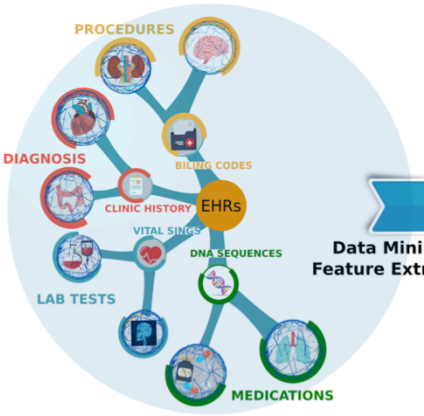
Scalability Analysis of Mini-Cluster Jetson TX2 for Training DNN Applied to Healthcare

Michel RIVEILL
John A. GARCÍA. H., Frédéric PRECIOSO, Pascal STACCINI

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis - I3S
Département d'Informations Médicales DIM – CHU Nice

Motivation: Health Care Decision-Making



Data Mining & Feature Extraction

General Patient Phenotype Representation

Predictive Modeling

Challenges

- *A common challenge in healthcare today is that physicians have access to massive amounts of data on patients, but have short time to analyze all of them.*
- *One limitation is that hospitals without robust computational systems for processing, storing and drawing conclusions requires to outsource the clinical tasks and that is a risk for privacy clinical data.*

Developing a Green Intelligence Medical System to derivate a patient representation for predict general medical targets and improving the computational resources usage.

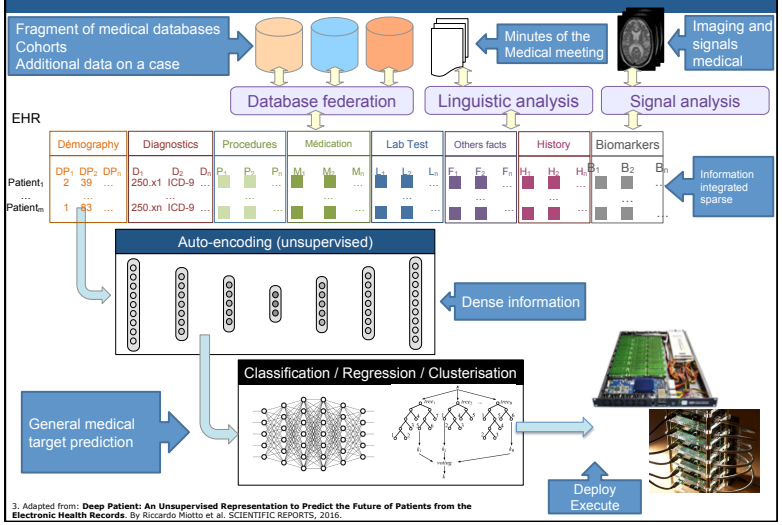
DiagnoseNET

Green Intelligence Medical System

Provides several high-level features:

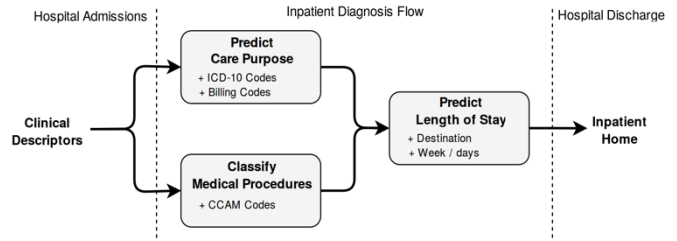
1) A framework for extracting the desired features from EHRs and encoding them
2) A framework to build full learning workflow (mainly related to DeepLearning)
3) A distributed processing building learning model on Jetson TX2 Mini-Clusters/ Array
4) An energy-monitoring tool for workload characterization.

DiagnoseNet theoritical workflow



3. Adapted from: **Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records**. By Riccardo Miotto et al. SCIENTIFIC REPORTS, 2016.

