CO-ATTENTIVE EQUIVARIANT NEURAL NETWORKS: FOCUSING EQUIVARIANCE ON TRANSFORMATIONS CO-OCCURRING IN DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariance is a nice property to have as it produces much more parameter efficient neural architectures and preserves the structure of the input through the feature mapping. Even though some combinations of transformations might never appear (e.g. a face with a horizontal nose) current equivariant architectures consider the set of all possible transformations in the transformation group while generating feature representations. Contrarily, the human visual system is able to attend to the set of relevant transformations occurring in the environment as to assist and improve object recognition. Based on this observation, we modify conventional equivariant feature mappings such that they are able to attend to the set of co-occurring transformations in data. Our experiments show that neural networks utilizing co-attentive equivariant feature mappings consistently outperform those utilizing conventional ones both for fully (rotated MNIST) and partially (CIFAR-10) rotational settings.

1 INTRODUCTION

Thorough experimentation in the fields of psychology and neuroscience has provided support to the intuition that our visual perception and cognition systems are able to identify familiar objects despite modifications in size, location, background, viewpoint and lighting (Bruce & Humphreys, 1994). Interestingly, we are not just able to recognize such modified objects, but are able to characterize which modifications have been applied to them as well. As an example, when we see a picture of a cat, we are not just able to tell that there is a cat in it, but also its position, its size, facts about the lighting conditions of the picture, and so forth. Such observations suggest that the human visual system is equivariant to a large transformation group containing translation, rotation, scaling, among others. In other words, the mental representation obtained by seeing a transformed version of an object, is equivalent to that of seeing the original object and transforming it mentally next.sdfgsdfg asdfasdf

These fascinating abilities exhibited by biological visual systems have inspired a large field of research towards the development of neural architectures able to replicate them. Among these, the most popular and successful approach is the Convolutional Neural Network (CNN) (LeCun et al., 1989), which incorporate equivariance to translation via convolution. Unfortunately, in counterpart to the human visual system, CNNs do not exhibit equivariance to other transformations encountered in visual data (e.g. rotations). Interestingly, if an ordinary CNN happens to learn rotated copies of the same filter, the stack of feature maps becomes equivariant to rotations even though individual feature maps are not (Cohen & Welling, 2016). Since ordinary CNNs must learn such rotated copies independently, they effectively utilize an important number of network parameters suboptimally to this end (see Fig. 3 in Krizhevsky et al. (2012)). Based on the idea that equivariance in CNNs can be extended to larger transformation groups by stacking convolutional feature maps, several approaches have emerged to extend equivariance to, e.g. planar rotations (Dieleman et al. [2016] Marcos et al., 2017; Weiler et al. [2018]; Li et al. [2018]), spherical rotations (Cohen et al. [2018]; Worrall & Brostow, 2018), scaling (Marcos et al. [2018]) and general transformation groups (Cohen & Welling [2016]), such that transformed copies of a single entity are not required to be learned independently.
Figure 1: The strength of context. Our visual system infers object identities according to their size, location and orientation in the scene. In this picture, observers describe the scene as containing a car and a pedestrian in the street. However, the pedestrian is in fact the same shape as the car, except for a $90^\circ$ rotation. The atypicality of this orientation for a car within the context defined by the street scene causes the car to be recognized as a pedestrian. Extracted from Oliva & Torralba (2007).

Although incorporating equivariance to two arbitrary transformation groups is conceptually and theoretically similar, evidence from real-world experiences motivating their integration might strongly differ. Several studies in neuroscience and psychology have shown that our visual recognition system does not react equally to all the transformations we encounter in visual data. Consider, for instance, the case of translation and rotation. Although we easily recognize objects independently of their position of appearance, a large corpus of experimental research has shown that this is not always the case for in-plane rotations. Yin (1969) showed that mono-oriented objects, i.e. complex objects such as faces which are customarily seen in one orientation, are much more difficult to be accurately recognized when presented upside-down. This behaviour has been reproduced, among others, for magazine covers (Dallett et al., 1968), symbols (Henle, 1942) and even familiar faces (e.g. from classmates) (Brooks & Goldstein, 1963). Intriguingly, Schwarzer (2000) found that this effect exacerbates with age (adults suffer from this effect much more than children), but, adults are much faster and accurate in detecting mono-oriented objects when appearing in usual orientations. Based on these studies, we can draw the following conclusions:

- The human visual system does not perform (fully) equivariant feature transformations to visual data. Consequently, it does not react equally to all possible input transformations encountered in visual data, even if they belong to the same transformation group (e.g. in-plane rotations).
- The human visual system does not just encode familiarity to objects but seems to learn through experience the poses in which these objects customarily appear in the environment to assist and improve object recognition (Freire et al., 2000; Riesenhuber et al., 2004; Sinha et al., 2006).

