GROUP ACTION EQUIVARIANCE AND GENERALIZED CONVOLUTION IN MULTI-LAYER NEURAL NETWORKS

Pan Zhong, Zhengdao Wang

Department of Electrical and Computer Engineering, Iowa State University, USA

ABSTRACT

Convolutional neural networks have achieved great success in speech, image, and video signal processing tasks in recent years. There have been several attempts to justify the convolutional architecture and to generalize the convolution operation for treatment of other data types such as graphs and manifolds. Based on group representation theory and noncommutative harmonic analysis, it has recently been shown that the so-called group equivariance requirement of a feed-forward neural network necessitates the convolutional architectures. In this paper, based on the familiar concepts of linear time-invariant systems, we develop an elementary proof of the same result. The nonlinear activation function, being a necessary components of practical deep neural networks, has been glossed over in previous analyses of the connection between equivariance and convolution. We identify sufficient conditions for the non-linear activation functions to preserve equivariance, and hence the necessity of the group convolution structure. Our analysis method is simple and intuitive, and holds the potential to be applied to more challenging scenarios such as non-transitive domains and multiple simultaneous equivariances.

Index Terms— group equivariance, convolutional neural network, algebraic convolution, nonlinear activation function

1. INTRODUCTION

In classification problems, *domain invariance*, which refers to the fact that the label is invariant to certain transformations of the input features, is present in many problems such as image classification and graph pattern recognition. Invariance is a special case of *equivariance* which means that when the input feature is transformed the output feature will transform in a similar way. In implementations that rely on multi-layer neural networks, domain invariance can be achieved by making every layer except the last one *equivariant*, and adding invariance only in the last layer. The equivariance property, which is present in networks such as convolutional neural networks (CNN) [1], greatly reduces the number of network parameters and as a result reduces chances of overfitting. In CNN, domain invariance is achieved in two ways. First, the convolution layer applies the same kernel to every part of the image. As a result, it yields translation equivariance which means when the input is translated, the output of a convolution layer will be translated in the same way. Second, the pooling layer will introduce local invariance. The receptive field increases as we reach deeper layers, so that the pooling layer in the higher level actually achieves global invariance. The good performance of CNN on image and video related tasks have triggered the study of generalized convolution networks [2, 3, 4, 5, 6] and alternatives in other tasks.

Several techniques can be used in deep neural networks to achieve equivariance [7]. Data augmentation is a simple method to inject soft or implicit constraints for equivariance [7]. It is a method orthogonal to parameter sharing. The network will benefit from data augmentation only if augmentation introduces transformation that is not explained by parameter sharing. So data augmentation generally requires more parameters than parameter sharing methods [8]. Another popular method is adding symmetry regularization to learn symmetry-adapted representations which are permutation invariant [9]. New network architectures have been proposed [10] which introduce new operations and combine the permuted features to make a model equivariant to rotation. Furthermore, [11] designed objective functions defined on sets which are invariant to permutations.

Several previous works apply group action or harmonic analysis to investigate equivariant neural networks [12, 13]. Steerable filters are studied in [14] to achieve equivariance in deep networks [15]. Necessary and sufficient architecture for a neural network to be equivariant to group actions has been explored in [8]. It is shown that when the transformation group acts discretely on the input and output of ϕ_w, a neural network layer ϕ_w is equivariant with respect to the G actions if and only if G explains the symmetries of the network parameter w [8]. Following that, it is proved in [16] that a neural network $\mathcal{N}$ is equivariant to the action of a compact group G if and only if $\mathcal{N}$ is a convolutional neural network that applies the generalized convolution on group G. Although the proposed framework is general and applicable to diverse transformations of the input/output of the neural layers, the treatment of nonlinearities is less than satisfactory. Specifically, the conditions on which nonlinear activation functions will keep the equivariance of convolutional layer have not been identified. Also, the proof for the necessary part of the equivalence between equivariance and convolution uses representation theory and noncommutative harmonic analysis, and is highly technical. It is our hope to provide a more elementary proof on the equivalence based on intuition obtained from familiar signal processing concepts. We also consider the effect on the equivalence of the nonlinearities and identify conditions on the nonlinearity that preserve the equivalence.

