THE UNREASONABLE EFFECTIVENESS OF MATHEMATICS / MACHINE LEARNING IN THE NATURAL SCIENCES OF THE ARTIFICIAL

Vijay Chandru
vijaychandru@iisc.ac.in
INAE Distinguished Technologist, IISc
Strand Life Sciences

CHAIR, Chalmers, March 2019
The Sciences of the Artificial
Herbert A Simon
1962-1968

• What of artificial sciences?
Knowledge about artificial objects and phenomena. Artificial as in “produced by art / man-made rather than natural”
  ▪ Artificial things are synthesized
  ▪ Artificial things may imitate the natural
  ▪ Artificial things have function, goals, adaptation
  ▪ Artificial things are designed by descriptors and imperatives
  ▪ Artificiality and the limits of rationality in administration
  ▪ Artificiality and Complexity: Systems Hierarchy

▪ Computing systems are rational artifacts, and like human brains, limitedly so.
Arthur Samuel in 1959 defined **Machine Learning as the field of study that gives computers the ability to learn without being explicitly programmed.**

- Valiant who introduced the framework of **PAC learning** (probably approximately correct learning)
- Vapnik casts the problem of `learning' as an **optimization problem** allowing people to use all of the theory of optimization that was already given.
- Nowadays machine learning is a combination of several disciplines such as **statistics, information theory, theory of algorithms, probability and functional analysis.**

**Optimization remains at the heart of all modern machine learning problems**

“I don't want to bore you,” Harvey said, “but you should understand that these heaps of wires can practically think - linear programming - which means that instead of going through all the alternatives, we have a hunch which is the right one.” *The Billion-Dollar Brain*, (1966) by Len Deighton
Example I

A Single Neuron - Perceptron
1 Layer, 2 Neurons
1 Layer, 3 Neurons
Example II: Harder
1 Layer, 4 Neurons
1 Layer, 8 Neurons
2 Layers, 8 Neurons Each
Searching for the right Curve

Chapter 37
Algorithms for the Optimal Loading of Recursive Neural Nets

V. Chandru*  A. Dattasharma†  S. S. Keerthi‡  N. K. Sancheti§  V. Vinay⁴

Abstract
We address the problem of choosing synaptic weights in a recursive (Hopfield) neural network so as to "optimize" the performance of the network on the recognition of binary strings. The problem has been called the net loading (or learning) problem in the literature [10]. The objective is to maximize the basins of attraction around the desired fixed points (binary strings) of the net. It is known that it is $\text{NP}$-hard to evaluate even the two-step radius of attraction of a recursive neural net [3]. We focus on the radius of direct (one-step) attraction and will refer to this as the loading problem. We have both theoretical and computational results on this problem, that we summarize below.

- A proof that the net loading problem can be solved in polynomial time using linear programming techniques. This resolves a standing problem in the complexity of recursive neural networks [10].
- An alternate formulation of the net loading problem as a proximity problem in high-dimensional convex geometry.
- The design and implementation of a hybrid algorithm for the said proximity problem.
- Successful solution of large scale test problems including the optimal solution to a 900 $\times$ 900 Hopfield net with approximately $4 \times 10^5$ synaptic weights.

It may be noted that our experiments indicate that the radius of direct attraction is actually a very good proxy of the intractable (multi-step) radius of attraction. In all the test problems that we have solved, the synaptic weights obtained as a solution to the maximum radius of direct attraction also maximize the radius of (multi-step) attraction. In this sense the message of this paper is that the effective design of large-scale associative memory based on recursive neural networks is possible.

ANNs

matrix. These networks are of interest because of their use as associative memory in classification and pattern recognition. Given a specified set of binary vectors $\{y^1, y^2, \ldots, y^n\}$, the network behaves as an associative memory if it eventually reproduces as its output one of the inputs, say $y^1$, when triggered at the input by a vector $z$ which is sufficiently close to $y^1$. Here "sufficiently close" means that the Hamming distance between $z$ and the fixed point $y^1$, i.e., the number of unequal components of the two bipolar strings, is no more than the radius of the basin of attraction around $y^1$.

The dynamical behaviour of a reservoir neural network makes it difficult to analyse the radius of attraction around a given pattern. It has been shown by Floreen and Orponen [3] that it is $\text{NP}$-hard to evaluate even the two-step radius of attraction of a recursive neural net. We present exact algorithms to synthesize recursive neural networks that optimize the radius of direct attraction.

(Net Loading) Given vectors $y^1, \ldots, y^n$, where each $y^i$ is an $n$-dimensional vector whose components are $\{+1, -1\}$, construct a weight matrix $W$, such that the resulting network has maximum radius of direct attraction about the $y^i$.

