Deep Learning for Computer Vision

UCA Master 2 Data Science

INRIA Sophia Antipolis – **STARS team**

S3.2 : **10 December / 25 February**

UNIVERSITÉ
CÔTE D'AZUR

STARS Inria Research Team

Objective: designing vision systems for the recognition of human activities

Challenges:

- Perception of Human Activities : **robustness**
  - Long term activities (from sec to months),
  - Real-world scenarios,
  - Real-time processing with high resolution.

- Semantic Activity Recognition : **semantic gap**
  - From pixels to semantics, uncertainty management,
  - Human activities including complex interactions with many agents, vehicles, …
  - Fine grained facial expressions, rich 3D spatio-temporal relationships.

- **Applications** : Safety & Health (CoBTeK from Nice Hospital : Behavior Disorder)

Toyota Smart-Home
Large scale daily living dataset

Related Courses @ UCA

MSc Data Science and Artificial Intelligence

http://univ-cotedazur.fr/en/idex/formations-idex/data-science/

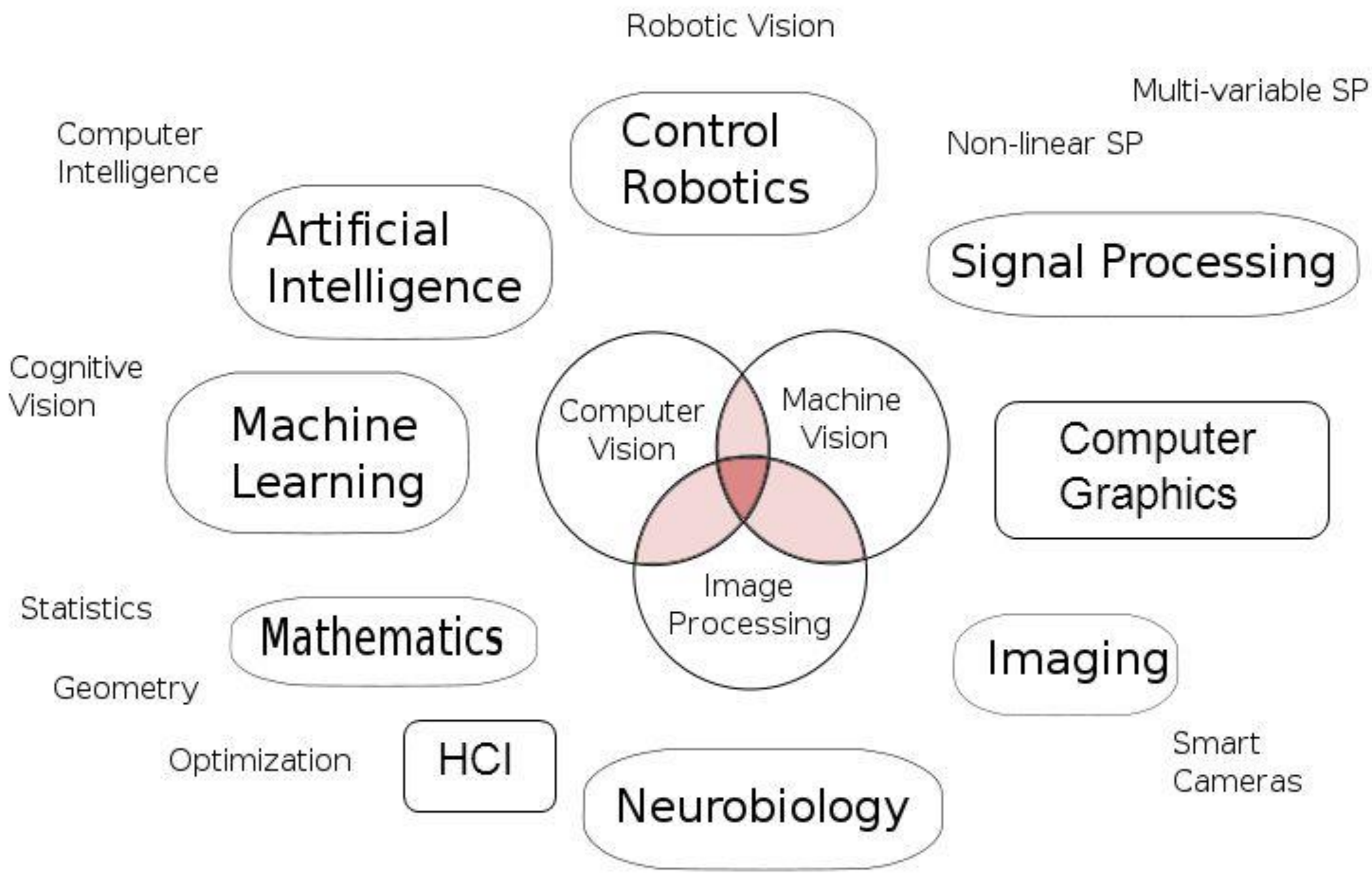
Master 1:

- Statistical inference: theory and practice I & II
- Processing large datasets with R
- Data visualization
- A general introduction to Data Mining
- Technologies for Big Data with Python
- Computer Vision: Foundations and Applications

Master 2:

- Computer Graphics
- Optimization for Data Science
- Medical Imaging
- Deep Learning
- **Computer Vision**

Vision is multidisciplinary



- Computer Vision is a subfield of artificial intelligence and machine learning.
- Techniques in machine learning and other subfields of AI (e.g. NLP) can be borrowed and reused in computer vision.

Computer Vision: many Tasks

Computer Vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from digital images or videos.

From the perspective of engineering, it seeks to automate tasks that the human visual system can do. [Wikipedia]

Computer Vision Tasks:

- Recognition : Objects or Events
  - Classification
  - Detection
  - Retrieval
- Motion analysis
  - Tracking
  - Optical flow
- Image synthesis
- Image restoration
- Biometrics
- etc...

Video Analytics (or VCA) applies CV & ML algorithms to extract/analysis content from videos

Video Analytics: many Domains

- Smart Sensors: Acquisition (dedicated hardware), thermal, omni-directional, PTZ, cmos, IP, tri CCD, RGBD Kinect, FPGA, DSP, GPU.
- Networking: UDP, scalable compression, secure transmission, indexing and storage.
- Image Processing/**Computer Vision**: feature extraction, 2D object detection, active vision, tracking of people using 3D geometric approaches
- Multi-Sensor Information Fusion: cameras (overlapping, distant) + microphones, contact sensors, physiological sensors, optical cells, RFID
- Event Recognition: CNN, Probabilistic approaches HMM, DBN, logics, symbolic constraint networks
- Reusable Systems: Real-time distributed dependable platform for video surveillance, OSGI, adaptable systems, Machine learning
- Visualization: 3D animation, ergonomic, video abstraction, annotation, simulation, HCI, interactive surface.

Video Analytics Applications

- Strong impact in transportation (metro station, trains, airports, aircraft, harbors)
- Traffic monitoring (parking, vehicle counting, street monitoring, driver assistance, self-driving car)
- Control access, intrusion detection and Video surveillance in public places, building
- Store monitoring, Retail, Aware House, Bank agency
- Health (HomeCare) patient monitoring,
- Video communication (Mediaspace, 3D virtual realty)
- Sports monitoring (Tennis coach, Soccer analytics, F1, Swimming pool monitoring)
- Other application domains : Robotics, Drones, Teaching, Biology, Animal Behaviors, Risk management …

➢ Creation of start-up

  ➢ Keeneo: http://www.keeneo.com/

  ➢ Ekinnox: https://www.ekinnox.com/

Video Analytics : Issues

Practical issues

Video Understanding systems have poor performances over time, can be hardly modified and do not provide semantics

Video Analytics : Issues

V1) Acquisition information:

- V1.1) Camera configuration: mono or multi cameras,
- V1.2) Camera type: CCD, CMOS, large field of view, colour, thermal cameras (infrared), Depth
- V1.3) Compression ratio: no compression up to high compression,
- V1.4) Camera motion: static, oscillations (e.g., camera on a pillar agitated by the wind), relative motion (e.g., camera looking outside a train), vibrations (e.g., camera looking inside a train),
- V1.5) Camera position: top view, side view, close view, far view,
- V1.6) Camera frame rate: from 25 down to 1 frame per second,
- V1.7) Image resolution: from low to high resolution,

V2) Scene information:

- V2.1) Classes of physical objects of interest: people, vehicles, crowd, mix of people and vehicles,
- V2.2) Scene type: indoor, outdoor or both,
- V2.3) Scene location: parking, tarmac of airport, office, road, bus, a park,
- V2.4) Weather conditions: night, sun, clouds, rain (falling and settled), fog, snow, sunset, sunrise,
- V2.5) Clutter: empty scenes up to scenes containing many contextual objects (e.g., desk, chair),
- V2.6) Illumination conditions: artificial versus natural light, both artificial and natural light,
- V2.7) Illumination strength: from dark to bright scenes,

