Visual Space, Neural Networks, and AI Algorithm Models

Taiwei Song

Shanghai Luyi New Energy Co., Ltd., Shanghai Riyue New Energy Co., Ltd.

January 6, 2026

This paper briefly discusses the concept of visual space discovered by the author [1-4], its transformation equation with the natural space-time, and points out that this transformation relationship is the key algorithm for AI embodied agents to automatically recognize the surrounding "world". It also briefly demonstrates that neural networks inherently possess the properties of "iterative convergence" and "self-learning evolution", and the "emergence of intelligence" in large AI models based on neural networks is inevitable.

1. Visual Space of the Natural World [1,2]

The coordinate form of the quasi-4D natural space-time is (x, y, z, ict) (abbreviated as (r, ict)). In the author's "Geometry of Spacetime Structures" [1-4], the distance differential dS of the natural space-time is defined as:

$$dS^2 = n_x^2 dx^2 + n_y^2 dy^2 + n_z^2 dz^2 - c^2 dt^2 = dr^2 - c^2 dt^2 \qquad (1)$$

where n(x,y,z) is the generalized refractive index tensor of the spatial medium. When dt=0, dS is the optical path differential of the isochronous 3D natural space.

For any observer, the scene of the natural world is the observed image, which is a scaled-down or enlarged detected image of nature. The image seen or detected by the observer is denoted as P′, and the actual natural scene is denoted as P. P→P′ is a one-to-one space-time transformation. The set of images P′ seen by the observer constitutes the visual space of the natural world. Assuming the coordinates of an arbitrary object point in the 3D natural space (referred to as "P-space") are (x, y, z), and its coordinates in the visual space (referred to as "P′-space") are (x′, y′, z′), the transformation (x, y, z) → (x′, y′, z′) is a one-to-one light induction transformation. The P′ coordinate system remains orthogonal, but the unit length shrinks as the coordinate values increase; only the distance differential dS′ within a tiny space still follows a vector operation logic similar to that of dS.

In general, the transformation T from the distance differential dS or (dx, dy, dz, icdt) to dS′ or (dx′, dy′, dz′, icdt′) is represented by a 4×4 matrix $T_{4\times4}$. Since the visual image P′ of the observer at any moment is simultaneous (dt′=0), the space-time transformation P→P′ is essentially a transformation from the quasi-4D space-time (x, y, z, ict) to the 3D space (x′, y′, z′), and the transformation matrix is simplified to a 3×4 matrix. The 3D visual image presents a dynamic natural scene. For any point (x, y, z) in the P-space at a distance s from the observer, the corresponding time is t = 0-s/c=-s/c, and its 4D coordinates are (x, y, z, -is). Therefore:

$$\begin{pmatrix} dx' \\ dy' \\ dz' \\ icdt' \end{pmatrix} = \begin{pmatrix} T_{11} & 0 & 0 & T_{14} \\ 0 & T_{22} & 0 & T_{24} \\ 0 & 0 & T_{33} & T_{34} \\ 0 & 0 & 0 & T_{44} \end{pmatrix} \begin{pmatrix} dx \\ dy \\ dz \\ icdt \end{pmatrix} = \begin{pmatrix} T_{11} & 0 & 0 & T_{14} \\ 0 & T_{22} & 0 & T_{24} \\ 0 & 0 & T_{33} & T_{34} \end{pmatrix} \begin{pmatrix} dx \\ dy \\ dz \\ -ids \end{pmatrix} \qquad (2)$$

Here, the zero matrix elements are determined by the mutual relationships of the orthogonal coordinate components corresponding to the space-time transformation; T_{i4} is proportional to $\partial x_i'/\partial t$, and $T_{44}=0$.