We shall see, from both theoretical and computational perspectives, that this problem is tractable.
The New AI World

- Handcrafted Intelligence vs Deep Learnt Intelligence
- Huge Training Sets (Digitization/Transduction)
- Huge Compute Power for Training (advances in hardware CPU/GPU/TPU)
DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman          Ming Yang          Marc’Aurelio Ranzato
Facebook AI Research  Tel Aviv University
Menlo Park, CA, USA    Tel Aviv, Israel
{yaniv, mingyang, ranzato}@fb.com  wolf@cs.tau.ac.il
Abstract

In modern face recognition, the conventional pipeline consists of four stages: detect $\rightarrow$ align $\rightarrow$ represent $\rightarrow$ classify. We revisit both the alignment step and the representation step by employing explicit 3D face modeling in order to apply a piecewise affine transformation, and derive a face representation from a nine-layer deep neural network. This deep network involves more than 120 million parameters using several locally connected layers without weight sharing, rather than the standard convolutional layers. Thus we trained it on the largest facial dataset to-date, an identity labeled dataset of four million facial images belonging to more than 4,000 identities. The learned representations coupling the accurate model-based alignment with the large facial database generalize remarkably well to faces in unconstrained environments, even with a simple classifier. Our method reaches an accuracy of 97.35% on the Labeled Faces in the Wild (LFW) dataset, reducing the error of the current state of the art by more than 27%, closely approaching human-level performance.
If you have a HAMMER, every problem is a nail!

- “If humans are doing this as a visual task, why not present this as a visual task?”
- Deep Learning with CNNs has been particularly effective in image classification
- DeepVariant (Google) calling SNP (point mutations) from DNAseq data. Compares well with GATK (gold standard from Broad Institute).
A real intelligence does not break when you slightly change the problem! A black box cannot explain why a solution found is good or why a prediction makes sense.

2015 Google’s Jeff Dean: Deep Mind on Atari Breakout

Vicarious 2017

Has Deep Learning reached its limits?
Can we create an AI with less Data?
Is it called Intelligent Systems?
Brain-inspired automated visual object discovery and detection, UCLA/Stanford, PNAS Dec 2018 Roychowdhury, Kailath

Can one **automatically browse internet image data**, as a human would, to construct brain-like and computationally tractable models of the visual world?

**Scalability and Robustness limitations of computer vision unleashed.** Less dependence on supervised frameworks requiring manually labeled exemplars

1. Access to **large-scale perceptual data** of the kind that humans experience,
2. **Flexible representations** - Structural Unsupervised Viewlets Model (SUVM)
3. **Efficient unsupervised learning algorithms** (MLE/Convex Programming).
Accolades at the KDD Cup

Challenge of Pulmonary Emboli Detection

Pulmonary embolism (PE) is a condition that occurs when an artery in the lung becomes blocked. In most cases, the blockage is caused by one or more blood clots that travel to the lungs from another part of your body. While PE is not always fatal, it is nevertheless the third most common cause of death in the US, with at least 650,000 cases occurring annually. The clinical challenge, particularly in an Emergency Room scenario, is to correctly diagnose patients that have a PE, and then send them on to therapy. This, however, is not easy, as the primary symptom of PE is dyspnea (shortness of breath), which has a variety of causes, some of which are relatively benign, making it hard to separate out the critically ill patients suffering from PE.

Task 3: One of the most useful applications for CAD would be a system with very high (100%?) Negative Predictive Value. In other words, if the CAD system had zero positive candidates for a given patient, we would like to be very confident that the patient was indeed free from PE's. In a very real sense, this would be the "Holy Grail" of a PE CAD system.

Winners of KDD Cup 2006: Task 3 - Negative Predictive Value

- **Best Student Entry:** Zhang Cas (IA, PKU)

- **First Place:** William Perrizo and Amal Shehan Perera (DataSURG Group, North Dakota State University)
- **Runner Up:** Nimisha Gupta and Tarun Agarwal (Strand Life Sciences Pvt. Ltd.)
- **Best Student Entry:** Karthik Kumara (team leader), Sourangshu Bhattacharya, Mehul Parsana, Shivramkrishnan K, Rashmin Babaria, Saketha Nath J, and Chiranjib Bhattacharyya (Indian Institute of Science)
Genomics in Precision Medicine

Exponential Laws in Technology

Since Illumina came on the scene, the cost of sequencing genes has dropped far faster than prices predicted by Moore’s Law.

Ref: Forbes, 2014
Analytics and Data Science in Precision Medicine
Our Impact on Precision Medicine

A way forward for several thousands of patients each year

Strand’s clinical exome test solves young girl’s rare condition

“Our daughter was suffering from seizures and hypoglycemia for close to a decade. Genetic testing at Strand helped us identify the cause, and take appropriate management measures.”

