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**.

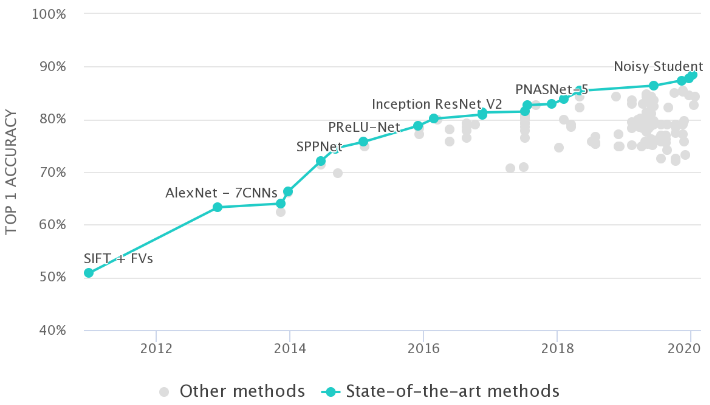
Old way

- 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

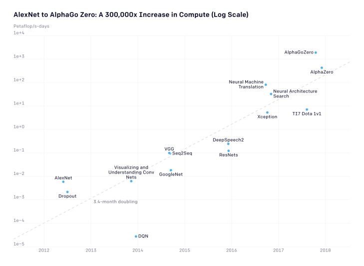
http://eyeriss.mit.edu/2019_neurips_tutorial.pdf and https://openai.com/blog/ai-and-compute/

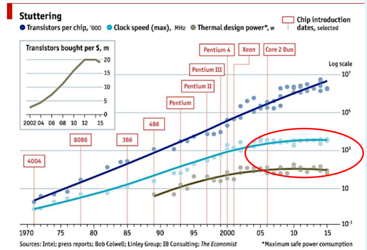
Why optimizing Deep Learning ?

AI on Embedded / Edge devices

- Privacy concerns, user customization
- Power consumption
- Latency

Power consumption for training and using large models





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

Deep Learning Optimization Challenges

Examples of challenges

- Micronet at NeurIPS 2019
- Low Power Computer Vision (since 2015)
- DCASE Task 1 challenges 2020 and 2021

MicroNet Challenge

Hosted at NeurIPS 2019

Leaderboard | Overview | Scoring & Submission

Announcements

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

Overview

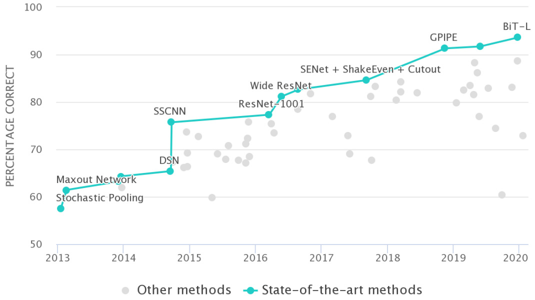
Contestants will compete to build the most efficient model that solves the target task to the specified quality level. The competition is focused on efficient inference, and uses a theoretical metric rather than measured inference speed to score entries. We hope that this encourages a mix of submissions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

source : `micronet-challenge.github.io`

Deep Learning Optimization Challenges

Examples of challenges

- Micronet at NeurIPS 2019
- Low Power Computer Vision (since 2015)
- DCASE Task 1 challenges 2020 and 2021

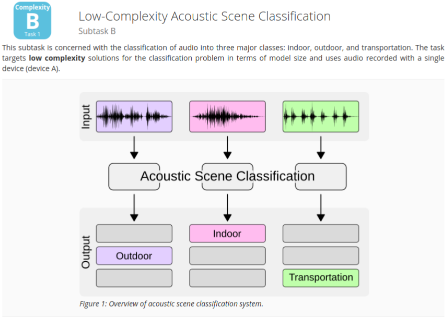


source : `micronet-challenge.github.io`

Deep Learning Optimization Challenges

Examples of challenges

- Micronet at NeurIPS 2019
- Low Power Computer Vision (since 2015)
- DCASE Task 1 challenges 2020 and 2021



source : dcase.community

Deep Learning Optimization Challenges

Examples of challenges

- Micronet at NeurIPS 2019
- Low Power Computer Vision (since 2015)
- DCASE Task 1 challenges 2020 and 2021

	Submission information		Evaluation dataset			Acoustic model				System
Rank	Submission label	Technical Report	Official system rank	Accuracy	Logloss	Parameters	Non-zero parameters	Sparsity	Size (KB) *	Complexity management
1	Koutini_CPJKU_task1b_2		1	96.5 %	0.101	345k	247k	0,284	483.5	pruning float16
2	Koutini_CPJKU_task1b_4		2	96.2 %	0.105	556k	249k	0,552	487.1	float16 smaller width/depth
3	Hu_GT_task1b_3		3	96.0 %	0.122	122k	122k	0	490.0	int8 quantization
4	McDonnell_USA_task1b_3		4	95.9 %	0.117	3M	3M	0	486.7	1-bit quantization
5	Hu_GT_task1b_1		7	95.8 %	0.357	94k	94k	0	375.0	int8 quantization
5	Hu_GT_task1b_4		5	95.8 %	0.131	125k	125k	0	499.0	int8 quantization
5	McDonnell_USA_task1b_4		6	95.8 %	0.119	3M	3M	0	486.7	1-bit quantization
6	Koutini_CPJKU_task1b_3		8	95.7 %	0.113	242k	242k	0	473.8	float16 smaller width/depth
7	Hu_GT_task1b_2		10	95.5 %	0.367	122k	122k	0	490.0	int8 quantization
7	McDonnell_USA_task1b_2		9	95.5 %	0.118	3M	3M	0	486.7	1-bit quantization

source : dcase.community

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.