Specifically, the contributions of this paper can be summarized as follows:

i) Using intuition from linear time-invariant systems, we prove that equivariance in a linear neural network layer is equivalent to the layer being convolutional. The convolution kernel is the unit impulse response of the layer.

ii) We present a careful analysis on the non-linear activation functions. Specifically, we show that a sufficient condition on the activation functions for the equivalence between equivariance and convolution is that the nonlinear activation function contain at least one segment that is one-to-one.

The rest of the paper is organized as follows: Section 2 introduces the system model and the basic definitions. Section 3 provides a proof of the equivalence of group equivariance and convolution operation for the shift-equivariant case and the more general case where the group action is transitive. Section 4 completes the analysis by identifying a sufficient condition for the nonlinear activation to preserve equivariance. Section 5 discusses the case of multi-layer networks and possible extensions of our analysis to nontransitive group equivariance and multiple equivariances. Finally, Section 6 concludes the paper.

2. SYSTEM MODEL AND DEFINITIONS

Let $\mathcal{N}$ be a feed-forward neural network with $L+1$ layers. We denote the output features of the layers as $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_L$, where $\mathbf{x}_0$ is the input of the entire neural network. For image processing, these are the images signals in the layers. The sets indexing the neurons of each layer are denoted as $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_L$, respectively. In the l-th layer, the output $\mathbf{x}_l$ is obtained by a map $\tilde{f}_l : \mathbf{x}_{l-1} \to \mathbf{x}_l$

$$\mathbf{x}_l = \tilde{f}_l(\mathbf{x}_{l-1}) = \sigma(\mathbf{w}_l \mathbf{x}_{l-1} + \mathbf{b}_l)$$

where $\mathbf{w}_l$ is a *weight matrix/tensor*, $\mathbf{b}_l$ is the *bias* term and $\sigma(z)$ is a nonlinear element-wise operator which is also called the *activation function*. This is a general definition for feed-forward neural network. Convolutional neural network is a special case where $\mathbf{w}_l$ has a special toeplitz structure.

2.1. Group Action and Equivariance

Definition 1. Let G be a group, $g \in G$, $\mathcal{X}$ and $\mathcal{Y}$ be two index sets with corresponding G-actions

$$T_g : \mathcal{X} \to \mathcal{X}, \quad T'_g : \mathcal{Y} \to \mathcal{Y} \tag{1}$$

that satisfy $T_{g_2 g_1} = T_{g_2} \circ T_{g_1}$ and $T'_{g_2 g_1} = T'_{g_2} \circ T'_{g_1}$. Let $L(\mathcal{U}, F)$ denote the vector space over a field F with vectors indexed by $\mathcal{U}$. A vector $\mathbf{x} \in L(\mathcal{U}, F)$ can be viewed as a mapping from $\mathcal{U}$ to F, such that $\mathbf{x}(u) \in F$, for $u \in \mathcal{U}$. For brevity, we may write $L(\mathcal{U})$ instead when the results hold for any specific F. We use $\circ$ to denote mapping composition, such that $\mathbf{T} \circ f(u) = [\mathbf{T} f](u)$. Let $\mathbf{x} \in L(\mathcal{X})$ and $\mathbf{y} \in L(\mathcal{Y})$. The transformations T_g and T'_g induces actions $\mathbf{T}_g$ and $\mathbf{T}'_g$ on $L(\mathcal{X})$ and $L(\mathcal{Y})$ through

$$\mathbf{T}_g \circ \mathbf{x}(u_x) = \mathbf{x}(T_g u_x), \quad \mathbf{T}'_g \circ \mathbf{y}(u_y) = \mathbf{y}(T'_g u_y). \tag{2}$$

We say that a map $f : L(\mathcal{X}) \to L(\mathcal{Y})$ is *equivariant* with the action of G (or *G-equivariant*) if

$$\mathbf{T}'_g \circ f \circ \mathbf{x} = f \circ \mathbf{T}_g \circ \mathbf{x}, \quad \forall g \in G. \tag{3}$$

