Generalizing Convolutional Neural Networks

Ezra Erives

6.s966, Spring 2023

Abstract

This project surveys progress towards generalizing conventional convolutional neural networks (CNN's). We will discuss how the conventional CNN can be realized as a cross-correlation of signals on Euclidean space with respect to the translation group, and how this perspective paves the way toward generalizing toward arbitrary groups (both in theory in practice), as developed in [1], [4], [7]. In light of this, we will also discuss progress in generalizing CNN's to more general homogeneous spaces on which groups can act, as developed in [5] [2] [6] [3].

Contents

1 Introduction

This project will survey progress towards generalizing conventional convolutional neural networks (CNN's). We will discuss how the conventional CNN can be realized as a cross-correlation of signals on Euclidean space with respect to the translation group, and how this perspective paves the way toward generalizing toward arbitrary groups (both in theory in practice), as developed in [1], [4], [7]. In light of this, we will also discuss progress in generalizing CNN's to more general homogeneous spaces on which groups can act, as developed in [5] [2] [6] [3]. As a proof of concept, we will discuss and implement methodology developed by [5] for the task of classifying signals of the sphere.

2 Background

In this section we include foundational concepts, including groups, representations, homogeneous spaces, and cross-correlations, among others, which the reader should take care to understand before proceeding to the next section.

2.1 Groups

Definition 1 (Group)**.** A *group* is a set G with an associative binary operation $\star$ which satisfies the following constraints:

(1) there exists an element $e \in G$, referred to as the *identity*, which satisfies $e \star g = g \star e = g$ for all $g = \in G$, and

(2) for every $g \in G$, there exists an element $g^{-1} \in G$ so that $g \star g^{-1} = g^{-1} \star g = e$.

Examples of groups include $\mathbb{T}(2)$, the group of translations of $\mathbb{Z}^2$, the group p4 of rototranslations of $\mathbb{Z}^2$ and the group SO(3) of 3D rotations.

When there is no danger of ambiguity, we drop the $\star$, and refer to the product of two group elements $g_1, g_2 \in G$ as $g_1 g_2 \in G$.

Definition 2 (Subgroup)**.** A subset H of a group G with binary operation $\star$ is said to be a *subgroup* of G if H is itself a group under $\star$.

Definition 3 (Coset)**.** Given a group G, a subgroup $H \subseteq G$, and an element $g \in G$, we denote by $gH = \{gh \mid h \in H\}$ the (left) g-coset of H.

Theorem 1 (Cosets Partition Group)**.** *Let G, H be as in the previous definition. Write $g_2 \sim g_1$ if $g_2 \in g_1 H$. Then*

(1) *the relation $\sim$ is an equivalence relation,*

(2) *the relation $\sim$ naturally partitions G into disjoint subsets, and*

(3) *these subsets are all of size $|H|$.*

Proof. Omitted for brevity. □

2.2 Representations

Definition 4 (General Linear Group)**.** Let V be a vector space. Then the set of invertible maps $T : V \to V$ is a group under composition. This group is referred to as the *general linear group* of V, and is denoted by $\mathrm{GL}(V)$.

Definition 5 (Representation)**.** Let G be a group and let V be a vector space. Then a map $\pi : G \to \mathrm{GL}(V)$ is said to be a *representation* of G if for all $g_1, g_2 \in G$, $\pi(g_1 g_2) = \pi(g_1)\pi(g_2)$.

Depending on the context, "representation" may refer to the map π, the vector space V on which the group G is acting on, or, less frequently, elements of V.

Figure 1: Equivariance.

Definition 6 (Irreducible Representation)**.** With the same notation as in the previous definition, the representation $\pi : G \to \mathrm{GL}(V)$ is said to be *reducible* if there exists a proper subspace $W \subset V$ which is invariant under action by G. A representation is said to be *irreducible* if it is not reducible.

It follows that all representations are direct sums of irreducible representations. As a result, the irreducible representations play a central role in the representation theory of a given group.

As will be clear later, it is often necessary to extend a representation π of a subgroup H of G to a representation ρ on the whole of G which is consistent with π. The canonical choice for such a representation ρ on G is known as the *induced representation*. To define the induced representation, we must first consider a representation $\rho : H \to \mathrm{GL}(V)$, describing action by the subgroup H of G on the vector space V. In extending ρ to the whole group G, notice that the cosets gH of H act as copies of H. Now let us choose representatives $g_1, \dots, g_n$ of the cosets of H, where $n = [G : H]$ is the number of cosets of H in G, and consider vector spaces $V_1, \dots, V_n \simeq V$. Then for any $g \in G$, and g_i, $gg_i = g_{i'}h$ for some i'. Letting $g_i v_i$ denote an element of V_i, we define the induced representation $\rho' = \mathrm{Ind}_H^G \rho : G \to \mathrm{GL}(\bigoplus_{i=1}^n V_i)$ by

$$\rho'(g) g_i v_i = g_{i'} \rho(h).$$

2.3 Equivariance

Definition 7 (Equivariance)**.** Let $\pi_1 : G \to \mathrm{GL}(V_1)$ and $\pi_2 : G \to \mathrm{GL}(V_2)$ be representations of a group G. A map $f : V_1 \to V_2$ is said to be *equivariant* if for all $g \in G$ and $v \in V_1$, $f(\pi_1(g)v) = \pi_2(g)f(v)$.

