Synaptic Field Theory

Jaeok Yi

Department of Physics, KAIST

October 23, 2025

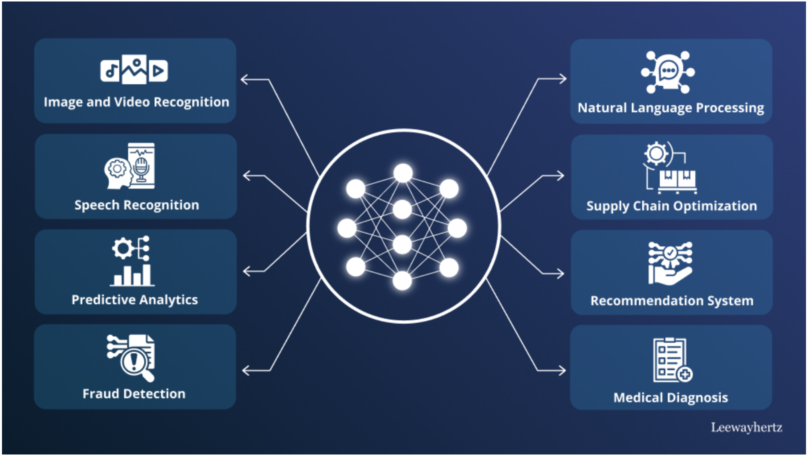
based on Phys. Rev. D **112**, L031902 [arXiv:2503.08827]
with Donghee Lee and Hye-Sung Lee

Overview

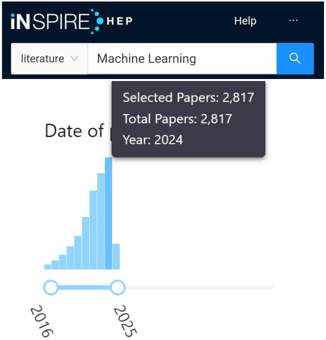
1. Introduction
2. Machine Learning 101
3. Synaptic Field Theory
4. Realization
5. Summary

I. Introduction

Machine Learning



Machine Learning and High Energy Physics



- Machine learning is also a topic of great interest in high energy physics.
  - Parton Distribution Function
  - Jet Classification
  - Constraining Effective Field Theories
  - Anomaly Detections
  - ...

Machine Learning is Still a Mystery

We have a rough idea of what it's doing,
but when it gets complicated,
we don't know what's going on,
similar to our understanding of the brain.

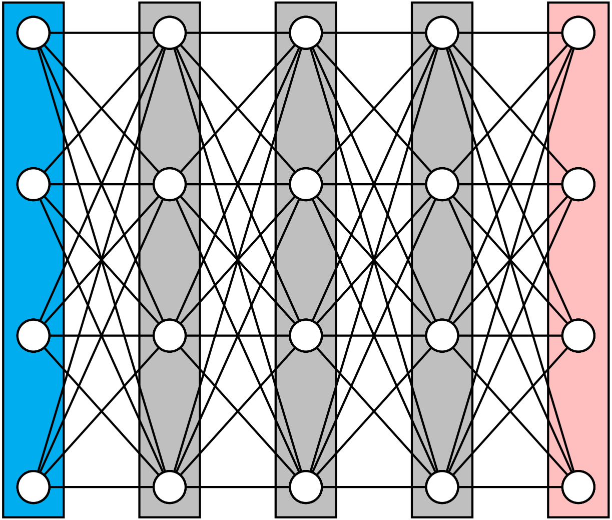
Geoffrey Hinton
(2024 Nobel Laureate in Physics)

