CRÉATION DU PÔLE

MACHINE LEARNING / DEEP LEARNING

AU CeSAM

Morgan Gray, Laurent Jorda, Jean-Charles Lambert,
Jean-Charles Meunier, Christian Surace, Didier Vibert

LAM, 9 octobre 2020

THÉMATIQUES ABORDÉES

- Missions du pôle
- Personnels du pôle
- Ressources en matériel informatique
- Brèves définitions de Machine Learning / Deep Learning
- Anatomie d'un réseau de neurones
- Deux exemples de développements utiles aux projets
- Besoins / Attentes (à compléter lors les présentations)

MISSIONS DU PÔLE

- **Apporter une expertise**

  sur les méthodes de Machine Learning (ML) et de Deep Learning (DL)
  pour les ***problématiques astrophysiques et instrumentales***,
  au sein des ***différents projets et équipes*** du LAM :
  - développement de codes, suivi & conseil,
  - ressources personnels (permanents & CDD au CeSAM),
  - ressources informatiques (GPU sur les clusters du LAM & Université)

- **Faciliter les interactions**

  entre les équipes *au sein même du LAM*,
  et **établir des liens** *avec les équipes d'autres laboratoires,*
  travaillant sur ces méthodes (QARMA du LIS, Univ. Montpellier, Trent Univ.,...)

- **Animer des rencontres**

  sur des **thématiques relatives au ML/DL**
  avec des *personnes du LAM* **et** des *personnes extérieures* :
  - un café-club au LAM
  - organisation d'une journée annuelle et de séminaires au LAM

PERSONNELS CeSAM

- **Morgan Gray** IR, responsable pôle ML / DL (***100 %***)
  Développement de codes numériques (Scikit-Learn / Keras / Tensorflow)
  APPLY (B. Neichel) ; EUCLID (V. Lebrun) ; BigSF (A. Zavagno) ; DEEPDIP (S. Arnouts)

- **Jean-Charles Lambert** IR, responsable pôle Infrastructure (***5 %***)
  Développement & Maintenance des GPUs du cluster de calcul du LAM

- **Jean-Charles Meunier** IR, technologies du BigData (***à terme 80 %***)
  Système de gestion de données distribuées, visualisation de données massives, MapReduce *pour la distribution massive de traitements*...

- **Didier Vibert** IR, responsable pôle TARDIS (***10 %***)
  EUCLID (V. Lebrun) ; DEEPDIP (S. Arnouts)

- **Christian Surace** IR, responsable du CeSAM (***à terme 50%***)

- **Laurent Jorda** Chercheur, responsable scientifique du CeSAM

- **Recrutement IR BAP E** *demande en cours*
  - simulations numériques sur les projets du LAM (**50 %)**
  - support aux activités ML / DL (**50 %**)
- **Collaborateurs extérieurs**

- François-Xavier Dupé (LIS-QARMA) - Sabine Mc Connel (Trent University)

RESSOURCES MATÉRIEL INFORMATIQUE

► Cluster du LAM

3 noeuds dédiés spécifiquement aux calculs sur GPUs (8 au total)

noeud 38

1 GPU **nVidia TitanXP**

1 GPU **nVidia RTX2080Ti** avec 20 coeurs + 128 GB de RAM

noeud 47 (*acheté* par le CeSAM + participation APPLY)

3 GPUs **nVidia RTX2080Ti**

avec 20 coeurs + 180 GB de RAM

noeud 50 (*acheté* par A*MIDEX DeepDIP)

3 GPUs **nVidia RTX2080Ti**

avec 20 coeurs + 180 GB de RAM

Bien respecter le protocole demandé pour lancer les jobs !

https://projets.lam.fr/projects/cluster-de-calcul-du-lam/wiki#GPU-partition

► Évolutions possibles

suivant **les besoins futurs en calcul** des différents projets

- **nouveau noeud (3 GPUs) sur le cluster du LAM**

pour le *développement de codes* (participation financières des équipes)

- **utilisation des GPUs du Mésocentre** (Univ. Aix-Marseille)

pour la *production de simulations* (demande de temps de calcul motivée)