Definition 8 (Homogeneous Space)**.** A set S is said to be a *homogeneous space* with respect to a group G if there exist functions $\{L_g : S \to S\}_{g \in G}$ so that for all $g_1, g_2 \in G$, $L_{g_1} \circ L_{g_2} = L_{g_1 g_2}$ and so that for all $x, y \in S$, there exists a $g \in G$ for which $L_g(x) = y$. We say that the maps L_g give a *group action* of G on S.

Note that from any representation $\pi : G \to V$ of a group G, we find that V is a homogeneous space by $L_g(v) = \pi(g)v$. One imporant example of a homogeneous space is that of $\mathbb{Z}^2$, with respect to the group $\mathbb{T}(2)$ of translations. The group action L_t for translation $t \in \mathbb{T}(2)$ is given by $L_t(x) = x + t$.

Definition 9 (Equivariance on homogeneous Spaces)**.** Let S_1, S_2 be two homogeneous spaces of a group G. A map $f : S_1 \to S_2$ is said to be *equivariant* if for all $g \in G$ and $s \in S_1$, $f(L_g v) = L_g f(v)$.

2.4 Convolution and Cross-Correlation

Definition 10 (Signal)**.** A signal f on a set S is a map from S to $\mathbb{R}^K$, for some positive integer K. Oftentimes, S has additional structure (e.g. is a homogeneous space). The individual coordinates, or *channels*, of the signal $f : S \to \mathbb{R}^K$ are themselves functions $f_k : S \to \mathbb{R}$.

For example, we can think of a colored image as a signal $f : \mathbb{Z}^2 \to \mathbb{R}^3$: f sends elements of $\mathbb{Z}^2$ (pixels) to $(r, g, b) \in \mathbb{R}^3$ corresponding to the RGB values at that pixel.

Given two signals $f : \mathbb{Z}^2 \to \mathbb{R}$ and $\Psi : \mathbb{Z}^2 \to \mathbb{R}$, the *convolution* of f and g, denoted by $f * g$, is given by

$$[f * g](x) = \sum_{y \in \mathbb{Z}^2} f(y)g(x - y). \tag{1}$$

Related is the *cross-correlation* of f and g, denoted by $f \star g$, and given by

$$[f \star g](x) = \sum_{y \in \mathbb{Z}^2} f(y)g(y - x). \tag{2}$$

In this project, as well as in practice, we will work instead with the cross-correlation, as it is more amenable to generalization. While the definitions presented here are intended for the homogeneous space $\mathbb{Z}^2$, they may be easily modified for other homogeneous spaces.

2.5 Fourier Bases

On many homogeneous spaces S we will concern ourselves with, the space of nicely behaved $\mathbb{R}$-valued functions defined on this space is in fact a vector space. The homogeneous space $S = \mathbb{R}$ has the standard Fourier basis $\{\sin(2\pi nx), \cos(2\pi nx)\}_{n=1}^{\infty}$. The sphere S^2 has the so-called *spherical harmonics*. Finally, the Lie group SO(3) of rotations has the so-called *Wigner D-matrices*. The process of rewriting a function in these so-called Fourier bases is known as the (generalized) *Fourier transform*.

2.6 Commonly Used Groups

We'll briefly define two symmetry groups of $\mathbb{Z}^2$ which we will make frequent use of in examples in later sections.

Definition 11 (p4)**.** The group p4 consists of all translations of $\mathbb{Z}^2$, rotations of multiples of 90° about grid points.

Definition 12 (p4m)**.** The group p4m is the group p4 with mirror symmetry. These are the isometries of $\mathbb{Z}^2$.

3 Related Work

As we travel from the conventional CNN to more general variants, we will pay special attention to the both **group(s)** involved, as well as the **underlying space** on which our signal live.

3.1 Conventional CNN's

Conventional convolutional neural networks (CNN's) work by learning a correlation filter Ψ, which is cross-correlated with a given input signal, on a discrete Euclidean space like $\mathbb{Z}^2$, to produce an output signal also on $\mathbb{Z}^2$. As we will see, it is misleading to label the output signal as being on $\mathbb{Z}^2$, as it is actually a signal on the group of translations $\mathbb{T}(2)$ of $\mathbb{Z}^2$.

