used to generate the channel attention coefficients $\alpha^{\mathcal{C}}$. However, $\varphi^{\mathcal{C}}$ *is* indeed translation equivariant. Recall that $\varphi^{\mathcal{C}}$ receives $s^{\mathcal{C}}$ as input, a signal obtained via global average pooling (a convolution-like operation). Resultantly, $s^{\mathcal{C}}$ can be interpreted as a $\mathbb{R}^{N_{\tilde{c}} \times 1 \times 1}$ tensor and hence, applying a fully connected layer to $s^{\mathcal{C}}$ equals a pointwise convolution between $s^{\mathcal{C}}$ and a filter $\psi_{\text{fully}} \in \mathbb{R}^{N_o \times N_{\tilde{c}} \times 1 \times 1}$ with N_o output channels[5].

3.1.2. GROUP EQUIVARIANT VISUAL ATTENTION

To the best of our knowledge, the only work that provides a group theoretical approach towards visual attention is that of Romero & Hoogendoorn (2020). Here, the authors consider affine groups G with elements $g = (x, h)$, $x \in \mathbb{R}^d$, $h \in H$ and cyclic permutation groups H. Consequently, they utilize a cyclic permutation equivariant map, $\varphi^{\mathcal{H}}(\cdot)$, to generate attention coefficients $\alpha^{\mathcal{H}}(h)$, $h \in H$, with which the corresponding elements h are modulated. As a result, their proposed attention strategy is H-equivariant. To preserve translation equivariance, and, hence, G-equivariance, $\varphi^{\mathcal{H}}$ is re-utilized at every spatial position $x \in \mathbb{R}^d$. This is equivalent to combining $\varphi^{\mathcal{H}}$ with a pointwise filter on $\mathbb{R}^d$. Romero & Hoogendoorn (2020) found that equivariance to cyclic groups H, can *only* be achieved by constraining $\varphi^{\mathcal{H}}$ to have a *circulant structure*. This is equivalent to a convolution with a filter ψ, whose group representations $\mathcal{L}_h$ induce cyclical permutations of itself (Fig. 4) and hence, resembles a special case of Thm. 1.

The work of Romero & Hoogendoorn (2020) exclusively performs attention on the h component of the group elements $g = (x, h) \in G$ and is only defined for (block) cyclic groups. Consequently, it does not consider spatial relationships during attention (Fig. 1) and is not applicable to general groups. Conversely, our proposed framework allows for simultaneous attention on both components of the group elements $g = (x, h)$ in a G equivariance preserving manner.

3.2. Efficient Group Equivariant Attention Maps

Attentive group convolutions impose the generation of an additional attention map $\alpha : G \to [0, 1]^{N_{\tilde{c}}}$, which is computationally demanding. To reduce this computational burden,

---

[4]In fact, conventional pooling operations (e.g. max, average) can be written as combinations of convolutions and pointwise non-linearities, which are translation equivariant, as well.

[5]This resembles a depth-wise separable convolution (Chollet, 2017) with the first convolution given by global average pooling.

we exploit the fact that visual data is defined on $\mathbb{R}^d$ and, hence, relevant groups are affine, to provide an efficient factorization of the attention map α.

Let G be an affine group with elements $g = (x, h)$, $x \in \mathbb{R}^d$, $h \in H$. In Sec. 2.3, we showed[6] for $G = \mathbb{R}^d$ that the attention coefficients $\alpha_{\tilde{c}}(g) = \alpha_{\tilde{c}}(x, h)$ can be described in a G-equivariant manner as the product of an *spatial attention map* $\alpha^{\mathcal{X}} : G \to [0, 1]$ and a *channel attention map* $\alpha^{\mathcal{C}} : H \to [0, 1]^{N_{\tilde{c}}}$, i.e., $\alpha_{\tilde{c}}(x, h) = \alpha^{\mathcal{X}}(x, h)\alpha_{\tilde{c}}^{\mathcal{C}}(h)$[7] as long as $\alpha^{\mathcal{X}}$ and $\alpha^{\mathcal{C}}$ are obtained in a G-equivariant manner. Consequently, we can rewrite Eq. 14 as:

