Introduction to course "Optimizing AI"

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Towards efficient deep learning

An overview of modern AI

What is AI?

- Next step towards **automation**:
  - Machines already good at **simple object manipulation** and **computing**.
  - Next steps are: **understanding the outside world** and **reasoning**.

Old way

- Let human experts code the machines,
  - Goods: we know what we are doing.
  - Bads: some problems we do not know how to solve (or how to solve efficiently).

Modern way

- Let machines teach themselves how to solve a problem.
  - Goods: machines do the work,
  - Bads: lack of understandability/robustness.
- Requires **training**.

An overview of modern AI

What is AI?

- Next step towards **automation**:
  - Machines already good at **simple object manipulation** and **computing**.
  - Next steps are: **understanding the outside world** and **reasoning**.

Old way

- Let human experts code the machines,
  - Goods: we know what we are doing.
  - Bads: some problems we do not know how to solve (or how to solve efficiently).

Modern way

- Let machines teach themselves how to solve a problem.
  - Goods: machines do the work,
  - Bads: lack of understandability/robustness.
- Requires **training**.

An overview of modern AI

What is AI?

- Next step towards **automation**:
  - Machines already good at **simple object manipulation** and **computing**.
  - Next steps are: **understanding the outside world** and **reasoning**.

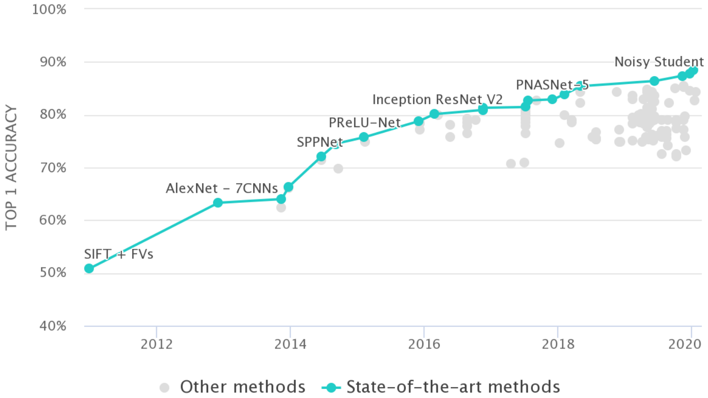
Old way

- Let human experts code the machines,
  - Goods: we know what we are doing.
  - Bads: some problems we do not know how to solve (or how to solve efficiently).

Modern way

- Let machines teach themselves how to solve a problem.
  - Goods: machines do the work,
  - Bads: lack of understandability/robustness.
- Requires **training**.

Modern Deep Learning



source : `https://paperswithcode.com/sota/image-classification-on-imagenet`

Why optimizing Deep Learning ?

AI on Embedded / Edge devices

- Privacy concerns, user customization
- Power consumption
- Latency

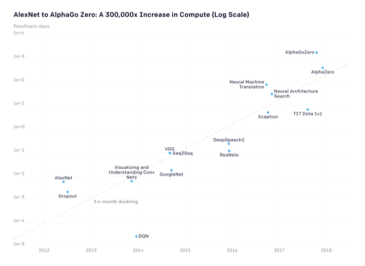
http://eyeriss.mit.edu/2019_neurips_tutorial.pdf and https://openai.com/blog/ai-and-compute/

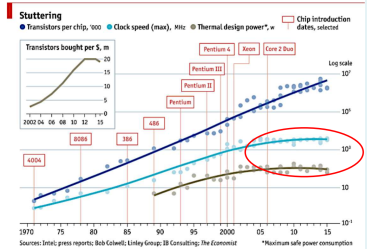
Why optimizing Deep Learning ?

AI on Embedded / Edge devices

- Privacy concerns, user customization
- Power consumption
- Latency

Power consumption for training and using large models





Sources: Intel; press reports; Bob Colwell; Linley Group; IB Consulting; *The Economist*

*Maximum safe power consumption

http://eyeriss.mit.edu/2019_neurips_tutorial.pdf and https://openai.com/blog/ai-and-compute/

Course organisation

Sessions

1. Deep Learning Essentials,
2. Quantification,
3. Pruning,
4. Factorization,
5. Distillation,
6. Operators and Architectures,
7. Embedded Software and Hardware for DL.

Lab Sessions and Challenge

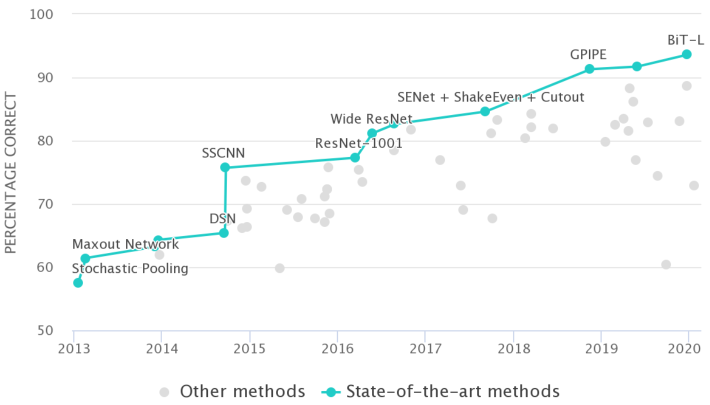
By groups of two, you are given a machine with complete access.

Sessions schedule

Each session has (roughly) the same structure:

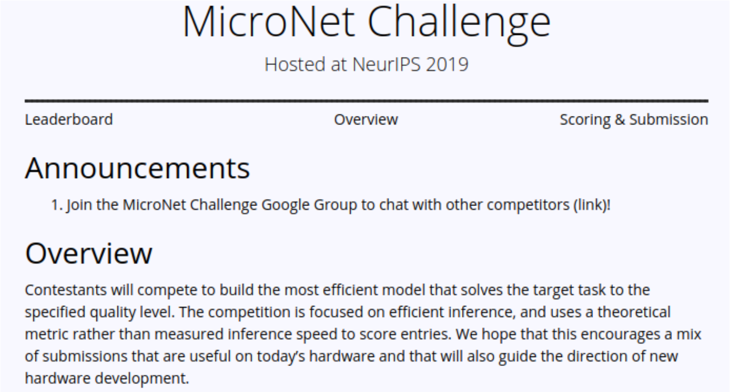
- **Short written eval** about the previous lesson (10 min),
- Short lesson (20 to 40 min),
- Lab Session,
- Project,
- Sessions 2, 4 and 6 include **students' presentations** before the lesson.

MicroNet Challenge - CIFAR100



source : `micronet-challenge.github.io`

MicroNet Challenge - CIFAR100



MicroNet Challenge

Hosted at NeurIPS 2019

Announcements

1. Join the MicroNet Challenge Google Group to chat with other competitors (link)!

Overview

Contestants will compete to build the most efficient model that solves the target task to the specified quality level. The competition is focused on efficient inference, and uses a theoretical metric rather than measured inference speed to score entries. We hope that this encourages a mix of submissions that are useful on today's hardware and that will also guide the direction of new hardware development.

source : `micronet-challenge.github.io`