Essay

Neural Networks through the Lens of Evolutionary Dynamics

Dan C. Baciu *

Posted Date: 8 April 2024

doi: 10.20944/preprints202404.0506.v1

Keywords: Evolutionary dynamics; Neural Networks; Self-attention ; Variation-selection; Frequency Dependent Selection

 doi:10.20944/preprints202404.0506.v1

Essay

Neural Networks through the Lens of Evolutionary Dynamics

Dan C. Baciu

baciu@ucsb.edu; https://orcid.org/0000-0002-0043-5616

Abstract: In this essay, I revisit Neural Networks (NNs) through the lens of evolutionary dynamics. From the two most important features of NNs, I recover the two most general equations of evolutionary dynamics. These two equations may thus serve as a connection between NNs and a body of knowledge that has been built over multiple centuries and has found application across a vast range of disciplines. This connection may therefore also help explain why NNs with the two features in question are applicable across a much broader range of domains than initially envisioned.

Keywords: evolutionary dynamics; Neural Networks; self-attention; variation-selection; frequency dependent selection

---

Introduction

Neural Networks engineers have followed an empirical approach. They have been creative in devising new architectures and selecting the most powerful ones through empirical testing. High-impact journals in the field of machine learning and artificial intelligence, such as "Nature Machine Intelligence," "IEEE Transactions on Neural Networks and Learning Systems," "Neural Networks," and "Journal of Machine Learning Research," often publish research that achieves a new state-of-the-art. On this basis, multiple highly significant breakthroughs have succeeded in providing increasingly powerful technologies. The present article highlights two generations of NNs, specifically: the initial generation and Transformers.

The first generation of NNs was greatly developed in the second part of the 20th century. This first generation became particularly powerful when Recurrent Neural Networks (RNNs) were introduced (Hopfield, 1982, Elman, 1990, Rumelhart, Hinton, & Williams, 1986, Hochreiter & Schmidhuber, 1997, Jordan, 1986). Transformers, on theater hand are primarily a product of the 21st century. This second generation of networks stands out through the development of self-attention layers (Bahdanau, Cho, & Bengio, 2014, Xu et al., 2015). Many researchers believe that an important stage was reached with the article "Attention is all you need" (Vaswani et al., 2017).

The present essay revisits RNNs and Self-attention through the lens of evolutionary dynamics. From the mathematical description of RNNs, I recover the quasispecies equation, which is a general equation of evolution, used to simulate variation-selection processes. This equation has been shown to describe evolution and creativity across a broad range of disciplines (Domingo & Schuster, 2016, Singh et al., 2023, Baciu 2023). Furthermore, from Self-attention layers, I recover the replicator equation, also known as generalized Lotka-Volterra equation, which is the general equation of evolutionary game theory (Hofbauer & Sigmund, 1998). This equation is used to describe frequency-dependent selection, which, in evolutionary dynamics, is the counterpart of variation-selection processes (Nowak 2006). The replicator equation, specifically builds on more than a century of science. It has been used to describe processes of play and diversification across a broad range of disciplines (Lotka 1910, Ross 1911, Baciu 2023). Thus, the article recovers the equation of variation-selection processes from RNNs and the equation of frequency dependent selection from Self-attention layers. This sets the stage for connecting NNs with a body of knowledge that has been developed over the centuries and applied to a vast array of disciplines.

 doi:10.20944/preprints202404.0506.v1

The essay ends with a question that I would like to open. Describing cultural change with the equations just mentioned, I have observed a process that the equations can describe very well (Baciu 2018). However, I am unaware whether this particular behavior is also found in the behavior or NNs.

1. NNs, RNNs, and Variation Selection Processes

The architecture of Neural Networks envisions layers of neurons which are connected such that each neuron in a layer y is connected with all neurons in the previous layer x. This architecture can be mathematically described as

$$y_j = \sum_{i=1}^{n} w_{ij} x_i + b_i.$$

where y_j is a neuron in the new layer y,while x_i is a neuron in the previous layer x, which consists of n neurons. The coefficient w_{ij} is the weight of the connection from x_i to y_j. The intercept b_i allows for more adjustments to the model and is often referred to as bias in the literature (Goodfellow, Bengio, & Courville, 2016).

Considering that the Neural Network is recurrent, the outputs that come out of the model become new inputs. This means that the values of the neurons on the layer y are re-inserted into the neurons on the layer x. Thus, we can denote layer x as x_t and layer y as x_{t+1}. For simplicity of notation and generality of the analysis, I use the notation for differential equations instead, setting y as $\dot{x}$. The equation then becomes

$$\dot{x}_i = \sum_{i=1}^{n} w_{ij} x_i + b_i.$$

where the layers x and y have each n neurons, and the layer y, now denoted as $\dot{x}$, is fed back into the equation, to account for the recurrent feature of the neural network.