Assume for the sake of simplicity that we wish to learn a map from a signal $f : \mathbb{Z}^2 \to \mathbb{R}$ with a single channel to a new signal $f' : \mathbb{Z}^2 \to \mathbb{R}$, again with only a single channel. A conventional convolutional neural network will parameterize this map as a cross-correlation between f, and a learnable signal $\Psi : \mathbb{Z}^2 \to \mathbb{R}$, referred to as the *convolution kernel* (although really it's a correlation kernel). In practice, this Ψ is compactly supported, meaning that it is zero in all but some compact subset of $\mathbb{Z}^2$ (e.g., a 3×3 grid within $\mathbb{Z}^2$). As a function f and Ψ, we then have

$$f'(x) = (f \star \Psi)(x) = \sum_{y \in \mathbb{Z}^2} f(y)\Psi(y - x). \tag{3}$$

Denote by $S_\Psi \subseteq \mathbb{Z}^2$ the support of Ψ. Then we may rewrite (3) as

$$f'(x) = (f \star \Psi)(x) = \sum_{y - x \in S_\Psi} f(y)\Psi(y - x). \tag{4}$$

Thus, the value of f' at x is given by shifting the kernel Ψ by x, and dotting the portion of f which aligns with the shifted support $S_\Psi + x$ with Ψ. It is not hard to adapt (4) for f and f' with arbitrary numbers of channels. Suppose now that $f : \mathbb{Z}^2 \to \mathbb{R}_1^c$ has c_1 channels and $f' : \mathbb{Z}^2 \to \mathbb{R}_2^c$ has c_2 channels. How does our kernel Ψ change? Our initial definition of $\Psi : \mathbb{Z}^2 \to \mathbb{R}$ had only a single channel encoding the relationship between the sole channel of f and the sole channel of f'. When f and f' now have c_1 and c_2 channels respectively, our kernel Ψ must still be capable of encoding the relationship between each input channel and each output channel, so that our adapted

$\Psi : \mathbb{Z}^2 \to \mathbb{R}^{c_1 c_2}$ must have $c_1 c_2$ channels. Explicitly, the j-th channel f'_j of f' is now given by

$$f'_j(x) = (f \star \Psi^j)(x) = \sum_{i=1}^{c_1} \sum_{y-x \in S_{\Psi_i^j}} f_i(y)\Psi_i^j(y-x). \tag{5}$$

where the signal $\Psi^j : \mathbb{Z}^2 \to \mathbb{R}^{c_1}$ is comprised of the channels of Ψ which affect the j-th channel of f', and where $\Psi_i^j : \mathbb{Z}^2 \to \mathbb{R}$ is i-th channel of Ψ^j, encoding the relationship between f_i and f'_j. Now, we encounter a **big idea**. In (5), whereas $y \in \mathbb{Z}^2$ is a point, x only ever acts as a translation, or in other words, as an element of the translation group $\mathbb{T}(2)$ of $\mathbb{Z}^2$. Thus, the signal f' is actually a signal on $\mathbb{T}(2)$. In practice, we get away with pretending its a signal on $\mathbb{Z}^2$ because of the isomorphism $\mathbb{Z}^2 \sim \mathbb{T}(2)$.

Big Idea 1. Given signals $f : \mathbb{Z}^2 \to \mathbb{R}^{c_1}$ and $\Psi : \mathbb{Z}^2 \to \mathbb{R}^{c_1 c_2}$, the output signal $f' = f \star \Psi$ is a signal on the translation group $\mathbb{T}(2)$ of $\mathbb{Z}^2$. In practice, it is possible to pretend that f' is a signal on $\mathbb{Z}^2$ because of the isomorphism $\mathbb{Z}^2 \sim \mathbb{T}(2)$.

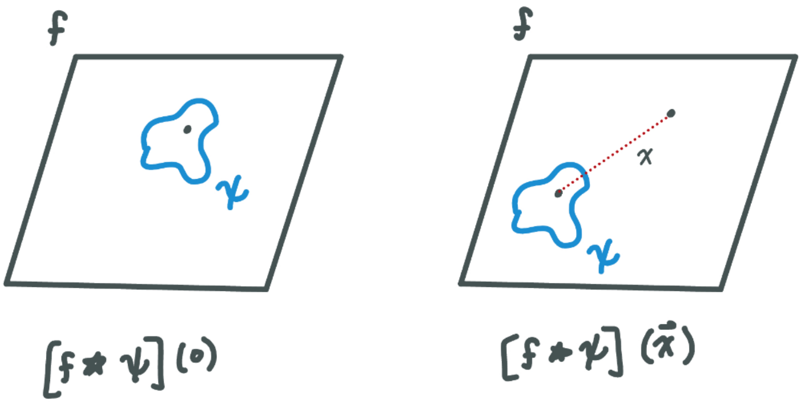

Figure 2: In illustration of the shift-and-dot nature of conventional cross-correlation.

The group $\mathbb{T}(2)$ is significant in another way: cross-correlations as defined in (5) are equivariant to $\mathbb{T}(2)$.