Definition 2. Let $\mathcal{N}$ be a feed-forward neural network. Let the feature of l-th layer be $\mathbf{x}_l = \sigma \circ f_l \circ \mathbf{x}_{l-1}$, $l = 0, 1, \ldots, L$. And let G be a group that acts on each index space $\mathcal{X}_0, \ldots, \mathcal{X}_L$. Let $\mathbf{T}^0, \ldots, \mathbf{T}^L$ be the corresponding actions on the features $\mathbf{x}_0, \ldots, \mathbf{x}_L$ lying in $L(\mathcal{X}_0), L(\mathcal{X}_1), \ldots, L(\mathcal{X}_L)$, respectively. We say that $\mathcal{N}$ is a *G-equivariant feed-forward neural network* if, when the inputs are transformed $\mathbf{x}_0 \to \mathbf{T}^0_g \mathbf{x}_0$ (for any $g \in G$), the activation of other layers transform correspondingly as $\mathbf{x}_l \to \mathbf{T}^l_g \mathbf{x}_l$, $l = 0, 1, \ldots, L$.

2.2. Generalized Group Convolution

Generalized group convolution was considered in [16] to build the correspondence between group equivariance and the convolution architecture. We adopt the same set of definitions as follows.

Definition 3. Let G be a compact group and f and g two functions $G \to \mathbb{C}$, then the *generalized convolution* of f and g is defined as:

$$(f * g)(u) = \int_G f(uv^{-1}) g(v) d\mu(v), \tag{4}$$

where uv^{-1} is a group action. And $\mu(v)$ is the Haar measure. The discrete counterpart of (4) for countable groups is

$$(f * g)(u) = \sum_{v \in G} f(uv^{-1}) g(v). \tag{5}$$

In neural network, a major challenge is that the index sets $\mathcal{X}_0, \ldots, \mathcal{X}_L$ are usually not isomorphic to G. Then the analysis in Theorem 1 (see Section 3.1) for linear time-invariant (LTI) system can not apply directly. However, we can adapt the proof and obtain a similar result as long as G acts on the index sets transitively. We say that G acts *transitively* on $\mathcal{X}$ if for any $u_{\mathbf{x}}, u'_{\mathbf{x}} \in \mathcal{X}$, there exists a $g \in G$ such that $u'_{\mathbf{x}} = T_g(u_{\mathbf{x}})$.

If G acts on $\mathcal{X}$ transitively, $\mathcal{X}$ is a *homogeneous space* of G. Take any $x_0 \in \mathcal{X}$, the set of group elements that map x_0 to itself form a subgroup H of G. And the set of group elements that map x_0 to x' is a *left coset* $gH := \{gh | h \in H\}$.We say G/H is a left quotient space. There is a one-to-one mapping between $\mathcal{X}$ and G/H. We represent it through an isomorphism $\psi : \mathcal{X} \longrightarrow G/H$, denoting as $\mathcal{X} \simeq G/H$. The projection of a function from G to its homogeneous space $\mathcal{X}$ is defined as mapping the average function value in the left coset gH to the corresponding $x = \psi^{-1}(gH)$. Lifting from $\mathcal{X} \simeq G/H$ to G is get by setting all the function value of elements in left coset equal to the function value of the coset representation in $\mathcal{X}$. Formally,

Definition 4. Given $f : G \to \mathbb{C}$, we define its *projection* to $\mathcal{X} = G/H$ as

$$f \downarrow_{\mathcal{X}} : \mathcal{X} \to \mathbb{C}, \quad f \downarrow_{\mathcal{X}} (x) = \frac{1}{|H|} \sum_{g \in \bar{x}H} f(g), \tag{6}$$

where $\bar{x}$ is a representative element of $\psi(x) = \bar{x}H$. Conversely, given $f : \mathcal{X} \to \mathbb{C}$, we define the *lifting* of f to G as

$$f \uparrow^G : G \to \mathbb{C}, \quad f \uparrow^G (g) = f(x) \tag{7}$$

where $x = \psi^{-1}(gH)$.

