Unifying transformers and convolutional networks as equivariant maps

Elias Nyholm eliasny@chalmers.se
Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden

Editors: Michael Bleher, Freya Jensen, Levin Maier, Diaaeldin Taha, and Anna Wienhard

Abstract

Motivated by the prevalence of equivariant machine learning models and the success of the framework of linear equivariant convolutional neural networks, we present in this work an extended framework that also includes non-linear equivariant models. More specifically, we represent these models as integral operators and derive conditions on the integrand for the operator to be equivariant. Further, we prove the generality of the proposed framework and show explicitly how common equivariant models, linear as well as non-linear, fit into the proposed formulation. This extended abstract summarises the central points of the preprint Nyholm et al. (2025), which is joint work together with Oscar Carlsson, Maurice Weiler and Daniel Persson.

1. Introduction

The inductive bias of *equivariance* ensures that a machine learning model respects symmetries of data and can often imply beneficial behaviour in training and inference (Gerken et al., 2022; Brehmer et al., 2024). Such symmetry properties can also be a powerful classification tool, an idea first made popular through Klein's Erlangen program (Klein, 1893) and recently picked up by the machine learning community (Bronstein et al., 2021). While linear machine learning layers have been completely classified in terms of their equivariance properties (Cohen et al., 2018a), the extension to non-linear layers has largely been out of reach of theoretical study apart from simple element-wise activation functions. The goal of the present work is to establish a framework that bridges this gap.

The assumptions on and formulation of machine learning models and features made here are in line with a long line of previous work on machine learning models (Cohen et al., 2018a,b; Romero et al., 2020; Gerken et al., 2021). Features are formalised as vector-valued maps $f : X \to V_\rho$ where the domain X is a homogeneous space acted on transitively by some symmetry group G. All stabilisers of the action are then isomorphic to some subgroup $H \leq G$, and we assume that H acts by some representation ρ on the vector space V_ρ. The homogeneous structure of X further implies that it is isomorphic to the quotient space G/H. The feature map $f : X \to V_\rho$ can then be *lifted* to a function on the full group $\hat{f} : G \to V_\rho$ by letting the representation ρ define the value of $\hat{f}$ on orbits of H via the so-called Mackey condition

$$\hat{f}(gh) = \rho(h)^{-1}\hat{f}(g), \quad \forall g \in G, h \in H. \tag{1}$$

We choose to consider this representation of data because the domain G is more convenient to work with than the space $X \cong G/H$. We denote the space of function lifted along the

representation ρ as $\mathcal{I}_\rho$, and note that there exist a canonical action of G on $\mathcal{I}_\rho$ given by precomposition $[k \cdot f](g) = f(k^{-1}g)$ for $k \in G$. For this action we use the simplified notation $gf \equiv g \cdot f$. With this formulation of data as functions living in $\mathcal{I}_\rho$, we represent models which map data to data as operators $\Phi : \mathcal{I}_\rho \to \mathcal{I}_\sigma$ taking maps lifted along ρ to maps lifted along some potentially different representation σ. Then Φ represents an equivariant model if $\Phi[gf] = g[\Phi f]$.

2. Our framework

By generalising the linear framework of equivariant convolutional neural networks (Cohen et al., 2018a,b), we propose a general family of neural network architectures given by maps $\Phi_\omega : \mathcal{I}_\rho \to \mathcal{I}_\sigma$ of the form

$$[\Phi_\omega f](g) = \int_G \omega(g^{-1}f, g')\mathrm{d}g' \tag{2}$$

where $\omega : \mathcal{I}_\rho \times G \to V_\sigma$ should satisfy the Mackey constraint

$$\omega(hf, g) = \sigma(h)\omega(f, g). \tag{3}$$

Note that the particular combination of arguments f, g ensures that the map Φ_ω is equivariant

$$\Phi_\omega[k \cdot f] = k \cdot \big[\Phi_\omega[f]\big] \tag{4}$$

for all $k \in G$ and all $f \in \mathcal{I}_\rho$, and that the condition eq. (3) is required for the output function $\Phi_\omega f$ to satisfy the Mackey condition (1) of lifted feature maps. The observant reader might notice that the argument g' and the integral over G in (2) do not go into either of these requirements – in fact, we could develop an equivalent framework with the integral and g'-dependence absorbed, as $[\phi_w f](g) = w(g^{-1}f)$ and with a condition similar to (3). Nonetheless, we choose to include the integral as it represents the message-passing perspective on machine learning models which is common for many architectures. Explicit examples of such message passing networks are given in Section 3.