$$[f \star_G^{\alpha} \psi](x, h) = \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_H \int_{\mathbb{R}^2} \mathcal{L}_x \mathcal{L}_h[\alpha^{\mathcal{X}}](\tilde{x}, \tilde{h}) \mathcal{L}_x \mathcal{L}_h[\alpha_{\tilde{c}}^{\mathcal{C}}](\tilde{h})$$
$$f_{\tilde{c}}(\tilde{x}, \tilde{h}) \mathcal{L}_x \mathcal{L}_h[\psi_{\tilde{c}}](\tilde{x}, \tilde{h}) \, d\tilde{x} \, d\tilde{h} \quad (16)$$

where $g = (x, h)$, $\tilde{g} = (\tilde{x}, \tilde{h}) \in G$, $x, \tilde{x} \in \mathbb{R}^d$ and $h, \tilde{h} \in H$. Conveniently, the attention coefficients $\alpha_k(\tilde{x}, \tilde{h})$ can be visualized as a function on $\mathbb{R}^d$, $\tilde{x} \mapsto \alpha_k(\tilde{h}), \forall \tilde{h} \in H, \forall k \in [N_c]$, which aids the interpretability of the learned concepts and the attended symmetries (Fig. 7).

Since visual attention is approximately invariant to changes in the target position (Sec. 2.3.1; Cao et al. (2019); Zhu et al. (2019)), Eq. 17 can be further reduced to:

$$[f \star_G^{\alpha} \psi](x, h) = \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_H \int_{\mathbb{R}^2} \mathcal{L}_h[\alpha^{\mathcal{X}}](\tilde{x}, \tilde{h}) \mathcal{L}_h[\alpha_{\tilde{c}}^{\mathcal{C}}](\tilde{h})$$
$$f_{\tilde{c}}(\tilde{x}, \tilde{h}) \mathcal{L}_x \mathcal{L}_h[\psi_{\tilde{c}}](\tilde{x}, \tilde{h}) \, d\tilde{x} \, d\tilde{h}. \quad (17)$$

3.2.1. THE ATTENTION OPERATOR $\mathcal{A}$

Recall that the attention map α is computed via an attention operator $\mathcal{A}$. In the most general case, α and, hence, the attention operator $\mathcal{A}$, is a function of both the input signal f and the filter ψ. In order to define $\mathcal{A}$ as such, we generalize the approach of Woo et al. (2018) such that: (1) equivariance to general symmetry groups is preserved and (2) the attention maps depend of the filter ψ as well.

Let $\phi^{\mathcal{C}} : f \mapsto s^{\mathcal{C}} = \{s_{\text{avg}}^{\mathcal{C}}, s_{\text{max}}^{\mathcal{C}}\}$, $s_i^{\mathcal{C}} : H \to \mathbb{R}^{N_{\tilde{c}}}$ and $\phi^{\mathcal{X}} : f \mapsto s^{\mathcal{X}} = \{s_{\text{avg}}^{\mathcal{X}}, s_{\text{max}}^{\mathcal{X}}\}$, $s_i^{\mathcal{X}} : G \to \mathbb{R}$ be functions that generate channel ($s^{\mathcal{C}}$) and spatial statistics ($s^{\mathcal{X}}$) from a vector valued signal $f : G \to \mathbb{R}^{N_{\tilde{c}}}$, respectively. Analogously to Woo et al. (2018), we compute spatial and channel statistics to reduce the dimensionality of the input. However, in contrast to their approach, we compute these statistics from intermediary convolutional maps (Fig. 5) rather than from the input signal f directly. As a result, take the influence of the filter ψ into account during the computation of

---

[6] $G = \mathbb{R}^d$ is an affine group with $H = \langle e \rangle$, the trivial group.