From this new equation, it is evident that the basic equation of variation-selection processes can be recovered easily. Variation selection processed can be formulated with the quasispecies equation, in its simplest form written as

$$\dot{x}_i = \sum_{i=1}^{n} w_{ij} x_i.$$

In the standard notation of the quasispecies equation, the weights here denoted as w are called mutation rates and denoted as q (Nowak 2006). Mathematically, they perform the same role, being coefficients in both cases. The mathematical function of the "weights" or "mutation rates", however one wishes to call them, is to specify how often x_i is transformed into y_j. Technically speaking, the variable specifies how much creativity there is in the system, transforming one variable into another.

The quasispecies equation is often written with an additional nonlinear term. With this term the equation is written as

$$\dot{x}_i = \sum_{i=1}^{n} w_{ij} x_i - \phi(\boldsymbol{x}).$$

Where $\boldsymbol{x}$ is the vector containing all values of x_i and the Greek letter Phi denotes a function of $\boldsymbol{x}$. The role of $\phi(\boldsymbol{x})$ is to avoid exponential growth if the largest eigenvalue is greater than one, which is the most common outcome (Nowak 2006). The history of this term of the equation goes back to Quetelet and Verhulst, who adjusted the Malthusian population growth model as follows

$$\dot{x} = wx - \phi(x).$$

The Malthusian population growth model is a linear model of exponential population growth. Its history goes back to antiquity and to compounding interest. In the equation, $\dot{x}$ is the value of the variable in the passage of time, w is the rate of change, x is the value of the variable before the change, and the component $\phi(\boldsymbol{x})$ performs the same function as later in the quasispecies equation. It is used

 doi:10.20944/preprints202404.0506.v1

to stop exponential growth (Verhulst, 1838, Quetelet, 1842). Specifically, Verhulst used empirical populations growth data, which allowed him to set $\phi(x)$ as wx^2 as follows:

$$\dot{x} = w_1 x - w_2 x^2.$$

The resulting function is a sigmoid growth function, also known as s-curve (Bejan & Lorente, 2012). The function grows exponentially in the beginning, but w_2x^2 eventually grows much faster than w_1x. Therefore, the growth is curbed a typical s-shape emerges.

This type of nonlinear behavior that came into the quasispecies equation through population studies has a matching counterpart in NNs. S-curves are a common activation function for the neuron layers of NNs (LeCun, Bottou, Bengio, & Haffner, 1998, Goodfellow, Bengio, & Courville, 2016). Thus, it can be concluded that even the additional nonlinear term $\phi(x)$ in the quasispecies equation has matching counterparts in NNs.

2. Self-Attention Layers and Frequency Dependent Selection

The architecture of Neural Networks envisions an additional way to connect neurons. In "Self-attention layers", the neurons in each layer x are directly connected not only to neurons on other layers, but also among themselves on the same layer. Many engineers believe that this type of architecture has led to a significant breakthrough that has advanced NNs from a rather obscure technology to a technology that spearheaded the explosion of Artificial Intelligence applications today (Uszkoreit, 2017). The architecture of these layers can be mathematically described as

$$y_i = x_i f_i(\boldsymbol{x}).$$

where x_i is a neuron on layer x, while $f(\boldsymbol{x})$ is a function of all neurons on layer x. The layer y follows after x with the same number of neurons. While the mathematical description given here is not the one usually used in textbooks, it describes what is required from all variants of Self-attention. There is a multiplication between x_i and a value obtained as a function of all variables in its own layer, which motivates the term "self" in Self-attention (Vaswani et al., 2017). Turning this architecture in a recurrent architecture, I set y as $\dot{x}$, as in the previous section. The following equation is obtained:

$$\dot{x}_i = x_i f_i(\boldsymbol{x}).$$

where $\dot{x}_i$ denotes the values as they change, passing through the system.

Evidently, this equation is the same as generalized Lotka-Volterra equations, also known as replicator equation (Hofbauer & Sigmund, 1998). Thus, it is demonstrated that self-attention layers can be described as frequency dependent selection. Admittedly, frequency dependent selection can allow for somewhat more flexibility because the function $f(\boldsymbol{x})$ can be any function, but the main idea of a multiplication between a variable and its own environment as part of the evolution of the dynamic system remains unchanged.

3. Discussion: In Search of an All-Encompassing Description of the World

The two equations that this article has recovered from NNs have a history spanning centuries. As part of this history, they have found application in a truly breathtaking array of sciences (Baciu 2023).

