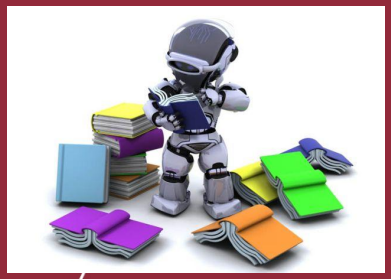


ATELIER 100% DATASCIENCE A PARIS LE 16 NOVEMBRE 2017

INSTITUT DES ACTUAIRES

LE DEEP LEARNING AU SERVICE DE L'ASSUREUR NON-VIE :

Automatisation de la chaîne de confiance lors de déclaration de sinistre auto

100% Datascience

INSTITUT DES ACTUAIRES

1/ Présentation : par Imen et Nathalie

- Création du Club Algo, objectifs et présentation du sujet (Nathalie)
- Principes des Réseaux de Neurones et introduction au CNN (Imen)

2/ Application des CNN à la reconnaissance marque et modèle de voiture : par Victor et Arthur

- Présentation de la méthodologie retenue
- Résultats obtenus

3/ Questions

100% Datascience

INSTITUT DES ACTUAIRES

1/ Présentation du Club Algo : par Imen et Nathalie

- Création, objectifs et présentation du sujet (Nathalie)
- Principes des Réseaux de Neurones et introduction au CNN (Imen)

2/ Application des CNN à la reconnaissance marque et modèle de voiture : par Victor et Arthur

- Présentation de la méthodologie retenue
- Résultats obtenus

3/ Questions

Présentation du Club Algo

- Lancé en juin 2016, dans la continuité de la formation DSA
- Le Club est un sous-groupe du « GT Big Data », de l'IA, créé en décembre 2013
- Objectif : Échanges et perfectionnement dans le domaine algorithmique pour l'assurance
- En 2016 : Cinq « uses cases »
- En 2017 : Études des algos de deep learning
- 32 membres

100% Datascience

INSTITUT DES ACTUAIRES

1/ Présentation du Club Algo : par Imen et Nathalie

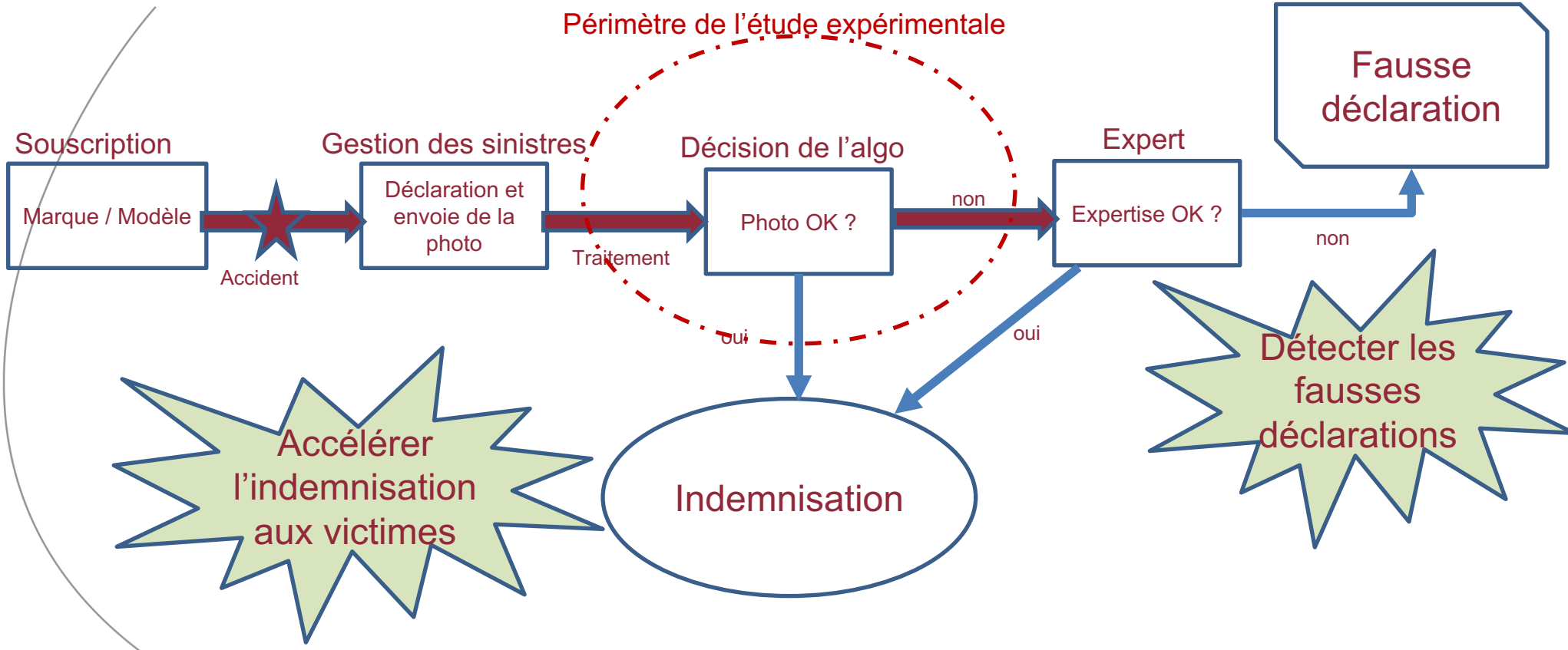
- Création, objectifs et présentation du sujet (Nathalie)
- Principes des Réseaux de Neurones et introduction au CNN (Imen)