Case Study: Predict the Medical Future of Hospitalized Patients



	Diagnosis-related Group	ICD-10 Codes	Definition
Patient 1	Morbidity Principal	R402	Unspecified coma
	Etiology	I619	Nontraumatic intracerebral hemorrhage, unspecified
Medical Target	Care Purpose	Z515	Encounter for palliative care
Label used	Clinical Major Category	20	Palliative care
Patient 2	Morbidity Principal	R530	Neoplastic (malignant) relate fatigue
	Etiology	C20	Malignant neoplasm of rectum
Medical Target	Care Purpose	Z518	Encounter for other specified aftercare
Label used	Clinical Major Category	60	Other disorders

Mining Electronic Health Records

Data-mining: Feature Extraction From Electronic Health Records

Serialized each patient record in a clinical document architecture schema

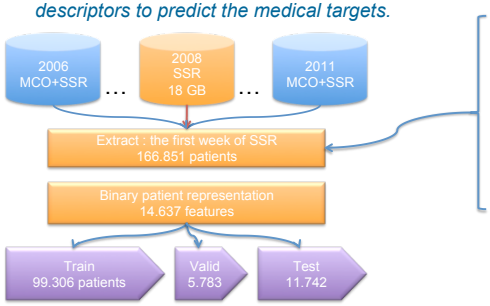
Patients	x1_demographics			x4_physical_dependance			x7_related_diagnoses		
	gender	...	age	feeding	...	displacement	Das1	...	Das 3
Patient 1	2	...	61	4	...	2	Z431	...	Z501
Patient 2	2	..	65	4	...	2	J459	...	F322
....	...	...	...	...	...	...	...	...	...
Patient m	1	...	95	1	...	2	C259	...	F322

Build a binary patient phenotype representation from their features selected

Patients	x1 :demographics			x4 :physical_dependance			x7 :related_diagnoses		
	[1 :male]	[2 :female]	60-74	[4 :Assistance]		[2 :normal_transfer]	Z431	...	F322
Patient 1	0	1	1	1	...	1	1	...	0
Patient 2	0	1	1	1	...	1	0	...	1
....	...	...	...	...	...	...	...	...	...
Patient m	1	0	0	0	...	1	0	...	1

PMSI-PACA Clinical Dataset

*As Input we are using **result features** that describe the patient clinical descriptors to predict the medical targets.*



PMSI SSR Clinical descriptors: (change over the year)

+ Demographics
+ Admission details
+ Hospitalization detail
+ Physical dependence
+ Cognitive dependence
+ Rehabilitation time
+ Morbidity
+ Comorbilities
+ Etiology

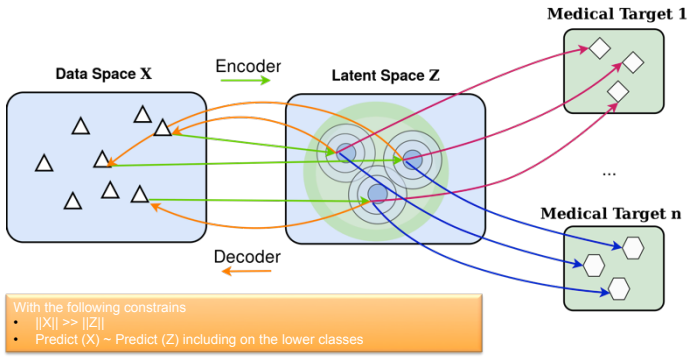
TO DO.
Manque la taille des train set en représentation binaire

Unsupervised Patient Phenotype Representation

Methodology: Unsupervised Patient Phenotype Representation

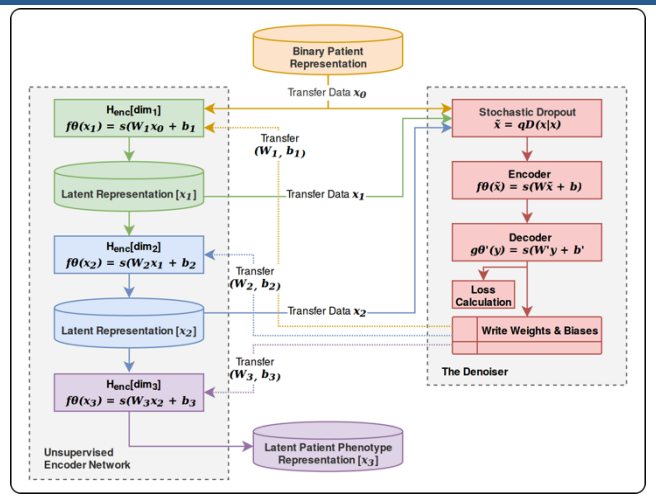
The task:

+ From a binary patient representation **{X}** derive a latent patient representation **{Z}.**
+ Using the general representation plus a supervised learning algorithms for predict different medical targets, eventually with different methods.



