Probabilistic symmetry and convolutional neural networks

Sean La

11 December 2019

1 Background

Deep learning has revolutionized the research field of machine learning due to its great accuracy and generalizability on large, high dimensional datasets. The convolutional neural network (CNN) architecture has become the mainstay for constructing computer vision platforms. The chief design principle of CNNs is the incorporation of convolutional layers that allow the network to detect certain features in the input image. In particular, these convolutional layers are equivariant under translation of the image, in the sense that translation of the input image causes the activations of the convolutional layers to translate in the same way. This design feature is the key reason why convolutional neural networks have achieved excellent generalizability on a variety of datasets.

In the last decade, there has been much work in expanding mathematical understanding of deep learning algorithms. Deep learning algorithms can be formalized under two separate (but not mutually exclusive) points of views: they can be viewed as deterministic functions $f : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X}$ is the set of input features and $\mathcal{Y}$ is the set of labels, or they can be viewed from a probabilistic point of view so that $Y = f(\eta, X)$, where X and Y are the input and output random variables, and is some outsourced noise random variable. Viewing deep learning algorithms deterministically, Kondor and Trivedi proved that f containing a convolution operation of the form $(h * g)(u) = \int h(uv^{-1})g(v)\, dv$ is both necessary and sufficient for the layers of a neural network $y = f(x)$ to be equivariant [KT18].

On the other hand, when viewed from a probabilistic point of view, Bloem-Reddy and Teh provided necessary and sufficient conditions for a neural network $Y = f(\eta, X)$ to be invariant or equivariant under the actions of a compact group [BT19]. The conditions posed by Bloem-Reddy and Teh on the structure of $Y = f(X, \eta)$ are quite general and do not, by themselves, restrict f to be a function involving convolution, as is the case in of the results of Kondor and Trivedi. Then, a natural question to ask is: *what further conditions on $Y = f(X, \eta)$ are necessary so that f is restricted to functions involving convolutions?*

In this project, we answer this question by providing a sufficient condition that is both simple and intuitive: linearity of f in the second argument (i.e. $f(\eta, X_1 + X_2) = f(\eta, X_1) + f(\eta, X_2)$). The project report proceeds as follows. First, we give a brief overview of mathematical concepts that will be useful for understanding the rest of the report. Then, we summarize the main results of [BT19]. After that, we provide key definitions and theorems from [KT18] that will prove useful for proving our main result. Penultimately, we prove that linearity of $Y = f(\eta, X)$ in the second argument is a sufficient condition for $Y = X * \chi(\eta)$, i.e. Y is a convolution involving X. Finally, we provide open questions and future research directions.

1.1 Mathematical background

The results of both [BT19] and [KT18] are very mathematics-heavy and involve results from a wide variety of fields of mathematics, namely group theory, fourier analysis, representation theory, and of course, probability theory. In this report, we assume the reader has a good working knowledge of probability theory, but it may be too much for us to ask for a background in these other subjects as well. So then, before we get into the thick of things, we give a brief overview of group theory and how it relates to symmetry. Note that we will not cover fourier analysis nor representation theory, as they are not necessary to understand the main definitions and theorems of [BT19] or [KT18].

1.1.1 A gentle introduction to group theory

Definition 1. *A group is a set of elements G along with a binary operator $\cdot : G \times G \to G$ so that for every $u, v \in G$, we have $u \cdot v = w \in G$. For succintness, we denote a group by the pair $(G, \cdot)$. Moreover, we assume the existence of two types of elements in G:*

- *there exists an element $e \in G$ (which we call the identity element) so that for every $u \in G$, $e \cdot u = u \cdot e = u$, and*
- *for every $u \in G$, there exists its inverse $u^{-1} \in G$ so that $u \cdot u^{-1} = u^{-1} \cdot u = e$.*

Example 1.1. An example of a group is the integers equipped with addition, i.e. $(\mathbb{Z}, +)$ is a group. In this case, 0 is the identity since $z + 0 = z$ for every $z \in \mathbb{Z}$, and every element in $\mathbb{Z}$ has an inverse $-z := z^{-1}$ so

that $z + (-z) = 0$. Note that the integers equipped with multiplication is not a group, since the only integer that has an inverse that is also an integer is 1.

1.1.2 Everything you need to know about fourier analysis and representation theory

1.2 Probabilistic symmetry in machine learning

1.3 Deterministic symmetry in machine learning

In the deterministic view of machine learning, a machine learning algorithm is considered to be a function $f : \mathcal{X} \times \Theta \to \mathcal{Y}$ which maps an input object $x \in \mathcal{X}$ and parameters of the algorithm $\theta \in \Theta$ to a label $y \in \mathcal{Y}$, so that $y = f(x; \theta)$. There is no randomness captured in this formulation of machine learning, and so we provide no distributional assumptions on $\mathcal{X}$ nor $\mathcal{Y}$.