Let me begin with variation-selection processes. The history of their discover stretches back to antiquity. To some extent, many early cultures may have known that constant growth rates lead to exponential growth. In addition, variation-selection processes, specifically, have been described by ancient philosophers such as Lucretius (Lucretius, 1st century BCE, Greenblatt, 2011).

In modern times, a fascinating, sparkling description of variation-selection processes can be found the work of Alfred Russel Wallace, specifically the article that, according to some historians, he has written after a life-threatening fever, before sending it to Darwin, who used it to write his own essay about natural selection, interpreting it as a variation-selection process (Wallace, 1858, Darwin

 doi:10.20944/preprints202404.0506.v1

& Wallace, 1858). Darwin and Wallace's work has inspired countless scientists, although it did not yet provide a mathematical description of variation-selection processes.

The mathematical description of variation-selection processes came somewhat later. It was advanced by Eigen and Schuster, and it has led to new perspectives in multiple fields, but in particular in virus dynamics and vaccines (Singh et al. 2023, Domingo & Schuster, 2016, Nowak, 2006, Nowak & May, 2000).

Building on this body of knowledge as well as additional experiments in the humanities, I have argued that variation-selection processes can be used to describe creativity in any kind of physical, biological, social, or cultural system (Baciu 2023). A 2023-article gives an overview of creative transformations that can be described with this math.

Frequency dependent selection, on the other hand, is not less important. It is a necessary counterpart to variation-selection processes, especially when describing nonlinear processes. In particular, frequency dependent selection has led to the formulation of the basic equations of ecology, virology, game theory, and is also a necessary ingredient of chaos theory (Vandermeer & Goldberg, 2013, Nowak & May, 2000, Hofbauer & Sigmund, 1998).

Evidently, frequency dependent selection works to describe processes in an immensely broad array of applications. In the 2023-article just mentioned, I have made the case that frequency dependent selection can describe diversification and interplay in any kind of physical, biological, social, or cultural system (Baciu 2023). The rationale and empirical support are found in the article.

The distinction between variation selection processes and frequency dependent selection is important to understand. It is the same distinction as that between additions and multiplication. Variation selection processes are expressed with linear operations and sums. Frequency dependent selection is described with multiplications between variables. This makes the distinction between these two processes the same as that between additions and multiplications in mathematics and Or and AND operations in logic. It is a fundamental distinction that permeates all human thinking (Baciu 2023).

Given that Neural Networks use both of these complementary modes of representation to describe the world, their broad applicability is no surprise. I hope that this perspectives inspires engineers to apply Neural Networks even more broadly.

Initially, the deployment of Self-attention layers was motivated through empirical success, specifically in the task of translating natural languages (Vaswani et al., 2017). The interpretation of the mathematics that was introduced by "attention" and later "self-attention" was motivated by theories that applied to natural language.

However, when engineers created larger and larger "Large Language Models", they soon observed that the same Neural Network Architecture can be applied across a much broader range of disciplines. This discovery came as a surprise to many, and debates emerged why this empirical observation predominated. Why could one use Attention layers in NNs that had little to do with phenomena of attention. Soon, NNs that contained self-attention layers were dubbed Transformers. The public-facing debate of their applicability continues.

With the interpretation of Neural Networks proposed here, the broad applicability is not only possible but greatly expected. It would be more surprising if an architecture as found in present-day Neural Networks did not apply to the broad range of subjects it does. In addition, the present interpretation links neural network to centuries of scientific thinking across all disciplines. I hope that this connection is inspiring in broadening the application of NNs even further.

Taken together, I interpret neuron layers in NNs as evolutionary models that can describe any kind of creative transformations, and I interpret attention layers as dynamical models that can describe any kind of play between independent elements. This interpretation provides a way of thinking about Neural Networks that renders justice to their broad applicability of Neural Networks and supports further application in the broadest imaginable range of subjects.

 doi:10.20944/preprints202404.0506.v1

4. An Open Question

In my previous work, I have developed mathematical descriptions of cultural change. These descriptions contain linear as well as nonlinear components, which describe creativity and play, respectively. Creativity comes through as variation selection processes, while play comes through as frequency dependent selection (Baciu 2018).

In a series of articles, I have tested these mathematical descriptions on massive data. This has been done as part of the project Everything Called Chicago School and What Every1 Says (Baciu 2018, 2019, 2020, 2023).

Already in the first of these projects, I used frequency dependent selection to describe fashions. A phenomenon that I predicted and observed was related to the spread of ideas.

The phenomenon begins with an idea that spreads, growing exponentially in popularity. However, in response to the spreading idea, it is commonly observed that boredom and opposition spreads as well, as a byproduct. Growing ideas inspired opposition.