Video Analytics : Issues

V3) Technical issues:

- V3.1) Illumination changes: none, slow or fast variations,
- V3.2) Reflections: reflections due to windows, reflections in pools of standing water, reflections,
- V3.3) Shadows: scenes containing weak shadows up to scenes containing contrasted shadows (with textured or coloured background),
- V3.4) Moving Contextual objects: displacement of a chair, escalator management, oscillation of trees and bushes, curtains,
- V3.5) Static occlusion: no occlusion up to partial and full occlusion due to contextual objects,
- V3.6) Dynamic occlusion: none up to a person occluded by a car, by another person,
- V3.7) Crossings of physical objects: none up to high frequency of crossings and high number of implied objects,
- V3.8) Distance between the camera and physical objects of interest: close up to far,
- V3.9) Speed of physical objects of interest: stopped, slow or fast objects,
- V3.10) Posture/orientation of physical objects of interest: lying, crouching, sitting, standing,
- V3.11) Calibration issues: little or large perspective distortion,

Video Analytics : Issues

V4) Application type:

- V4.1) Tool box : primitive events, enter/exit zone, change zone, running, following someone, getting close,
- V4.2) Intrusion detection: person in a sterile perimeter zone, car in no parking zones,
- V4.3) Suspicious behaviour: violence, fraud, tagging, loitering, vandalism, stealing, abandoned bag,
- V4.4) Monitoring: traffic jam detection, counter flow detection, activity optimization, homecare,
- V4.5) Statistical estimation: people counting, car speed estimation, data mining, video retrieval,
- V4.6) Simulation: risk management,
- V4.7) Biometry and object classification: fingerprint, face, iris, gait, soft biometry, license plate, pedestrian.
- V4.8) Interaction and 3D animation: 3D motion sensor (Kinect), action recognition, serious games.
- V4.9) Robotics, Drones

Video Analytics : Issues

Successful application: right balance between

- Structured scene: constant lighting, low people density, repetitive behaviours,
- Simple technology: robust, low energy consumption, easy to set up, to maintain,
- Strong motivation: fast payback investment, regulation,
- Cheap solution: 120 to 3000 euros per smart camera.
- Availability of Knowledge or large video datasets with annotation

Commercial products:

- Intrusion detection: ObjectVideo, Keeneo, Evitech, FoxStream, IOimage, Acic,…
- Traffic monitoring: Citilog, Traficon,…
- Swimming pool surveillance: Poseidon,…
- Parking monitoring: Ivisiotec,…
- Abandoned Luggage: Ipsotek,…
- Biometry: Sagem, Sarnof,…,SenseTime, MegVii (face++),
- Integrators: Honeywell, Thales, IBM, Siemens, GE, …, CVTE, Huawei,
- Camera providers: Bosh, Sony, Panasonic, Axis, …, HIK Vision,
- Game industries: Microsoft, Nitendo,...
- Retail: Amazon,… Tencent YouTu Lab, CloudWalk, Baidu, Alibaba, Tencent
- Self-driving Cars: Tesla, Google, Uber,…Argo AI,

Video Analytics : Issues

Performance: robustness of real-time (vision) algorithms

Bridging the gaps at different abstraction levels:

- From sensors to image processing
- From image processing to 4D (3D + time) analysis
- From 4D analysis to semantics

Uncertainty management:

- uncertainty management of noisy data (imprecise, incomplete, missing, corrupted)
- formalization of the expertise (fuzzy, subjective, incoherent, implicit knowledge)

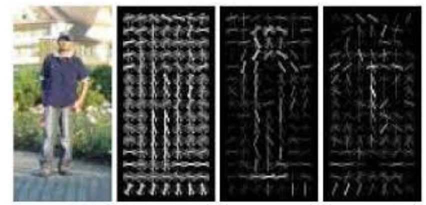
Independence of the models/methods versus:

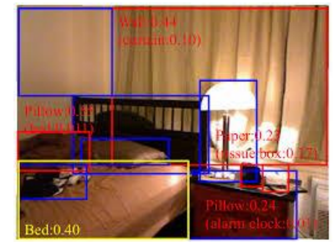
- Sensors (position, type), scenes, low level processing and target applications
- several spatio-temporal scales

Knowledge management :

- Bottom-up versus top-down, focus of attention
- Regularities, invariants, models and context awareness
- Knowledge acquisition versus ((none, semi)-supervised, incremental) learning techniques
- Formalization, modeling, ontology, standardization