[7] Functions on H are also functions of G with spatial dimension 1×1, similar to the case of channel attention in SE-Nets (Sec. 3.1).
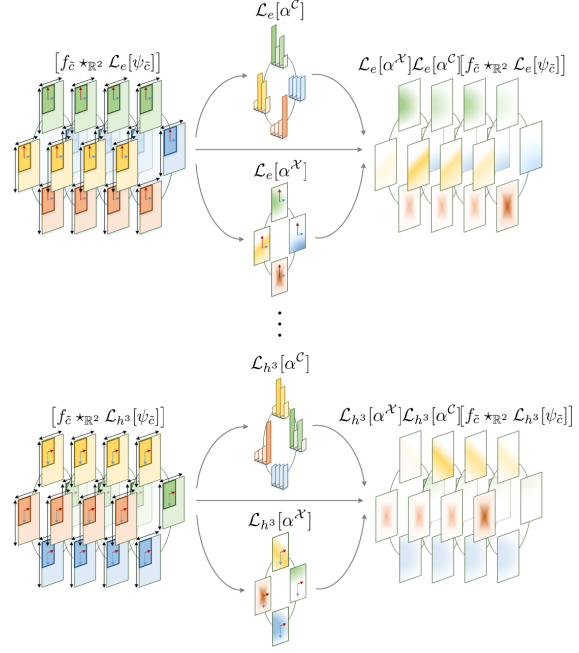

Figure 5. Attentive group convolution on the roto-translation group $SE(2)$. In contrast to group convolutions (Fig. 2, Eq. 13), attentive group convolutions utilize channel $\mathcal{L}_g[\alpha^{\mathcal{C}}]$ and spatial $\mathcal{L}_g[\alpha^{\mathcal{X}}]$ attention to modulate the intermediary convolutional responses $[f_{\tilde{c}} \star_{\mathbb{R}^2} \mathcal{L}_h[\psi_{\tilde{c}}]]$ before pooling them along the $\tilde{c}$ and $\tilde{h}$ axes.

the attention maps. Note that the group representation $\mathcal{L}_h$ must be applied to $\phi^{\mathcal{C}}$ and $\phi^{\mathcal{X}}$ during the attentive group convolution (Eq. 17). However, since both $\phi^{\mathcal{C}}$ and $\phi^{\mathcal{X}}$ are pooling functions (i.e., convolutions with a large filter of equal values everywhere), one has that $\mathcal{L}_h[\phi^{\mathcal{C}}] = \phi^{\mathcal{C}}$ and $\mathcal{L}_h[\phi^{\mathcal{X}}] = \phi^{\mathcal{X}}$ and hence, $\mathcal{L}_h$ can be omitted.

Channel Attention. Let $\varphi^{\mathcal{C}} : s^{\mathcal{C}} \mapsto \alpha^{\mathcal{C}}$ be a function that generates a channel attention map $\alpha^{\mathcal{C}} : H \to [0, 1]^{N_{\tilde{c}}}$ from a vector of channel statistics $s^{\mathcal{C}}$. Similarly to Woo et al. (2018), our channel attention map $\alpha^{\mathcal{C}}$ is defined as:

$$\alpha^{\mathcal{C}} = \varphi^{\mathcal{C}}(s^{\mathcal{C}}) = \sigma \Big(\big[[s_{\text{avg}}^{\mathcal{C}} \star_G w_1]^+ \star_G w_2\big] + \big[[s_{\text{max}}^{\mathcal{C}} \star_G w_1]^+ \star_G w_2\big] \Big) \quad (18)$$

with $(\cdot)^+$ the ReLU function, σ the sigmoid function, r a reduction ratio and $w_1 : H \to \mathbb{R}^{\frac{N_{\tilde{c}}}{r} \times N_{\tilde{c}}}$, $w_2 : H \to \mathbb{R}^{N_{\tilde{c}} \times \frac{N_{\tilde{c}}}{r}}$ filters defined on H.[8]

Spatial Attention. Let $\varphi^{\mathcal{X}} : s^{\mathcal{X}} \mapsto \alpha^{\mathcal{X}}$ be a function that generates a spatial attention map $\alpha^{\mathcal{X}} : G \to [0, 1]$ from a vector of channel statistics $s^{\mathcal{X}}$. Similarly to Woo et al. (2018), our spatial attention map $\alpha^{\mathcal{X}}$ is defined as:

$$\alpha^{\mathcal{X}} = \varphi^{\mathcal{X}}(s^{\mathcal{X}}) = \sigma \left([s^{\mathcal{X}} \star_G \psi^{\mathcal{X}}]\right) \quad (19)$$