Projection and lifting to/from right quotient spaces and double quotient spaces are defined analogously. With the definition of lifting from quotient space to G we then define convolution on quotient space.

Definition 5. Let G be a finite or countable group, $\mathcal{X}$ and $\mathcal{Y}$ be (left or right) quotient space of G, $f : \mathcal{X} \to \mathbb{C}$ and $g : \mathcal{Y} \to \mathbb{C}$ we then define the *convolution* of f and g as

$$(f * g)(u) = \sum_{v \in G} f \uparrow^G (uv^{-1}) g \uparrow^G (v) \tag{8}$$

$$= \sum_{v \in G} f \uparrow^G (v) g \uparrow^G (v^{-1}u), \quad u \in G. \tag{9}$$

With these definitions, we are ready to state and provide an elementary proof the main theorem of [16] when the index sets are quotient space of G.

3. EQUIVARIANCE AND GENERALIZED CONVOLUTION

3.1. Translation Equivariance and Traditional Convolution

We will first provide a proof of the equivalence of translation equivariance and traditional convolution to illustrate the main proof ideas we use in our paper. The main idea of proof is the same as showing that convolution is the necessary and sufficient operation for a LTI system in signal processing [17].

Definition 6. Given a linear mapping $f : L(\mathcal{X}_x) \to L(\mathcal{X}_y)$, where $\mathcal{X}_x = \mathcal{X}_y = \mathbb{Z}^2$. Let $\mathbf{T}_v$ denote the translation operator such that

$$\mathbf{T}_v \mathbf{x}(u) := \mathbf{x}(u+v). \tag{10}$$

We say the mapping $f : L(\mathcal{X}_x) \to L(\mathcal{X}_y)$ is *translation equivariant* if

$$\mathbf{T}_v \circ f \circ \mathbf{x} = f \circ \mathbf{T}_v \circ \mathbf{x}, \quad \forall v \in \mathbb{Z}^2. \tag{11}$$

Theorem 1. A linear mapping $f : L(\mathcal{X}_x) \to L(\mathcal{X}_y)$ is translation equivariant if and only if it implements a convolution

$$f \circ \mathbf{x}(u) = (\mathbf{x} * \mathbf{h})(u) = \sum_{v \in \mathbb{Z}^2} \mathbf{x}(v)\mathbf{h}(u-v) \tag{12}$$

where $\mathbf{h}$ is the convolution kernel.

Proof. The sufficient part is straight forward which can be shown as follows. For any $u_0 \in \mathbb{Z}^2$,

$$\begin{aligned}
\mathbf{T}_{u_0} \circ f \circ \mathbf{x}(u) &= f \circ \mathbf{x}(u+u_0) \\
&= \sum_{v \in \mathbb{Z}^2} \mathbf{x}(v)\mathbf{h}(u+u_0-v) \\
&= \sum_{v' \in \mathbb{Z}^2} \mathbf{x}(v'+u_0)\mathbf{h}(u-v') = f \circ \mathbf{T}_{u_0} \circ \mathbf{x}(u)
\end{aligned}$$

To prove the necessary part, we first define the *unit-impulse* input:

$$\delta(u) = \begin{cases} 1 & \text{if } u = (0,0), \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

then the kernel $\mathbf{h}$ is chosen as *impulse response* $\mathbf{h}(u) = f \circ \delta(u)$, and $\mathbf{h}(u-v) = \mathbf{T}_{-v} \circ f \circ \delta(u) = f \circ \mathbf{T}_{-v} \circ \delta(u)$ due to the equivariance. And the input $\mathbf{x}(u)$ can be represented as a convolution as follows:

$$\mathbf{x}(u) = \sum_{v \in \mathbb{Z}^2} \mathbf{T}_{-v} \circ \delta(u)\mathbf{x}(v) = \sum_{v \in \mathbb{Z}^2} \mathbf{x}(v)\delta(u-v). \tag{14}$$

So the operator $\mathbf{x} = \sum_{v \in \mathbb{Z}^2} \mathbf{x}(v)\mathbf{T}_{-v} \circ \delta$.