With the following constrains
- ||X|| >> ||Z||
- Predict (X) ~ Predict (Z) including on the lower classes

Unsupervised Learning Representation



4. Adapted from **Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion.** By Pascal Vincent et al. 2010

Experiment Analysis

Number of Gradient Updates as Factor to Early Model Convergence.

Number of Gradient Updates as Factor to Early Model Convergence.

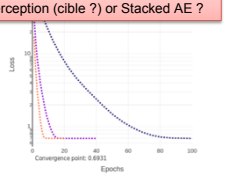
GPU NVIDIA GTX Titan X

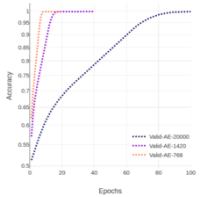
Trainset: 99.306 patients * 14.637 features → XX Go
Valset : 5.783 patients * 14.637 features → XX Go

MLP Perceptron (2048 / 2048 / 768 + relu + adam + lr=0.0001 + drop=0.5)

TO DO
On est sur l'auto encodeur ?
Quel est le nombre de feature cible ?
MLP Perception (cible ?) or Stacked AE ?

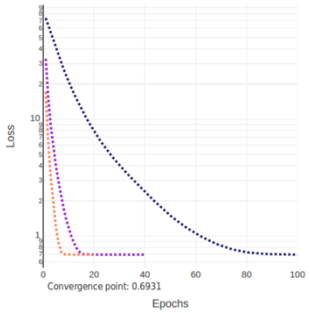
Batch size To reach convergence point	20 000	1 420	768
Epochs	100	20	10
Gradient updates	400	1 050	1 100
Times	2176 s	476 s	266 s
Energy	137.65 Kj	41.26 Kj	21.87 Kj
Power	63.25 W	86.61	82.21

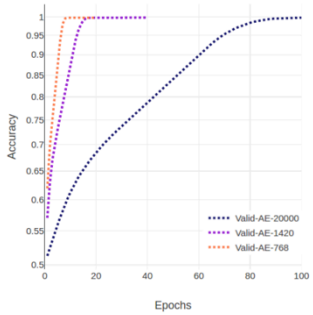




1) Number of Gradient Updates as Factor to Early Model

- *Network convergence using batch partitions of [20000, 1420, 768] records to generate [4, 59, 110] gradient updates by epoch respectively.*

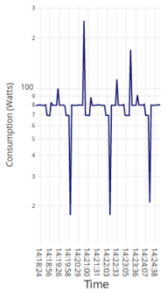




	1-Layer1	2-Layer2	3-Layer3	4-Activation_funct	5-GD_Optimizer	6-Learning_rate	7-Dropout-rate
0	2048	2048	768	relu	adam	0.0001	0.5

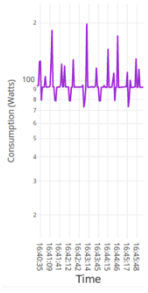
Power consumption in a window of 6 minutes

BF: 20.000
100 Epochs
EC: 137.65 Kj



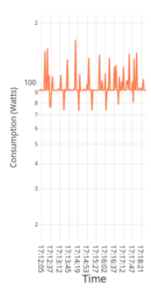
63.35 *Watts* in average to process 68 gradient updates in 17 epochs.

BF: 1420
40 Epochs
EC: 41.26 Kj



86.61 *Watts* in average to process 885 gradient updates in 15 epochs.

BF: 768
20 Epochs
EC: 21.87 Kj



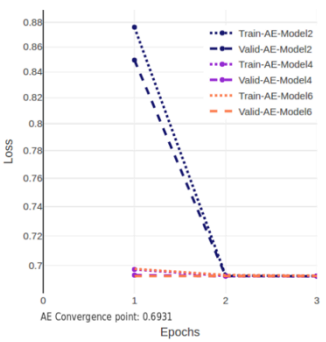
82.21 *Watts* in average to process 1540 gradient updates in 14 epochs.