---

[8] The filters w_1, w_2 correspond to the G-convolution version of the matrices W_1, W_2 in the channel attention of Woo et al. (2018).
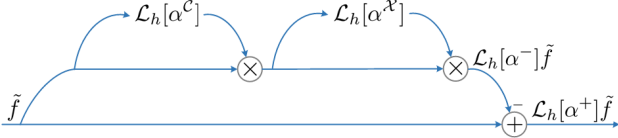
Figure 6. Sequential channel and spatial attention performed on a residual attention branch (Sec. 3.3).

with $\psi^{\mathcal{X}} : G \to \mathbb{R}^2$ a vector valued filter on G.

Full Attention. Woo et al. (2018) carried out extensive experiments to find the best performing configuration to combine channel and spatial attention maps for the $\mathbb{R}^d$ case (e.g., in parallel, serially starting with channel attention, serially starting with spatial attention). Based on their results we adopt their best performing configuration, i.e., *serially starting with channel attention*, for the G case.

Let $\tilde{f}_{\tilde{c}}(x, h, \tilde{h}) = \left[f_{\tilde{c}} \star_{\mathbb{R}^d} \mathcal{L}_h[\psi_{\tilde{c}}] \right](x, \tilde{h})$ be the intermediary result from the convolution between the input f and the h-transformation of the filter ψ, $\mathcal{L}_h[\psi]$ before pooling over $\tilde{c}$ and $\tilde{h}$ (Fig. 5, Eq. 13). We perform attention on top of $\tilde{f}$ in a sequential manner as depicted in Fig. 6, where $\alpha^{\mathcal{C}}$ and $\alpha^{\mathcal{X}}$ are computed by Eqs. 18, 19, respectively. Conclusively, our attentive group convolution is computed as:

$$[f \star_G^\alpha \psi](x, h) = \sum_{\tilde{c}=1}^{N_{\tilde{c}}} \int_H \left[\mathcal{L}_h[\alpha^{\mathcal{X}}] \mathcal{L}_h[\alpha^{\mathcal{C}}] \right. $$
$$\left. \left[f_k \star_{\mathbb{R}^d} \mathcal{L}_h[\psi_k] \right] \right] (x, \tilde{h}) \, d\tilde{h}. \quad (20)$$

3.3. The Residual Attention Branch

Based on the findings of He et al. (2016), several visual attention approaches propose to utilize residual blocks with direct connections during the course of attention (Hu et al., 2018; Park et al., 2018; Woo et al., 2018; Wang et al., 2018; Cao et al., 2019) to facilitate gradient flow. However, these approaches calculate the final attention map α^+ as the sum of the direct connection $\mathbf{1}$ and the attention map obtained from the attention branch α, i.e., $\alpha^+ = \frac{1+\alpha}{2}$. Consequently, the obtained attention map $\alpha^+ : G \to [0.5, 1]^{N_c}$ is *restrained* to the interval $[0.5, 1]$ and the network loses the ability to fully suppress input components.

Inspired by the aforementioned works, we propose to calculate attention in what we call a *residual attention branch* (Fig. 6). Specifically, we utilize the attention branch to calculate a *residual attention map* $\alpha^- = \alpha = \mathbf{1} - \alpha^+$. Next, we subtract the residual attention map α^- from the direct connection $\mathbf{1}$ to obtain the resultant attention map α^+, i.e., $\alpha^+ = \mathbf{1} - \alpha^-$. Resultantly, we are able to produce attention maps α^+ that span the entire $[0, 1]$ interval while preserving the benefits of the direct connections of He et al. (2016).