$$\begin{aligned}
\mathbf{y}(u) = f \circ \mathbf{x}(u) &= (f \circ \sum_{v \in \mathbb{Z}^2} \mathbf{x}(v)\mathbf{T}_{-v} \circ \delta)(u) \qquad (15) \\
&= \sum_{v \in \mathbb{Z}^2} \mathbf{x}(v) f \circ \mathbf{T}_{-v} \circ \delta(u) \\
&= \sum_{v \in \mathbb{Z}^2} \mathbf{x}(v)\mathbf{T}_{-v} \circ f \circ \delta(u) \\
&= \sum_{v \in \mathbb{Z}^2} \mathbf{x}(v)\mathbf{h}(u-v) \\
&= (\mathbf{x} * \mathbf{h})(u)
\end{aligned}$$

□

To generalize Theorem 1 to any other group action equivariance when the index sets are isomorphic to G, we just need to change the translation operation to the group operation of G with small adjustment.

3.2. Equivalence in Transitive Case

In this subsection, we will analyze the necessary and sufficient condition for a linear mapping to be G-equivariant when the input index set and output index set are isomorphic to quotient space of G . As we usually have discrete index set in real application, all of the analysis in this section are based on discrete index sets and when G is a countable group for simplicity. Generalization to continuous case is possible. Also we identify the index set with the appropriate quotient group of G for brevity.

Theorem 2. Let G be a compact group and $\mathcal{X} = G/H_{\mathbf{x}}$ and $\mathcal{Y} = G/H_{\mathbf{y}}$ be two sets with corresponding G-actions

$$T_g : \mathcal{X} \to \mathcal{X}, \quad T'_g : \mathcal{Y} \to \mathcal{Y}. \tag{16}$$

A linear mapping $f : L(\mathcal{X}) \to L(\mathcal{Y})$ is G-equivariant if and only if it is a generalized convolution on quotient space defined in Definition 5. In particular,

$$f \circ \mathbf{x}(u) = (\mathbf{x} * \mathbf{h})(u) = \sum_{v \in G} \mathbf{x}\uparrow^G(v)\mathbf{h}\uparrow^G(v^{-1}u), \quad u \in \mathcal{Y}.$$

Proof. The sufficient part is straight forward and similar to the proof of Theorem 1. If f is a convolution defined as Definition 5, for any $u \in \mathcal{Y}$,

$$\begin{aligned}
\mathbf{T}'_g \circ f \circ \mathbf{x}(u) &= (f \circ \mathbf{x})(T'_g u) \\
&= \sum_{v \in G} \mathbf{x}\uparrow^G(v)\mathbf{h}\uparrow^G(v^{-1}(T'_g u)) \\
&= \sum_{v \in G} \mathbf{x}\uparrow^G(v)\mathbf{h}\uparrow^G(v^{-1}gu) \\
&= \sum_{v' \in G} \mathbf{x}\uparrow^G(T_g v')\mathbf{h}\uparrow^G(v'^{-1}u) \\
&= f \circ \mathbf{T}_g \circ \mathbf{x}(u).
\end{aligned}$$

Then we will provide proof for necessary part. Let $\delta(u)$ denote the unit-impulse function. Define a kernel $\mathbf{h}$ as the unit-impulse response $\mathbf{h}(u) = f \circ \delta(u)$. We have

$$\mathbf{h}(vu) = \mathbf{T}_v \circ \mathbf{h}(u) = \mathbf{T}_v \circ f \circ \delta(u) = f \circ \mathbf{T}_v \circ \delta(u) \tag{17}$$

thanks to the equivariance. The input $\mathbf{x}\uparrow^G(u)$ can be represented as a convolution as follows:

$$\begin{aligned}
\mathbf{x}\uparrow^G(u) &= \sum_{v \in G} \mathbf{x}\uparrow^G(v)\delta\uparrow^G(v^{-1}u) \\
&= \sum_{v \in G} \mathbf{x}\uparrow^G(v)\mathbf{T}_{v^{-1}} \circ \delta\uparrow^G(u).
\end{aligned}$$

Then we have $\mathbf{x}\uparrow^G = \sum_{v \in G} \mathbf{x}\uparrow^G(v)\mathbf{T}_{v^{-1}} \circ \delta\uparrow^G$.

