ML & AI: A SWOT Analysis

Volkan Cevher, Associate Professor EPFL

https://lions.epfl.ch
Preface

My research:
- Machine Learning (ML)
- Optimization
- Signal Processing
- Information Theory
- Statistics

My courses (2019-20):
- Mathematics of Data
- Reinforcement Learning
- Advanced Topics in ML

This talk
Group:
14 PhDs & 3 postdocs
14 nationalities
Faculty@Rice, NUS, Umea, Zhejiang, UNC, Linkoping, AIMS, UoCB, VNU, NTU, Technion
Postdoc@ETHZ (3), MIT (2), McGill, Turing
Others@Kandou bus, SwissRE, TUM

ML & AI | Volkan Cevher |
https://lions.epfl.ch

Preface
My research:
Machine Learning (ML)
Optimization
Signal Processing
Information Theory
Statistics

My courses (2019-20):
Mathematics of Data
Reinforcement Learning
Advanced Topics in ML
Strengths

A SWOT Analysis

Machine Learning (ML)

Neural Networks (NN)

Multilayer NNs
Machine Learning (ML)

- ML is an interdisciplinary study of algorithms, statistical models, and error functions jointly with computer systems to perform specific tasks

  “Only a fool learns from his own mistakes. The wise man learns from the mistakes of others” - Otto von Bismarck

- ML makes you wiser
The ingredients via a simplified supervised learning example

- ML is an interdisciplinary study of algorithms, statistical models, and error functions jointly with computer systems to perform specific tasks.

Task: Learn a mapping from image to disease

\[ y = \text{function}_x(a) = f(a'x) \]

Task: Learn a mapping from control inputs to walking
ML is an interdisciplinary study of algorithms, statistical models, and error functions jointly with computer systems to perform specific tasks.

Supervised ML: Use algorithms to learn “model”

\[
\min_x \text{Error} (y, f(a'x))
\]
Academic theory vs industrial practice

Conventional wisdom in ML until 2010:

Simple models + simple errors

optimization landscapes
Enter neural networks: Universal approximation

\[ f(x; \beta, W, b) = \beta^T \sigma(Wx + b) \]

Challenges:
1. too big to optimize!
2. did not have enough data
3. could not find the optimum via algorithms
Multilayer neural networks: Tractable & nearly universal

Subquadratic overparameterization for shallow neural networks

Song et al.

Neural Network Analysis

Algorithm

Initialization

Network Structure

: First-order method (Gradient descent...)

: Appropriate initialization For weight matrices W

: Sufficiently large network (Overparameterization)

Venture Investments in Artificial Intelligence Surge

Total investment, in billions of dollars, and number of deals for each year.

* - 2019 data for nine months

Chart: WIRED · Source: Pitchfork

ML & AI | Volkan Cevher | https://lions.epfl.ch
Weaknesses

A SWOT Analysis

Robustness
Bias
Reproducibility
Malicious data
Interpretability
Robustness
Robustness is an active research area

- Madry, Aleksander and Makelov, Aleksandar and Schmidt, Ludwig and Tsipras, Dimitris and Vladu, Adrian. Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR.
Adversarial examples are inevitable!

\[ f(x; \beta, W, b) \]

\[ f(x) \approx f(x) + \langle \epsilon, \nabla f(x) \rangle \]

\[ \|f(x + \epsilon) - f(x)\| \leq \|\epsilon\| \|\nabla f(x)\| \]

- Concentration-of-measure phenomenon
- Lipschitz constant is important

[Shafahi et al. ICLR 2019]
Progress towards robustness

\[ |f(x + \epsilon) - f(x)| \leq \|\epsilon\| \left\| \nabla f(x) \right\| \sup_{x \rightarrow L(f)} \]

**NP-hard** for NNs [Scaman et al. NeurIPS 2018]

Lipschitz Constant Estimation of Neural Network via Sparse Polynomial Optimization.
Interpretability

Humans

↑ inform

Interpretability Methods

↑ extract

Black Box Model

↑ learn

Data

↑ capture

World

Interpretability

- Linear Regression
- Decision Tree
- K-Nearest Neighbors
- Random Forest
- Support Vector Machines
- Neural Nets

Accuracy
Interpretability in ML is an active research field


- Sundararajan, Mukund and Taly, Ankur and Yan, Qiqi. Axiomatic Attribution for Deep Networks. ICML'17.