Technology and Physics

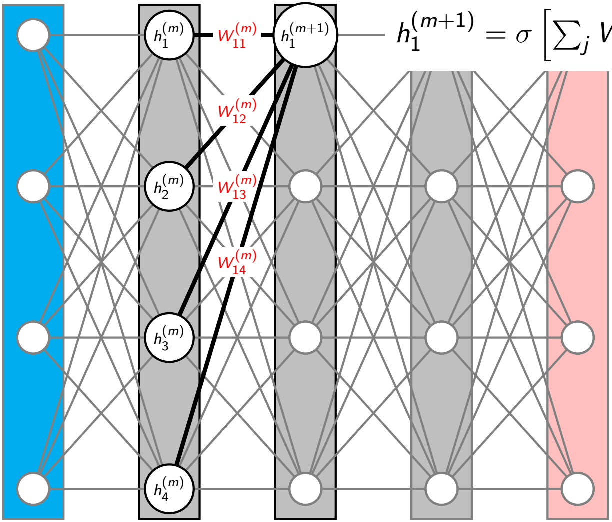
- Technological development sometimes comes before full theoretical understanding. Once the physics is clear, progress tends to accelerate.
  - Steam Engine & Thermodynamics:
    Invention (18C) $\rightarrow$ Thermodynamics (19C)
    $\Rightarrow$ Steam locomotive and First industrial revolution.
  - Electromagnetic Phenomena & Maxwell's Theory:
    Static Electricity, Compass (Ancient) $\rightarrow$ Maxwell's Theory (19C)
    $\Rightarrow$ Powerplant, Telephone and Second Industrial Revolution.
  - Transistor & Semiconductor Physics:
    Invention (1947) $\rightarrow$ Semiconductor Theory (1950s)
    $\Rightarrow$ Computer, Internet and Third Industrial Revolution.
- Understanding the physics behind machine learning could drive its future progress.

II. Machine Learning 101

Neural Networks

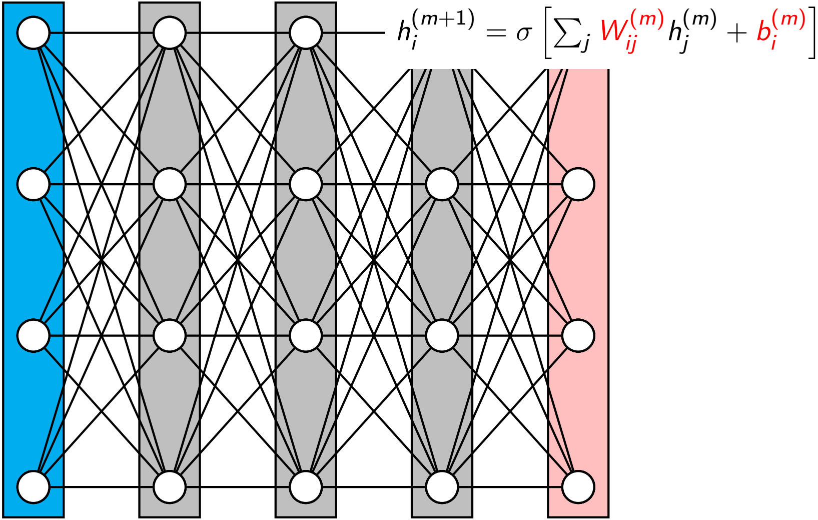


Neural Networks



$$h_1^{(m+1)} = \sigma\left[\sum_j W_{1j}^{(m)} h_j^{(m)} + b_1^{(m)}\right]$$

Neural Networks



Universal Approximation Theorem

- For any arbitrary continuous function, there exists a set of synaptic weights such that a neural network can approximate it.
  - Infinite width cases: proved

> **Universal approximation theorem**—Let $C(X, \mathbb{R}^m)$ denote the set of continuous functions from a subset X of a Euclidean $\mathbb{R}^n$ space to a Euclidean space $\mathbb{R}^m$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$. Note that $(\sigma \circ x)_i = \sigma(x_i)$, so $\sigma \circ x$ denotes σ applied to each component of x.
>
> Then σ is not polynomial if and only if for every $n \in \mathbb{N}$, $m \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K, \mathbb{R}^m), \varepsilon > 0$ there exist $k \in \mathbb{N}$, $A \in \mathbb{R}^{k\times n}$, $b \in \mathbb{R}^k$, $C \in \mathbb{R}^{m\times k}$ such that
>
> $$\sup_{x\in K} \|f(x) - g(x)\| < \varepsilon$$
>
> where $g(x) = C \cdot (\sigma \circ (A \cdot x + b))$

  - Infinite depth or bounded depth and width cases: partially proved
- The universal approximation theorem guarantees the existence of a solution but it does not provide a method for finding the solution.
  - "We are not guaranteed, however, that the training algorithm will be able to learn that function."

