Quantum Deep Learning and Renormalization

– A Group-Theoretic Approach to Hierarchical Feature Representations –

WASP PHD PROJECT
MATHEMATICS FOR AI

RESEARCH TEAM

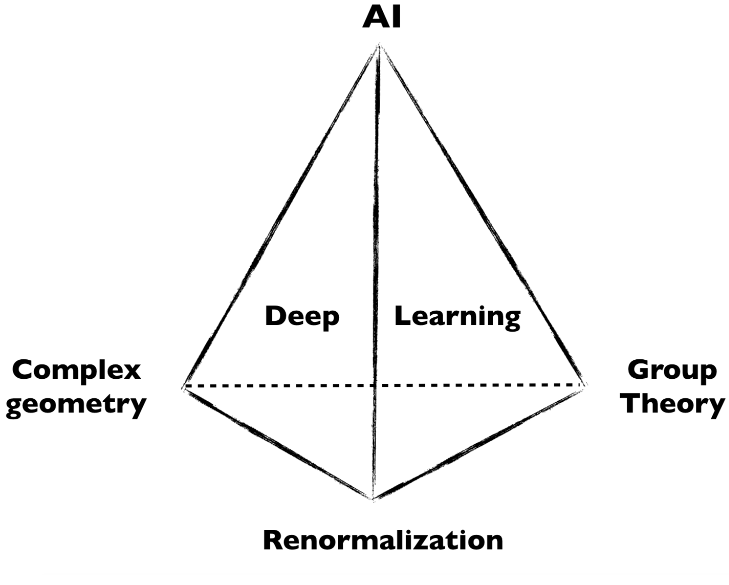
Main supervisor: Daniel Persson, *Associate Professor, Dept. of Mathematical Sciences, Chalmers*

Assistant supervisors:

Robert Berman, *Professor, Dept. of Mathematical Sciences, Chalmers*

Christoffer Petersson, *Deep Learning Research Engineer at Zenuity and Docent in Theoretical Physics*

Computer support: Martin Raum, *Associate Professor, Dept. of Mathematical Sciences, Chalmers*



Abstract

Deep learning is an approach to machine learning that uses multiple transformation layers to extract hierarchical features and learn descriptive representations of the input data. It has been enormously successful in tasks such as computer vision, speech recognition and language processing. However, despite these successes we are lacking a fundamental understanding of *why it works*. It is therefore imperative to develop an understanding of the mathematical structures underlying deep learning. The PhD project will address this from the point of view of group theory combined with insights from theoretical physics. The project also has intriguing relations with gradient flows in complex geometry.

Research project description

Group theory and deep learning. The basic idea of deep learning (DL) is that learning processes takes place in multi-layer networks known as Deep Neural Networks (DNN) of "artificial neurons", where each layer receives data from the preceding layer and processes it before sending it to the subsequent layer. Suppose one wishes to categorize some data sample X according to which class Y it belongs to. As a simple example, the input sample X could be an image and the output Y could be a binary classification of whether a dog or a cat is present in the image. The first layers of a (convolutional) DNN would learn some basic low-level features, such as edges and contours, which are then transferred as input to the subsequent layers, which would learn more sophisticated high-level features, such as combinations of edges, legs and ears. The learning process takes place in the sequence of hidden layers, until finally producing an output $\hat{Y}$, which is to be compared with the correct image class Y. The better the learning algorithm, the closer the DNN predictions $\hat{Y}$ will be to Y on new data samples it has not trained on. In short, one wishes to minimize the *loss function*, which measures the difference between the output $\hat{Y}$ and the class Y.

Concretely, one can represent the transformation from one layer to the next as a linear map $\mathbb{R}^n \to \mathbb{R}^k$, combined with a non-linear transformation $f : \mathbb{R} \to \mathbb{R}$, called the *activation function*. Let the vector $\mathbf{x} \in \mathbb{R}^n$ represent the input data and let W be the weight matrix of learnable parameters which realizes the linear transformation from one layer to the next. Then the output from each layer is given by $f(W\mathbf{x} + \mathbf{b}) \in \mathbb{R}^k$, where f acts component-wise on vectors and $\mathbf{b}$ is a "bias"-vector of learnable parameters which, together with W, combines into an affine transformation of $\mathbf{x}$.

Once some important feature has been identified, or *learned*, this information may propagate unchanged through the subsequent layers, and hence be preserved throughout the correspondning transformations. By analogy with terminology from group theory a feature is thus learned if it is *invariant* under a transformation. So, one can interpret the learning process of a particular feature F as searching for transformations T that stabilizes it, i.e. such that $T(F) = F$. In group theory one can study the orbits of some element s of a set S with respect to a group G, i.e one studies the orbit $\mathcal{O}_x = \{gxg^{-1} \,|\, g \in G\}$. The set S will be decomposed into unions of such obits according to a partial ordering into successively larger and larger orbits $\mathcal{O}_0, \mathcal{O}_1, \ldots,$ where $\mathcal{O}_0$ is the trivial orbit, $\mathcal{O}_1$ the smallest non-trivial orbit (the minimal orbit), etc. From this perspective, we can associate simpler features with smaller orbits; hence, the process of learning the simplest features is analogous to searching for *minimal orbits* [1]. Although the DNNs do not form groups in the strict sense, they still exhibit an approximate group structure, called *shadow group* [1], in which the above analogy can be utilized.

The upshot of these observations is that learning algorithms can be mapped to the group theory problem of finding minimal orbits. This is promising since it provides steps towards a mathematical framework underlying deep learning. Minimal orbits also play an important role in representation theory and thus this opens up a vast “tool box” for further studies, which the project aims to exploit.