2/ Application des CNN à la reconnaissance marque et modèle de voiture : par Victor et Arthur

- Présentation de la méthodologie retenue
- Résultats obtenus

3/ Questions

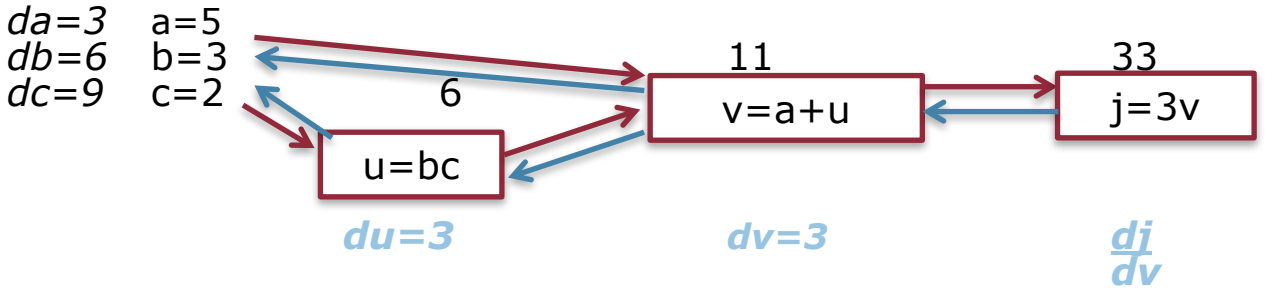
Automatisation de la chaîne de confiance



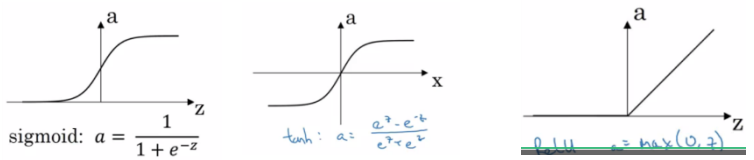
Bases de la programmation d'un réseau de neurones
(source Coursera « NN and Deep Learning d'Andrew NG)

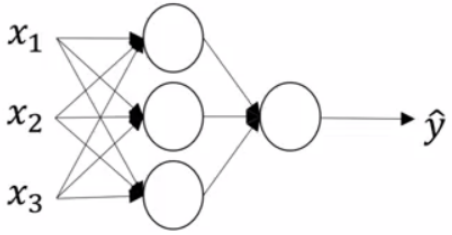
- Notions :
  - Classification binaire, régression logistique et fonction de perte : descente de gradient, calcul des dérivées et fonction d'activation

  - Propagation et rétropropagation :