Moreover, complementary studies (Tarr & Pinker, 1989; Oliva & Torralba, 2007) suggest that our visual system encodes atypicality of object orientations relative to their context rather than on an absolute manner (Fig. 1). Motivated by the aforementioned observations we state the co-occurrence envelope hypothesis:

**The Co-occurrence Envelope Hypothesis.** By allowing equivariant feature mappings to dynamically learn the extent of transformations exhibited in the data, and utilize this information such that learned feature representations are optimal in the set of transformations that co-occur with one another (i.e. the co-occurrence envelope of the data), we are able to develop more adequate feature representations of the data, and, consequently, improve the recognition capability of neural networks utilizing such feature mappings. We refer to one such feature mapping as co-attentive equivariant.

In this work, we introduce co-attentive equivariant feature mappings and apply them in the context of equivariant neural networks. To this end, we leverage the concept of attention (Bahdanau et al., 2014) to modify existing mathematical frameworks for equivariance, such that equivariant neural networks are able to attend to the co-occurrence envelope of the data. It is critical not to disrupt equivariance during the attention process, such that subsequent layers are able to take advantage of it as well. To this end, we introduce cyclic equivariant self-attention, a novel equivariance preserving attention mechanism with which equivariance to the entire group is not disrupted.

\[^1\] It is achieved by developing feature mappings that utilize the transformation group in the feature mapping itself (e.g. translating a filter in the course of a feature transformation is used to obtain translation equivariance).
Figure 2: Effect of multiple attention strategies for the prioritization of relevant pattern orientations in rotation equivariant networks for the task of face recognition. Given that all attention strategies are learned exclusively from upright faces, we show the set of relevant directions for the recognition of faces in two orientations (Fig. 2a) obtained by: no attention (Fig. 2b), attending to the pattern orientations of appearance independently (Fig. 2c) and, attending to the pattern orientations of appearance relative to one another (Fig. 2d). Built upon Figure 1 from Schwarzer (2000).

Identifying the co-occurrence envelope. Consider a rotation equivariant network receiving two copies of the same face as shown in (Fig. 2a). A conventional rotation equivariant neural network is required to perform inference and learning on the set of all possible orientations for all visual pattern constituting a face regardless of the input orientation (Fig. 2b). However, by virtue of its rotation equivariance, it is able to recognize rotated faces even if it is trained on upright faces only. A possible strategy to simplify the task at hand could be to restrict the network to react exclusively to upright faces (Fig. 2c). In this case, the set of relevant visual pattern orientations becomes much smaller, simplifying both learning and inference. Unfortunately, this comes at the cost of disrupting equivariance to the rest of the rotation group. Resultantly, the network would risk becoming unable to detect faces in any other orientation than those it is trained on. A better strategy results from restricting the set of relevant pattern orientations by defining them relative to one another (e.g. mouth orientation w.r.t. the eyes) as opposed to absolutely (e.g. upright mouth) (Fig. 2d). In such a way, we are able to exploit information about the co-occurrence of orientations present in the data without disrupting equivariance to the rotation group. The set of co-occurent orientations corresponds to the co-occurrence envelope of the samples in Fig. 2a for the transformation group defined by rotations.

Experiments and results. We explore the effects of co-attentive equivariant feature mappings in the context of rotation equivariant networks. Specifically, we allow the rotation equivariant networks introduced by Cohen & Welling (2016) (p4-CNNs) and Li et al. (2018) (DRENS) to attent to the co-occurrence envelope of the data by replacing conventional rotation equivariant feature mappings with co-attentive ones. We show that co-attentive rotation equivariant neural networks consistently outperform their conventional counterparts for the task of object recognition in fully (rotated MNIST) and partially (CIFAR-10) rotational settings. Subsequently, we extend cyclic equivariant self-attention to attent to multiple similarity groups and apply it to p4m-CNNs (Cohen & Welling, 2016) (equivariant to rotation and mirror reflections). Our results are consistent with those obtained for single transformations groups. Our results support the stated hypothesis.

Contributions.

- We introduce the concept of co-attentive equivariant feature mapping and apply it in the context of equivariant neural networks, such that they attent to the co-occurrence envelope of the data.

- We demonstrate that co-attentive equivariant networks consistently outperform their conventional equivariant counterparts for fully and partially rotational settings.

- We show that the notion of co-attentive equivariant feature mappings is generalizable to multiple symmetry groups, such that we are able to attend to the co-occurrence envelope of the data for multiple symmetries simultaneously.

- To the best of our knowledge, this is the first work to utilize attention in equivariant neural networks, others than CNNs, for a purpose other than the generation of invariant feature representations (Kuzminykh et al., 2018).
The remainder of this document is organized as follows: In Section 2, we introduce key preliminary concepts that serve as foundations for our approach. In Section 3 we introduce co-attentive feature mappings in the context of rotational equivariant networks to, subsequently, generalize the idea of co-attentive feature mappings to larger symmetry groups. We evaluate our approach in Section 4. In Section 5 we discuss possible future work directions to conclude our work in Section 6.

## 2 Preliminaries

In this section we present key preliminary concepts that serve as foundations for our approach. We begin by introducing the notion of equivariance and motivate its importance for neural network architectures. Subsequently, we illustrate the construction and training procedures of conventional equivariant neural networks to conclude with some key implications that motivate our approach.