MACHINE LEARNING / DEEP LEARNING

- **Paradigme CLASSIQUE** de programmation

  Règles (programme) **+** ***Données*** → ***Réponses*** à une problématique

- **NOUVEAU paradigme** de programmation

  1) ***Données*** **+** ***Réponses*** (connues / attendues) **+** ***Mesure*** *(math.)* → **Règles**

  2) ***Nouvelles Données*** **+** ***Règles précédentes*** → **Réponses** originales

- Algorithme de **Machine Learning**

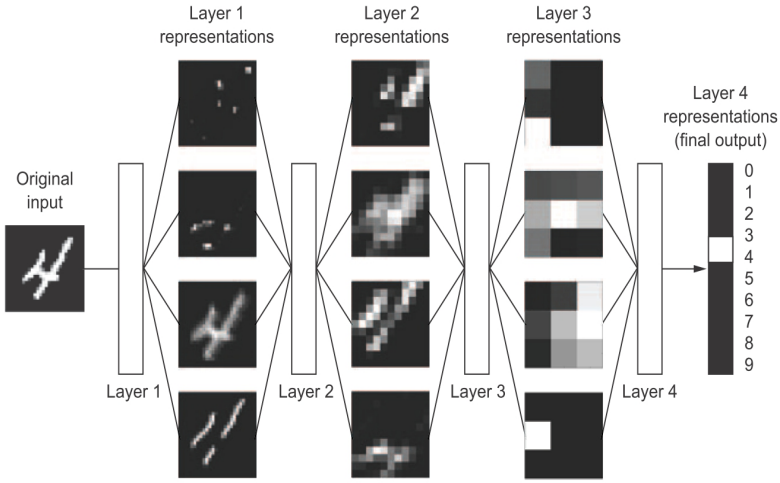
est entraîné
- *produire des structures statistiques* dans des données d'apprentissage
- *permettre de définir des règles* pour une future automatisation de taches

utilise
- *transformations (linéaires ou non)* effectuées ***sur les données***,
- *représentations plus adéquates* des données ***dans un autre espace*** *(math.)*

Apprentissage **Supervisé** *versus* Apprentissage **Non Supervisé**

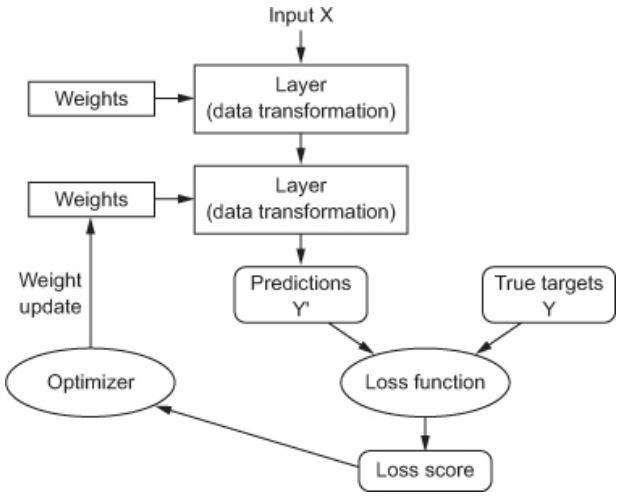
- **Deep Learning** (sous-ensemble du ML)

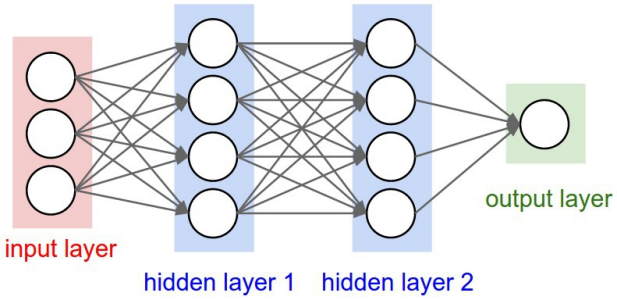
utilise des représentations extraites des données, reposant sur un apprentissage par *des couches successives permettant d'extraire des informations de plus en plus 'significatives'*