3.4. The Attentive Group Convolution as a Sequence of Group Convolutions and Pointwise Non-linearities

CNNs are usually organized in layers and hence, the input f is usually convolved in parallel with a set of N_o filters $\{\psi_o\}_{o=1}^{N_o}$. As outlined in the previous section, this implies that the attention maps can change as a function of the current filter ψ_o. One assumption broadly utilized in visual attention is that these maps do not depend on the filters $\{\psi_o\}_{o=1}^{N_o}$, and, hence, that α is a sole function of the input signal f. Then, one is able to shift the generation of the attention maps $\alpha^{\mathcal{C}}$, $\alpha^{\mathcal{X}}$ to the input and, hence, reduce the attentive group convolution to a sequence of multiple conventional group convolutions and point-wise non-linearities:

$$[f \star_G^\alpha \psi] = [f^{\mathcal{A}} \star_G \psi] = [(\alpha^{\mathcal{X}} \alpha^{\mathcal{K}} f) \star_G \psi] \quad (21)$$

where $\alpha^{\mathcal{C}}$ and $\alpha^{\mathcal{X}}$ are computed directly from f via Eqs. 18 and 19. Resultantly, the computational cost of attentive group convolutions is roughly reduced to that of computing multiple group convolutions in a sequential manner.

4. Experiments

We validate our approach by exploring the effects of using attentive group convolutions in contrast to conventional ones. We compare the conventional group equivariant networks $p4$-CNNs and $p4m$-CNNs (Cohen & Welling, 2016a) with their corresponding attentive counterpart α-$p4$-CNNs and α-$p4m$-CNNs. We exclude (Romero & Hoogendoorn, 2020) from the experiments since it is not applicable to general groups. Furthermore, we explore the effects of only applying channel attention (e.g., α_{CH}-$p4$-CNNs), spatial attention (e.g., α_{SP}-$p4$-CNNs) and applying attention directly on the input (e.g., α_F-$p4$-CNNs). We replicate as close as possible the training and evaluation strategies of Cohen & Welling (2016a) and initialize any additional parameter in the same way as the corresponding baseline.

4.1. rot-MNIST

The rotated MNIST dataset (Larochelle et al., 2007) contains $62k$ gray-scale 28x28 handwritten digits uniformly rotated for $[0, 2\pi]$. The dataset is split into training, validation and test sets of $10k$, $2k$ and $50k$ images respectively. We compare $p4$-CNNs with all the corresponding attention variants mentioned above. For our attention models, we utilize a filter size of 7 and a reduction ratio r of 2 on the attention branch. Since attentive group convolutions impose the learning of additional parameters, we also instantiate bigger $p4$-CNNs by increasing the number of channels uniformly at every layer to roughly match the number of parameters of the attentive versions. Our results show that attentive versions consistently outperform non-attentive ones (Tab. 1).

Table 1. Test error rates on rot-MNIST (with standard deviation under 5 random seed variations).

NETWORK	TEST ERROR (%)	PARAM.
$p4$-CNN	2.048 ± 0.045	24.61K
BIG$_{19}$-$p4$-CNN	1.796 ± 0.035	77.54K
α-$p4$-CNN	$\mathbf{1.696 \pm 0.021}$	73.13K
BIG$_{15}$-$p4$-CNN	1.848 ± 0.019	50.42K
α_{CH}-$p4$-CNN	$\mathbf{1.825 \pm 0.048}$	48.63K
α_{SP}-$p4$-CNN	$\mathbf{1.761 \pm 0.027}$	49.11K
BIG$_{11}$-$p4$-CNN	1.996 ± 0.083	29.05K
α_{F}-$p4$-CNN	$\mathbf{1.795 \pm 0.028}$	29.46K

Table 2. Test error rates on CIFAR10 and augmented CIFAR10+.

NETWORK	TYPE	CIFAR10	CIFAR10+	PARAM.
ALL-CNN	$p4$	9.32	8.91	1.37M
	α_{F}-$p4$	**8.8**	**7.05**	1.40M
	$p4m$	7.61	7.48	1.22M
	α_{F}-$p4m$	**6.93**	**6.53**	1.25M
RESNET44	$p4m$	15.72	15.4	2.62M
	α_{F}-$p4m$	**10.82**	**10.12**	2.70M