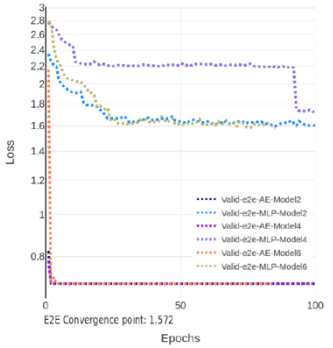
Experiment Analysis

Model Dimensionality as Factor to Generate Quality Latent Representation

Model Dimensionality as Factor to Generate Quality Latent Representation

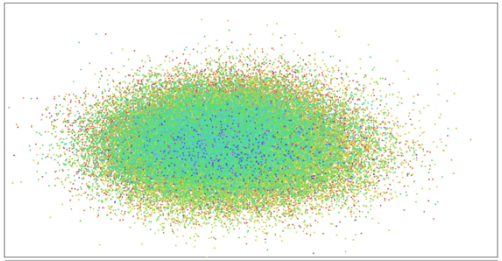
- *Comparison of different model dimensionality using relu as function to generate the latent representation.*



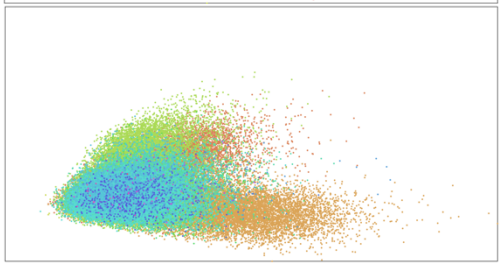


Model Dimensionality as Factor to Generate Quality Latent Representation

Autoencoders:

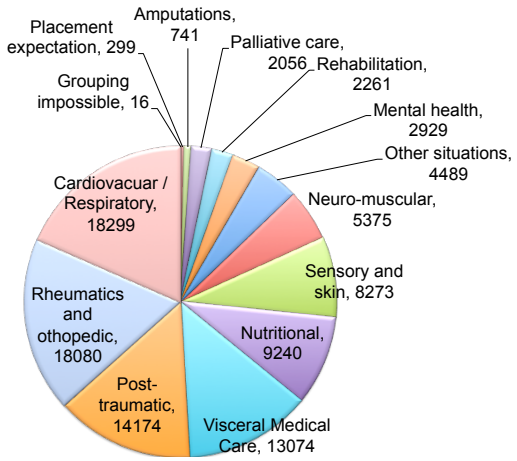


End to End:

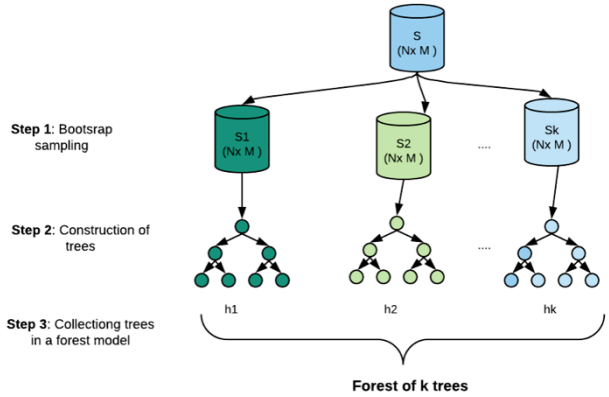


Supervised Learning

Medical Target 1: Care Purpose Description Labels

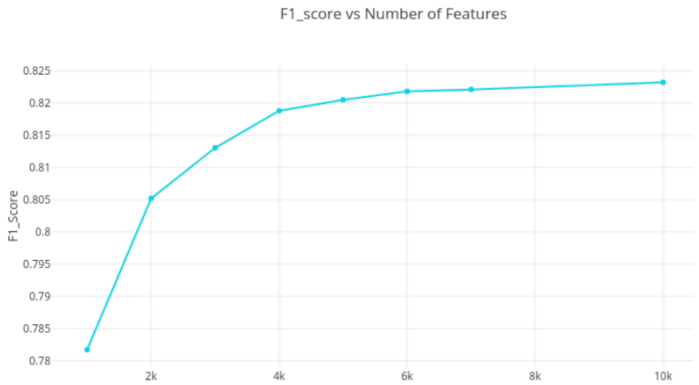


Machine Learning Algorithm: Random Forest



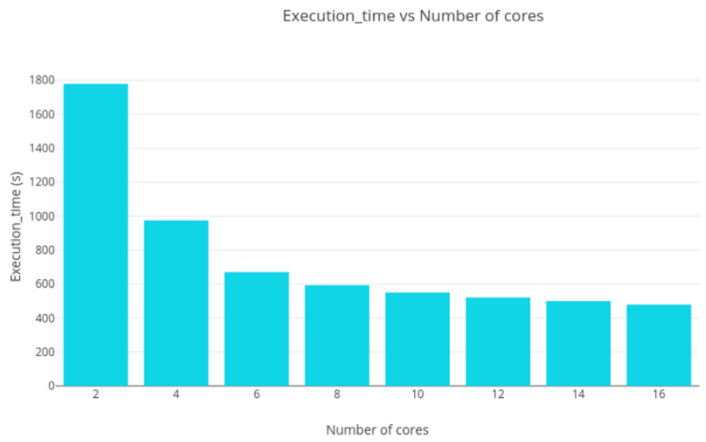
Ensemble Algorithm Based on Decision Tree Model