Theorem 2. The cross correlation defined by (5) is equivariant to the group $\mathbb{T}(2)$. That is, given some translation $t \in \mathbb{T}(2)$, the following equality holds:

$$[L_t\,(f \star \Psi^j)](x) = (L_t\, f \star \Psi^j)](x).$$

Proof. Observe that

$$[L_t\,(f \star \Psi^j)](x) = (f \star \Psi^j)(t^{-1}x) = \sum_{i=1}^{c_1} \sum_{y \in \mathbb{Z}^2} f_i(y)\Psi_i^j((x^{-1}t)y).$$

And permuting the order of summation with $\sum_{y\in\mathbb{Z}^2} f_i(y)\Psi_i^j((x^{-1}t)y) = \sum_{y\in\mathbb{Z}^2} f_i(t^{-1}y)\Psi_i^j(x^{-1}y)$, we obtain

$$\sum_{i=1}^{c_1}\sum_{y\in\mathbb{Z}^2} f_i(y)\Psi_i^j((x^{-1}t)y) = \sum_{i=1}^{c_1}\sum_{y\in\mathbb{Z}^2} f_i(t^{-1}y)\Psi_i^j(x^{-1}y) = (L_t f \star \Psi^j)(x). \tag{6}$$

□

The cross-correlation described in (5) is a simplified version of convolution in practice. Oftentimes, the cross-correlation is taken with *stride*, and consecutive cross-correlation layers are separated by *max pooling* layers, and non-linearities. Stride effectively amounts to restricting $f' : \mathbb{T}(2) \to \mathbb{R}^2$ to some subgroup of $\mathbb{T}(2)$. Letting S be some neighborhood of the origin in $\mathbb{T}(2)$, the max pooling operator P_S is given by

$$[P_S f'](g) = \max_{g'\in gS} f'(g'). \tag{7}$$

In practice S is taken to be a $n \times n$ grid centered at the origin, for some small value of n, such as two or three.

In the next section, we will explore what happens when $\mathbb{T}(2)$ is replaced with a more complex group.

3.2 Group Equivariant CNN's

In this section, we'll build off of Big Idea 1 to generalize (5) to more complex groups, following work done in [1], and will conclude by generalizing striding and max-pooling, all as described by [1].

3.2.1 Group Equivariant Cross-Correlation

Recall that given a signal $f : \mathbb{Z}^2 \to \mathbb{R}^{c_1}$ and a desired number of output channels c_2, a conventional CNN will learn a correlation kernel $\Psi : \mathbb{Z}^2 \to \mathbb{R}^{c_1 c_2}$, which together with f produces the signal $f' : \mathbb{T}(2) \to \mathbb{R}^{c_1}$ given by

$$f_j'(x) = (f \star \Psi^j)(x) = \sum_{i=1}^{c_1} \sum_{y-x\in S_{\Psi_i^j}} f_i(y)\Psi_i^j(y-x). \tag{8}$$

Let us now write $G = \mathbb{T}(2)$, and let g be the group element corresponding to addition of x, then we obtain $f' : G \to \mathbb{R}^{c_2}$ defined on G by

$$f_j'(g) = (f \star \Psi^j)(g) = \sum_{i=1}^{c_1}\sum_{y\in\mathbb{Z}^2} f_i(y)\Psi_i^j(g^{-1}y). \tag{9}$$

Here, G can be any subgroup of the group symmetries of $\mathbb{Z}^2$.

Big Idea 2. Let p4m be the group of symmetries (isometries) of $\mathbb{Z}^2$. If $f : \mathbb{Z}^2 \to \mathbb{R}^{c_1}$ is a signal on $\mathbb{Z}^2$, and $\Psi : \mathbb{Z}^2 \to \mathbb{R}^{c_1 c_2}$ is a correlation kernel, then for any subgroup G of $E(\mathbb{Z}^2)$, we may define the G cross-correlation as

$$f'_j(g) = (f \star \Psi^j)(g) = \sum_{i=1}^{c_1} \sum_{y \in \mathbb{Z}^2} f_i(y) \Psi^j_i(g^{-1}y).$$

The signal f' obtained is *a signal an G*. This idea applies as well when $\mathbb{Z}^2$ is swapped with $\mathbb{R}^2$.

To see that the map $f \to f'$ is equivariant to some choice of G, we follow along the lines of (6) to find that

$$\begin{aligned}
[L_h(f \star \Psi^j)](g) &= [f \star \Psi^j](h^{-1}g) \\
&= \sum_{i=1}^{c_1} \sum_{y \in \mathbb{Z}^2} f_i(y) \Psi^j_i((h^{-1}g)^{-1}y) \\
&= \sum_{i=1}^{c_1} \sum_{y \in \mathbb{Z}^2} f_i(y) \Psi^j_i(g^{-1}hy) \\
&= \sum_{i=1}^{c_1} \sum_{y \in \mathbb{Z}^2} f_i(h^{-1}y) \Psi^j_i((g^{-1}y) \\
&= [[L_h f] \star \Psi^j](g)
\end{aligned}$$

where $h \in G$ is arbitrary. Implementing G-convolutions in this way is exactly the approach by in [1], in which neural networks built from these G-convolutions are referred to as G-CNN's. We'll adopt this terminology as well. In addition to providing the G-equivariant group convolution which we've just described, this paper additionally discusses non-linearities, sub-sampling, and pooling layers, all of which are important for realizing G-CNN's as true generalizations of conventional CNN's.