4.2. CIFAR-10

The CIFAR-10 dataset (Krizhevsky et al., 2009) consists of $60k$ real-world 32x32 RGB images uniformly drawn from 10 classes. The dataset is split into training, validation and test sets of $40k$, $10k$ and $10k$ images, respectively. We compare the $p4$ and $p4m$ versions of the All-CNN (Springenberg et al., 2014) and the Resnet44 (He et al., 2016) in (Cohen & Welling, 2016a) with attentive variations. For all our attention models, we utilize a filter size of 7 and a reduction ratio r of 16 on the attention branch. Unfortunately, attentive group convolutions impose an unfeasible increment on the memory requirement for this dataset[9]. Resultantly, we are only able to compare the α_{F} variations of the corresponding networks. Our results show that attentive α_{F} networks consistently outperform non-attentive ones (Tab. 2). Furthermore, we demonstrate that our proposed networks focus on relevant parts of the input and that the predicted attention maps behave equivariantly for group symmetries (Fig. 7).

5. Discussion and Future Work

Our results show that attentive group convolutions can be utilized as a drop-in replacement for standard and group equivariant convolutions that simultaneously facilitates the interpretability of the network decisions. Similarly to convolutional and group convolutional networks, attentive group convolutional networks also benefit of data augmentation. Interestingly, however, we also see that including additional symmetries reduces the effect of augmentations given by group elements. This finding supports the intuition that

---

[9]the α-$p4$ All-CNN requires approx. 72GB of CUDA memory, as opposed to 5GBs for the $p4$-All-CNN. This is due to the storage of the intermediary convolution responses required for the calculation of the attention weights.
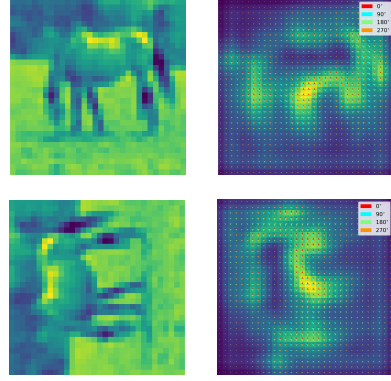

Figure 7. Equivariant attention maps on the roto-translation group $SE(2)$. The predicted attention maps behave equivariantly for group symmetries. The arrows depict the strength of the filter responses at the corresponding orientations throughout the network.

symmetry variants of the same concept are learned independently for non-equivariant networks (see Fig. 2 in (Krizhevsky et al., 2012)). The main shortcoming of our approach is its computational burden. As a result, the application of α-networks is computationally unfeasible for networks with several layers or channels. We believe, however, by extrapolation of our results on rot-MNIST, that further performance improvements are to be expected for α variations should hardware requirements suffice.

In future work, we want to explore the effect of attentive group convolutions on 3D symmetries. Group convolutional networks have been very successful in medical imaging applications (Winkels & Cohen, 2018). Since explainability plays a crucial role here, we believe that our attentive maps will be of high relevance to aid the explainability of the network decisions in a similar manner to that of Ilse et al. (2018). Moreover, since our attention maps are guaranteed to be equivariant, it is ensured that the predicted attention maps will be consistent across group symmetries, which is of major importance as well (e.g. a malignant tissue will generate the attention response regardless of its orientation).

6. Conclusions

We introduced attentive group convolutions, a generalization of the group convolution in which attention is utilized to explicitly highlight meaningful relationships among symmetries. We provided a general mathematical framework for group equivariant visual attention and indicated that prior work on visual attention can be perfectly described as special cases of the attentive group convolution. Our experimental results indicate that attentive group convolutional networks consistently outperform conventional group convolutional ones and additionally provide equivariant attention maps that behave predictively for symmetries of the group, with which learned concepts can be visualized.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bekkers, E. J. B-spline {cnn}s on lie groups. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gBhkBFDH.

Bekkers, E. J., Lafarge, M. W., Veta, M., Eppenhof, K. A., Pluim, J. P., and Duits, R. 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.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. Attention augmented convolutional networks. *arXiv preprint arXiv:1904.09925*, 2019.

Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94 (2):115, 1987.

Blake, R. and Lee, S.-H. The role of temporal structure in human vision. *Behavioral and cognitive neuroscience reviews*, 4(1):21–42, 2005.

Bruce, V. and Humphreys, G. W. Recognizing objects and faces. *Visual cognition*, 1(2-3):141–180, 1994.