Random Forest: F1-Score for Different Number of Features Scales

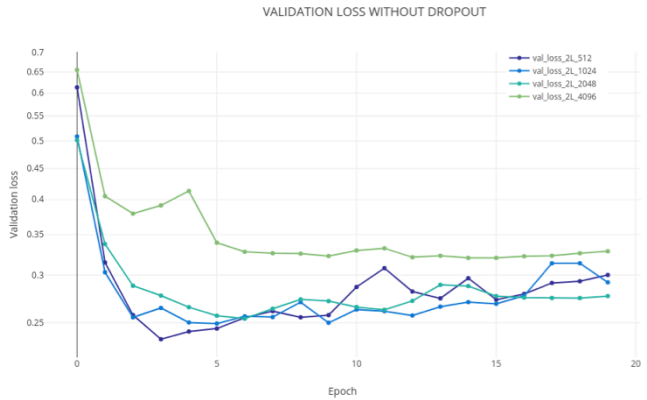


Random Forest: F1-Score for Different Number of Features Scales

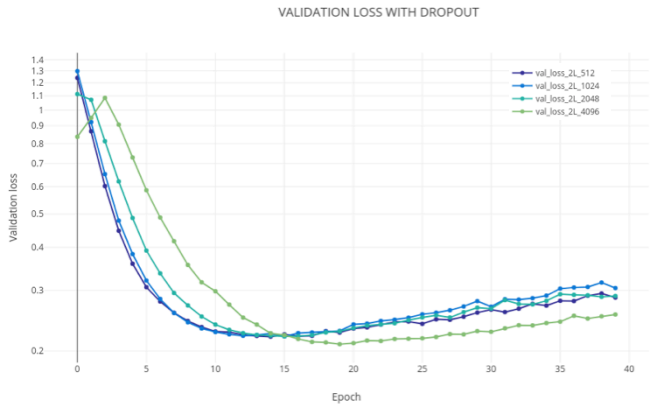
Random Forest: F1-Score for Different Number of Features Scales



Experiments: Feed-forward Multilayer Perceptron



Experiments: Feed-forward Multilayer Perceptron



Classification using a Feed-forward Multilayer Perceptron

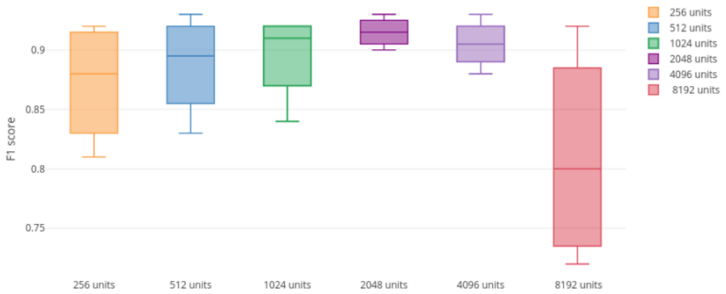
For similar F1 score, generally
- The energy consumption is increasing
- when the number of units increase

Architecture	F1 Score	Exec. time sec	Energy Kj
256 units on 2 layers	0.92	686	59.97
2048 units on 8 layers	0.91	654	66.49
8192 units on 2 layers	0.92	1,108	238.06

For the same number of neurones in the hidden layers (here 9,192 neurones)
- When the number of layer is increasing
- The F1 score decrease
- And the energy consumtion decrease also

Nb units: 9,192 Distributed on	F1 Score	Exec. time sec	Energy Kj
2 layers	0.92	1,108	238.06
4 layers	0.85	934	161.74
8 layers	0.72	793	124.04
16 layers	0.75	693	90.74

Classification using a Feed-forward Multilayer Perceptron



The lower value of F1, is generally for bigger number of layer.