### Equivariance

We say that a feature mapping \( f : \mathcal{X} \to \mathcal{Y} \) is equivariant to a transformation group \( \Phi \) if for any transformation \( \phi \in \Phi \), \( \phi : \mathcal{X} \to \mathcal{X} \) of an input \( x \in \mathcal{X} \), there exists a transformation \( \psi \in \Psi \), \( \psi : \mathcal{Y} \to \mathcal{Y} \) in the feature space \( \mathcal{Y} \) to which \( \phi \) is associated. Formally:

\[
f(\phi(x)) = \psi(f(x))
\]

In other words, the ordering in which we apply the transformation and the feature mapping is inconsequential. There are multiple reasons as to why equivariant feature representations are advantageous for learning systems. Since the transformations group \( \Phi \) produces predictable and interpretable transformations \( \Psi \) in the feature space, the learning process is simplified (Worrall et al., 2017). Moreover, equivariance allows the construction of \( L \)-layered networks by stacking several feature mappings \( \{f^1, \ldots, f^L\} \) without disrupting the input structure as regarded by the transformation group \( \Phi \) (e.g. CNNs and input translations). As a result, an intermediate representation \((f^1 \circ \ldots \circ f^l)(x)\) preserves the input structure and allows the subsequent layers to take advantage of it as well. Invariance is an special case of equivariance in which \( \Psi = \{1\} \), the identity, and thus transformed inputs are mapped to the same feature representation.

### Equivariant Neural Networks

The integration of equivariance to arbitrary transformation groups \( \Phi \) in neural architectures has been achieved by developing feature mappings \( f^l \) that utilize the transformation group \( \Phi \) in the feature mapping itself. Interestingly, equivariant feature mappings encode equivariance in form of parameter sharing with respect to \( \Phi \), i.e. the same weights are reused for every \( \phi \in \Phi \), which makes the incorporation of larger transformation groups extremely appealing in the context of parameter efficient networks.

Formally, the \( l \)-th layer of a neural network receives a 3-D tensor \( \mathbf{F}^l \in \mathbb{R}^{I^l \times R^l \times C^l} \) with \( I^l \) channels, \( R^l \) rows and \( C^l \) columns, and applies the feature mapping \( f^l : \mathbb{R}^{I^l \times R^l \times C^l} \to \mathbb{R}^{I^{l+1} \times R^{l+1} \times C^{l+1}} \) to generate the output feature representation \( \mathbf{F}^{l+1} \). In the context of convolutional layers, the feature mapping \( f^l := f^l_T \) is defined by a convolution between the input tensor \( \mathbf{F}^l \), with element \( \mathbf{F}^l_{i,j,k} \) giving the input unit within channel \( i \) at row \( j \) and column \( k \), and a 4-D learnable kernel tensor \( \mathbf{K}^l \), with element \( K^l_{i,j,k,m} \) giving the connection strength between an output unit in the channel \( i \) and an input unit in the channel \( j \) with an offset of \( k \) rows and \( m \) columns between the output and input units. Formally:

\[
\mathbf{F}^{l+1}_{i,j,k} = f^l_T(\mathbf{F}^l)_{i,j,k} = \sum_{p,m,n} \mathbf{F}^l_{p,j+m-1,k+n-1} K^l_{i,p,m,n}
\]

By sliding \( \mathbf{K}^l \) across \( \mathbf{F}^l \), CNNs preserve the spatial structure of the input through the feature mapping \( f^l_T \) and successfully provide equivariance to the translation group \( \Phi_T \). The underlying idea for the inclusion of additional equivariances in CNNs is conceptually equivalent to the strategy utilized by LeCun et al. (1989) for translation equivariance. Consider, for instance, the inclusion of equivariance to the set of rotations by \( \theta_r \) degrees: \( \Theta = \{\theta_r = r \frac{2\pi}{r_{\text{max}}} \mid r \in \{1, \ldots, r_{\text{max}}\}\} \). To this end, we modify the feature mapping \( f^l := f^l_R \) to include the rotations defined by \( \Theta \). Let \( \mathbf{K}^l_{i,j,:,:} \) be a \( \theta_r \)-rotated instance of \( \mathbf{K}^l \) in the plane \( \mathbf{K}^l_{i,j,:,:} \). The rotational convolution \( f^l_R \) is defined as the

---

*Derivation taken from Goodfellow et al. (2016)*
concatenation of the convolutions $f^{l,\theta_r}_T$ of the input and the rotated versions of the kernel $K^{l,\theta_r}$:

$$
F^{l+1} = f^{l}_R(F^l) = [f^{l,0}_T(F^l),...,f^{l,\theta_{max}}_T(F^l)]
$$

$$
F^{l+1}_{i,j,k} = f^{l,\theta_r}_T(F^l)_{i,j,k} = \sum_{p,m,n} F^l_{p,j+m-1,k+n-1} K^{l,\theta_r}_{i,p,m,n}
$$

For the sake of clarification we have simplified the actual mathematical derivations for rotation equivariance provided by [Cohen & Welling (2016), Li et al. (2018) and other authors]. Note that $f^{l}_R$ produces $(\text{dim}(\Theta) = r_{\text{max}})$ more output feature maps than $f^{l}_T$. Consequently, we need to learn much smaller kernel tensors $K$ to produce the same number of output feature channels.