What happens next is that the interplay between idea and boredom or opposition become increasingly frequent. This is a consequence of the fact that both the idea and the opposition become widespread. Clashes between idea and opposition must therefore become even more frequent, at some point.

As such clashes continue, they have a negative effect on the growth of the idea. In consequence, the idea begins to disappear being faced by increasingly fierce opposition or increasingly persistent boredom. Overall a fashion wave is observed in which the idea becomes fashionable, only to go out of fashion again (Baciu 2018, 2019, 2020, 2023).

This phenomenon can be described as a very basic case of frequency dependent selection, and the same phenomenon is found in ecology, virology, epidemiology, and other fields of science (Baciu 2023).

In addition, in human culture, there are also circumstances in which multiple fashions can work together, which can be described with a more complex model of frequency dependent selection. The overlap of multiple nested processes then leads to larger waves of growth and reform (Baciu 2018, 2019, 2020, 2023). Similar phenomena are also observed in virology (Nowak and May 2000).

I am mentioning this example here because I would like to open a question. I am unaware of any Neural Network that is able to detect fashions or is behaving in ways that resemble fashions. The question is: Has anyone trained a network that detects how fashions that come, go, and return, as a result of the interplay between a group of spreading ideas and opposed responses to it? Or has anyone trained a Neural Network that behaves as if it was going through fashions or mood swings, giving output of one kind for some time, only to eventually switch to opposing its own output with new divergent outputs?

References

Baciu, D. C. (2018). From Everything Called Chicago School to the Theory of Varieties. Dissertation. Restricted Access. In Copyright. Retrieved from http://hdl.handle.net/10560/4376

Baciu, D. C. (2019). The Chicago School: Large-Scale Dissemination and Reception. In Chicago Schools: Authors, Audiences, and History, Prometheus 2. Published on 2019-11-16, within the section Keynotes.

Baciu, D. C. (2020). Cultural life: Theory and empirical testing. Biosystems, 197, 104208. https://doi.org/10.1016/j.biosystems.2020.104208.

Baciu, D. C. (2023). Causal models, creativity, and diversity. Humanities and Social Sciences Communications, 10, Article 134. Link to the article.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

Bejan, A., & Lorente, S. (2012). The S-Curves are Everywhere. Mechanical Engineering, 134(5), 44–47. https://doi.org/10.1115/1.2012-MAY-5

Darwin, C., & Wallace, A. R. (1858). On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. Journal of the Proceedings of the Linnean Society of London. Zoology, 3(9), 45-62.

Domingo, E., & Schuster, P. (Eds.). (2016). Quasispecies: From Theory to Experimental Systems. Springer. Part of the book series: Current Topics in Microbiology and Immunology, volume 392.

 doi:10.20944/preprints202404.0506.v1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Greenblatt, S. (2011). The Swerve: How the World Became Modern. W.W. Norton & Company.

Hofbauer, J., & Sigmund, K. (1998). Evolutionary Games and Population Dynamics. Cambridge University Press. Online ISBN: 9781139173179. DOI: 10.1017/CBO9781139173179.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Lotka, A. J. (1910). Contribution to the theory of periodic reactions. The Journal of Physical Chemistry, 14(3), 271-274.

Lucretius. (1st century BCE). De Rerum Natura.

Nowak, M. A. (2006). Evolutionary Dynamics: Exploring the Equations of Life. Harvard University Press.

Nowak, M. A., & May, R. M. (2000). Virus Dynamics: Mathematical Principles of Immunology and Virology. Oxford University Press.

Quetelet, A. (1842). A Treatise on Man and the Development of His Faculties. Edinburgh: William and Robert Chambers.

Ross, R. (1911). The Prevention of Malaria. John Murray.

Singh, K., Mehta, D., Dumka, S., Chauhan, A. S., & Kumar, S. (2023). Quasispecies Nature of RNA Viruses: Lessons from the Past. Vaccines, 11(2), 308. This article reviews the quasispecies nature of RNA viruses, addressing the theoretical and mathematical origins of quasispecies and their dynamics.

Uszkoreit, J. (2017, August 31). Transformer: A Novel Neural Network Architecture for Language Understanding. Google Research Blog. Retrieved from https://blog.research.google/2017/08/transformer-novel-neural-network.html.

Vandermeer, J., & Goldberg, D. (2013). Population Ecology: First Principles. Princeton University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Verhulst, P. F. (1838). Notice sur la loi que la population suit dans son accroissement. Correspondance mathématique et physique, 10, 113-121.

Wallace, A. R. (1858). On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. Journal of the Proceedings of the Linnean Society of London. Zoology, 3(9), 53-62.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).