Brain and Cognition 123 (2018) 154–164

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

Modeling cognitive deficits following neurodegenerative diseases and traumatic brain injuries with deep convolutional neural networks

Bethany Lusch*, Jake Weholt, Pedro D. Maia, J. Nathan Kutz

Department of Applied Mathematics, University of Washington, United States

ARTICLE INFO

Keywords:
Traumatic brain injury
Neurodegenerative disease
Focal axonal swellings
Convolutional neural networks

ABSTRACT

The accurate diagnosis and assessment of neurodegenerative disease and traumatic brain injuries (TBI) remain open challenges. Both cause cognitive and functional deficits due to focal axonal swellings (FAS), but it is difficult to deliver a prognosis due to our limited ability to assess damaged neurons at a cellular level *in vivo*. We simulate the effects of neurodegenerative disease and TBI using convolutional neural networks (CNNs) as our model of cognition. We utilize biophysically relevant statistical data on FAS to damage the connections in CNNs in a functionally relevant way. We incorporate energy constraints on the brain by pruning the CNNs to be less over-engineered. Qualitatively, we demonstrate that damage leads to human-like mistakes. Our experiments also provide quantitative assessments of how accuracy is affected by various types and levels of damage. The deficit resulting from a fixed amount of damage greatly depends on which connections are randomly injured, providing intuition for why it is difficult to predict impairments. There is a large degree of subjectivity when it comes to interpreting cognitive deficits from complex systems such as the human brain. However, we provide important insight and a quantitative framework for disorders in which FAS are implicated.

1. Introduction

The mathematical architecture of convolutional neural networks (CNNs) was originally inspired by the Nobel prize-winning work of Hubel and Wiesel on the primary visual cortex of cats (Hubel & Wiesel, 1962). Their seminal experiments were the first to suggest that neurons in the visual system organize themselves in hierarchical layers of cells for processing visual stimulus. The first quantitative model of the CNN, termed the Neocognitron by Fukushima (1980), already displayed many of the characteristic features of today's deep CNNs, including a multi-layer structure, convolution, max pooling and nonlinear dynamical nodes. The connection between neuroscience and CNN theory, although clearly a conceptional abstraction (Poggio, 2016a), has since been instrumental to improving quantitative models of how the brain integrates neuro-sensory information for stimulus classification and decision making. Given that CNNs mimic many of the important cognitive features of the brain, we use it as a model for understanding how neurodegenerative diseases and traumatic brain injuries (TBI) can compromise an array of recognition tasks. Specifically, by using well-established biophysical data on the statistics (distribution and size) of focal axonal swellings (FAS), which are among the primary symptoms of neurodegeneration and TBI, we evaluate the progress of impairments on a CNN-based model of cognition. Our model provides quantitative metrics for understanding how cognitive deficits are accumulated as a function of FAS development, allowing for potentially new diagnostics for the evaluation of brain disorders due to neurodegenerative diseases and/or TBI.

Understanding how neurodegenerative diseases and TBI affect cognitive function remains a critically important challenge for societal mental health. TBI alone is one of the major causes of disability and mortality worldwide, which in turn, dramatically jeopardizes society in several socioeconomic ways (Menon & Maas, 2015). Not only is it the signature injury of the wars in Afghanistan and Iraq (Jorge et al., 2012), it is also the leading cause of death among young people (Faul, Xu, Wald, & Coronado, 2010). While many survive the events that induce TBI, persistent cognitive, psychiatric, and physiological dysfunction often follows from the mechanical impact (see Section 2). Likewise, neurodegenerative diseases are responsible for an overwhelming variety of functional deficits, with common symptoms including memory loss or behavioral/cognitive impairments which are related to an inability to correctly process multi-modal information for decision-making tasks. The majority of brain disorders have a complex cascade of pathological effects spanning multiple spatial scales: from cellular or network levels to tissues or entire brain areas. Unfortunately, our limited ability to diagnose cerebral malfunctions *in vivo* cannot detect several anomalies that occur on smaller scales. FAS, however, are

* Corresponding author at: University of Washington, Lewis Hall #202, Box 353925, Seattle, WA 98195-3925, United States.
E-mail address: herwaldt@uw.edu (B. Lusch).

https://doi.org/10.1016/j.bandc.2018.02.012
Received 13 December 2016; Received in revised form 17 November 2017; Accepted 27 February 2018