Before proving the main theorem, we first show several useful property of lifting and function composition. First, for $u \in \mathcal{X}_{\mathbf{y}}$

$$(f\uparrow^G \circ \mathbf{x}\uparrow^G)(u) = (f \circ (\mathbf{x}\uparrow^G)\downarrow_{\mathcal{X}})(u) = (f \circ \mathbf{x})(u).$$

Second, since $\mathbf{h}\uparrow^G(u) = f\uparrow^G \circ \delta\uparrow^G(u)$, we have

$$\begin{aligned}
\mathbf{h}\uparrow^G(v^{-1}u) &= \mathbf{T}_{v^{-1}} \circ f\uparrow^G \circ \delta\uparrow^G(u) \\
&= f\uparrow^G \circ \mathbf{T}_{v^{-1}} \circ \delta\uparrow^G(u).
\end{aligned}$$

Next, we will prove the necessary part using similar ideas as used in Theorem 1.

$$\begin{aligned}(f \circ \mathbf{x})(u) &= (f \uparrow^G \circ \mathbf{x} \uparrow^G)(u) \\ &= f \uparrow^G \circ \sum_{v \in G} \mathbf{x} \uparrow^G (v) \mathbf{T}_{v^{-1}} \circ \delta \uparrow^G (u) \\ &= \sum_{v \in G} \mathbf{x} \uparrow^G (v) f \uparrow^G \circ \mathbf{T}_{v^{-1}} \circ \delta \uparrow^G (u) \\ &= \sum_{v \in G} \mathbf{x} \uparrow^G (v) \mathbf{T}'_{v^{-1}} \circ f \uparrow^G \circ \delta \uparrow^G (u) \\ &= \sum_{v \in G} \mathbf{x} \uparrow^G (v) \mathbf{h} \uparrow^G (v^{-1}u) = (\mathbf{x} * \mathbf{h})(u) \qquad \square\end{aligned}$$

Theorem 2 recovers the main theorem of [16], without resorting to group representation theory and harmonic analysis results. Our proof shows that the convolutional kernel $\mathbf{h}$ is the impulse response of the convolutional layer.

4. NONLINEAR ACTIVATION FUNCTION

We have obtained the necessary and sufficient condition for a linear mapping to be G-equivariant in last section. To extend it to a neural network layer, we need to include nonlinear activation function and explore the G-equivariant transitive property when we have a nonlinear function after a linear mapping. The following lemma will identify conditions under which G-equivariance with nonlinear activation is equivalent to G-equivariance without nonlinear activation.

Lemma 1. Consider a nonlinear activation function (pointwise operator) $\sigma : \mathcal{C} \to \mathcal{C}$ and a linear mapping $f : L(\mathcal{X}) \to L(\mathcal{Y})$. The set $\mathcal{C}$ can be $\mathbb{R}$ or $\mathbb{C}$, for example. If there exists a ball $S := B_\epsilon(\mathbf{y}_0) = \{||\mathbf{y} - \mathbf{y}_0|| \leq \epsilon\}, \epsilon > 0$ in $\mathcal{C}$ such that σ forms an injection from S to $\mathcal{C}$, then $\sigma \circ f$ is G-equivariant if and only if f is G-equivariant.

Proof. The sufficient part is trivial. As σ is a pointwise operator, the equivariance of f leads to the equivariance of $\sigma \circ f$.

For the reverse direction, take $S_y = \{\mathbf{y}|\mathbf{y}(u_y) \in S, \forall u_y \in \mathcal{Y}\}$, $\sigma(S_y) = \{\sigma \circ \mathbf{y}|\mathbf{y} \in S_y\}$, $f^{-1}(S_y) = \{\mathbf{x}|f \circ \mathbf{x} \in S_y\}$. Then σ is a bijection (one to one mapping) from S_y to $\sigma(S_y)$ and $f^{-1}(S_y)$ forms a set that has non-zero volume in $L(\mathcal{X})$.

