Cracking the genetic code with neural networks

Marc Joiret[1], Marine Leclercq[2], Gaspard Lambrechts[3], Francesca Rapino[2], Pierre Close[2], Gilles Louppe[3], Liesbet Geris[1]

[1]Biomechanics Research Unit, GIGA In Silico Medicine; [2]Cancer Signaling, GIGA Stem Cells; [3]Department of Electrical Engineering and Computer Science, Artificial Intelligence and Deep Learning, Montefiore Institute, Liège University, Liège, Belgium.

LIÈGE université GIGA institute

biomech UNIVERSITY OF LIÈGE RESEARCH UNIT

INTRODUCTION

AI: Machine learning and Deep learning

- Both Machine learning ML and Deep Learning DL are part of the broad field of Artificial Intelligence AI
- ***ML*** first requires ***features extraction*** for classification or regression purposes
- ***DL*** skips the feature extraction and directly uses the raw data to learn from them by training a so called ***neural network***

A pedagogical showcase

- Deep learning holds great promise for ***biomedical research*** using ***omics data***
- Applying DL technologies to omics research still faces two difficulties: (i) the '***black box***' problem and (ii) the ***data quality and availability*** problem.
- Our study is a ***pedagogical contribution*** to address the black-box problem

Deep learning toy project

- The genetic code is ***textbook scientific knowledge*** established without resorting to AI
- Can DL architectures ***crack the code*** and unravel the correct knowledge ***from a training dataset*** ?
- The self-learning algorithm will lead to a deciphered code that should not be perceived as a black-box. ***How much data is needed*** to decipher the complete genetic code table?

MATERIALS & METHODS

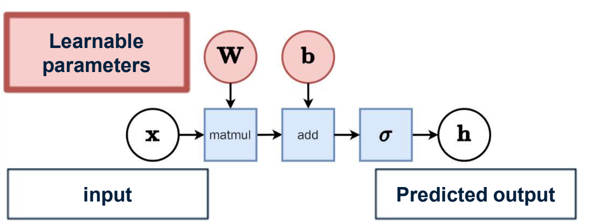
Basic architectures in Deep Learning: the Multi-Layer Perceptron (MLP)

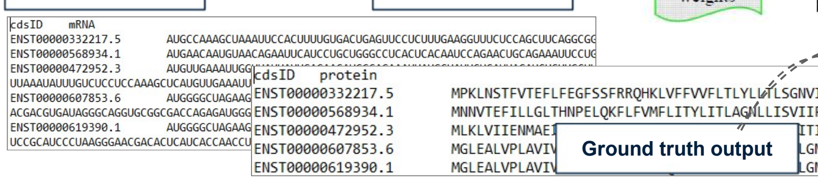
- ✓ The primitive of all neural networks is the ***perceptron***, invented by McCulloch and Pitts and improved by Rosenblatt (1943, 1958).
- ✓ A perceptron mimics the behavior of a ***brain neuron*** by combining the mathematical properties of ***linear algebra with an activation function***.
- ✓ The perceptron receives several input data, multiplies them ***with weights (learnable parameters)*** and produces one or several ***firing signal(s)***.
- ✓ The firing signals may serve as input data to a second layer of perceptrons. The ***larger the number of layers, the deeper*** the network (this is ***DL***).
- ✓ The output signals may be turned into ***a set of probabilities (simplex vector) for classification purposes*** and then compared with a ***ground truth*** vector.
- ✓ An objective function ***(loss function)*** is computed and optimized by updating the learnable parameters iteratively through a **stochastic gradient descent method** (automatic differentiation, backpropagation). The loss function keeps decreasing during learning on the ***training dataset***.