Definition 2. *Let $\mathcal{X}_0, \dots, \mathcal{X}_L$ be a sequence of index sets, $V_0, \dots, V_L$ vector spaces, $\phi_1, \dots, \phi_L$ linear maps*

$$\phi_\ell : L_{V_{\ell-1}}(\mathcal{X}_{\ell-1}) \to L_{V_\ell}(\mathcal{X}_\ell)$$

and $\sigma_\ell : V_\ell \to V_\ell$ pointwise nonlinear functions. A multi-layer feed-forward neural network (MFF-NN) is a sequence of maps $f_0 \mapsto f_1 \mapsto f_2 \mapsto \dots, \mapsto f_L$, where

$$f_\ell(x) = \sigma_\ell(\phi_\ell(f_{\ell-1})(x))$$

Definition 3. *Let G be a group and $\mathcal{X}_1, \mathcal{X}_2$ be two sets with corresponding G-actions*

$$T_g : \mathcal{X}_1 \to \mathcal{X}_1, \; T'_g : \mathcal{X}_2 \to \mathcal{X}_2.$$

Let V_1 and V_2 be vector spaces, and $\mathbb{T}$ and $\mathbb{T}'$ be the induced actions on G on $L_{V_1}(\mathcal{X}_1)$ and $L_{V_2}(\mathcal{X}_2)$. We say that a map $\phi : L_{V_1}(\mathcal{X}_1) \to L_{V_2}(\mathcal{X}_2)$ is equivariant with the action of G (or G-equivariant for short) if

$$\phi(\mathbb{T}_g(f)) = \mathbb{T}'_g(\phi(f)) \; \forall f \in L_{V_1}(\mathcal{X}_1)$$

for any group element $g \in G$.

Definition 4. *Let $\mathcal{N}$ be a feed-forward neural network as defined in Definition 2, and G be a group that acts on each index space $\mathcal{X}_0, \dots, \mathcal{X}_L$. Let $\mathbb{T}^0, \dots, \mathbb{T}^L$ be the corresponding group actions on $L_{V_0}(\mathcal{X}_0), \dots, L_{V_L}(\mathcal{X}_L)$. $\mathcal{N}$ is a G-equivariant feed-forward network if, when the inputs are transformed $f_0 :\mapsto \mathbb{T}^0_g(f_0)$ for any $g \in G$, the activations of the other layers correspondingly transform as $f_\ell \mapsto \mathbb{T}^\ell_g(f_\ell)$.*

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

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

for some $u \in G$

Definition 6. *Let G be a compact group and $\mathcal{N}$ an $L + 1$ layer feed-forward network in which the ith index set is G/H_i for some subgroup H_i of G. We say that $\mathcal{N}$ is a G-convolutional neural network (or G-CNN) if each of the linear maps $\phi_1, \dots, \phi_L$ in $\mathcal{N}$ is a generalized convolution (see definition 5) of the form $\phi_\ell(f_{\ell-1}) = f_{\ell-1} * \chi_\ell$ with some filter $\chi_\ell \in L_{V_{\ell-1} \times V_\ell}(H_{\ell-1} \setminus G / H_\ell)$.*

Theorem 1. *Let G be a compact group and $\mathcal{N}$ be an $L + 1$ layer feed-forward neural network in which the ℓth index set is of the form $\mathcal{X}_\ell = G/H_\ell$, where H_ℓ is some subgroup of G. Then $\mathcal{N}$ is equivariant to the action of G in the sense of definition 3 if and only if it is a G-CNN.*

2 A sufficient condition for convolution structure of Y

First, we give an essential definition.

Definition 7. *A function $h : E \to F$ is linear if for $x_1, x_2 \in E$, we have*

$$h(x_1 + x_2) = h(x_1) + h(x_2)$$

for all $x_1, x_2 \in E$.

In this section, we show that equivariant $Y = f(\eta, X)$ being linear in its second argument, i.e.

$$Y = f(\eta, X_1 + X_2) = Y = f(\eta, X_1) + f(\eta, X_2),$$

is a sufficient condition for Y to be a convolution of X, i.e.

$$Y \overset{\text{a.s.}}{=} X * \chi(\eta).$$

Theorem 2. *Consider random vectors (X, Y) with index sets $\mathcal{I}_X$ and $\mathcal{I}_Y$ and a compact group G which acts on $\mathcal{I}_X$ and $\mathcal{I}_Y$. Moreover, suppose that $P_{X,Y}$ is G-equivariant. Lastly, suppose that the action of G on both $\mathcal{I}_X$ and $\mathcal{I}_Y$ is transitive. Then if, conditioned on X, $Y = f(\eta, X)$ is a linear function almost surely (i.e. for some almost sure subset of the range of η), then $Y \overset{\text{a.s.}}{=} X * \chi(\eta)$ in the sense of definition 5. More generally, $(X, Y) \overset{\text{a.s.}}{=} (X, X * \chi(\eta))$.*

Proof. Let $F(X) = f(\eta, X)$ be some realization of Y so that $F(\cdot) = f(\eta, \cdot)$ is linear and $gY = f(\eta, gX)$ (the set of all such realizations has probability one since Y is linear almost surely and $gY = f(\eta, gX)$ almost surely). We prove this theorem by constructing an appropriate MFF-NN as in definition 2, and then invoking theorem 1.