A brief history of Computer Vision





David Lowe, 1999
SIFT

Sivic & Zisserman, 2003
Bags of words

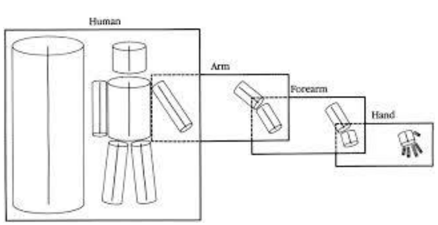
Dalal & Triggs, 2005
HOG

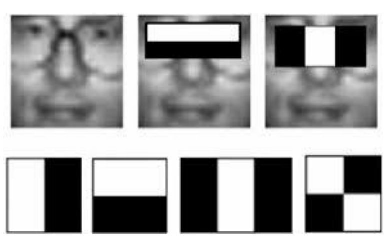
Everingham, 2012
PASCAL Challenge

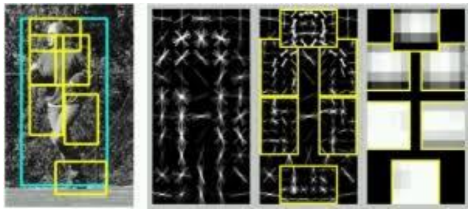
David Marr, 1970s
from images to geometric
blobs, edges, 3-D models

Viola & Jones, 2001
Face Detection

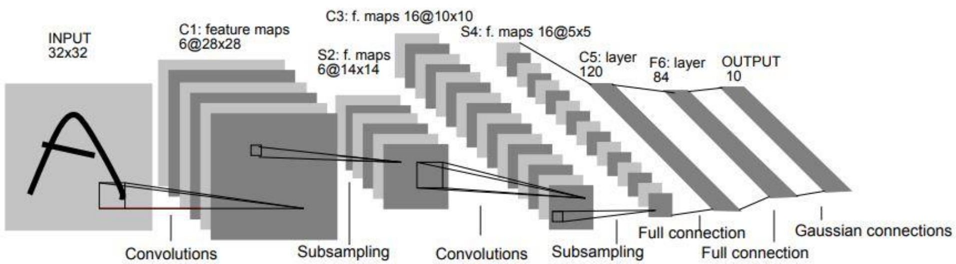
Felzenswalb & Ramanan, 2009
Deformable Part Model

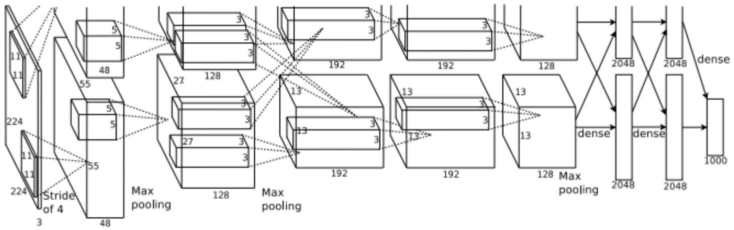






A brief history of Deep Learning





Minsky & Papert, 1969
perceptron

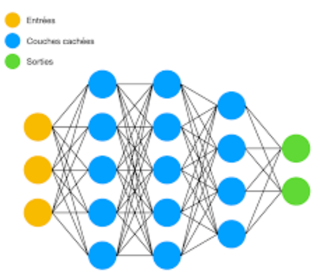
LeCun, 1990
convolutional networks

LeCun, Bengio, 1998
LeNet-5
Gradient-based learning

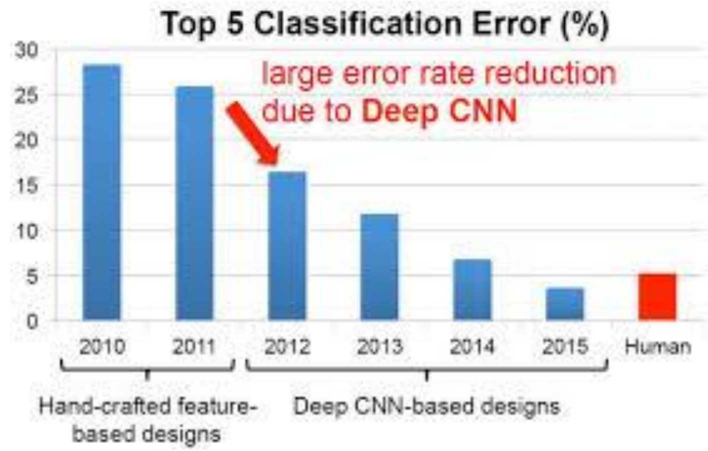
Li Fei-Fei, 2009
Image-net
22K categories and 15M images