**Learning Equivariant Neural Networks.** In general, a parameter $k^l$ from the $l$-th layer is updated proportionally to its gradient with respect to an arbitrary objective function $E$:

$$
\frac{\partial E}{\partial k^l} = \frac{\partial E}{\partial F^{l+1}} \frac{\partial F^{l+1}}{\partial k^l} = \frac{\partial E}{\partial f^l(F^l)} \frac{\partial f^l(F^l)}{\partial k^l}
$$

where the term $-\frac{\partial E}{\partial k^l}$ is the gradient of the objective function backpropagated from the following layer. Now, consider the feature mapping of an ordinary convolutional layer $f^l_T$ (Eq. 2) and a rotation equivariant convolutional layer $f^l_R$ (Eqs. 3-4). The corresponding gradients with regard to $E$ are given by Eq. 5 and Eq. 7 respectively:

$$
\frac{\partial E}{\partial K^l_{i,j,k,p}} = \sum_{m,n} \frac{\partial E}{\partial F^{l+1}_{i,m,n}} F^l_{j,(m-1)+k,(n-1)+p}
$$

$$
\frac{\partial E}{\partial K^l_{i,j,k,p}} = \sum_{r,m,n} \frac{\partial E}{\partial F^{l+1}_{i,m,n}} (f^{l,\theta_r}_T)_{j,(m-1)+k,(n-1)+p}
$$

The superscript $\theta_r$ refers to a $\theta_r$-rotated version of a tensor in the opposite direction. When contrasting the update mechanism of both layers, we clearly observe that additionally to getting feedback from every output planar position $(m, n)$, rotation convolutional layers do so on all of the $r$ feature maps generated by the rotational convolution. Consequently, just as ordinary convolutional layers tend to learn features that perform well on the entire image, hence for the entire translation group $\Phi_T$, their rotation equivariant extensions do so for features that perform well on the entire image and for all rotations, thus for the set composed by the translation and rotation groups $\{\Phi_T, \Phi_R\}$.

## 3 CO-ATTENTIVE EQUIVARIANT NEURAL NETWORKS

In this section we define co-attentive feature mappings and apply them in the context of equivariant neural networks. To this end, we introduce cyclic equivariant self-attention and utilize it to construct co-attentive rotation equivariant neural networks. Subsequently, we show that cyclic equivariant self-attention is extendable to larger symmetry groups and make use of this fact to construct co-attentive neural networks equivariant to rotation and mirror reflections.

### 3.1 CO-ATTENTIVE ROTATION EQUIVARIANT NEURAL NETWORKS

To allow rotation equivariant networks to utilize and learn co-attentive rotation equivariant representations, we introduce an attention operator $A$ on top of the rotational convolution operator with which discernment among the $r_{\text{max}}$ feature responses $f^{l,\theta_r}_T(F^l)$ generated by the $\theta_r$-rotated variants of the kernel tensor $K^l$ is possible. Formally, our co-attentive rotation equivariant feature mapping $f^l_R$ is defined as follows:

$$
F^{l+1} = f^{l}_R(F^l) = A(f^{l}_R(F^l)) = A\left([f^{l,0}_T(F^l),...,f^{l,\theta_{max}}_T(F^l)]\right)
$$

Theoretically, we could apply an arbitrary attention operator $A$ globally over $f^{l}_R(F^l)$ as depicted in Eq. 8. However, we apply attention locally as to (1) grant the algorithm enough flexibility
to attend locally to the co-occurrence envelope of feature representations and (2) utilize attention exclusively along the axis of rotated responses, such that our contributions are clearly separated from those possibly emerging from spatial attention. To this end, we apply attention at a pixel level on top of the rotational convolution responses (Eq. [9]). Furthermore, we assign a single attention instance $A_i$ to each learned feature representation, i.e. each output channel of the kernel tensor $K_i$ to handle feature representations independently. Subsequently, we utilize the attention instances $A_i$ correspondingly across the spatial dimension of the output feature maps:

$$[F_{i,j,k}^{l+1,1}, ..., F_{i,j,k}^{l+1,r_{max}}] = A_i\left(\left[f_{T}^{l,\theta_i}(F_i)_{i,j,k}, ..., f_{T}^{l,\theta_i}(F_i)_{i,j,k}\right]\right)$$

(9)

**Attention and self-attention.** Consider a source vector $x = (x_1, ..., x_n)$ and a target vector $y = (y_1, ..., y_m)$. In general, an attention operator $A$ leverages information from the source vector $x$ (or multiple feature mappings thereof) to estimate an attention matrix $A \in [0, 1]^{n \times m}$, such that (1) the element $A_{i,j}$ provides an importance assessment of the source element $x_j$ with reference to the target element $y_j$ and (2) the sum of importance over all $x_j$ is equal to one: $\sum_i A_{i,j} = 1$. Subsequently, the matrix $A$ is utilized to modulate the original source vector $x$ as to mainly “attend” to a subset of relevant source positions with regard to $y_j$: $\hat{x}_j = (A_{i,j})^T \odot x$. A special case of attention is that of self-attention (Cheng et al., 2016), in which the target vector is equal to the source vector and thus, the attention mechanism relates the entire input sequence $x$ to each element $x_j$ for its weighting.