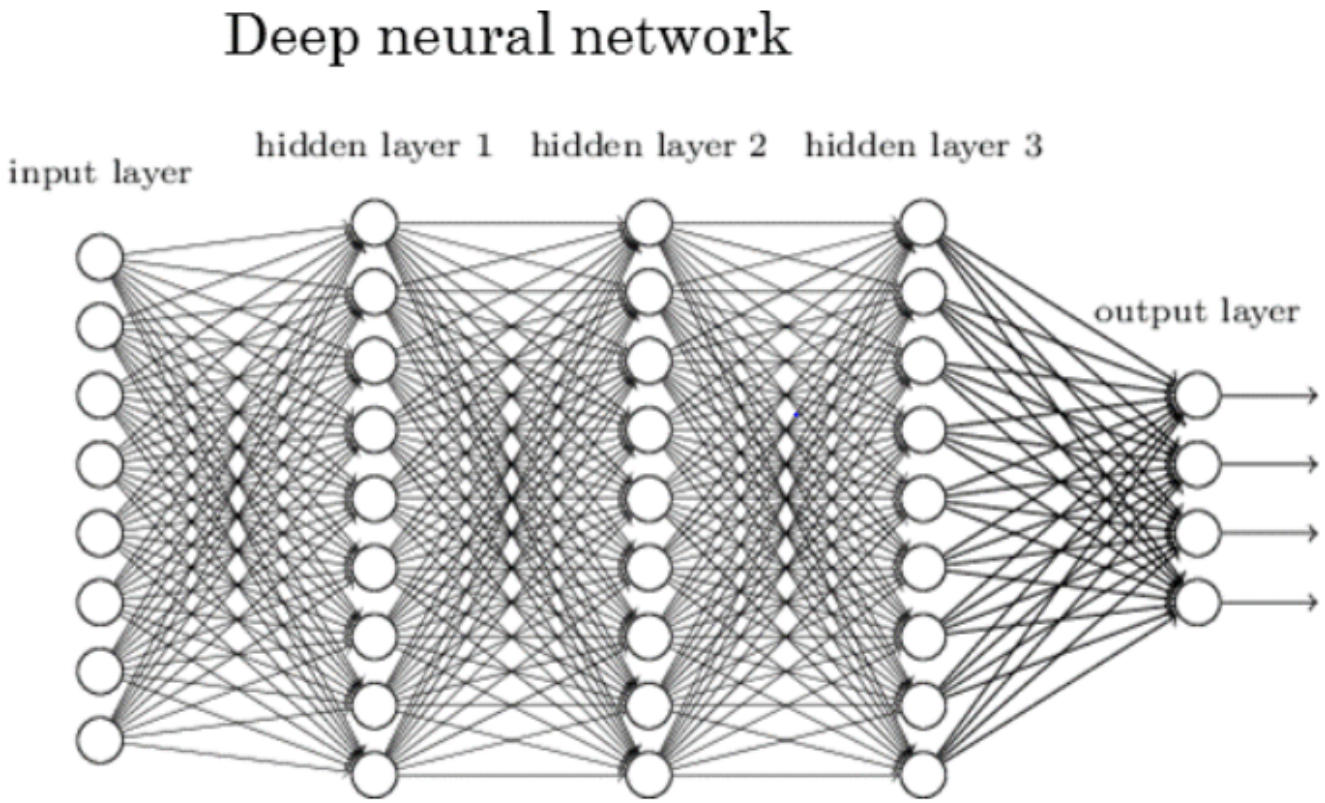
  - Fonctions d'activation : Sigmoïd, tanh, reLu





  - Quelles fonctions d'activation pour quel usage ?

Cas de plusieurs couches d'apprentissage « profond »



100% DataScience

INSTITUT DES ACTUAIRES

1/ Présentation : par Imen et Nathalie

- Création du Club Algo, objectifs et présentation du sujet (Nathalie)
- Principes des Réseaux de Neurones et introduction aux CNN (Imen)

2/ Application des CNN à la reconnaissance marque et modèle de voiture : par Victor et Arthur

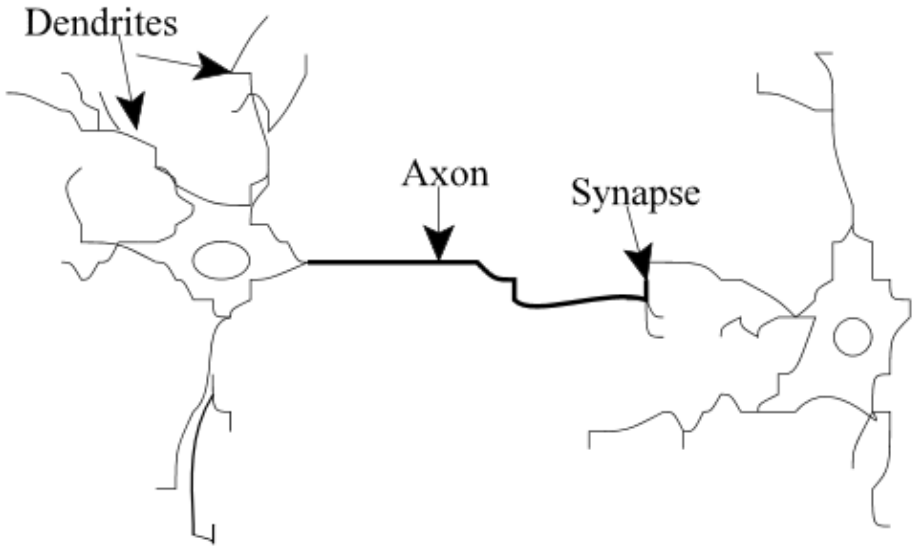
- Présentation de la méthodologie
- Résultats obtenus