Deep Learning and Renormalization. There are also close analogies between the hierarchical learning algorithm, in which low-level feature representations are transformed and combined into semantic meaningful high-level feature representations, and the concept of renormalization group flow in theoretical physics [2]. This implements the idea of *course graining* which is key to deep learning. In statistical mechanics, renormalization provides a way to “integrate out” degrees of freedom that are irrelevant at large scales. In other words, it is a process that isolates the most important features of a physical system, while throwing away the less important information. This mirrors the process of DL where the input could be a picture involving millions of numbers in terms of pixel values, while the output could be a simple binary classification (“dog” or “cat”).

This heuristic comparison between deep learning and renormalization was made concrete in a seminal paper by Mehta and Schwab [2]. They were able to provide an exact mapping between a specific model of renormalization known as *block spin renormalization* and a DNN based on so called *restricted Boltzmann machines.* In restricted Boltzmann machines the neurons in each layer only communicate with the some of the neurons in the next layer, and the interactions are controlled by a Boltzmann distribution. In this correspondence the problem of minimizing the difference between the object Y and the output $\hat{Y}$ can be phrased as the optimization problem of minimizing the difference between the free energies of the initial system and the renormalized one.

The theory of renormalization is also an enormously powerful technical and conceptual tool in quantum field theory, playing a crucial role in the standard model of particle physics. The results of Mehta and Schwab only employs a very simple model for renormalization and it is natural to seek a deeper connection between DL and renormalization in quantum field theory. It would also be very interesting to explore potential relations to *entanglement renormalization* and the associated tensor networks in quantum information theory [4]. The aim of the PhD project is to combine the insights from group theory and representation theory, as well as the renormalization group in statistical physics and quantum theory to further our understanding of the mathematical structures behind deep learning.

Connections with complex geometry. The optimization problem in DL can be formulated in terms of a gradient descent flow on the parameter space of the DNN, the goal being to minimize the loss function. This opens the window towards intriguing connections with other parts of current mathematics, in particular complex geometry. In recent years a formalism has been developed that allows to understand certain classes of complex geometries (Kähler-Einstein geometries) as macroscopic limits of a microscopic random point processes in statistical mechanics [5]. The process of passing from the microscopic point process to the macroscopic geometry is analogous to the course graining that takes place in passing to higher layers in the DNN. Moreover, the Kähler-Einstein geometry corresponds to a minimum of a certain functional (the K-energy), which is reminiscent of the loss functions in DNNs. There also appear to be close connections between the (stochastic) gradient flows that appear both in the complex geometric and in the DNN settings. A secondary aim of the PhD project will therefore be to explore these connections.

Applications to AI

Artificial Intelligence (AI) refers to intelligence exhibited by machines, as opposed to the natural intelligence of humans. These are machines demonstrating cognitive functions that we normally associate with humans, like solving problems or learning new skills. The notion of *machine learning* refers to the development of computer algorithms that can mimic the brain's ability to learn and improve automatically through experience. Machine learning is often classified into two main categories: *supervised* and *unsupervised* learning. In supervised learning the learning process is driven by known input-output pairs, while unsupervised learning corresponds to the search for some unknown patterns.

An important application of AI is self-driving vehicles, which are desired for numerous reasons, in particular for increased safety and reduced environmental impact. To this end one needs a variety of sensors (cameras, radar, lidar etc) mapping the area around the car, and an efficient algorithm that can assess the data in real time. To implement this one is using precisely the types of deep learning algorithms that have been discussed above. However, as already stressed, as of yet there is no good understanding of why certain specific models work better than others. Recently it was proposed that the key mechanism is the *Information Bottleneck Principle* (IBP) [3] which states that deep learning should be understood as a method for efficiently compressing data, or, equivalently, throwing away irrelevant information. In other words, it solves the problem of finding the maximally compressed mapping of the input variable, while preserving as much as possible of the information about the output. This tradeoff between compression of data and information content is exactly the same form of course graining which lies at the core of renormalization. The results of the PhD project will therefore have direct applications to the DL algorithms which form an key part of AI.

The research team

The PhD project will be supervised by Daniel Persson, who has a background in theoretical physics, notably string theory and quantum field theory. He is an expert in group theory and representation theory, with extensive experience with the study of group orbits. He has recently co-authored a book [6] which in particular employs these techniques. One of the assistant supervisors will be Robert Berman who is an expert on complex geometry and has developed an approach to Kähler-Einstein metrics using techniques from statistical mechanics and optimal transport [5]. To maintain the connection with applications to AI, the second assistant supervisor will be Christoffer Petersson, docent in theoretical physics from Chalmers with expertise in particle physics, currently working as Deep Learning Research Engineer at the newly formed company Zenuity (`https://www.zenuity.com`), which is a joint venture between Volvo Cars and Autoliv, developing software for self-driving vehicles. Finally, to implement deep learning algorithms on the computer the team will benefit from the programming skills of Martin Raum, who is an expert on C++, Haskell and Python, as well as in representation theory.

References

[1] Paul, A. and Venkatasubramanian, S. [arXiv:1412.6621 [cs.LG]]

[2] Mehta, P. and Schwab, D. J., [arXiv:1410.3831 [stat.ML]]

[3] Tishby, N. and Zaslavsky, N., [arXiv:1503.02406 [cs.LG]]

[4] Swingle, B., [arXiv:0905.13.17 [cond-mat.str-el]]

[5] Berman, R. J., [arXiv:1609.05422 [hep-th]]

[6] Fleig, P., Gustafsson, H. A. P., Kleinschmidt, A, and Persson, D., [arXiv:1511.04265 [math.NT]]