A robustness & interpretability result

On Certifying Non-Uniform Bounds against Adversarial Attacks.
Liu, Chen and Tomioka, Ryota and Cevher, Volkan. ICML’19.
Further evidence: Robust training <> interpretability

Robust fundus classification & dataset bootstrapping via interpretable features
Krawczuk et al.
Reproducibility

**IS THERE A REPRODUCIBILITY CRISIS?**

- 7% Don't know
- 52% Yes, a significant crisis
- 38% Yes, a slight crisis
- 3% No, there is no crisis

1,576 researchers surveyed

---

**Grad student descent**

Posted on 2014/01/25 by science@dryad

_A popular method for designing deep learning architectures is GDGS (gradient descent by grad student). This is an iterative approach, where you start with a straightforward baseline architecture (or possibly an earlier SOTA), measure its effectiveness; apply various modifications (e.g. add a highway connection here or there), see what works and what does not (i.e. where the gradient is pointing) and iterate further on from there in that direction until you reach a (local?) optimum._

Share  Report  Save
Reproducibility challenge: Non-convexity

Lagrangian perspective: New theory for nonlinear optimization with nonlinear constraints

\[ L^* \text{ reconstruction error (per pixel)} \]

\[ 10^{-3} \]

\[ 6 \times 10^{-4} \]

\[ 10^1 \quad 10^3 \]

iAL
Sahin M. F. et. al. [NeurIPS 2019]

ADMM
Latorre F. et. al. [NeurIPS 2019]

AL^2
Eftekhar A. et. al. [Under review]
Extending reproducibility via universality in convex optimization

One algorithm to rule them all!

<table>
<thead>
<tr>
<th>Type</th>
<th>Complexity</th>
<th>Smooth</th>
<th>Stochastic</th>
<th>Nonsmooth</th>
<th>Strongly convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smooth</td>
<td>$O\left(\frac{1}{k^2}\right)$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Stochastic</td>
<td>$O\left(\frac{1}{\sqrt{k}}\right)$</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmooth</td>
<td>$O\left(\frac{1}{\sqrt{k}}\right)$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Strongly convex</td>
<td>$O\left(\rho^k\right)$, $\rho &lt; 1$</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

$k$ is the iteration counter.

✓ Universal primal-dual, Yurtsever et al.
✓ UniXGrad, Kavis et al. Accelegrad, Levy et al.
✓ Random extrapolation, Alacaoglu et al.
Many other weaknesses

1. Bias
2. Malicious data
3. Privacy
4. ...

The New York Times  ONE NATION, TRACKED  NATIONAL SECURITY

Opinion | THE PRIVACY PROJECT

Twelve Million Phones, One Dataset, Zero Privacy

By Stuart A. Thompson and Charlie Warzel
DEC. 19, 2019
A geometric perspective on bias
Opportunities

A SWOT Analysis

Generative Adversarial Networks

- Automation
- Financial
- Design
- Medical
Generative Adversarial Networks

Progressive Growing of GANs for Improved Quality, Stability, and Variation
Karras et al. [ICLR 2018]

High-Fidelity Image Generation With Fewer Labels
Challenge: Limit cycles (minimax)

\[
\min_{\theta \in \mathbb{R}} \max_{\omega \in \mathbb{R}} f(\theta, \omega) = \theta \omega + \phi(\theta) - \phi(\omega), \quad \phi \text{ non-convex}
\]

The limits of min-max optimization algorithms: Convergence to spurious non-critical sets, Hsieh, Mertikopoulos, and Cevher 2020.
Training GANs via mixed Nash equilibria (minimax)

\[ \max_{\theta \in [-2,2]} \min_{\omega \in [-2,2]} f(\theta, \omega) = \theta^2 \omega^2 - \theta \omega \]

Minimax formulations and robust RL

Robust Reinforcement Learning with Langevin Dynamics. Kamalaruban et al.

https://lions.epfl.ch
Minimax formulations and robust BO

Adversarially robust Gaussian Process Optimization.
Bogunovic et al. NeurIPS 2018
New opportunities via GANs

Closed loop deep Bayesian inversion: Uncertainty driven acquisition for fast MRI.
Sanchez et al.
New opportunities in RL

Interactive Teaching Algorithms for Inverse Reinforcement Learning. Kamalaruban et al. IJCAI 2019

Interaction-limited Inverse Reinforcement Learning. Troussard et al.
New opportunities in deep learning

Generalization <-> Robustness

- **Network Structure**: Sufficiently large network (Overparameterization)
- **Neural Network Analysis**: First-order method
  - **Algorithm**: Stochastic gradient
  - **Initialization**: Appropriate initialization for weight matrices W

Convergence of SGD for neural networks without heavy overparameterization. Song & Cevher

Efficient proximal mapping of the 1-path-norm of shallow networks. Latorre et al.

