Position Paper
Preliminary Steps Towards Isotropic Deep Learning

George Bird

Department of Computer Science
University of Manchester
Manchester, UK
george.bird@postgrad.manchester.ac.uk

May 1, 2025

Abstract

This position paper proposes a new paradigm to be termed *Isotropic Deep Learning.* It has been shown that functional forms in deep learning influence the activation distribution. Broken symmetries in functional forms induce broken symmetries in embedded representations through training. This is an artefact in the human choice of functional forms leaving this artificial signature in the representations. There appears no strong a-priori reason why such a representation is desirable and therefore justified, and several reasons why it is undesirable are listed in this paper. Therefore, it will be explored how unconstraining such representations may allow a network to acquire a more optimal representation matching the computational demand of the current task. Due to the substantial functional form overhaul required, it will be justified why this exploration could be considered a new form of Deep Learning, in analogy to the separation seen in radial basis function networks. In upcoming works regularisers, normalisations and optimisers will be added to expand the branch of Isotropic Deep Learning.

This position paper is a brief outline of the authors area of interest for their PhD. This paper will be continuously updated to include additional details and is shared for purposes of feedback and collaboration. The author has decided to produce this position paper outlining the concept since empirical validation is been delayed due to developments of a suitable optimiser. This work is currently considered incomplete and only outlines the basic principles proposed.

1 Introduction

Current activation functions have an elementwise functional form. Due to this, their non-linear transformation differ angularly in effect. Therefore, this is termed an *anisotropic* function, due to this rotational asymmetry. This assymetry is uneven particularly about the standard (Kronecker) basis vectors and its negative $\{+\hat{e}_i, -\hat{e}_i\}_{\forall i}$, due to the elementwise nature. Therefore it can be said to distinguish the standard basis - a *distinguished basis*[1]. This can be considered a choice to break rotational symmetry, and reduce it to a discrete permutation symmetry about this standard basis. Since non-linearities are pivotal to the networks ability to be a universal approximator, then it is an important aspect in achieving its desired computation.

The network is then expected to adapt its computation due to these stand-out directions. Through training the broken symmetry in functional forms is transferred to a broken symmetry in the activations. Non-linearities may locally stretch or compress the manifold, grouping or separating representations in a desirable way. Since these non-linear zones are centred around the distinguished bases, it is expected that the embedded representations then move to these useful angular regions to achieve computations. Therefore

[1] This is generalised from a *privileged basis.* The change to *distinguished basis* reflects that the basis may be more-or-less aligned to the representation; whereas, privlidged basis will be used to suggest a basis more aligned to the embedded activations. The term 'basis' will be retained even though the set of *distinguished vectors* may also be under/over complete for spanning the whole space.

they may move to the non-linearities's extremums, aligned, anti-aligned or other aligned geometries, through training. This may correspond to a local, dense or sparse coding respectively, where the latter includes the superposition phenomena.

Therefore the network has adapted its representations through training due to these functional form choices. This is the causal hypothesis proposed. This tendency of neural alignment is observed in several studies - this hypothesis is proposed to explain why a neuron-alignment, or more generally privileged basis alignment, is tended to be observed. This is a human caused representational collapse onto the privileged basis - not a task necessitated collapse. There appears little justification as to why this is desirable. This effect on representations will no doubt influence computation. Without a-priori justification, this inductive bias may be detrimental.

Therefore, this position paper encourages a departure from this functional form paradigm towards the isotropic paradigm, which can be thought to unconstrain the representations into more optimal arangments for a task. This tenants of this paradigm are suggested for all architectures on general tasks.

2 Problems With Anisotropy

I posit there at least the following disadvantages with anisotropic functional forms which would justify the departure into using the isotropic paradigm. Many of these are centred around activation functions as a starting point but should be generalised and added to. For this brief first position paper these will be left as mostly qualitative arguments and updated shortly to more quantative ones.

2.1 Neural Refractive Problem

Describes how large magnitude activations may converge or diverge from their initial trajectories, causing multiple semantic meanings to be undesirably intertwined or generaliation to fail. Particularly, the network, through training, may carefully bring activations for training samples into a 'zone of sensibility' where computations and semantic interactions operate in an expected and predictable way to the network. However, out-of-training-distribution data, may fall outside this zone and the interference by inseparable semantics or unpredictable non-linear mappings result in the network dramatically failing to extrapolate.