Krizhevsky, Hinton, 2012
AlexNet

Ross Girshick, 2016
Faster RCNN



ImageNet Large Scale Visual Recognition Challenge
Russakovsky et al. IJCV 2015



informatics mathematics Inria

UNIVERSITÉ CÔTE D'AZUR

Components for Deep Learning

3 Components for Deep Learning:

- Software: Deep Learning Algorithms, Libraries
- Data : Images, Videos, Annotation
- Hardware: High Computation

Deep Learning Software

Libraries

- Caffe — (Berkeley Vision Lab)
- **TensorFlow** — (Google)
- CNTK — (Microsoft) - discontinued
- Torch — (Facebook) - discontinued
  - **PyTorch** — (Facebook)
- Theano — (MILA) – discontinued
- MXNet – Apache Software Foundation
- built on top of other libraries:
  - **Keras** — (Individual initiative + Google push)

Models

A complete end-to-end system performing a well-defined vision task

- FRCNN, Mask-RCNN; SSD, YOLO, RetinaNet (detection/segmentation),
- FCNN (Fully Convolutional, segmentation)
- RNN, GRU, LSTM

Networks

A neural network consisting of convolutional or recurrent layers or both, which extracts features from an image.

- VGG16, Alexnet,
- Siamese,
- ResNet, Inception, Inception-Resnet, DenseNet

Data : machine learning

Machine Learning : Data-Driven Approach

- Collect a dataset of images and labels
- Use Machine Learning to train a classifier
- Evaluate the classifier on new images

Machine Learning : Few Approaches

- supervised learning
  - Learn to map an input (data) to a target output (representation), which can be discrete (classification) or continuous (regression)
- unsupervised learning
  - Learn a compact representation of the data that can be useful for other tasks, e.g. density estimation, clustering, sampling, dimension reduction, manifold learning
  - but: in many cases, labels can be obtained automatically, transforming an unsupervised task to supervised
- semi-supervised
  - weakly supervised, ambiguous/noisy labels, self-supervised etc.
- reinforcement learning
  - learn to select actions, supervised by rewards.