Lipschitz controlled polynomial
New opportunities in scalable optimization

Randomization <> Scalability

Towards stochastic SDP & LP’s with stochastic constraints

\[
\min_{x \in \mathcal{X}} \mathbb{E} [f(x, \xi)] \quad \text{s.t.} \quad A(\xi)x = b(\xi) \text{ almost surely}
\]

- Conditional gradient methods for stochastically constrained convex minimization. Vladarean et al.
- Ex: scalable solutions (sparsest) cut problems and their variants
- Ex: robustness certifications for NNs
New opportunities in engineering applications

Chemical machine learning with kernels: The impact of loss functions. Van Nguyen et al. [Quantum Chemistry 2019]

EDA Gym. Krawczuk et al.
Threats

A SWOT Analysis
The AI hype vs the ML revolution

A surprising number of firms are jumping on the artificial-intelligence bandwagon—without actually investing in any AI.
Talent pool: Missing the top talent vs the needed talent
## Sustainability:

### The estimated costs of training a model

<table>
<thead>
<tr>
<th>Model</th>
<th>Date of original paper</th>
<th>Energy consumption (kWh)</th>
<th>Carbon footprint (lbs of CO2e)</th>
<th>Cloud compute cost (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer (65M parameters)</td>
<td>Jun, 2017</td>
<td>27</td>
<td>26</td>
<td>$41-$140</td>
</tr>
<tr>
<td>Transformer (213M parameters)</td>
<td>Jun, 2017</td>
<td>201</td>
<td>192</td>
<td>$289-$981</td>
</tr>
<tr>
<td>ELMo</td>
<td>Feb, 2018</td>
<td>275</td>
<td>262</td>
<td>$433-$1,472</td>
</tr>
<tr>
<td>BERT (110M parameters)</td>
<td>Oct, 2018</td>
<td>1,507</td>
<td>1,438</td>
<td>$3,751-$12,571</td>
</tr>
<tr>
<td>Transformer (213M parameters) w/ neural architecture search</td>
<td>Jan, 2019</td>
<td>656,347</td>
<td>626,155</td>
<td>$942,973-$3,201,722</td>
</tr>
<tr>
<td>GPT-2</td>
<td>Feb, 2019</td>
<td>-</td>
<td>-</td>
<td>$12,902-$43,008</td>
</tr>
</tbody>
</table>

Note: Because of a lack of power draw data on GPT-2’s training hardware, the researchers weren’t able to calculate its carbon footprint.

Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper
Sustainability:
Dennard scaling & Moore’s law vs Growth of data

FORECASTED AMOUNT OF DATA GENERATED,
2016-2025

DART Consulting

Number of Each Year’s Best GPUs Needed to Process Incoming YouTube Data by Year

ML & AI | Volkan Cevher | https://lions.epfl.ch

Andy Burg | 90 nm | 28 nm

TODAY

Time

Technology Benefits
Requirements
Sustainability:
Energy constraints / Time constraints

Learning-based compressive sensing + hardware design.
Baldassarre et al., Gozcu et al., Aprile et al. [IEEE TMI, IEEE TSP, IEEE CnS, IEEE TCAS]

IBM Thesis Award 2019
Sustainability:

Energy constraints of recording neural data

\[
\begin{align*}
\mu\text{-electrode input} & \rightarrow \text{AFE} \rightarrow \text{ADC} \rightarrow \text{DSP} \rightarrow \text{TX} \\
& > 30 \text{ dB quality} \\
& \text{AFE + ADC} \\
& \text{DSP} \\
& \text{TX} \\
& 10 \mu\text{W} \\
& 80 \mu\text{W} \\
& \approx 2.5 \mu\text{W} \\
& 50 \mu\text{W} \\
& \approx 2.5 \mu\text{W} \\
& \approx 3 \mu\text{W}
\end{align*}
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Compression rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBCS</td>
<td>38.90</td>
</tr>
<tr>
<td>SHS</td>
<td>33.67</td>
</tr>
<tr>
<td>BERN</td>
<td>33.57</td>
</tr>
<tr>
<td>MCS</td>
<td>33.22</td>
</tr>
</tbody>
</table>

ML & AI | Volkan Cevher | [https://lions.epfl.ch](https://lions.epfl.ch)
Sustainability:

Time constraints of MRI

- Accelerate the MRI scan 5 times.
- Pick the most relevant data only for your method.

Rethinking Sampling in Parallel MRI: A Data-Driven Approach. Gözcü B. et al. [EUSIPCO 2019]
Sustainability:

**Time constraints of MRI**

- Time drastically increases the dimensionality of data
- Reduce computations by a factor 200: from a month to 4 hours without losing performance.

Scalable learning-based sampling optimization for compressive dynamic MRI. Sanchez T., et al. [IEEE ICASSP 2020]
Sustainability: Resource constrained optimization

Truncated variance reduction: A unified approach to Bayesian optimization and level-set estimation.
Bogunovic et al. NIPS 2017
Conclusions

● Are you wiser?
  ○ time-data-power and other trade offs

● Existential threats = “Opportunities”
  ○ talk to me offline

● ML - AI: Mathematical understanding
  ○ Hype protection

volkan.cevher@epfl.ch
https://lions.epfl.ch
Twitter: @CevherLIONS