3/ Questions

INSTITUT DES ACTUAIRES

Réseaux de neurones : retour sur les principes

- Un réseau de neurone est défini par :
  - Des nodules (node) : entité de computation
  - Des connections entre les nodules : flux d'informations entre les nodules



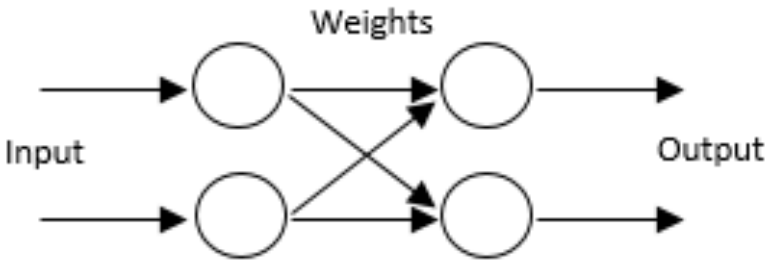
INSTITUT DES ACTUAIRES

Réseaux de neurones : retour sur les principes

- Les réseaux de neurones artificiels reprennent le même principe :
  - Fonction d'activation
  - Fonction output

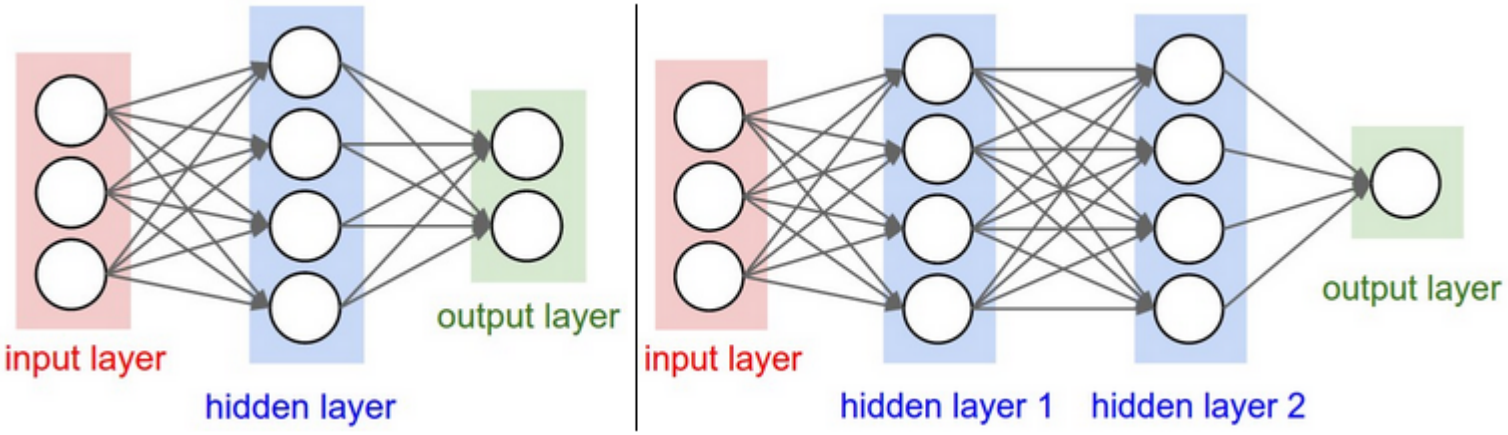
- Exemple basique : the backpropagation algorithm (Rumelhart and McClelland, 1986) Ajuster les poids pour diminuer l'erreur par la méthode du *gradient descendent*

$$\begin{cases} A_j(\bar{x},\bar{w}) = \sum_{i=0}^{n} x_i w_{ji} \\ O_j(\bar{x},\bar{w}) = \frac{1}{1+e^{A_j(\bar{x},\bar{w})}} \\ E(\bar{x},\bar{w},\bar{d}) = \sum_j (O_j(\bar{x},\bar{w}) - d)^2 \end{cases}$$



Réseaux de neurones : retour sur les principes

- Exemple d'architecture d'un réseau : Fully-connected layer



- 26 paramètres pour le premier et 41 pour le deuxième !