The strategy for saving energy

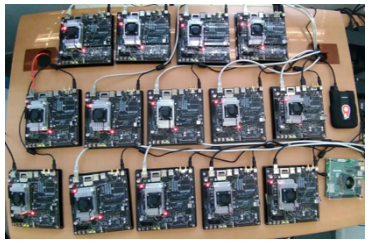
- Choose the good number of neurones in hidden layer by using many layer (8 or 16)
- When the number of neurones are fixed choose the number of layer to have a good compromise between F1 score and energy consumed

Performance Results By class to Classify the Medical Target 1

Classes	True positives	False positives	False negatives	precision	recall	f1 score	occurence de la classe
0	424	54	122	0.89	0.78	0.83	546
1	2089	136	59	0.94	0.97	0.96	2148
2	1382	98	79	0.93	0.95	0.94	1461
3	598	72	34	0.89	0.95	0.92	632
4	211	73	153	0.74	0.58	0.65	364
5	861	136	141	0.86	0.86	0.86	1002
6	2086	96	105	0.96	0.95	0.95	2191
7	1574	115	74	0.93	0.96	0.94	1648
8	76	9	10	0.89	0.88	0.89	86
9	101	74	122	0.58	0.45	0.51	223
10	36	1	3	0.97	0.92	0.95	39
11	275	31	20	0.90	0.93	0.92	295
12	1088	44	16	0.96	0.99	0.97	1104
13	0	2	3	0.0	0.0	0.0	3

Distributed Processing for Training DNN on Jetson TX2 Mini-Clusters

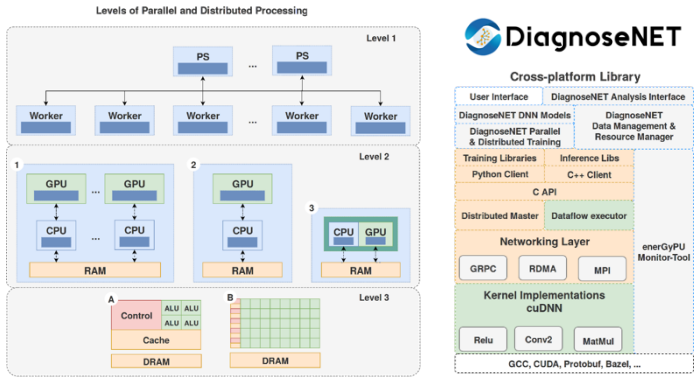
Computational Resources



Mini-Cluster with 14 Jetson TX2
(Distributed Memory)

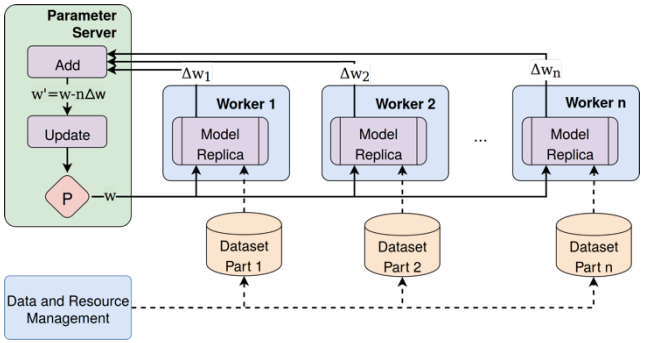
Array Node with 24 Jetson TX2
(Hybrid Memory)

Develop DiagnoseNET for Training Large-Scale DNN on Distributed Systems



6. Adapted from Snap Machine Learning. By IBM Research et al. 2018
7. Adapted from TensorFlow Architecture. By Google Research.

Task-Based Data Parallelism: Synchronous

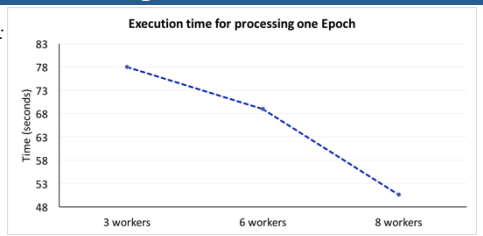


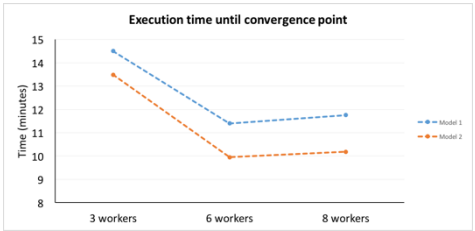
(Data Size):
+ Setting the number of workers and micro batch.
+ Fine-tuning DNN hyperparameters.
+ Speeds up the training.
+ I/O Intensive.