Denote $\mathbf{T}'_g$ as the transformation operator on $\sigma \circ f$ which is also a transformation operator on f since σ is a pointwise operator. And we denote $\mathbf{T}_g$ as transformation operator on $\mathbf{x}$. We will first show that $\sigma \circ f$ is G-equivariant lead to f to be G-equivariant on $f^{-1}(S_y)$. Then we will extend it to that f is G-equivariant on the whole domain.

As the transformations are defined on the whole domain, we need to show that S_y and $\sigma(S_y)$ are closed on the transformation of $\mathbf{T}'_g,\ \forall g \in G$ and $f^{-1}(S_y)$ is closed on the transformation of $\mathbf{T}_g,\ \forall g \in G$. The first part is true by the definition of S_y so we just need show the second part.

Take any $\mathbf{x} \in f^{-1}(S_y)$, then $f \circ \mathbf{x} \in S_y$ and $\sigma \circ f \circ \mathbf{x} \in \sigma(S_y)$. By the definition of S_y , $\mathbf{T}'_g \circ f \circ \mathbf{x} \in S_y, \forall g \in G$. So $\sigma \circ \mathbf{T}'_g \circ f \circ \mathbf{x} = \mathbf{T}'_g \circ \sigma \circ f \circ \mathbf{x} \in \sigma(S_y), \forall g \in G$. As $\sigma \circ f$ is G-equivariant, $\mathbf{T}'_g \circ \sigma \circ f \circ \mathbf{x} = \sigma \circ f \circ \mathbf{T}_g \circ \mathbf{x}$, so $f \circ \mathbf{T}_g \circ \mathbf{x} \in S_y,\ \forall g \in G$. And $\mathbf{T}_g \circ \mathbf{x} \in S_x,\ \forall g \in G$. That is, take any $\mathbf{x} \in f^{-1}(S_y)$, we have $\mathbf{T}_g \circ \mathbf{x} \in S_x,\ \forall g \in G$.

Since $\sigma \circ f$ is G-equivariant, then take any $\mathbf{x} \in f^{-1}(S_y)$

$$\mathbf{T}'_g \circ \sigma \circ f \circ \mathbf{x} = \sigma \circ f \circ \mathbf{T}_g \circ \mathbf{x} = \sigma \circ \mathbf{T}'_g \circ f \circ \mathbf{x}.$$

So $\mathbf{T}'_g \circ f \circ \mathbf{x} = f \circ \mathbf{T}_g \circ \mathbf{x}\ \forall g \in G, \forall \mathbf{x} \in f^{-1}(S_y)$ as σ is a bijection from S_y to $\sigma(S_y)$.

We can then extend the result to the whole domain. For any $\mathbf{x} \in L(\mathcal{X})$ there exists a constant r satisfying $|r| > 0$ such that $\mathbf{x} = r\mathbf{x}'$ and $\mathbf{x}' \in f^{-1}(S_y)$ as $f^{-1}(S_y)$ has non-zero volume. Because f is a linear mapping, we have $\mathbf{T}_g \circ f \circ \mathbf{x} = r\mathbf{T}_g \circ f \circ \mathbf{x}'$ and $f \circ \mathbf{T}_g \circ \mathbf{x} = rf \circ \mathbf{T}_g \circ \mathbf{x}'$. It follows from $\mathbf{T}_g \circ f \circ \mathbf{x}' = f \circ \mathbf{T}_g \circ \mathbf{x}'$,that $\mathbf{T}_g \circ f \circ \mathbf{x} = f \circ \mathbf{T}_g \circ \mathbf{x}, \forall \mathbf{x} \in L(\mathcal{X})$. Thus f is necessarily G-equivariant in the whole space when $\sigma \circ f$ is G-equivariant. $\square$

Remark 1. With Lemma1, we can see that almost all popular activation functions satisfy the condition in Lemma1; one exception being the step-function activation. As a result, these nonlinear activation functions can inherit the G-equivariance from linear mapping as well as passing the G-equivariance to the linear mapping which they operate on. The unit-step function does not satisfy the condition because there is no neighborhood of non-zero volume in the domain that is injection mapped to $\mathbb{R}$. Rectified Linear Unit (ReLU) $x^+ := \max(x, 0)$ maps $[0, 1]$ to $[0, 1]$ one-to-one. Sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ also injectively maps $[0, 1]$ to $\mathbb{R}$.