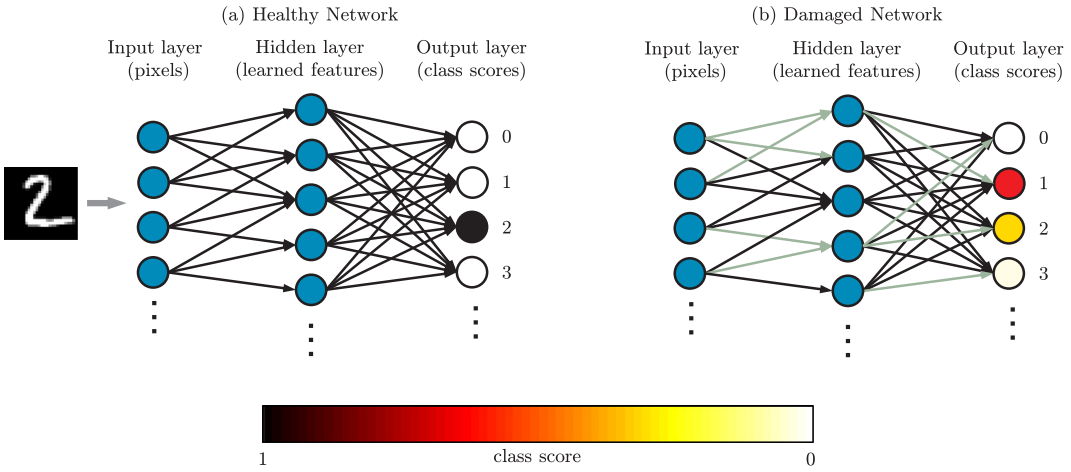

Fig. 1. Damaging a Convolutional Neural Network (CNN). (a) We start with a "healthy" CNN that accepts an image of a handwritten digit as an input and outputs scores for each possible digit, 0–9. We classify the image as the digit with the highest score. (b) We then damage the weights on the network in a biophysically-relevant way. In this figure, the healthy network correctly classifies the image as a 2, but the damaged network classifies it as a 1.

ubiquitous to TBI and most leading and incurable disorders that dramatically affect signaling properties of neurons, such as Multiple Sclerosis, Alzheimer's and Parkinson's diseases.

Given the currently available wealth of data on FAS morphology from TBI studies and from almost every leading neurodegenerative disease, significant progress can be made towards understanding qualitatively how FAS impacts cognitive function. In this work, we consider a set of deep CNN models as an *abstraction* for functioning brains. Our goal is to understand how the processing of input data (classification) is compromised as a function of increasing injury and/or disease progression. Of course, it is obvious that the system's performance will be compromised as the CNN is injured, but the *manner* in which the cognitive impairments arise is quite illustrative and informative, providing intuitively appealing results about how cognitive deficits can develop and evolve as a neurodegenerative disease progresses.

Fig. 1 illustrates our approach. We begin with the original (healthy) CNN, which is trained to perform a classification task. In Fig. 1, the specific task is to label a handwritten digit. We then expose the CNN to different injury protocols based upon biophysical observations of FAS statistics and morphological parameters. In particular, we use statistical distributions of FAS from a recent experiment consisting of TBI-induced damage in the visual cortex of rats (Wang, Hamm, & Povlishock, 2011). To impose these injury statistics on the original CNN, we assume that each neuronal connection has a biophysically plausible probability to malfunction; while mild axonal injury may simply weaken a connection, severe cases may break it permanently (i.e., an axomtotomy occurs so that the connection strength goes to zero). Ultimately, the severity of the injury and re-weighting of connections is also determined by biophysical data and the statistical distribution of the size of the FAS. We can then progressively monitor the deleterious effects of the injury on the functionality of the CNN, providing metrics for cognitive deficits that arise.

The paper is outlined as follows: In Section 2 we provide key background material on the two primary fields integrated into this work: convolutional neural networks and neural disorders in which FAS are implicated. We describe our methodology in Section 3 and present results in Section 4. We summarize our conclusions in Section 5. For full details, all MATLAB and Python codes used for this paper are available online at http://github.com/BethanyL/damaged_cnns.

2. Background

2.1. Convolutional neural networks

Deep learning is transforming almost every field of science involving big data. The success of the method has been enabled by two critical components: (i) the continued growth of computational power (e.g. GPU and networked computing), and (ii) exceptionally large labeled data sets capable of taking advantage of the full power of a multilayer architecture (Goodfellow, Bengio, & Courville, 2016).

Deep learning has experienced a resurgence in popularity since 2006. However, it was a topic of intensive research long before. Indeed, neural networks were highly successful in a wide range of applications and machine learning architectures Goodfellow et al. (2016). By the early 1990s, they were studied as standard textbook material (Bishop, 1995), with the focus typically on a small number of layers. Importantly, there were a number of critical innovations which established multilayer feedforward networks as a class of universal approximators. Specifically, Hornik, Stinchcombe, and White (1989) rigorously established that standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions were capable of approximating any Borel measurable function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units were available. Thus, multilayer feedforward networks could be thought of as a class of universal approximators (Hornik et al., 1989).

There is a recent but very large body of papers in which CNNs are used to study human brain function. Yamins, Hong, Cadieu, and DiCarlo (2013), for instance, used CNNs to model the ventral stream as this series of cortical areas are thought to subserve object recognition. By extending their model class to contain mixtures of deeper CNN networks, corresponding intuitively to specialized subunits in the ventral visual system, they could predict spiking responses from the inferior temporal cortex with high accuracy (Yamins et al., 2014). In fact, the same brain region was investigated by a wide range of computational model representations (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014). DNNs were also used to model the neural representation of music across the superior temporal gyrus (Güçlü, Thielen, Hanke, van Gerven, & van Gerven, 2016). See Kriegeskorte (2015), Poggio (2016b), and Yamins and DiCarlo (2016) for recent reviews. Overall, CNNs provide a powerful platform that can be tailored to simulate a broad array of cognitive functions within different brain circuits.