[Goodfellow, I., Bengio, Y., & Courville, A. (2018). Deep learning. MITP.]

Gradient Descent

- Prepare the training set $(X_i^{[l]}, Y_i^{[l]})$ and the define the cost function:

$$C = \sum_{i,l} (Y_i^{[l]} - Z_i^{[l]})^2$$

where $Z_i^{[l]}$ is the result of the neural network for $X_i^{[l]}$.

- Update the synaptic weights and biases using gradient descent:

$$\Delta W_{ij}^{(m)} = -\eta \frac{\partial C}{\partial W_{ij}^{(m)}}, \qquad \Delta b_i^{(m)} = -\eta \frac{\partial C}{\partial b_i^{(m)}},$$

with the step size η.

Issues on Gradient Descent

- It is still unknown whether training algorithms actually find the solutions guaranteed by the universal approximation theorem.

- Training Dataset -

$$7 + 2 = 9$$
$$5 + 3 = 8$$
$$4 + 2 = 6$$
$$3 + 1 = 4$$

- Test Artificial intelligence -

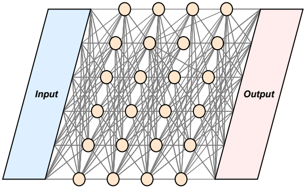
$$5 + 4 = ?$$

- Almost all training algorithms are based on gradient descent.
  - Nearly all of deep learning is powered by one very important algorithm: stochastic gradient descent. Stochastic gradient descent is an extension of the gradient descent algorithm.

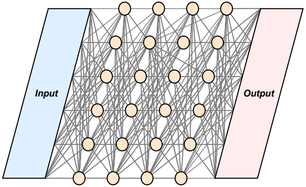
[Goodfellow, I., Bengio, Y., & Courville, A. (2018). Deep learning. MITP.]