5. GROUP EQUIVARIANT NEURAL NETWORK

It is difficult to give a necessary and sufficient condition for a neural network to be G-equivariant as a single system. It is obvious that if each layer of a deep neural network is equivariant, then the whole network is equivariant. However, the converse is not true: it is possible for the whole network to be equivariant, whereas the layers are not equivariant. We omit the counter example here but only remark that it is possible to permute the neurons in the hidden layers, together with all the connection coefficients without altering the end-to-end mapping of the whole network.

Practical convolutional neural networks are G-equivariant layer-wise, if each layer contains only one feature map. However, most convolutional neural networks contain multiple feature maps in each hidden layer, in which case our results (and existing results in the literature) do not apply directly because the group actions are not transitive: the shift operation does not map an index in one feature map (i.e., "channel") to an index of a different feature map. Rotation action, which acts only transitively on the sub-domain of concentric circles is also non-transitive. It is possible to extend our result to these non-transitive cases (see our coming work[18]).

6. CONCLUSIONS

Our work provides an elementary analysis of group equivariance in neural network. We showed that a necessary and sufficient condition for neural network to be G-equivariant when G acts transitively on the index sets of neural network function is that the layer implements a group convolution. The proof method is elementary, relying on familiar concepts of linear time-invariant systems. Our analysis is complete with the analysis of the nonlinear activation function. We identified a sufficient condition for the nonlinear function to preserve equivariance. Our analysis method is intuitive and can be extended to deal with cases where the group actions are non-transitive, and where there are multiple equivariances.

7. REFERENCES

[1] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990, pp. 396–404.

[2] Robert Gens and Pedro M Domingos, "Deep symmetry networks," in *Advances in neural information processing systems*, 2014, pp. 2537–2545.

[3] Mikael Henaff, Joan Bruna, and Yann LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[4] Martin Simonovsky and Nikos Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proc. CVPR*, 2017.

[5] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling, "Spherical CNNs," *arXiv preprint arXiv:1801.10130*, 2018.

[6] Benjamin Chidester, Minh N Do, and Jian Ma, "Rotation equivariance and invariance in convolutional neural networks," *arXiv preprint arXiv:1805.12301*, 2018.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos, "Equivariance through parameter-sharing," *arXiv preprint arXiv:1702.08389*, 2017.

[9] Fabio Anselmi, Georgios Evangelopoulos, Lorenzo Rosasco, and Tomaso Poggio, "Symmetry regularization," Tech. Rep., Center for Brains, Minds and Machines (CBMM), 2017.

[10] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu, "Exploiting cyclic symmetry in convolutional neural networks," *arXiv preprint arXiv:1602.02660*, 2016.

[11] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola, "Deep sets," in *Advances in Neural Information Processing Systems*, 2017, pp. 3391–3401.

[12] Taco Cohen and Max Welling, "Group equivariant convolutional networks," in *International conference on machine learning*, 2016, pp. 2990–2999.

[13] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia, "Rotation equivariant vector field networks.," in *ICCV*, 2017, pp. 5058–5067.

[14] Taco S Cohen and Max Welling, "Steerable CNNs," *arXiv preprint arXiv:1612.08498*, 2016.

[15] William T. Freeman and Edward H Adelson, "The design and use of steerable filters," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 9, pp. 891–906, 1991.

[16] Risi Kondor and Shubhendu Trivedi, "On the generalization of equivariance and convolution in neural networks to the action of compact groups," in *ICML*, 2018.

[17] Alan V Oppenheim, Alan S Willsky, and S Hamid Nawab, *Signals & Systems*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[18] Pan Zhong and Zhengdao Wang, "Characterization and design of generalized convolutional neural network," in *Conference in Information Sciences and Systems*, 2019.