3.2.2 Striding and Max Pooling for G-CNN's

As discussed for conventional CNN's, strides can be thought of a way of downsampling a signal to reduce its dimensionality. The $\mathbb{T}(2)$ cross-correlation with a stride of two on $\mathbb{Z}^2$ corresponds to first computing the cross-correlation on $\mathbb{T}(2)$, and then downsampling onto the subgroup H of translations with even displacements. Thus, we may view cross-correlation with stride as the composition of two operations: cross-correlation over $\mathbb{T}(2)$ and downsampling onto H. However, one issue is that downsampling in this manner is equivariant only to H, and not to the full group $\mathbb{T}(2)$.

Recall from (7) that the conventional max-pooling operator P is given by $[P_S f'](g) = \max_{g' \in gS} f'(g')$, where $S \subset \mathbb{T}(2)$ is some neighborhood of the origin. There is no reason that pooling in this manner would be equivariant to $\mathbb{T}(2)$. The authors of [1] propose to ensure equivariance by replacing S with a subgroup H of p4m. The corresponding max pooling operator P_H is then given by

$$[P_H f'](g) = \max_{g' \in gH} f'(g'). \tag{10}$$

where $g \in \mathrm{p4m}$, and $f' : \mathrm{p4m} \to \mathbb{R}^c$ is arbitrary. The sets gH are known as the *left cosets* of H with respect to p4m, and it follows from Theorem 1 that $P_H f'$ is well-defined as function on the quotient space $\mathrm{p4m}\,/H$ of cosets H.

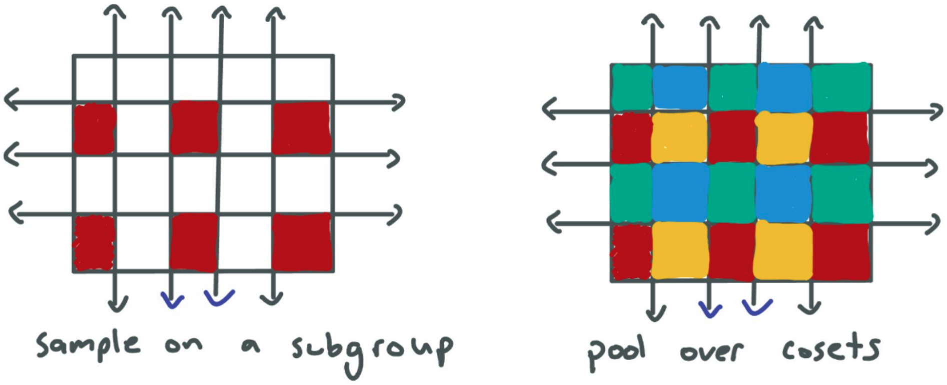

Figure 3: Group-theoretic generalizations of striding and max-pooling layers.

Big Idea 3. The equivalent of a strided cross-correlation for a G-CNN is to first apply a normal cross-correlation, and then subsample on a subgroup H of G. The equivalent of pooling for a G-CNN is given by pooling over the cosets of a subgroup H. Formally, for choice of H, the pooling operator P_H is given by

$$[P_H f'](g) = \max_{g' \in gH} f'(g').$$

3.2.3 Limitations of G-CNN's

The cost of computing a group cross-correlation as in (9) scales with the size of the group, and isn't well suited for continuous groups. In the next section, we'll explore how to overcome these challenges.

3.3 Steerable CNN's

In this section we will discuss an alternative to G-CNN's for constructing equivariant cross-correlations, surveying the perspective taken in [4] to explore so-called *steerable CNN's*. We will draw a distinction between the *base space* $\mathbb{Z}^2$, and the *fibers* $\mathbb{R}^k$ attached to each grid point, and use this distinction to construct cross-correlations equivariant to representations G which act not just on the base space, but the fibers as well.

3.3.1 Fibers and Base Spaces

With the same notations as before, recall that the G-CNN cross-correlation of a signal $f : \mathbb{Z}^2 \to \mathbb{R}^{c_1}$ with the kernel $\Psi : \mathbb{Z}^2 \to \mathbb{R}^{c_1 c_2}$ is given by

$$f'_j(g) = (f \star \Psi^j)(g) = \sum_{i=1}^{c_1} \sum_{y \in \mathbb{Z}^2} f_i(y) \Psi_i^j(g^{-1}y).$$

Up until now, the kernel Ψ has been arbitrary, and it is only by the taking this cross-correlation to produce a signal on the group G itself are we able to ensure equivariance. Explicitly, the equivariance is summarized by

$$[L_h(f \star \Psi^j)](g) = ([L_h f] \star \Psi^j)(g)$$

where L_h denotes action by some group element $h \in G$. Again, what this says is that if we apply group element h to f, and then take the correlation, we get the same result as first taking the correlation, and then applying the group action on resulting signal on G. Importantly, such L_h act only on $\mathbb{Z}^2$, seemingly ignoring the our signal $f : \mathbb{Z}^2 \to \mathbb{R}^{c_1}$ is in fact a stack of $c_1 \geq 1$ maps $f_j : \mathbb{Z}^2 \to \mathbb{R}$. Somehow, there is important distinction to be made between the $\mathbb{Z}^2$-dimension of our signal f, and the $\mathbb{R}^k$-dimension of f.