Cao, Y., Xu, J., Lin, S., Wei, F., and Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*, 2019.

Cassirer, E. The concept of group and the theory of perception. *Philosophy and phenomenological research*, 5(1): 1–36, 1944.

Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., and Kalantidis, Y. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 433–442, 2019.

Cheng, J., Dong, L., and Lapata, M. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

Cheng, X., Qiu, Q., Calderbank, R., and Sapiro, G. Rotdcf: Decomposition of convolutional filters for rotation-equivariant deep networks. *arXiv preprint arXiv:1805.06846*, 2018.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

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

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

Cohen, T. S. and Welling, M. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical cnns. *CoRR*, abs/1801.10130, 2018. URL http://arxiv.org/abs/1801.10130.

Cohen, T. S., Weiler, M., Kicanaoglu, B., and Welling, M. Gauge equivariant convolutional networks and the icosahedral cnn. *arXiv preprint arXiv:1902.04615*, 2019.

Delahunt, C. B. and Kutz, J. N. Insect cyborgs: Bio-mimetic feature generators improve ml accuracy on limited data. 2019.

Dieleman, S., De Fauw, J., and Kavukcuoglu, K. Exploiting cyclic symmetry in convolutional neural networks. *arXiv preprint arXiv:1602.02660*, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Ilse, M., Tomczak, J. M., and Welling, M. Attention-based deep multiple instance learning. *ICML*, 2018.

Kalchbrenner, N. and Blunsom, P. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709, 2013.

Kondor, R. and Trivedi, S. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pp. 473–480. ACM, 2007.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Lenssen, J. E., Fey, M., and Libuschewski, P. Group equivariant capsule networks. In *Advances in Neural Information Processing Systems*, pp. 8844–8853, 2018.

Li, J., Yang, Z., Liu, H., and Cai, D. Deep rotation equivariant network. *Neurocomputing*, 290:26–33, 2018.

Lin, X., Ma, L., Liu, W., and Chang, S.-F. Context-gated convolution, 2019.

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

Marcos, D., Kellenberger, B., Lobry, S., and Tuia, D. Scale equivariance in cnns with vector fields. *arXiv preprint arXiv:1807.11783*, 2018.

Park, J., Woo, S., Lee, J.-Y., and Kweon, I. S. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

Pashler, H. *Attention*. Psychology Press, 2016.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., and Shlens, J. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

Romero, D. W. and Hoogendoorn, M. Co-attentive equivariant neural networks: Focusing equivariance on transformations co-occurring in data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1g6ogrtDr.

Schmidt, F., Spröte, P., and Fleming, R. W. Perception of shape and space across rigid transformations. *Vision research*, 126:318–329, 2016.

Schwarzer, G. Development of face processing: The effect of face inversion. *Child development*, 71(2):391–401, 2000.

Sosnovik, I., Szmaja, M., and Smeulders, A. Scale-equivariant steerable networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgpugrKPS.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. Tensor Field Networks: Rotation-and Translation-Equivariant Neural Networks for 3D Point Clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Venkataraman, S. R., Balasubramanian, S., and Sarma, R. R. Building deep equivariant capsule networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgNJgSFPS.

von Helmholtz, H. Über die Tatsachen, die der Geometrie zugrunde liegen. 1868.

Wang, X., Girshick, R., Gupta, A., and He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

Weiler, M. and Cesa, G. General e (2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, pp. 14334–14345, 2019.

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

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

Woo, S., Park, J., Lee, J.-Y., and So Kweon, I. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

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

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

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

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.

Zhang, R., Zou, Y., and Ma, J. Hyper-{sagnn}: a self-attention based graph neural network for hypergraphs. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryeHuJBtPH.

Zhaoping, L. *The V1 hypothesiscreating a bottom-up saliency map for preattentive selection and segmentation*, pp. 189–314. 05 2014. ISBN 9780199564668. doi: 10.1093/acprof:oso/9780199564668.003.0005.

Zhu, X., Cheng, D., Zhang, Z., Lin, S., and Dai, J. An empirical study of spatial attention mechanisms in deep networks. *arXiv preprint arXiv:1904.05873*, 2019.