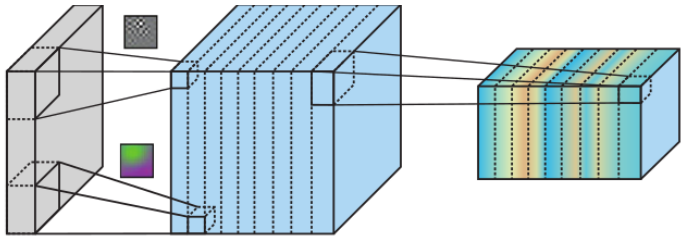
EXTRACTION AUTOMATIQUE DES DESCRIPTEURS PERTINENTS

ANATOMIE D' UN RÉSEAU DE NEURONES





Couches Denses



Couche de Convolution

- **Couche (dense / de convolution)** Tenseur → Tenseur
  # neurones, poids, # filtres, kernels, strides, padding, fonctions d'activation, initialiseurs…
- **Fonction de Perte** avec Labels "Prédits" & Labels "Vrais"
  calcul d'une Distance / Mesure de perte
- **Optimizer** Algorithme de rétroprogation du gradient des erreurs
  permet de *déterminer* les valeurs des paramètres du réseau *minimisant* la Fonction de Perte
- **Métrique** évaluation de l'évolution des performances
- **Modèle** architecture du réseau de neurones & choix des différents paramétrages

ENTRAÎNER le modèle d'un réseau de neurones

Répéter un "nombre suffisant'' de fois (# d'époques) **la boucle d'apprentissage**,
- *en propageant* des batchs (petites parties) d'un ensemble d'apprentissage (Training Set)
- *en minimisant* la fonction perte,

⇒ **détermination les valeurs des paramètres du réseau**

DEUX EXEMPLES UTILES AUX PROJETS

▶ Traitement des données à la volée

Problème Taille & Nombre des images (plusieurs centaines de milliers) × différentes bandes
+ paramètres du réseau de neurones (forward & back propagations) **+** codage 32 bits...
- **toutes** ces données doivent être **dans la RAM** pour l'étape d'apprentissage
- ***quantité nécessaire de RAM trop importante*** pour certaines simulations

Solution ***Ne charger*** dans la RAM ***que les données de chaque batch***
puis itérer sur tous les batchs composant l'ensemble d'apprentissage

Développement d'un système de chargement itératif & de gestion des images de chaque batch, sur la totalité de l'ensemble d'apprentissage réalisant une époque

▶ Parallélisation des calculs sur plusieurs GPUs d'un même nœud

Problème Même configuration que l'exemple précédent
- ***temps de calcul trop long*** pour les simulations successives nécessaires à la recherche itérative des meilleurs hyper-paramètres du réseau de neurones

Solution **Image miroir** du réseau sur plusieurs GPUs + calculs des différents batchs sur chaque GPU + **mise à jour synchrone** des paramètres à chaque itération

Développement d'une parallélisation synchrone des calculs avec les batchs sur plusieurs GPUs d'un même noeud

SUJETS POUR LA DISCUSSION

Bibliographie sommaire (***mon*** top 3)

- Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow
  d'Aurélien Géron (O'Reilley editor, second edition)
- Deep Learning with Python de François Chollet (Manning editor)
- Deep Learning de Goodfellow, Bengio, Courville (MIT press book editor)

Vos attentes auprès du pôle ML / DL du CeSAM : idées issues des discussions

- Animations au sein du LAM
  - **Formations** pour des connaissances de base
  - Support pour des encadrement des personnes recrutées
  - **Suivi des projets**
  - Suivi des participations aux encadrements de thèse entre les différentes personnes aux « langages » différents
  - **Nouvelles ressources** sur le cluster
  - **Journal club** pour partager les problématiques sur des sujets qui peuvent se recouper

- Interactions
  - Avec les laboratoires extérieurs suivant les thématiques des différents projets

- Connaissances / Expertises
  - **Ne pas perdre les expertises développées** par les thésards / Postdocs pour une utilisation ultérieure
  - Pouvoir proposer les outils développés pour d'autres projets...
  - Points de vue extérieurs sur les algorithmes