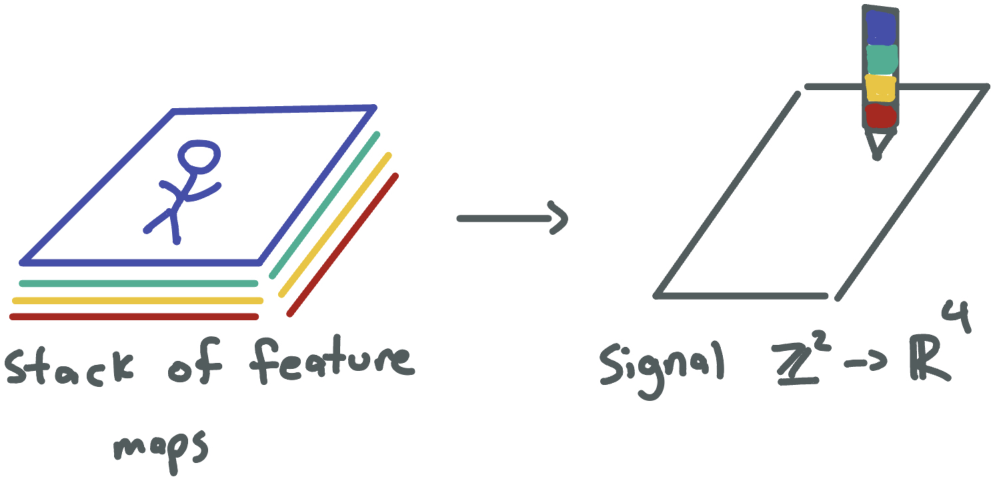

Figure 4: The objects we care about are *signals* rather than *stacks of feature maps*.

To properly motivate this distinction, let's first make explicit what might be traditionally called a multi-channel feature map. Such a mapping is given by a stack of c (the number of channels) feature maps $\{f_j : \mathbb{Z}^2 \to \mathbb{R}\}_{i=1}^c$. Of course, we can rewrite these c signals as a single, combined signal $f : \mathbb{Z}^2 \to \mathbb{R}^c$. This rewritten version is the perspective taken up to this point in our survey. We refer to the space $\mathbb{Z}^2$ here as the *base space*, and the vector space $\mathbb{R}^c$ as a *fiber*. We consider the fiber $F_x \simeq \mathbb{R}^c$ at each point to be distinct. Thus, such a signal f is defined on the base space $\mathbb{Z}^2$, and at each point $x \in \mathbb{Z}^2$ takes values $f(x) \in F_x \simeq \mathbb{R}^k$.

3.3.2 Fiber Equivariant Kernels

One issue with G-CNN's is that they consider only group actions on the base space $\mathbb{Z}^2$, leaving the fibers intact. We lose out on a significant opportunity to inject invariance by doing so, and must instead rely on an expensive group cross-correlation. We can simultaneously solve the problem of this expensive group cross-correlation and our neglect of the fibers by instead forcing Ψ to satisfy an equivariance property which involves a group action on the fibers. Considering again the standard translational cross-correlation given by

$$f'(x) = [f \star \Psi](x) = \sum_{y \in \mathbb{Z}^2} f(y)\Psi(y - x) \in \mathbb{R}^{c_2} \tag{11}$$

and letting π_0 and π_1 respectively be representations of some subgroup H of G acting on $\mathbb{Z}^2$ and on the fiber $F_x \simeq \mathbb{R}^{c_2}$, we seek Ψ so that $\pi_1 \circ \Psi = \Psi \circ \pi_0$, or in other words if

$$\pi_1[[f \star \Psi](x)] = [[\pi_0 f] \star \Psi](x) \tag{12}$$

for all $x \in \mathbb{T}(2)$, and signals $f : \mathbb{Z}^2 \to \mathbb{R}^{c_2}$. The reason for only considering equivariance with respect to a subgroup H, rather than the whole group G, is that for many choices of G, such as p4, it is impossible to achieve such this equivariance condition on the whole group. For example, Ψ must be compactly supported in practice, so that it cannot possibly satisfy the desired equivariance property with respect to translations, as we could simply translate out of the support of Ψ. In [4], H is taken to be the subgroup of G composed of rotations about the origin.