As the integrand ω can take any possible form, the expression (2) can represent a wide range of different models and mapping, all of which are equivariant. In fact, we prove that if we allow ω to be chosen from a sufficiently large space of maps (more specifically the space of distributions, also known as generalised functions), we can represent **any** equivariant map on this form:

Theorem 1: For any equivariant map $\lambda : \mathcal{I}_\rho \to \mathcal{I}_\sigma$ there exist a choice of vector-valued distribution $\omega : \mathcal{I}_\rho \times G \to V_\sigma$ such that $\lambda = \Phi_\omega$.

This result ensures that we do not restrict ourselves in picking the specific form (2) for our map – any result we show inside our particular framework should hold for any arbitrary instance of equivariant map, since any equivariant map is a special case. We take this point further in the next section by giving explicit expressions for ω that yield well-known equivariant architectures from the literature, from both the CNN and attention family of models.

3. Deriving architectures from the literature

Here we give a brief overview of some of the instances of equivariant models that we explicitly recover inside our framework in the main article (Nyholm et al., 2025). An overview of the full range of instances given in the article is illustrated in fig. 1.

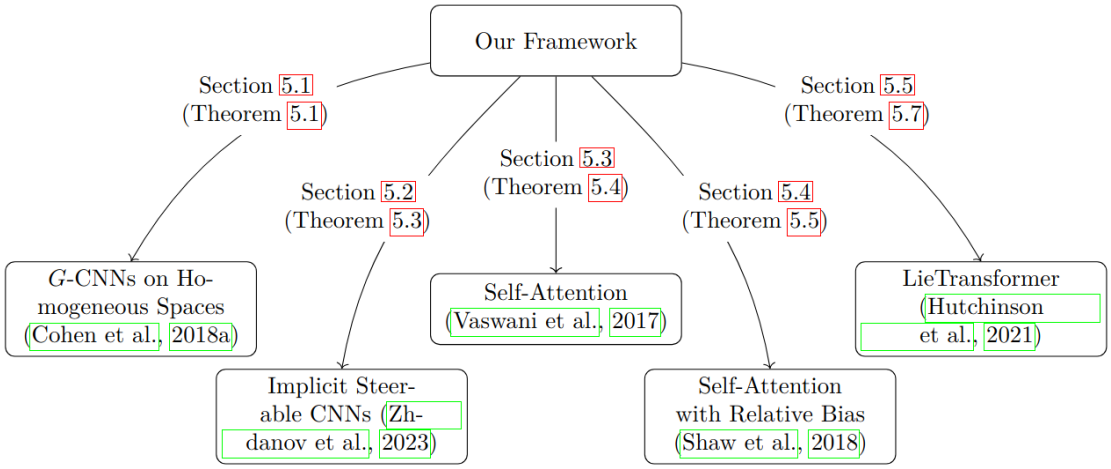

Figure 1: Instances of linear and non-linear equivariant models studied as special cases of our proposed framework in Nyholm et al. (2025). References to sections and theorems refer to those in the aforementioned article.

The transformer model consist of a self-attention layer which, when not augmented using positional embeddings, is equivariant under the symmetric group $G = S_n$ acting by permutations on the sequence of input and output tokens (Vaswani et al., 2017). By considering a finite index set $X = \{0, 1, \dots, n-1\}$ and noting that the group of all permutations S_n acts homogeneously on X, we can represent tokens as vector-valued functions on X which can be lifted to feature maps $f : S_n \to V_\rho$ living in the representation induced from the trivial representation $\rho(g) \equiv \mathbb{1}_{V_\rho}$. Acting on such lifted feature maps, the vanilla self-attention layer without positional embeddings is given by

$$\omega(g^{-1}f, g') = \frac{\exp\left([g^{-1}f](e)^\top W_Q^\top W_K [g^{-1}f](g')\right)}{\mathcal{Z}} W_V [g^{-1}f](g') \tag{5}$$

where $\mathcal{Z}$ is allowed to depend on global properties of the function $g^{-1}f$ and W_Q, W_K, W_V are matrices of tunable parameters. Note the interpretation of the first fraction as a weight to the message $W_V[g^{-1}f](g')$ associated to site g'. In this way the integral in (2) can be interpreted as a cumulation of messages in the sense of message passing networks.

It is also possible to incorporate relative positional embeddings (Shaw et al., 2018) in our framework, with slight modification due to the changed symmetry group. More specifically, we can consider an infinite sequence of tokens indexed by $X = \mathbb{Z}$, and the symmetry group

$G = (\mathbb{Z}, +)$ acting by addition. Then the relative positional embedding term is given by some choice of function $\psi : \mathbb{Z} \to \mathbb{R}$ which can be added or appended as $\psi(g')$ to the expression (5) depending on how one wishes to implement the embedding.