According to the Geometry of Space-time Structures, the squared geometric distance differential dS^2 of the quasi-4D natural space-time space characterizes material energy. With light as the propagation medium, the corresponding quantity observed by the observer satisfies $dS'^2 \propto dS^2/S^2$, i.e., $dS'^2 = \frac{\lambda^2}{s^2} dS^2$, where S is the optical path and λ is a constant. Using formula (1), setting $dt' = 0$, the fixed time differential $-ids = 0$, and letting $\lambda=1$, equation (2) becomes:

$$\begin{pmatrix} dx' \\ dy' \\ dz' \end{pmatrix} = \frac{1}{S} \begin{pmatrix} n_x & 0 & 0 & T_{14} \\ 0 & n_y & 0 & T_{24} \\ 0 & 0 & n_z & T_{34} \end{pmatrix} \begin{pmatrix} dx \\ dy \\ dz \\ 0 \end{pmatrix} \quad (3)$$

Equation (3) is the space-time transformation differential equation from the real scene of the natural world to the image in the visual space, which can be applied to AI 3D dynamic recognition algorithms such as autonomous driving. Through equation (3), we can establish the corresponding relationship set of (x, y, z) → (x′, y′, z′), and endow the visual space with information meaning, i.e., associating the (x′, y′, z′) coordinates with visual sensory information. We can use AI visual embodiment to conduct a large number of practical scenario training and learning, obtain relatively accurate and complete visual matrix sets at different times and with different object states, as well as the direct correspondence relationships between visual images and real scenes, so as to achieve real-time recognition of surrounding and distant scenes. This is the ultimate natural recognition algorithm for autonomous driving.

2. Neural Networks

The basic structure of a neural network is shown in Fig. 1 (Michael Nielsen [5]). It is a feedforward neural network where signals are only transmitted forward. In applications, the basic form of the relationship (input-output) between network units (a_{ij}, b_{ij}) of adjacent layers [5,6,8] is: $z_{ij} = \sum_k w_{ijk} a_{i-1,k} + b_{ij}$, $a_{ij} = \sigma(z_{ij})$, where σ is the activation function (only taking positive values, setting negative values to 0), b_{ij} is the bias (simulating the excitation threshold of neurons), i is the layer number, j denotes the sequence number of elements within the same layer, and w_{ijk} is the front-feed weight matrix element between adjacent layers; By continuously learning and iteratively approximating, determine w_{ijk} and b_{ij}. This basically reflects the synaptic connection and signal transmission properties of numerous neural units in the brain's nervous system. The brain's nervous system is composed of 86 billion neurons organized in an orderly manner according to functional modules. Each neuron has one axon and hundreds or even more dendrites. Nerve cells form a complex and orderly three-dimensional network connection structure in 3D space through directional synaptic connections. Signal transmission between neurons of the same type is unidirectional, i.e., from the axon of the previous neuron to the dendrite of the next neuron; numerous adjacent synaptic connections of neurons can also form closed loops, and synaptic connection loops between neurons of different types and layers have a feedback effect.

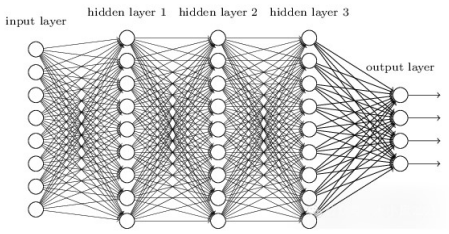

Fig. 1 LLM neural network, Michael Nielsen, Neural Networks and Deep Learning[5]

Consciousness is an immediate set of information generated by the brain's nervous system that can be perceived by an individual's physical and mental state. It is a meaningful knowledge information flow synthesized and processed by the brain's nervous system, associated with the individual's entire body and the external world, and is a non-embodied manifestation of a living organism. Life is an irreversible self-organization process driven by time, and autonomous learning and progress are one of the essential attributes of life. The brain's nervous system fully reflects this natural attribute; through training and learning, the brain's nerves form a definite cognition of objective things, and naturally "output" the corresponding cognitive consciousness when stimulated or "questioned". Therefore, intelligent agents based on neural networks (including models, embodiments, and even chips) will naturally exhibit human-like intelligence through learning. With the aforementioned AI neural network algorithm and appropriate iterative optimization constraints, the results will always "converge". If the model parameters and training volume are sufficiently large, "emergence" is inevitable [6,7].

3. Large Models and Embodied Models of AI

Large AI models mainly include Large Language Models (LLMs) and World Models (WWs) [8]. Large language models mainly adopt the Transformer architecture [6], whose key innovation is the introduction of the attention function [6] ($\mathrm{Attention}(Q,K,V)= \mathrm{softmax}(QK^T/\sqrt{d_k})V$) to solve the problem of long-range correlation of token vectors. In fact, this is the dot product relationship of vectors ($A \cdot B$), which is exactly the definition of the correlation degree of space units [3], and this is the reason why the Transformer architecture is highly applicable to large models.