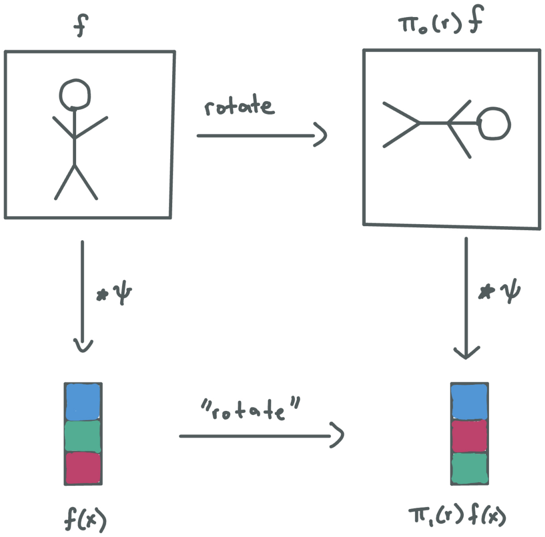

Figure 5: Realizing the fibers $F_x \sim \mathbb{R}^3$, $x \in \mathbb{R}^2$ as a representation of the subgroup H of rotations about the origin. In this case, the rotation acts as a permutation matrix, although this need not be the case in general.

Notice that for fixed π_0, π_1 of the form described above (with π_1 acting on the fibers), the set of Ψ which satisfy the desired equivariance property $\pi_1 \circ \Psi = \Psi \circ \pi_0$ is a linear subspace of the space of the possible signals $\Psi' : \mathbb{Z}^2 \to \mathbb{R}^c$. The approach taken by [4] is to precompute a basis for this subspace, and parameterize Ψ as a learnable linear combination of these basis elements.

3.3.3 The Induced Representation

In order to obtain equivariance with respect to the whole group $G = \mathrm{p4}$, we must somehow construct a representation for G, which extends the representation we have chosen for H. The canonical choice for such a representation is known as the *induced representation* $\rho_1 = \mathrm{Ind}_H^G \pi_1$ [8]. Every group element $g \in \mathrm{p4}$ can be written uniquely as a product $g = tr$, where t is a pure translation, and r, a pure translation. Then, letting ρ_0 denote the standard action of p4 on $\mathbb{Z}^2$, we have that.

$$\begin{aligned}[\rho_0(tr)f \star \Psi](x) &= [L_{tr}f \star \Psi](x) \\ &= \rho_1(tr)[L_{tr}f \star \Psi](x) \\ &= \pi_1(r)[f \star \Psi]((tr)^{-1}x)\end{aligned}$$

establishing equivariance. In other words, denoting by $F = \bigoplus_{x \in \mathbb{Z}^2} F_x$, the induced representation described how p4 acts on the space F. The mental picture is that the group element $g = tr$ first applies the map $\pi_1(r) : F_x \to F_x$, shuffling the contents of each fiber F_x. The contents of each F_x are then replaced by the contents of $F_{(tr)^{-1}(x)}$, or, in other words, the fibers are translated by $(tr)^{-1}$.

Big Idea 4. Steerable CNN's first establish fiber equivariance with respect to the stabilizer H (rotations about the origin) of p4. When the group action on the the channel-space is realized as a represenation of H, it can then be extended to the induced representation $\mathrm{Ind}_H^G \pi_1$, providing full G-equivariance.

Another reason it is clear to choose H to be the group of rotations is that the resulting signal $f \star \Psi$ is defined on the the quotient $G/H = \mathbb{T}(2) \simeq \mathbb{Z}^2$.

3.4 Spherical CNN's

Thus far, our discussion has been concerned with the flat (and discrete) space $\mathbb{Z}^2$. However, results from the past few sections apply equally well to $\mathbb{R}^2$ in theory, if not in practice. In this section, we will turn away from $\mathbb{Z}^2$ and $\mathbb{R}^2$ to consider more general homogeneous spaces, in doing so following along with [5]. Recall that on $\mathbb{Z}^2$, the G-cross correlation $f' : G \to \mathbb{R}^{c_2}$ of a signal $f : \mathbb{Z}^2 \to \mathbb{R}^{c_1}$ with a kernel $\Psi : \mathbb{Z}^2 \to \mathbb{R}^{c_1 c_2}$ is given by

$$f_j'(g) = \sum_{i=1}^{c_1} \sum_{y \in \mathbb{Z}^2} f_i(y)\Psi_i^j(g^{-1}y). \tag{13}$$

where $1 \leq j \leq c_2$. When $\mathbb{Z}^2$ is replaced with $\mathbb{R}^2$, the summation becomes an integral:

$$f_j'(g) = \sum_{i=1}^{c_1} \int_{y \in \mathbb{R}^2} f_i(y)\Psi_i^j(g^{-1}y). \tag{14}$$

Then, to switch to the sphere S^2, we need only swap $\mathbb{R}^2$ for $\mathbb{S}^2$ in the equation above, and take G to be the group SO(3), so that $R \in \mathrm{SO}(3)$ is a rotation:

$$f_j'(R) = \sum_{i=1}^{c_1} \sum_{y \in \mathbb{S}^2} f_i(y)\Psi_i^j(R^{-1}y). \tag{15}$$

The signal f' is defined on SO(3). Further convolutional layers must therefore involve SO(3)-convolutions. One challenging complication which therefore emerges is that SO(3) is three dimensional, so that the time complexity of such a convolution becomes $\mathcal{O}(n^6)$. S^2 convolutions, which are taken with respect to the group $G = \mathrm{SO}(3)$, are similarly expensive, and would naively require $\mathcal{O}(n^5)$ time to compute.