informatics mathematics Inria
UNIVERSITÉ CÔTE D'AZUR

Data : machine learning

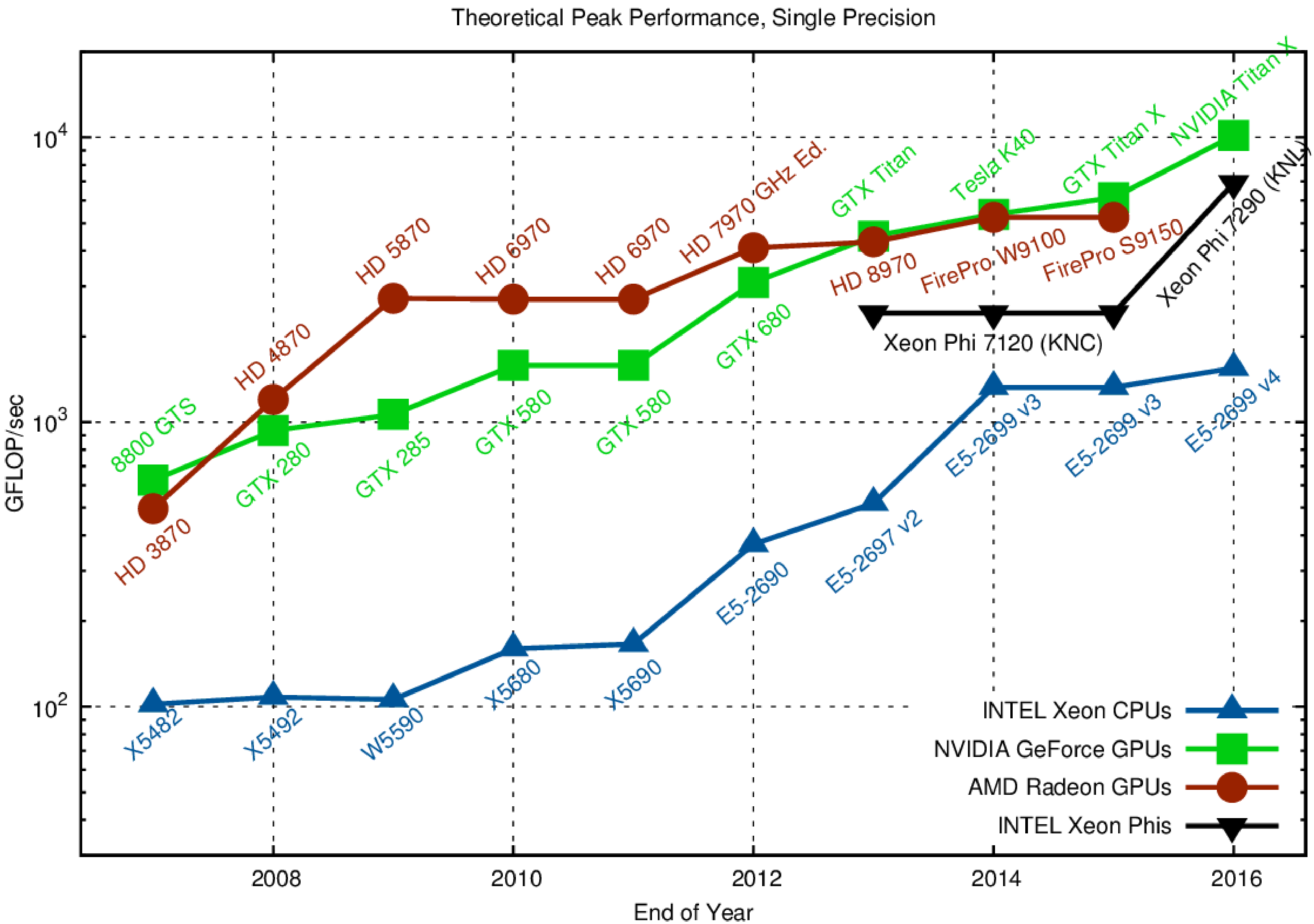
Image DataSets

- CIFAR10 (CIFAR100, MNIST) [1998 - 2006]
  - 10 classes/ 50,000 training images/ 10,000 testing images
- Image-net [2009 – 2012]
  - 22K categories and 15M images; (subset) 1K categories and 1.2M images
- Pascal VOC [2006 - 2012]
  - 20 object categories, 11.5K images, detection + segmentation
- MS COCO [2014]
  - 90 object categories, 183 K images, detection + segmentation + keypoints
- OpenImages
  - 600 object categories, 1.7 M images, detection – weakly annotated