The attention matrix $A$ is conventionally constructed via nonlinear space transformations $f_A : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ on the source vector $x$, on top of which the softmax function is applied to obtain the aforementioned properties: $A_{i,j} = \text{softmax}(f_A(x_i))$. Typically, attention mappings $f_A$ found in literature take pairs of feature transformations of $x$ as input (e.g. $\{s, H\}$ in RNNs (Luong et al., 2015), $\{Q, K\}$ in self-attention networks (Vaswani et al., 2017)), and perform (non)-linear mappings on top of it ranging from multiple feed-forward layers (Bahdanau et al., 2014) to several operation types between the transformed pairs (Luong et al., 2015; Vaswani et al., 2017; Mishra et al., 2018). Due to the computational complexity of such approaches and the fact that we do extensive pixel-wise usage of $f_A$ on every network layer, we modify the usual self-attention formulation as to provide more descriptive power in a much more compact setting.

**Compact local self-attention.** Initially, we relax the range of values of $A$ from $[0, 1]^{n \times n}$ to $\mathbb{R}^{n \times n}$. This allows us to encode much richer relationships between the pairs of elements $(x_i, x_j)$ at the cost of less interpretability. Subsequently, we define $A = x^T \odot \hat{A}$, where $\hat{A} \in \mathbb{R}^{n \times n}$ is a matrix of learnable parameters. Instead of applying column-wise softmax directly to $\hat{A}$, we first sum over the contributions of each element $x_j$ to obtain a vector $a = (\sum_i A_{i,1}, ..., \sum_i A_{i,n})$, which is then passed to the softmax function. Following Vaswani et al. (2017), we prevent the softmax function from reaching regions of low gradient by scaling its argument by $(\sqrt{\text{dim}(\hat{A})})^{-1} = (1/n): \tilde{a} = \text{softmax}(1/n a)$. Lastly, we counteract the softmax contractive behaviour by normalizing $\tilde{a}$ before weighting $x$ to conserve the magnitude range of the input values. This allows us to use $\hat{A}$ in deep architectures. Formally, our compact modified self-attention mechanism is defined as:

$$\tilde{a} = \text{softmax}(1/n \ x \hat{A})$$

(10)

$$\hat{x} = \hat{A}(x) = (\tilde{a} / \text{max}(\tilde{a})) \odot x$$

(11)

**The cyclic equivariant self-attention operator $A_C$.** Consider the vector $[F_{i,j,k}^{l+1,1}, ..., F_{i,j,k}^{l+1,r_{max}}]$ of responses generated by the rotational convolution $f_R^{l}$ between an input tensor $F_i$ and a kernel tensor $K_i$ in the channel $i$ at row $j$ and column $k$. By applying self-attention, we obtain an importance matrix $A \in \mathbb{R}^{r_{max} \times r_{max}}$ relating the response of a particular $\theta_i$-rotated response to the entire rotational group response at a certain position. We refer to this attention mechanism as full self-attention ($A_F$). Although $A_F$ is able to encode arbitrary linear source-target relationships for each target position, it is not restricted to conserve equivariance to the rotational group. Resultantly, we risk incurring into the behavior outlined in Fig. [2c]. Before we elaborate on the problem at hand, we introduce the cyclic permutation operator $P^i$, which produces a cyclic shift of $i$ positions on its input.

Consider an input pattern $p$ to which the rotational convolution $f_R^{l}$ produces a strong response in the $j$-th feature map $F_j^{l+1,j}$ only (corresponding to $K_j^{l,\theta_i}$). During learning, only the corresponding attention coefficients $A_{i,j}$ would be significantly increased. Subsequently, consider the presence of the input pattern $p^{\theta_i}$, a $\theta_i$-rotated variant of $p$. By virtue of the rotational equivariance property of
the feature mapping $f_{\tilde{R}}^l$, we obtain (locally) an exactly equal response to that of $p$ up to a cyclic permutation of $i$ positions, producing a strong activation in the output feature map $P^i(F^{l+1}) = F^{l+1, (j+i) \mod r_{\text{max}}} (\text{corresponding to } K^{l, (j+i) \mod r_{\text{max}}})$. We encounter two problems in this setting: $A_C$ would not be able to detect that $p$ and $p^\theta$ correspond to the exact same input pattern and, as each but the attention coefficients $A_{i,j}$ is small, the network might considerably damp the response generated by $p^\theta$. As a result, the network might (1) squander important feedback information during learning and (2) induce learning of repeated versions of the same pattern for different orientations.