Big Idea 5. Let S be a homogeneous space with respect to a group G. Then cross-correlation on S is realized as the G cross-correlation on S. Examples of this include $\mathbb{T}(2)$ cross-correlations on $\mathbb{Z}^2$ and SO(3) cross-correlations on S^2.

3.4.1 Efficient Continuous Cross-Correlation

To circumvent such expensive convolution, the authors of [5] instead perform the convolution in frequency space after taking the generalized Fourier transform of the signal and kernel. The necessary identity is

$$\widehat{f \star g} = \hat{f}\hat{g} \tag{16}$$

which essentially says that the Fourier transform of the cross-correlation of two functions is the same as the product of the respective Fourier transforms of the two functions. Taking the inverse Fourier transform then yields the desired cross-correlation. Performing the fast Fourier transform and inverse Fourier transform in practice in the context of spherical CNN's is explored and elaborated on in great detail in [5].

3.5 CNN's on Homogeneous Spaces

In this final section, we briefly survey the work of [3], which brings together ideas discussed previously, including the base space/fiber formalization from Section 3.3.

Big Idea 6. Start with a Homogeneous space B with respect to a group G, and fix an origin $o \in B$. Let $H \leq G$ be the stabilizer of o, so that $B \simeq G/H$. Setting B to be our base space, and a representation of ρ of H as our fibers, we obtain the *associated bundle.* By ensuring that our kernel is ρ-equivariant, we obtain via cross-correlation a signal on $B \simeq G/H$ which transforms as the induced representation Ind_H^G. This construction is illustrated in Figure 6.

3.6 Future Work

While we touched on many novel ideas, this survey is by no means complete. Several notable omissions I would like to point out (perhaps to be included in later drafts of this same document) are *3d Steerable CNN's* [7], which extends steerable CNN's to three dimensions, and *Gauge Equivariant CNN's* [2], which generalize CNN's beyond the global symmetries of homogeneous spaces to arbitrary manifolds, in which local symmetry is realized through guage equivariance.

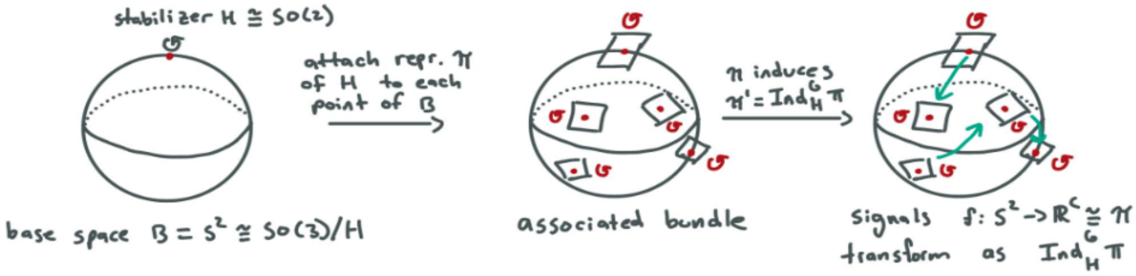
stabilizer H ≅ SO(2)
attach repr. π of H to each point of B
base space B = S² ≅ SO(3)/H
associated bundle
π induces π' = Ind_H^G π
Signals f: S² -> R^C ≅ π transform as Ind_H^G π

Figure 6: A general formula for CNN's on homogenous spaces.

4 Acknowledgements

I'd like to thank Xiaochen Du, Zack Ankner, Mark Takken, and Peter Pao-Huang for helpful discussion and feedback during the writing of this survey paper. I'd also like to thank Professor Tess Smidt, and the other course staff of 6.s966 for a wonderful semester.

References

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

[2] Taco Cohen et al. "Gauge equivariant convolutional networks and the icosahedral CNN". In: *International conference on Machine learning*. PMLR. 2019, pp. 1321–1330.

[3] Taco S Cohen, Mario Geiger, and Maurice Weiler. "A general theory of equivariant cnns on homogeneous spaces". In: *Advances in neural information processing systems* 32 (2019).

[4] Taco S Cohen and Max Welling. "Steerable cnns". In: *arXiv preprint arXiv:1612.08498* (2016).

[5] Taco S Cohen et al. "Spherical cnns". In: *arXiv preprint arXiv:1801.10130* (2018).

[6] Pim De Haan et al. "Gauge equivariant mesh CNNs: Anisotropic convolutions on geometric graphs". In: *arXiv preprint arXiv:2003.05425* (2020).

[7] Maurice Weiler et al. "3d steerable cnns: Learning rotationally equivariant features in volumetric data". In: *Advances in Neural Information Processing Systems* 31 (2018).

[8] Wikipedia contributors. *Induced representation — Wikipedia, The Free Encyclopedia*. [Online; accessed 10-May-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Induced_representation&oldid=1144264252.