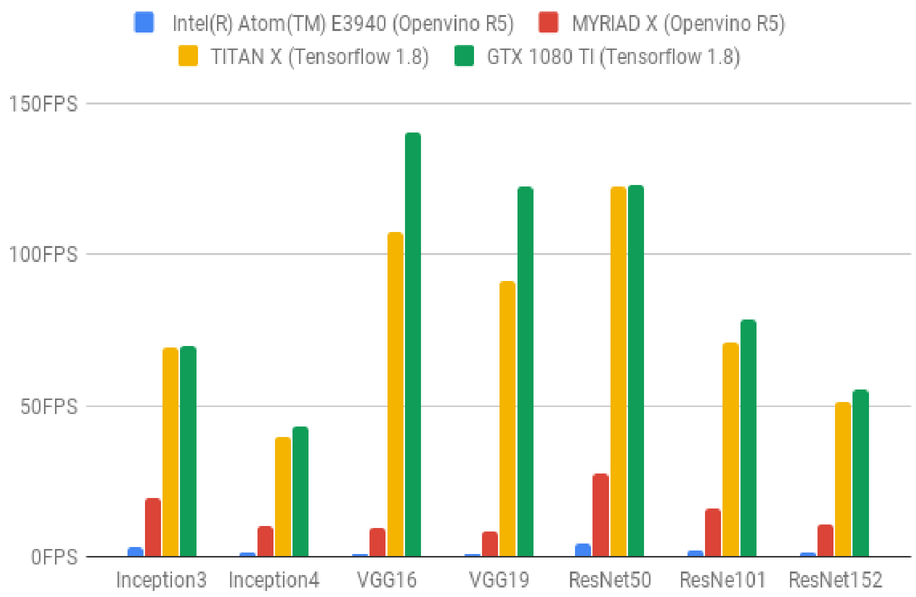
Video DataSets

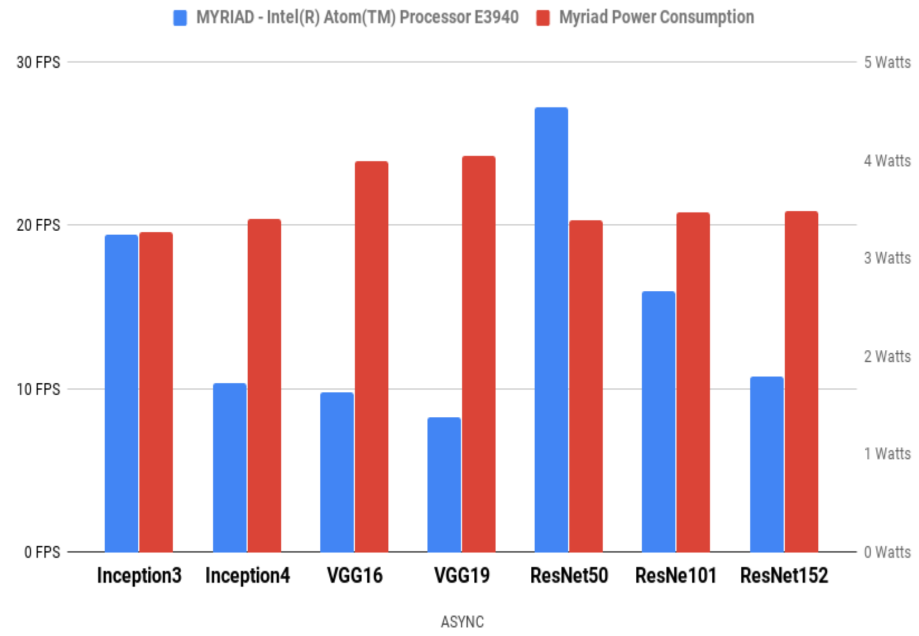
- Kinetics
  - 400-600-700 action classes, 325-650K video clips [2017-2019]
- ActivityNet-200
  - 200 action classes, 20K untrimmed videos, 31K action instances [2016]
- MSRDailyActivity3D:
  - 16 action classes, 320 video clips [2012]
- NTU RGB+D
  - 60 action classes, 56880 videos [2016], 120 action classes, 120K videos [2019]
- Toyota Smarthome
  - 31 action classes, 16129 videos [2019], 53 action classes, 536 videos , 41K action instances [2020]

Deep Learning Hardware



Deep Learning Hardware





Limitations on Nvidia Deep learning on Embedded hardware

- Power consumption : GTX 1080: 250 W > Myriad X: 5 W
- Only 3 years of Warranty (at least 8 needed)

Educational Objectives:

- Discuss well-known methods from low-level description to intermediate representation, and their dependence on the end task
- Study a data-driven approach where the entire pipeline is optimized jointly in a supervised fashion, according to a task-dependent objective
- Study deep learning models in details
- Interpret them in connection to conventional models
- Focus on recent, state of the art methods and large scale applications

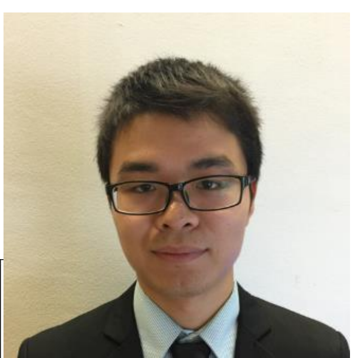
Course Planning

Each session : lecture (theoretical) + practice

- **Lecture 1:** Introduction to CV : Francois + Hao
  - Traditional and modern Computer Vision & Artificial Intelligence [FB]
  - Neural Networks for CV : one neuron, activation, loss function, BP [HC]
  - Practice: Back Propagation with Python
- **Lecture 2:** Image Classification : Hao
  - CNN : convolution, pooling, receptive field, normalization [HC]
  - Practice: LeNet-5 for digit recognition with Pytorch
- **Lecture 3,4,5:** Object Detection : Ujjwal
  - Object detection techniques will include Faster-RCNN, SSD and Feature Pyramid Networks.
  - Each will be deeply described and compared.
- **Lecture 6:** Video Classification, RNNs (Vanilla network), LSTM : Srijan
- **Lecture 7:** Action Recognition: Srijan
  - Dense Trajectories, different video aggregation techniques, two-streams, LSTMs for AR, 3D ConvNets
- **Lecture 8:** Attention Mechanism : Srijan
  - spatial attention for image classification, spatio-temporal attention for action recognition.
- **Lecture 9:** GAN and VAE : Yaohui
- **Lecture 10:** Article presentation : all