Arpita’s Parents

Fighting recurring ovarian cancer

“I love life too much to give up! Multiple rounds of failed chemotherapy did not make me lose hope. Genetic testing helped identify a new generation drug to keep my cancer in remission.”

Meena Nandak, 56

Battling an aggressive enemy

“Triple negative breast cancer can be quite an opponent. However, genetic testing empowered my doctor and me to take the right treatment decisions and battle the disease.”

Dimple Bawa, 34
Informatics (2000 - )

- Products for Data Analytics:
  - GeneSpring, Strand NGS, Strand Omics
- Software & Curation Services:
  - Agilent, Illumina, Roche

Diagnostics (2012 - )

- Germline Tests:
  - Hereditary Cancer, Clinical Exome
- Somatic Tests:
  - 152-gene, Liquid biopsy

In the Clinic (2018 - )

- Hospital system integration:
  - Largest Cancer chain
  - 28 Labs, 900 Employees, 1.5M Tests/yea
- Pharma Services:
  - Biobank, Clinical Studies,
Cover the entire range of high-throughput data, including Microarrays, Mass Spectrometry, Next-Generation Sequencing, Microscopy etc

GeneSpring, Mass profiler Pro, Strand NGS, Sarchitect, ...
24K citations
Data Crunching

Storage

Accuracy

<table>
<thead>
<tr>
<th>Sample</th>
<th>Library</th>
<th>Read length</th>
<th>Precision/Recall (%)</th>
<th>InDel Precision/Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA12878</td>
<td>Whole Exome (PE)</td>
<td>100 bp</td>
<td>99.43/98.0</td>
<td>97.8/91.9</td>
</tr>
<tr>
<td>NA12878</td>
<td>Whole Genome (PE)</td>
<td>151 bp</td>
<td>98.5/99.5</td>
<td>97.2/96.2</td>
</tr>
</tbody>
</table>

*As measured against BWA + GATK HC restricted to GIB high confidence regions
Interpretation Work flow

VCF file

- Variant count- ~13500

Filter for clinically relevant genes for the given cancer type. E.g., >100 genes (>1000 variants)

Gene Disease association

- Conservation scores (43 species)
- ~15 QC parameters: (SR, BQ, SB...)
- 7 In-silico prediction

Variant Quality & Conservation

- Disease databases: ClinVar, HGMD, ARUP, OMIM, Emory, LOVD....
- Strand curation database
- PPDB
- Population databases- dbSNP, ExAC, 1000 Genomes, EVS, GnomAD....

Literature and databases

- Disease causing variant(s)
Ramanujan Augmented Intelligence

- **Data Analytics** for extracting more out of data
- **Natural Language Processing** for extracting knowledge from vast amounts of literature

---

**Diagram**

- Clinicians and Patients
- Clinical Interpretation and Reporting
- Variant Annotation
- Variant Detection
- Bioinformatics and Data Curation

**Lists**

**Dictionary Building**
- Building comprehensive dictionaries for relevant document search

**Text Reading**
- Format handling, sentence splitting, tokenization, POS tagging, phrase extraction

**Entity Recognition**
- Dictionary-driven robust identification of entities

**Syntax Parsing**
- Grammar-driven deep parsing of sentences

**Semantic Inferencing**
- Understanding meaning from parsing output

**Fact Extraction**
- Extracting associations between entities via shallow parsing
Combining Assays Optimizations with *Strandomics*, our platform comprising Algorithms, Databases, Curated Content and various Dashboards helps us deliver NGS-based testing at industry-lowest costs.
The StrandOmics Platform

- Algorithms, databases, curated content, dashboards for variant interpretation and reporting
- Honed over 10s of thousands of reports
- In constant usage by an experienced team of 30+
- Drives our costs of reporting down to 10% or less of costs in comparable labs.

Strand IRIS

A powerful and easy-to-use web-based tool for automated clinical reporting of NGS-based diagnostic tests in solid tumors.

- AMP compliant variant categorisation
- HIPAA compliant
- Prioritized clinical trials
- Customisable templates
- Option to request manual review
- Drug class based reporting
- Report generation in minutes.

www.strandls.com
- NGS Data Quality
  - Learn error patterns
- Sequence Alignment
  - Variant calling (extensions of the DeepVariant SNP calling)
  - Small RNA classification
- Digital pathology
  - Phillips Solution for ER/PR/HER2 (breast cancers)
  - Classification based on our training sets created from our test volumes leading to new models for other cancers.
Acknowledgements

Mentor: Prof Roddam Narasimha

Partners: The Gang of Four – van Gogh Bot Renderings