From the perceptron (primitive of all neural networks)…

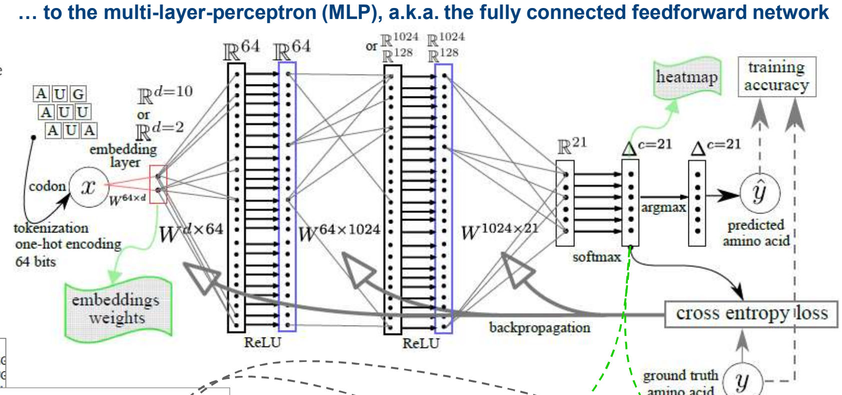
$$\mathbf{h} = \sigma(\mathbf{W}^T\mathbf{x} + \mathbf{b})$$

where $\mathbf{h} \in \mathbb{R}^q, \mathbf{x} \in \mathbb{R}^p, \mathbf{W} \in \mathbb{R}^{p\times q}, b \in \mathbb{R}^q$ and where $\sigma(\cdot)$ is upgraded to the element-wise sigmoid function.





… to the multi-layer-perceptron (MLP), a.k.a. the fully connected feedforward network

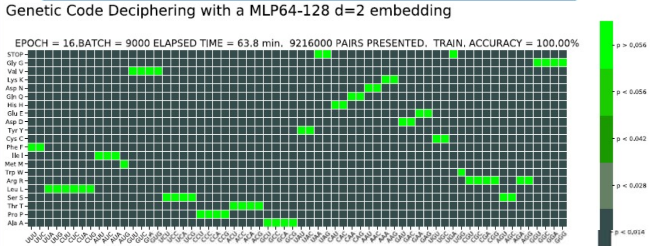


Cross Entropy loss =

$$L_{CE} = -\sum_{i=1}^{21} y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)$$

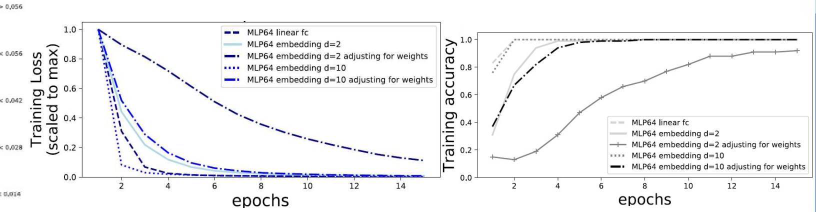
RESULTS AND CONCLUSIONS

Genetic Code Deciphering with a MLP64-128 d=2 embedding



Training performance and data efficiency of the learning process

Loss function and training accuracy evolution during learning on the training dataset:



Conclusions

- ✓ Resorting to AI and Deep Learning to gain data-driven knowledge requires a huge amount of high quality data for training the neural network.
- ✓ The wide generic capacities and modularity of DL networks allow them to be customized easily to learn the deciphering task of the genetic code.
- ✓ The biomedical research community is confronted to a trade-off between model complexity (or understandability) and data efficiency (amount of data needed to produce the inferred rules with a chosen accuracy).

REFERENCES

Joiret M, Leclercq M, Lambrechts G, Rapino F, Close P, Louppe G, Geris L. Cracking the genetic code with neural networks. Front. Artif. Intell. 2023 Apr 6;6:1128153. doi: 10.3389/frai.2023.1128153. PMID: 37091301; PMCID: PMC10117997.

CONTACT

GIGA in silico medicine, University of Liège

M. Joiret: marc.joiret@uliege.be; **L. Geris:** liesbet.geris@uliege.be; http://www.biomech.ulg.ac.be/

This poster was presented at the 'Neural networks: real vs man-made' GIGA day conference, Liège, Belgium, 4th September, 2023. We acknowledge the European Research Council under the European Union's Horizon 2020 Framework Program (H2020/2014-2020) /ERC grant agreement n°772418 (INSITE).