How to Contact Us

- Course Website:
  - http://www-sop.inria.fr/members/Francois.Bremond/MSclass/deepLearningWinterSchool/index.html
  - Syllabus, lecture slides, schedule, etc
- Emails:
  - Hao Chen: hao.chen@inria.fr
  - Ujjwal: ujjwal.ujjwal@inria.fr
  - Srijan Das: srijan.das@inria.fr
  - Yaohui Wang: yaohui.wang@inria.fr
  - Antitza Dantcheva: antitza.dantcheva@inria.fr
  - Francois Bremond: francois.bremond@inria.fr



Evaluation Policy

- Engagement while attending class (oral) : 30%
  - Answering questions
  - Practical training

- Article presentation: 70%
  - 5 groups of 2 – 3 students
  - Select 1 article out of 10
  - Last day: slide presentation : 20 min + 10 min questions
    - Motivation (P1)
    - State-of-the-art (P1)
    - Proposed approach (P2)
    - Performance/limitations (P3)
    - Future directions (P3)

Proposed articles

- **Visual explanation [Hao]**
  - Learning Deep Features for Discriminative Localization (CVPR 2016)
- **Object Detection [Ujjwal]**
  - Read the papers of YOLOv1, YOLOv2 and YOLOv3 and make a consolidated presentation explaining the YOLO pipeline along with its historical development
    - YOLOv1 Paper : https://pjreddie.com/media/files/papers/yolo_1.pdf
    - YOLOv2 Paper: https://pjreddie.com/media/files/papers/YOLO9000.pdf
    - YOLOv3 Paper: https://pjreddie.com/media/files/papers/YOLOv3.pdf
  - Read the following two papers and discuss their relative and comparative ideas.
    - http://openaccess.thecvf.com/content_iccv_2017/html/Bodla_Soft-NMS_--_Improving_ICCV_2017_paper.html
    - http://openaccess.thecvf.com/content_CVPR_2019/papers/Liu_Adaptive_NMS_Refining_Pedestrian_Detection_in_a_Crowd_CVPR_2019_paper.pdf
- **Re-ID [Hao]**
  - Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline) (ECCV 2018)
- **Action recognition [Srijan]**
  - An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data (AAAI 2017)
    - Project (optional) - To implement the above framework and validate on a small dataset like MSRdailyActivity3D (Skeleton data will be provided)
    - Expected results - Classification accuracy, ablation studies, attention visualization
  - Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? (CVPR 2018)
    - Project (optional) - To implement the above framework and validate on a small dataset like MSRdailyActivity3D (RGB data will be provided)
    - Expected Results - Classification accuracy, analysis of the network (comparison with ResNet, DenseNet)
- **GANs [Yaohui]**
  - Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks (ICCV 2017)
  - Image-to-Image Translation with Conditional Adversarial Networks (CVPR2017)
- **Biometry [Antitza]**
  - Rössler et al. "Faceforensics++: Learning to detect manipulated facial images." ICCV, 2019.
  - Shi and Jain. DocFace+: ID Document to Selfie Matching. IEEE TRANS. ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE, 2019

References

- Marr, David. “ Vision", The MIT Press, 1982.
- Lowe, David. « Three-dimensional object recognition from single two-dimensional images » Artificial Intelligence, 1987.
- Viola, Paul and Michael, Jones. « Rapid object detection using a boosted cascade of simple features. » Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001.
- Lowe, David. « Distinctive image features from scale-invariant key points. » International Journal of Computer Vision, 2004.

References

- Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.
- Felzenszwalb, Pedro, David Mc Allester, and Deva Ramanan. "A discriminatively trained, multiscale, deformable part model." IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.
- Everingham, Mark, et al. "The pascal visual object classes (VOC) challenge." International Journal of Computer Vision 88.2 (2010):303-338.
- Deng, Jia, et al. "Image net : A large-scale hierarchical image database." IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.
- Lin, Yuanqing, e tal. "Large-scale image classification: fast feature extraction and SVM training." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Image net classification with deep convolutional neural networks."Advances in neural information processing systems. 2012.
- Szegedy, Christian, et al. "Going deeper with convolutions." arXiv preprint arXiv:1409.4842 (2014).
- Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition.“ arXiv preprint arXiv:1409.1556 (2014).
- LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11(1998) : 2278-2324.
- Fei-Fei, Li, et al. "What do we perceive in a glance of a real-world scene?” Journal of vision 7.1 (2007):10.