Interestingly, we are able to introduce prior-knowledge into the attention model by restricting the learning and (2) induce learning of repeated versions of the same pattern for different orientations.

Interestingly, we are able to introduce prior-knowledge into the attention model by restricting the learning and (2) induce learning of repeated versions of the same pattern for different orientations.

A circulant matrix specified by a learnable attention vector $\tilde{\theta}$. Formally, a circulant matrix $C = p_{\tilde{\theta}}$, such that its $j$-th column is a cyclic permutation of $j - 1$ positions of $c$: $C_{c,j} = p^{-1}(c)^T$. We construct our cyclic equivariant self-attention operator $A_C$ by defining $\tilde{A}$ as a circulant matrix specified by a learnable attention vector $\tilde{\theta} = (\tilde{a}_1, \ldots, \tilde{a}_{r_{\text{max}}})$:

$$A = [(\alpha^c)^T, P^1(\alpha^c)^T, \ldots, P^{r_{\text{max}}-2}(\alpha^c)^T, P^{r_{\text{max}}-1}(\alpha^c)^T]$$

(12)

and subsequently applying Eq. [10] and Eq. [11]. Resultantly, $A_C$ is able to assign the responses generated by $f_{\tilde{R}}^l$ to rotated versions of an input pattern $p$ to a unique entity: $f_{\tilde{R}}^l(p^\theta) = P^l(f_{\tilde{R}}^l(p))$, and dynamically adjust its output to the angle of appearance $\theta$, such that the attention operation does not disrupt its propagation downstream the network: $A_C(f_{\tilde{R}}^l(p^\theta)) = P^l(A_C(f_{\tilde{R}}^l(p)))$. Consequently, the attention weights $\alpha^c$ are updated equally regardless of specific values of $\theta$. Due to these properties, $A_C$ does not incur in any of the problems outlined earlier in this section. Conclusively, our co-attentive rotation equivariant feature mapping $f_{\tilde{R}}^l$ is defined as follows:

$$F^{l+1} = f_{\tilde{R}}^l(F^l) = A_C(f_{\tilde{R}}^l(F^l)) = A_C(f_{\tilde{T}}^l(F^l)) = \tilde{A} \left( \left[ f_{\tilde{T}}^l(F^l)_{i,c}, \ldots, f_{\tilde{T}}^l(F^l)_{i,c} \right] \right)$$

(13)

Note that a co-attentive equivariant feature mapping $f_{\tilde{R}}^l$ corresponds to a conventional equivariant feature mapping $f_{\tilde{R}}^l$ if $\tilde{A} = \alpha I$ for any $\alpha \in \mathbb{R}$.

### 3.2 Extending $A_C$ to Multiple Symmetry Groups

The self-attention mechanisms outlined in the previous section are easily extendable to larger transformation groups consisting of multiple symmetries. Consider, for instance, the transformation group $\theta, m$ consisting of rotations by $\theta$ degrees and mirror reflections $m$ as defined by Cohen & Welling (2016). Let $\theta, m$ be a feature mapping equivariant to the $\theta, m$ group. The group convolution $f_{\theta, m}$ produces two times as many output channels ($2r_{\text{max}}$) as those generated by the rotational convolution $f_{\tilde{R}}$ (Eq. [3][4][5]), which correspond to rotated and mirrored rotated versions of a learned kernel tensor $K$. Full self-attention $A_F$ can be integrated directly by modulating the output of the feature mapping $f_{\theta, m}$ by a learnable matrix $\tilde{A} \in \mathbb{R}^{2r_{\text{max}} \times 2r_{\text{max}}}$ that relates the group convolution responses with one another as depicted in Section [5.1]. However, just as for the rotation group, $A_F$ disrupts the equivariance property of the feature mapping to the entire transformation group.