Mathematically, supplying a vector $\alpha\hat{b}$ to an elementwise non-linear activation function, $\sigma : \mathbb{R}^n \to \mathbb{R}^n$, usually causes skewing of directions, particularly at large magnitude inputs there results in limit points. If elementwise-tanh is used: $\sigma = \sum_{i=1}^{n} \frac{\hat{e}_i}{1+e^{-\vec{x}\cdot\hat{e}_i}}$, then the limiting behaviour is $\lim_{\alpha\to\infty} \sigma\left(\alpha\vec{b}\right) = \left(\pm 1, \cdots, \pm 1\right)^T$ - thus semantic separability is lost, particularly if directions correspond to interpretable meanings (the linear features hypothesis). In other activation functions like ReLU, this is only a subset of $\hat{b} \in \mathcal{S}^{n-1}$. This may also be a skew at large magnitudes as opposed to a limit point.

A general fix, under the linear features hypothesis, would be $\Phi\left(\alpha\hat{b}\right) = \phi\left(\alpha\right)\hat{b}$, which is identical to the isotropic equivariant functional form proposed in this work $\left[\Phi, \mathbf{R}\right] = \Phi\mathbf{R} - \mathbf{R}\Phi = 0$ for $\mathbf{R} \in \mathrm{SO}\left(\mathrm{n}\right)$.

2.2 Weight Locking - Optimisation Energy Barriers

Weight locking is an emergent consequence of broken rotationally symmetric functional forms. As discussed, a causal link between anisotropic functional forms and privilidged basis alignment is suggested through training. Therefore, after sufficient amount of training, activations are already privilidged basis aligned, and semanticicty is given to each vector of this privilidged basis. Any purtabation to a parameter which changes an activation from a privilidged-aligned state to a privilidged-unaligned state, results in the activation becoming semantically disjoint from the networks interpretation. Therefore, it may become semantically ambigious and one may expect a resultant rise in the loss.

Consequently, an optimisation threshold is formed around every parameter - the optimisation step must be sufficiently large to overcome this threshold such that the parameter correction produces a transformation which maps the activations to an existing semantic direction. The optimisation barrier may be some function of the angular separation of the semantic directions. This may create a plethora of architectural local-minima about the space.

This is deeply linked to connectivity of basins in the loss. The continuous special orthogonal rotation group means in the mathematical sense, similar solutions sit within connected basins since they can be continously connected with constant loss. In effect, parameters may be rotated, and following parameters anti-rotated to result in consistent function and therefore constant loss. This is not the case for the discrete permutation symmetry, where such a continuous transform will not guarantee consistent loss, so therefore connected basins are not guaranteed. Despite this, the success of ensemble models would suggest that sets of distinct basins may exist in either case. This would mean that representational alignment can be more consistently established through rotation of latent spaces between models if they occupy analogous basins.

2.3 Semantic Interpolatability and Optimised Representations

A continuous symmetry may also enable continuously interpolatable semantics. Representations may then occupy intermediate unaligned states which are semantically interpolatable, rather than discrete semantic directions suggested by the linear features hypothesis. This may advantage the model by allowing

As mentioned, there is little justification that privilidged-basis alignment is desirable to begin with. Allowing continous semantic interpolation may allow the representations to become unaligned from any privilidged basis and produce a more effective distribution - one which suits the data or the task, rather than a human imposition.

2.4 Representational Capacity Limits

Discretising aligment to vectors of the distinguished basis limits representational capacity. If the privilidged basis is neuron-aligned then we get $2n$ semantics concepts for an $\mathbb{R}^n$ space, whilst sparse coding may be $\binom{\sum_{i=1}^{n} o(i)n}{i}$, with occupation function o, or dense 2^n. The interference and discrete representation capacity must be balanced through training.

However, under a continuous embedding the representation capacity can in-principle be driven to infinity and is only limited by semantic interferences. Since semantic interpolatability may emerge, then representational capacity is a distinctly discrete term which may not be applicable to a more continuous networks.

2.5 Non-linear Corrections

To some extent this is an emergent phenomena of the aforementioned issues, when considering discrete semantic directions and weight-locking; however, the network may learn to compensate for undesirable aspects of an elementwise mapping. For example, training time may be needlessly expended on bringing activations within the zone-of-sensibility, or undoing unpredictable reshaping of manifolds by the anisotropic network. This may be more common in anisotropic functional forms due to the angular

3 Elevating Isotropy

As mentioned, elementwise functional forms are defined by a permutation equivariance in functional form. The permutation group is a discrete subset of the continuous special orthogonal (rotation) group; therefore, to produce a continuous symmetry one may use this equivariance relation to construct new functions.

This generalises to continuous rotational symmetry which is then extended into a basis-independent framework. this can be generalised to many other functions Geometric Deep Learning

4 Functional Form Outlines

The basic framework requires equivariance or invariance to $\mathrm{SO}(n)$ symmetries in functional forms. This is superficially similar to equivariant networks; however,

5 Initial Proposed Functional Forms

5.1

6 Proposed Applications

Immediate applications where one may consider benefit.

6.1 Representational Alignment

6.2 Contrastive Learning

Weight locking

6.3 Attention

7 Conclusion

References