Critical Neural Networks in Atari Games

Thomas Pluck[1] and Aaron McAfee[1]
[1]Department of Electronic Engineering
Maynooth University
Maynooth, Ireland
Email: {thomas.pluck, aaron.mcafee}@mu.ie

Abstract—

I. INTRODUCTION

II. BACKGROUND

III. REINFORCEMENT LEARNING

A. Criticality in Statistical Mechanics

Criticality in statistical mechanics is the study of phase transitions, most classically thought of as transitions between states of matter, such as from water to ice. At the boundary between such phases, are phase boundaries where the system is said to be at criticality. At criticality, systems exhibit "scale-free" correlations in both space and time, where unlike systems not at criticality the microscopic and macroscopic are difficult to distinguish - hence, the system is scale-free.

Of particular relevance to the arguement presented in this paper is the notion of criticality in a model of magnetic materials called the Ising model. In an Ising model, atoms are arranged typically in a 3D lattice where each atom s_i can either be spin up $+1$ or spin down -1 with some probability $\mathbb{P}(s_i)$, each atom s_i is coupled to every other atom s_j with the coupling parameter J_{ij} and locally biased by the transverse magnetic field given by h_i. The energy E of a given configuration of spins in the spin vector $\vec{s}$ is then given by:

$$E(\vec{s}) = \sum_{i \neq j} J_{ij} s_i s_j + \sum_i h_i s_i$$

Which in turn defines the Gibbs distribution:

$$\mathbb{P}(\vec{s}) = \frac{1}{Z} \exp(-\beta E(\vec{s}))$$

Where Z is the normalizing partition function and β is the inverse pseudotemperature.

In the study of criticality in neural systems, special attention has been paid to the mean-field approximation of the Ising model connected in graphs with high degree N. In the mean-field approximation, we replace the exact interactions between spins with an interaction between each spin and the average magnetization of the system. This simplification becomes increasingly accurate as the connectivity of the system grows, making it particularly relevant for highly connected neural networks.

Under the mean-field approximation, the expected value of each spin s_i can be expressed as:

$$\langle s_i \rangle = \tanh(\beta(\sum_j J_{ij} \langle s_j \rangle + h_i))$$

At criticality, which occurs at a specific inverse temperature β_c (or equivalently, specific coupling strengths J_{ij}), the system exhibits maximal sensitivity to perturbations. Small changes in input can propagate throughout the entire network, leading to large-scale reconfiguration of the system state. This critical point represents a delicate balance between order and disorder - below criticality, the system tends toward fixed point attractors, while above criticality, dynamics become chaotic.

B. Criticality in Neural Systems

This equation bears a striking resemblance to the activation function of artificial neurons that use hyperbolic tangent activation. Indeed, if we consider a neural network with tanh activation functions, the state of each neuron can be written as:

$$\sigma_i = \tanh(\sum_j w_{ij} x_j + b_i)$$

Where w_{ij} are the connection weights, x_j are the inputs, and b_i is the bias. The formal correspondence between these two equations reveals a deep connection: neural networks with tanh activation functions can be interpreted as simulating the dynamics of an Ising model at a specific temperature.

In the context of biological neural systems, criticality has been observed and studied particularly in the context of avalanche statistics and in a variety of different settings including in characterizing the convergence of learned parameters, disappearing and vanishing gradients in RNN systems and optimal information transport.

Taking cues from biological neurons which are believed to be slightly subcritical state at all times allowing them maximal information processing with minimum energy expenditure. Critical neural systems are a valuable target to emulate to learn about

IV. METHODOLOGY

A. Critical Neural Network Architecture

The proposed network architecture relies on the clean identification between tanh activated neurons and the Ising model. The

ACKNOWLEDGMENT

The authors would like to thank...