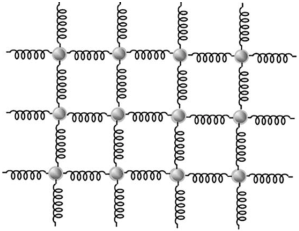
III. Synaptic Field Theory

Motivation

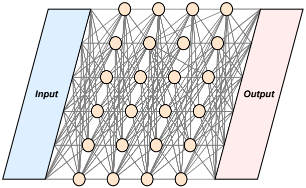


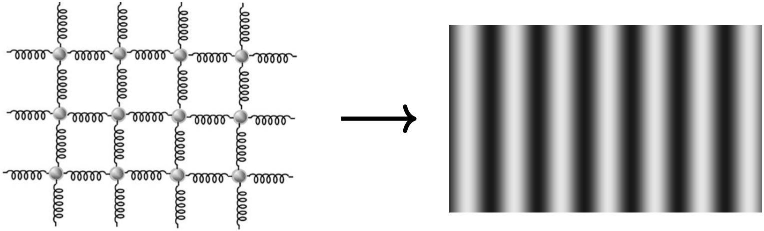
Motivation



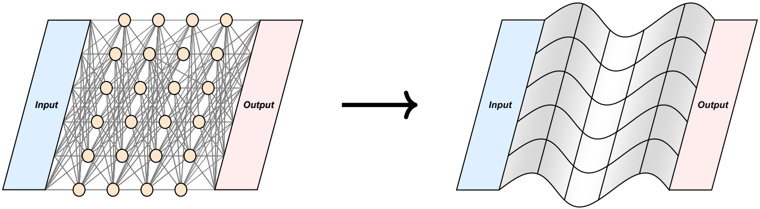


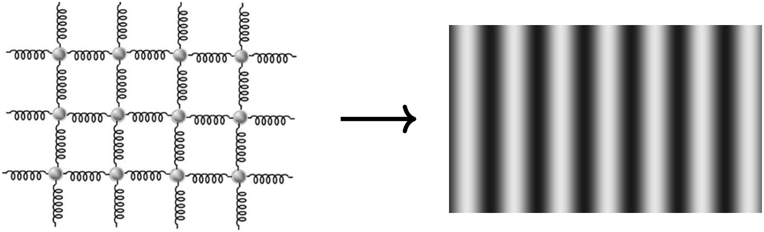
Motivation





Motivation





Lagrangian Approach to Gradient Descent

- In the continuum limit, the equation for gradient descent becomes

$$\dot{W} = -\eta \frac{\partial C}{\partial W}.$$

- It can be considered as the high-viscosity limit ($\gamma = \eta^{-1} \gg 1$) of

$$\ddot{W} + \gamma \dot{W} + \frac{\partial C}{\partial W} = 0.$$

- This equation can be derived from the action given as

$$S = \int dt\ e^{\gamma t} \left[\frac{1}{2} \dot{W}^2 - C \right] = \int dt\ \sqrt{-g} L_W$$

with $\sqrt{-g} = e^{\gamma t}$ and $L_W = \frac{1}{2} \dot{W}^2 - C$.

Gradient Descent as the de Sitter Dynamics

- Assume that L_W admits a continuum limit, meaning it can be expressed as an integral of a Lagrangian density composed of fields:

$$L_W = \int d^d\mathbf{x}\ \mathcal{L}_w[w(t,\mathbf{x})].$$

- The action has the form of the action of fields in the curved spacetime:

$$S = \int d^{d+1}x\ \sqrt{-g}\mathcal{L}_w.$$

- In particular, $\sqrt{-g} = e^{\gamma t}$ matches that of a universe dominated by a positive cosmological constant, a typical example of de Sitter space.

Synaptic Field Theory

- L_W includes a sum over the indices of synaptic weights and biases.
- By taking the continuum limit of this summation to a spatial integral, we can develop a field theory in de Sitter spacetime.
- The training dataset behaves as the external sources $J(\mathbf{x})$, $K(\mathbf{x})$ in the synaptic field theory.
- The resulting synaptic field theory would be a familiar framework to those who study high energy physics or cosmology.

IV. Realization

Nonlocality of Neural Networks

- Series expansion of cost function:

$$C = \sum J_{1\ i_1 j_1}^{(m_1)} W_{i_1 j_1}^{(m_1)} + \sum J_{2\ i_1 j_1 i_2 j_2}^{(m_1 m_2)} W_{i_1 j_1}^{(m_1)} W_{i_2 j_2}^{(m_2)} + \cdots .$$

  The coefficients $J_{1\ i_1 j_1}^{(m_1)}$ and $J_{2\ i_1 j_1 i_2 j_2}^{(m_1 m_2)}$ depend on the data set.
- Note that there are terms involving different indices.
- Taking the continuum limit,

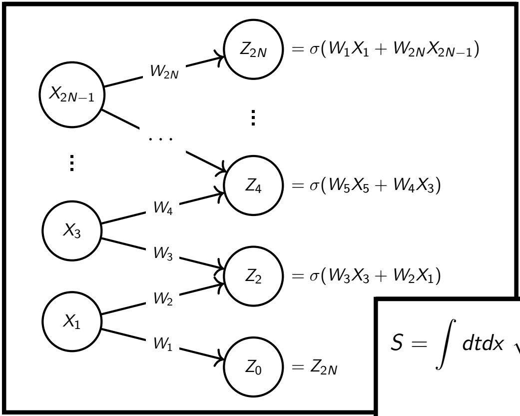
$$L \supset \int d^3\mathbf{x}\ J_1(\mathbf{x}) w(t, \mathbf{x}) + \int d^3\mathbf{x} d^3\mathbf{y}\ J_2(\mathbf{x}, \mathbf{y}) w(t, \mathbf{x}) w(t, \mathbf{y}) + \cdots .$$

  Here, J_1 and J_2 act as external sources given by the training examples.
- Since the second term involves two spatial coordinates $\mathbf{x}$ and $\mathbf{y}$, this Lagrangian is not local.