8. Adapted from TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. By Google Research. 2015.

2) Preliminary Results to Scale the Unsupervised Representation Learning

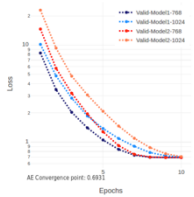
Preliminary results using: 10.000 records and 11.466 features.



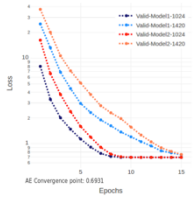


3) Number of Workers and Task Granularity as Factor to Early Model Convergence

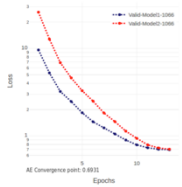
- *Early convergence comparison between different groups of workers and task granularity for distributed training with 10.000 records and 11.466 features.*



1.30 mins in avergange for processing one epoch on 1 PS 3 workers.



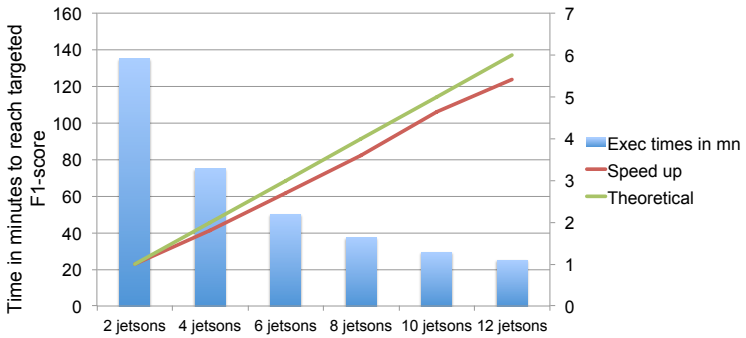
1 min in avergange for processing one epoch on 1 PS 6 workers.



50.6 secs in avergange for processing one epoch on 1 PS 8 workers.

Preliminary Results to Scale the Feed-forward Multilayer Perceptron

- **Jetson:** 8 Gb of disk, direct transert from disk to GPU memory
- **Network architecture:** 8-Layers Model with 256 neurons per layer
- **Task:** classification from binary representation → 6 Go
  (116,851 patients / 14,637 features)



Conclusion

Latent representation:
- Reduces the number of sparse features without loos of precision in future classification
- Reduces training time (41 %)

Use the unsupervised embedding stage to create a new lower dimensional patient representation, reduces the number of sparse features to classify at stage 3. In which, the execution time for training is minimized by 41% with regard to BPPR and the precision to classify the first medical target is almost equal.

Data partitioned on different Jetsons + small batch =
→ frequent gradient number update
→ early model convergence
→minimizes energy consumption

DiagnoseNET: Green Intelligence System

Process

1. Select optimal computational resources and make good mapping of task granularity for training one model in less time and less power consumption give a mini-batch size factor.
2. Minimize the number of different trained models to converge the optimal generalization-accuracy model.
3. Management the queue of models to training and determine optimal combination of computational resources to use in each model training.

Next work:
- Distibute others kind of DL architecture (CNN or recurrent neurones) or random forest architecture
- Compare several architecture :
  - multi-GPU (share memory)
  - vs Cluster (distributed memory)
  - vs Array (hybrid memory)
- For several task
  - **MT-1:** Predict the '**Major Clinical Category** of patients' (coarse grain CMC / fin grain GHJ) from inpatients features recorded at the admission time
  - **MT-2:** Predict the ***'Clinical Procedures'*** from inpatients features recorded at the admission time and the Primary Morbidity
  - **MT-3:** Predict the ***'Inpatient Destination'*** (home, transfer, death) and length of hospitalization stay from inpatients features recorded at the admission time and Primary Morbidity and Clinical Procedures

Gracias
Por su
Atención

Scalability Analysis of Mini-Cluster Jetson TX2 for Training DNN Applied to Healthcare

Michel RIVEILL
John A. GARCÍA. H., Frédéric PRECIOSO, Pascal STACCINI

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis - I3S
Département d'Informations Médicales DIM – CHU Nice

Future Work

Evaluate the DNN approaches using the different platform such as, cluster Jetson TX2, a multiGPU Node with 8 GPUs and the array Node with 24 Jetson TX2.

1. Port the framework DiagnoseNET to array Node.
2. Integrate the communication measures with the energy monitor on distributed and Hybrid platform.
3. Perform the different experiments to evaluate the case studies on the different platform