Convolutions are linear in the feature map f, and linearity in our framework (2) implies

$$\omega(g^{-1}f, g') = \kappa(g')[g^{-1}f](g') \tag{6}$$

for some linear map $\kappa : G \to \mathrm{Hom}(V_\rho, V_\sigma)$. Given bases for the vector spaces V_ρ, V_σ the output of κ will then be a matrix. This is in fact exactly the kernel of the equivariant CNN layer introduced in Cohen and Welling (2016), see also (Cohen et al., 2018a), which further specialises to the original CNN layer (LeCun et al., 1989) when one chooses translations as the symmetry group. The more standard expression for equivariant CNN kernels is recovered by a change of variable $g' \mapsto g^{-1}g'$ under the integral sign in (2). The observation that linear equivariant operators all represent convolutional models is well-known in the geometric deep learning literature (Kondor and Trivedi, 2018). One can note here that the convolutional kernel $\kappa(g')$ and the relative positional bias $\psi(g')$ from the previous paragraph are almost the same object in our framework, up to the fact that κ is matrix-valued while ψ only outputs a single scalar.

The LieTransformer is a family of machine learning models which generalise the vanilla transformer architecture to equivariant models with symmetries given by an arbitrary Lie group (Hutchinson et al., 2021). The architecture is set by fixing a map $\alpha : V_\rho \times V_\rho \times G \to \mathbb{R}$, and in its most general form depends additionally on an embedding matrix W_V. This family of equivariant models fits into our framework by a choice of integrand in (2) as

$$\omega(g^{-1}f, g') = \frac{\alpha\big([g^{-1}f](e), [g^{-1}f](g'), g'\big)}{\mathcal{Z}} W_V [g^{-1}f](g') \tag{7}$$

where $\mathcal{Z}$ is allowed to depend on global properties of the function $g^{-1}f$ as well as on the choice of α. The work of Hutchinson et al. (2021) also outlines several specialisations of this family, in particular to models that strongly resemble standard self-attention.

4. Conclusion

We present a fully general framework to represent equivariant machine learning models as integral operators which extends the linear formulation to the non-linear regime. We also show explicitly how specific linear and non-linear equivariant models from the literature fit into the framework, more specifically linear CNN models and non-linear attention models. At this point the work is exclusively of a theoretical nature, and we have not implemented any version of the general framework. Instead, we see the main point of the present framework to provide general building blocks for equivariant models. Future directions of the project include studying possible avenues to make our findings more concrete in terms of implementations, and to develop general ways to efficiently implement such equivariant non-linear models in hardware. Theoretical extensions to our framework are also of interest, for example to manifold domains or by extending the global symmetry group G to one acting by local gauge transformations.

References

Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale?, 2024. URL https://arxiv.org/abs/2410.23179.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/cohenc16.html.

Taco Cohen, Mario Geiger, and Maurice Weiler. A General Theory of Equivariant CNNs on Homogeneous Spaces. In *Neural Information Processing Systems*, 2018a. URL https://api.semanticscholar.org/CorpusID:53248796.

Taco S. Cohen, Mario Geiger, and Maurice Weiler. Intertwiners between Induced Representations (with Applications to the Theory of Equivariant Neural Networks), 2018b. URL https://arxiv.org/abs/1803.10743.

Jan Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Equivariance versus augmentation for spherical images. In *International Conference on Machine Learning*, pages 7404–7421. PMLR, 2022.

Jan E. Gerken, Jimmy Aronsson, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Geometric deep learning and equivariant neural networks. *Artificial Intelligence Review*, 56:14605 – 14662, 2021. URL https://api.semanticscholar.org/CorpusID:235248075.

Michael Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. LieTransformer: Equivariant Self-Attention for Lie Groups. In *ICML*, pages 4533–4543, 2021. URL http://proceedings.mlr.press/v139/hutchinson21a.html.

Felix Klein. Vergleichende betrachtungen über neuere geometrische forschungen. *Mathematische Annalen*, 43(1):63–100, Mar 1893. ISSN 1432-1807. doi: 10.1007/BF01446615. URL https://doi.org/10.1007/BF01446615.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2747–2755. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kondor18a.html.

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*,

volume 2. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.

Elias Nyholm, Oscar Carlsson, Maurice Weiler, and Daniel Persson. Equivariant non-linear maps for neural networks on homogeneous spaces, 2025. URL https://arxiv.org/abs/2504.20974.

David W. Romero, Erik J. Bekkers, Jakub M. Tomczak, and Mark Hoogendoorn. Attentive group equivariant convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-2074/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, 30, 2017. URL https://dl.acm.org/doi/10.5555/3295222.3295349.