Spacetime Geometry and Neural Network Architecture

- This naive approach expects the nonlocal Lagrangian.
- The locality is related to the spacetime geometry.
- In the synaptic field theory, the spacetime is given as the continuum limit of the indices of parameters.
- The spatial geometry depends on how to construct the architecture and how to index the parameters.
- We may construct a neural network possessing locality.

Toy Neural Network



$$S = \int dtdx \, \sqrt{-g} \left[\frac{1}{2} [\partial_t w(t,x)]^2 - \frac{1}{2} K(x) [\partial_x w(t,x)]^2 - \frac{1}{2} J(x) w(t,x)^2 \right]$$

Discussions on Toy Neural Network

- These examples may be too simple to behave as a practical artificial intelligence.
- However, it is an interesting example that shows the locality.
- This locality comes from the architecture of the neural network and the indexing convention.
- One may attempt to build practical neural networks whose architecture and indexing convention enable locality to emerge.
- Once such examples are established, we can study them by using the various tools in field theory.

V. Summary

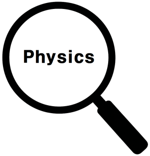
Summary Table

- The synaptic field theory suggests a friendly framework for physicists to study machine learning.

Neural Network	**Synaptic Field Theory**
Parameters $W_{ij}^{(m)}$	Field $w(t, x)$
Training examples (X, Y)	External sources $K, J, \cdots$
Indices i, j, m	Space $\mathbf{x}$
Training step T	Time t
Cost function C	Lagrangian L
Step size η	Hubble parameter H

Take-home Message

- Understanding the nature of deep learning is the mission of physicists so more physics is warranted.



Thank you for listening

Back-up Slides

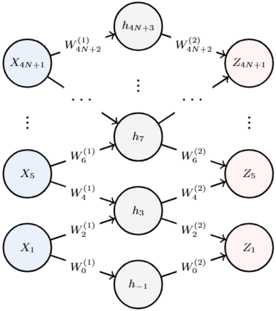
Discussions

- The linear activation is used and results in the bilinear action.
  - If we use non-polynomial activations, then higher order interaction terms should be considered.
- The previous example is not Lorentz invariant.
  - With a time-dependent training algorithm such as stochastic gradient and carefully designed training datasets, a Lorentz-invariant theory may emerge.
- The previous example results in the local action due to the simple structure and the practical indexing convention.
  - For the finite depth neural networks, the naive approach of continuum limit will result in nonlocal terms like

$$\int d^d\mathbf{x}d^d\mathbf{y}\ A(\mathbf{x},\mathbf{y})w(t,\mathbf{x})w(t,\mathbf{y})$$

  - As shown in the example, the structure and the indexing convention will be related to the geometry and locality of the space of the theory.

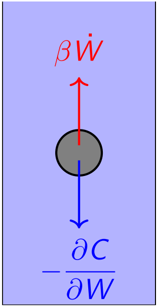
Toy Neural Network



$$
\begin{aligned}
S = \int dtdx\, \sqrt{-g}\Big[&\frac{1}{2}(\partial_t w_1)^2 + \frac{1}{2}(\partial_t w_2)^2 - \frac{1}{2}m^2 w_2^2 \\
&-J_1 - J_2 w_2 - J_3 w_2 w_1 - K_1 w_2 \partial_x w_1 \\
&-K_2 w_2 \partial_x^2 w_2 - K_3 w_2 \partial_x^2 w_1 - K_4 \partial_x w_2 \partial_x w_1 \\
&-K_5 w_1 \partial_x w_2 - K_6 \partial_x^2 w_2 - K_7 w_1 \partial_x^2 w_2 + \cdots\Big].
\end{aligned}
$$