Similarly, the cyclic equivariant self-attention operator $A_C$ can be extended to multiple symmetry groups as well. Before we continue, we introduce the cyclic permutation operator $P^{\theta}$, which produces a cyclic shift of $i$ positions on its input along the permutation axis of the transformation $t$. Consider the input patterns $p$ and $p^\theta$ outlined in the previous section and $p^\theta$, a mirrored instance of the input pattern $p$. Let $\{ F_{i,j,k}^{m}, F_{i,j,k}^{m}^{\theta}, \ldots, F_{i,j,k}^{m}^{\theta}, \ldots \} = \{ r, m \} = f_{\theta, m}(p)$ be the vector of responses generated by the group convolution $f_{\theta, m}$ in the channel $i$ at row $j$ and column $k$ for the input pattern $p$. The first $r_{\text{max}}$ values, $r$, correspond to the responses of the rotated versions of $K$ and the remaining $r_{\text{max}}$ values, $m$, to those of the mirrored rotated versions of $K$. By virtue of the
rotation equivariance property of the feature mapping, the response generated to $p^{θ_0}$ is equivalent to that of $p$ up to a cyclic permutation of $i$ positions along the dimension of rotations: $f_{θ,m}(p^{θ_0}) = \mathcal{P}^i,θ_{θ_0}(f_{θ,m}(p)) = [\mathcal{P}^i(r), \mathcal{P}^i(r^m)]$. Similarly, by virtue of the mirror equivariance property of the feature mapping $f_{θ,m}$, the response generated to $p^m$ is equivalent to that of $p$ up to a cyclic permutation of one position along the dimension of mirroring: $f_{θ,m}(p^m) = \mathcal{P}^1,θ_{θ_0}(f_{θ,m}(p)) = [\mathcal{P}^1(r, r^m)] = [r^m, r]$. Note that if we have two symmetry transformations $g$ and $h$ and we compose them, the result $(g \circ h)$ is another symmetry transformation (Cohen & Welling [2016]). Resultantly, $f_{θ,m}(p^{θ_0}, m) = (\mathcal{P}^1,θ_{θ_0} \circ \mathcal{P}^i,θ_{θ_0})(f_{θ,m}(p)) = \mathcal{P}^{1,m}([\mathcal{P}^i(r), \mathcal{P}^i(r^m)]) = [\mathcal{P}^i(r^m), \mathcal{P}^i(r)]$.

In simpler terms, in order to extend $A_c$ to multiple symmetry groups, we are required to restrict the structure of $A$ to that of a block matrix, such that: (1) the composing blocks permute internally as defined by one of the transformation groups (here $\mathcal{P}^z,θ_{θ_0}$) and (2) the blocks themselves permute with one another as defined by the the remaining transformation group (here $\mathcal{P}^{1,m}$). For the case of the group convolution $f_{θ,m}$ outlined here $A$ is defined as follows:

$$A = \begin{bmatrix} \tilde{A_1} & \tilde{A_2} \\ \tilde{A_2} & \tilde{A_1} \end{bmatrix}$$

(14)

where $\{\tilde{A_i} \mid i \in \{1, 2\} \in \mathbb{R}^{r_{\text{max}} \times r_{\text{max}}}$ are circulant matrices (Eq. 12). Note that the ordering of the permutation directions are interchangeable as long as the input vector is modified correspondingly.

By following this same strategy, cyclic self-attention can be extended to additional symmetry groups.

4 Experiments

Experimental Setup. We validate our approach by extending the equivariant architectures provided by Cohen & Welling [2016] (G-CNNs) and Li et al. [2018] (DRENs). We evaluate both strategies for classification in fully rotational (rotated MNIST) and partially rotational settings (CIFAR-10).

To this end, we modify all of the G-CNNs and the DRENs proposed in the corresponding works by replacing rotation equivariant layers with co-attentive rotation equivariant layers as to attend to the co-occurrence envelope of the data. Unless specified otherwise, we utilize the same data processing, initialization strategies, hyperparameter values and evaluation strategies utilized by the baselines for the corresponding neural architectures in our experiments. Note that the goal of this paper is to study and evaluate the relative effects obtained by co-attentive equivariant networks with regard to their conventional counterparts. Accordingly, we do not perform any additional tuning relative to the baselines. We believe that improvements on our reported results are feasible by performing further parameter tuning (e.g. on structure or hyperparameters) of the co-attentive equivariant networks.

The additional learnable parameters, i.e. those associated to the cyclic self-attention operator ($\tilde{A}$) are initialized identically to the rest of the layer. Subsequently, we replace the values of $\tilde{A}$ along the diagonal by 1 (i.e. $\text{diag}(\tilde{A}_{\text{init}}) = 1$) such that $\tilde{A}_{\text{init}}$ approximately resembles the identity $I$ and, hence, co-attentive equivariant layers are initially approximately equal to equivariant ones.

Rotated MNIST. The rotated MNIST dataset [Larochelle et al. 2007] contains 62000 gray-scale 28x28 handwritten digits uniformly rotated on the entire circle $[0, 2\pi]$. The dataset is split into training, validation and tests sets of 10000, 2000 and 50000 samples, respectively.

We replace rotation equivariant layers in P4CNN (Cohen & Welling [2016], DREN and DREN-MaxPooling (Li et al. [2018]) with co-attentive rotation equivariant layers. Our results show that co-attentive rotation equivariant networks consistently outperform rotation equivariant ones without any additional parameter tuning (see Table I).

CIFAR-10. The CIFAR-10 dataset [Krizhevsky et al. 2009] consists of 60000 real-world 32x32 RGB images uniformly drawn from 10 classes. Contrarily to the rotated MNIST dataset, this dataset does not exhibit rotation symmetry. The dataset is split into training, validation and tests sets of 40000, 10000 and 10000 samples, respectively.

We replace equivariant layers in the $p4$ and $p4m$ variations of the All-CNN (Springenberg et al. 2014) and the ResNet44 (He et al. 2016) proposed by Cohen & Welling [2016] with co-attentive

Our proposed architectures are signalized with the prefix $a$, e.g. $a-p4m$-All-CNN.
Table 1: Comparison between conventional, equivariant and co-attentive equivariant networks on rotated MNIST (left) and CIFAR-10 (right). Test set error rates and number of parameters are reported. Values between parenthesis correspond to relevant results obtained from our own implementation.