The introduction of token long-range correlation by Transformers significantly improves the efficiency of accurate generation of long sentences. However, the global value correlation algorithm of Transformers not only limits the possibility of large language models (LLMs) becoming world models (WWs) but also gives them the disadvantage of wasting a lot of computing power and being dependent on it.The current development of the AI technology industry is characterized by continuously expanding demand for computing power and significant resource wastage due to global competition in establishing computing centers—this is precisely the root cause. Humanity is still far from learning the wisdom of nature.

Within information space, including consciousness space, information content units have timeliness and relative independence, and only space-time adjacent correlations may exist. Space-time causal relationships can be directly identified through comparison of adjacent states, which is actually the "world law" that machines need to recognize, and can be fully realized through iterative learning. This is the logical basis of world models.

Finding internal laws from the colorful world and artificially generating a "real" world from scratch are two completely different logical processes. The former is from concrete to abstract and from complex to simple, while the latter is from concrete to concrete and from simple to complex. The latter does not require advanced abstract thinking or profound mathematical and physical knowledge, but only needs to continuously do it simply according to personal experience and imitating real scenes. This is the real reason why a three-year-old child is smarter (more human-like) than a humanoid robot with super computing power and massive knowledge. For a humanoid robot to do housework, it should move as simply as a human being, and there is no need to know "Newtonian mechanics" or "vector analysis", nor why the world is the way it is. This is a misunderstanding of Artificial General Intelligence (AGI).

A real scene is a set of images of various objects changing, growing, and arranging in time. As long as we know the short-range relationships of various objects in the time and space dimensions (physically, it is the rate of change; visually, it is a very intuitive amount of change, which can be realized through simple and direct algorithm iteration), it can also be "generated" step by step.

Embodied intelligence mainly involves software and hardware issues of programs such as recognition, decision-making, action, and adjustment. The difficulty of humanoid robots in home environments lies in how to move as simply as humans. It does not need to understand the internal laws of related things, but only needs to learn to recognize and act like a child. In addition to hardware, the software must be dynamic like the human brain, with decision-making changes closely following space-time changes. The learning algorithm itself and its iteration speed must be premised on adapting to the training of a large number of application scenarios, rather than being developed in isolation.

The machine actions of autonomous driving basically only include standardized actions within the vehicle, such as braking, accelerating, and steering. The technical difficulty lies in accurate and rapid recognition and decision-making. Its application also faces the problem of public acceptance, after all, car accidents are major issues related to life. The visual space logic of the Geometry of Space-time Structures is a direct sensory transformation relationship of humans over time, and is a natural training and iterative algorithm.

Acknowledgments

The author would like to thank Mr. Song Yecheng for his strong support! Mr. Song Yecheng often communicates and discusses issues related to AI and the brain with the author, which has been very helpful in broadening the author's research ideas on related issues.

References

- E-mail: caisheng99@sina.com

[1] Taiwei Song, Physical Foundations and Mathematical Logic of the Natural World, https://www.vixra.org/abs/2512.0009, 2024.

[2] Taiwei Song, Space Warp, Space-Time Transformation, and Cosmic Redshift: The Application of the Geometry of Space-Time Structures on Large Spatial Scales, 2024. http://viXra.org/abs/2510.0087

[3] Taiwei Song, Strong Correlation Function between Particles in the Low Dimension Structures , http://www.luyipower.com/Business.aspx?id=43, 2020.

[4] Taiwei Song, Basic Logic of Brain Nerves, Consciousness, and Artificial General Intelligence , Advances in Neurology and Neuroscience, Vol 9(1), 01-08, 2026.

[5] Michael Nielsen, Neural Networks and Deep Learning, 2019-2023. http://neuralnetworksanddeeplearning.com/index.html

[6] Ashish Vaswani, et al., Attention Is All You Need, arXiv:1706.03762v7, 2023. https://arxiv.org/abs/1706.03762

[7] Tongtong Feng, Xin Wang, et al., Embodied AI: From LLMs to World Models, 2025. https://arxiv.org/pdf/2509.20021

[8] David Ha, Jürgen Schmidhuber, World Models, 2018, https://arxiv.org/abs/1803.10122

[9] Jelassi, S., Li, Y., et al., A Mathematical Perspective on Transformers, 2024. https://doi.org/10.1002/cpa.22192