$m^2 = 8a^{-1}N_l r^2$	
$J_1(x) = a^{-1}\sum_l (Y^{[l]})^2$	$J_2(x) = 4ra^{-1}\sum_l Y^{[l]}$
$J_3(x) = 8qa^{-1}\sum_l (X^{[l]} + 4a^2\partial_x^2 X^{[l]})Y^{[l]}$	
$K_1(x) = 48aq\sum_l Y^{[l]}\partial_x X^{[l]}$	$K_2 = 4aN_l r^2$
$K_3(x) = 20qa\sum_l X^{[l]}Y^{[l]}$	$K_4(x) = 16aq\sum_l X^{[l]}Y^{[l]}$
$K_5(x) = 16aq\sum_l Y^{[l]}\partial_x X^{[l]}$	$K_6(x) = 2ra\sum_l Y^{[l]}$
$K_7(x) = 4qa\sum_l X^{[l]}Y^{[l]}$	

High-viscosity Limit



- High Viscosity Medium
- Large Drag Force
- Terminal Velocity
- $\ddot{W} = 0$

High Viscosity Limit

- In the high-viscosity limit, $\eta = 1/\gamma$ is small, allowing a perturbative expansion:
$$W = \mathbb{W}^{(0)} + \eta \mathbb{W}^{(1)} + \mathcal{O}(\eta^2).$$
- The equation from the action becomes
$$\frac{1}{\eta}\dot{\mathbb{W}}^{(0)} + \left(\ddot{\mathbb{W}}^{(0)} + \dot{\mathbb{W}}^{(1)} + \left.\frac{\partial C}{\partial W}\right|_{W=\mathbb{W}^{(0)}}\right) + \mathcal{O}(\eta) = 0.$$
- At $\mathcal{O}(\eta^{-1})$, we find $\dot{\mathbb{W}}^{(0)} = 0$ at any t, which implies $\ddot{\mathbb{W}}^{(0)} = 0$. Therefore, at $\mathcal{O}(\eta^0)$, we have
$$\dot{\mathbb{W}}^{(1)} + \left.\frac{\partial C}{\partial W}\right|_{W=\mathbb{W}^{(0)}} = 0.$$

It is the equation of motion for the training of neural networks.

Field Theoretic Approach to Neural Networks

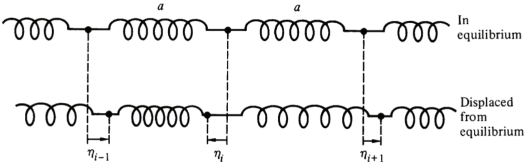
- There exist some previous works trying to apply field theory to neural networks.
- Krippendorf and Spannowsky attempted to develop an effective theory of outputs of neural network and proposed a relationship between neural networks and cosmology.

  [S. Krippendorf and M. Spannowsky Mach.Learn.Sci.Tech. 3 (2022) 3, 035011]
- To do so, they considered the limit where the effect from synaptic weights and biases becomes a constant.
- Since weights and biases are fundamental building blocks, their effects should not be neglected.
- The theory dealing with fields developed by the continuum limit of weights and biases is worth studying.

Continuum Limit

- Here is a typical example of taking continuum limit.

[H. Goldstein, C. Poole, J. Safko (2002). Classical Mechanics, Pearson.]



$$L = \frac{1}{2}\sum_i [m\dot{\eta}_i^2 - k(\eta_{i+1} - \eta_i)^2] \quad \Rightarrow \quad L = \frac{1}{2}\int \left[\mu\dot{\eta}^2 - Y\left(\frac{d\eta}{dx}\right)^2\right] dx$$

- This example gives a local Lagrangian because every term involves only variables with the same index.

Future Directions

- The stochasticity is an important component to train neural network.
- It is implemented by the time dependent training algorithm such as the stochastic gradient descent.
- This time depedence would be interpreted as the time dependent sources.
- Investigating this possibility would further enrich the study of synaptic field theory.