<table>
<thead>
<tr>
<th>Rotated MNIST</th>
<th>CIFAR-10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network</strong></td>
<td><strong>Test Error (%)</strong></td>
</tr>
<tr>
<td>Z2CNN</td>
<td>5.03 ± 0.002</td>
</tr>
<tr>
<td>P4CNN</td>
<td>2.28 ± 0.0004</td>
</tr>
<tr>
<td>a-P4CNN</td>
<td><strong>2.06 ± 0.0429</strong></td>
</tr>
<tr>
<td>DREN</td>
<td>1.78 (1.99)</td>
</tr>
<tr>
<td>a-DREN</td>
<td><strong>1.674</strong></td>
</tr>
<tr>
<td>DRENMxPool.</td>
<td>1.56 (1.60)</td>
</tr>
<tr>
<td>a-DRENMxPool.</td>
<td><strong>1.34</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* We were not able to replicate the results reported in Cohen & Welling (2016) for any of the ResNet44 architectures based on the online implementation.

Equivariant ones. Likewise, we modify the $r \times 4$-variations of the NIN (Lin et al., 2013) and ResNet20 (He et al., 2016) models proposed by Li et al. (2018) in the same manner. Our results show that co-attentive rotation equivariant networks consistently outperform rotation equivariant in this setting as well (see Table 1).

Training convergence of equivariant networks. Li et al. (2018) reported that adding too many rotational equivariant (isotonic) layers decreased the performance of their models on CIFAR-10. As a consequence, they did not report results of fully rotational equivariant networks for this setting and attributed this behaviour to the non-symmetry of the data. We noticed that with equal initialization strategies rotational equivariant CNNs were much more prone to divergence than ordinary CNNs. This behaviour can be traced back to the additional feedback resulting from rotational convolutions (Eq. 7) compared to ordinary ones (Eq. 6). After further analysis, we noticed that the data preprocessing strategy utilized by Li et al. (2018) leaves some very large outlier values in the data ($||x|| > 100$), which strongly contribute to the behaviour outlined before. In order to evaluate the relative contribution of co-attentive equivariant neural networks we constructed fully DREN equivariant architectures based on their implementation. Although the obtained results were much worse than those originally reported in Li et al. (2018), we were able to stabilize training such that the same hyperparameters could be kept equal across network types by clipping input values outside of the 99 percentile of the data ($||x|| \leq 2.3$) and reducing the learning rate to 0.01. The obtained results (see Table 1) signalize that DREN networks are comparatively better than CNNs both in fully and partially rotational settings, contradictorily to the conclusions drawn in Li et al. (2018). This behaviour elucidates the fact that although the inclusion of equivariance to larger transformation groups is beneficial for neural architectures both in terms of accuracy and parameter efficiency, one must be aware that such benefits are directly associated to an increase of the susceptibility of the network to divergence during training. This is caused due to an increase of the information flow relative to the number of parameters in the network.

5 Discussion and Future Work

Our results show that co-attentive equivariant feature mappings can be used to improve results obtained by conventional equivariant ones. Interestingly, attending to the co-occurrence envelope of the data is beneficial for fully rotational settings as well. We attribute this to the fact that a set of co-occurring orientations between patterns can be easily defined (and exploited) in both settings.
In future work, we want to utilize and extend more complex attention strategies (e.g. Bahdanau et al. (2014); Luong et al. (2015); Vaswani et al. (2017); Mishra et al. (2017)) such that they can be applied to large transformation groups without disrupting equivariance and possibly capture richer relationships than that of co-occurrence. This becomes very challenging from the computational perspective as well, as it requires extensive usage of the corresponding attention mechanism. Furthermore, we want to extend co-attentive equivariant feature mappings to continuous (e.g. Worrall et al. (2017)) and 3D space (e.g. Cohen et al. (2018); Worrall & Brostow (2018)) groups, and for applications other than visual data (e.g. speech recognition). Finally, we believe that our approach could be refined and extended to a first step towards dealing with the problem of enumeration of transformations in large groups (Gens & Domingos 2014) by dynamically attending (and possibly restricting) transformation groups to the set of co-occurring transformations in data.

6 Conclusion

We have introduced the concept of co-attentive equivariant feature mapping and applied it in the context of equivariant neural networks. By attending to the co-occurrence envelope of the data, we are able to improve the performance of conventional equivariant ones on fully (rotated MNIST) and partially (CIFAR-10) rotational settings. We developed cyclic equivariant self-attention, an attention mechanism able to attend to the co-occurrence envelope of the data without disrupting equivariance to a large set of transformation groups (i.e. all the transformation groups that produce cyclic permutations on their responses). Based on our results, we validate the co-occurrence envelope hypothesis.

Acknowledgments

Omitted for the sake of the double-blind review.

References