We may consider X to be a function of its index set so that $X(i)$ gives the value of X at component i for $i \in \mathcal{I}_X$. Likewise, we view Y as a function of its index set $\mathcal{I}_Y$ in the same way. Then the linear function F can be seen as a function from the space of functions on $\mathcal{I}_X$ to the space of functions on $\mathcal{I}_Y$. Lastly, we set the function σ to be simply the identity function. Then we define our MFF-NN $\mathcal{N}$ by the sequence of maps

$$X := f_0 \mapsto f_1 =: F(X) = Y$$

where

$$Y(i) = F(X)(i) = \sigma(F(X)(i)).$$

for $i \in \mathcal{I}_Y$. Clearly, this construction matches the description of a MFF-NN given in definition 2.

Now, define the action of group G on X to be the action of G on its index set, so that

$$gX := \mathbb{T}_g(X) = X(g \cdot i)$$

for $i \in \mathcal{I}_X$ and $g \in G$. Similarly, define the action of G on Y as the action of G on the its index set $\mathcal{I}_Y$, so that

$$gY := \mathbb{T}'_g(Y) = Y(g \cdot i)$$

for $i \in \mathcal{I}_Y$ and $g \in G$. By assumption, we have

$$\mathbb{T}'_g(Y) = gY = F(gX) = F(\mathbb{T}_g(X))$$

for all $g \in G$ since $P_{X,Y}$ is G-equivariant. Therefore, the layers of $\mathcal{N}$ is equivariant under the action of G.

Because the actions of G on both $\mathcal{I}_X$ and $\mathcal{I}_Y$ are transitive, we may identify the index sets of X and Y by quotient groups G/H_X and G/H_Y, where H_X and H_Y are the stabilizer subgroups of elements chosen to be the origins of $\mathcal{I}_X$ and $\mathcal{I}_Y$, respectively.

Therefore, by theorem 1, our MFF-NN $\mathcal{N}$ is a G-CNN, i.e. we have

$$F(X) = X * \chi$$

for some filter $\chi \in L_{V_{\ell-1} \times V_\ell}(H_{\ell-1} \setminus G / H_\ell)$. Letting the value of η to vary over the subset of its range that gives $f(\eta, X)$ to be linear in X, it is easy to see that

$$f(\eta, X) = X * \chi(\eta).$$

Take $\Omega_0 \subseteq \Omega$ such that $f(\eta, X)$ is both linear and equivariant under the action of G. Because every $\omega \in \Omega_0$ gives that $Y(\omega) = X(\omega) * \chi(\eta(\omega))$ and $\mathcal{P}(\Omega_0) = 1$, we have

$$(X, Y) \overset{\text{a.s.}}{=} (X, X * \chi(\eta))$$

which finishes the proof. □

3 Open questions and research directions

A Exercises

1. Consider a measure μ on a compact group $\mathcal{G}$ that acts on a measureable space $\mathcal{X} = \mathcal{G}$. Define the $\mathcal{G}$-measurable function $f' : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ defined by $f'(x, g) = f(gx)$ for some function f. Compute $\int_{\mathcal{G}} f(x, g) d\mu(g)$ under the following conditions.
   (a) f is $\mathcal{G}$-invariant.
   (b) f is $\mathcal{G}$-equivariant.
2. Prove the following:
   (a) $P_{X,Y}$ is $\mathcal{G}$-equivariant if and only if $(X, Y) \overset{\text{d}}{=} (gX, gY)$.
   (b) $P_{X,Y}$ is $\mathcal{G}$-invariant if and only if $(X, Y) \overset{\text{d}}{=} (gX, Y)$.
3. Suppose $Y \perp\!\!\!\perp X$ and Y is uniformly distributed. Provide f so that $Y \overset{\text{d}}{=} f(\eta, X)$ conditioned on X for some uniformly distributed η.

References

[BT19] B. Bloem-Reddy and Y. W. Teh. *Probabilistic symmetry and invariant neural networks*. 2019. arXiv: 1901.06082 [stat.ML].

[KT18] R. Kondor and S. Trivedi. *On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups*. 2018. arXiv: 1802.03690 [stat.ML].