INSTITUT DES ACTUAIRES

Réseaux de neurones : retour sur les principes

- Défi : comment paramétrer le réseau ?
  - Nombre de layers
  - Taille des layers
- Risque : sur apprentissage
- Réponse classique : diminuer le nombre de layers
- Spécificité des réseaux de neuronnes : autres méthodes pour traiter le bruit
  - Le dropout : probabilité de récupérer l'output à 50%
  - Permet de changer l'architecture de l'input

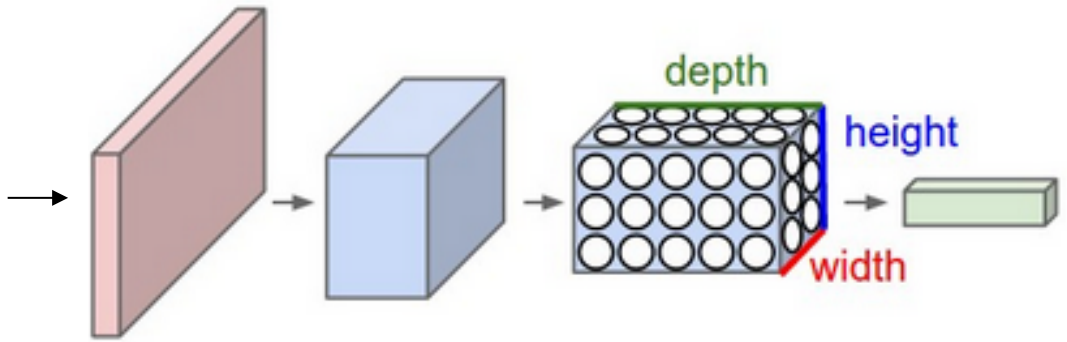
CNNs/ ConvNets :

- Même structure qu'un réseau de neurone classique : poids, fonction output, fonction d'activation, fonction de perte, optimisation par gradient descendent
- Hypothèse principale d'un CNN : l'input est une image
- L'image est un tableau 3D : width x height x RGB : explosion du nombre de paramètres pour un réseau classique
- Dans le CNN, la neurone est 3D

➔ L'importance de l'architecture de la donnée (tensorflow et theano)

- Schéma d'un réseau ConvNet :

Les dimensions de l'input correspondent aux dimensions de l'image



- Paramètres :
  - Dimension de l'image (input)
  - Nombre de filtres
  - Nombre de layers

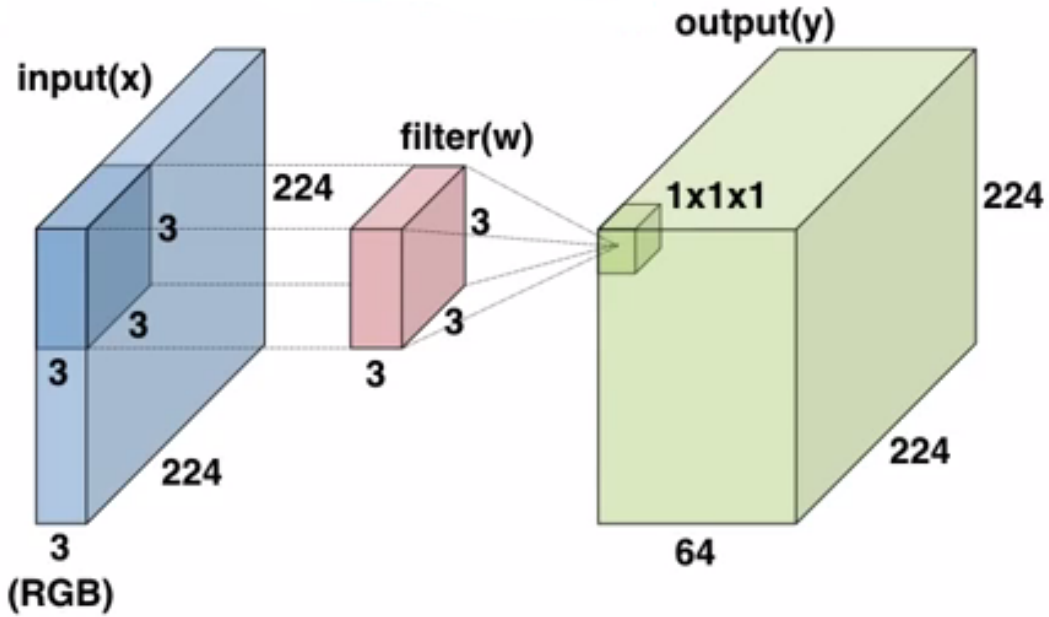
CNNs/ ConvNets :

- Input : image représentée avec 3 dimensions + nombre de filtres à appliquer

Exemple : dimensions 224x 224 x RGB (3) +12 filtres

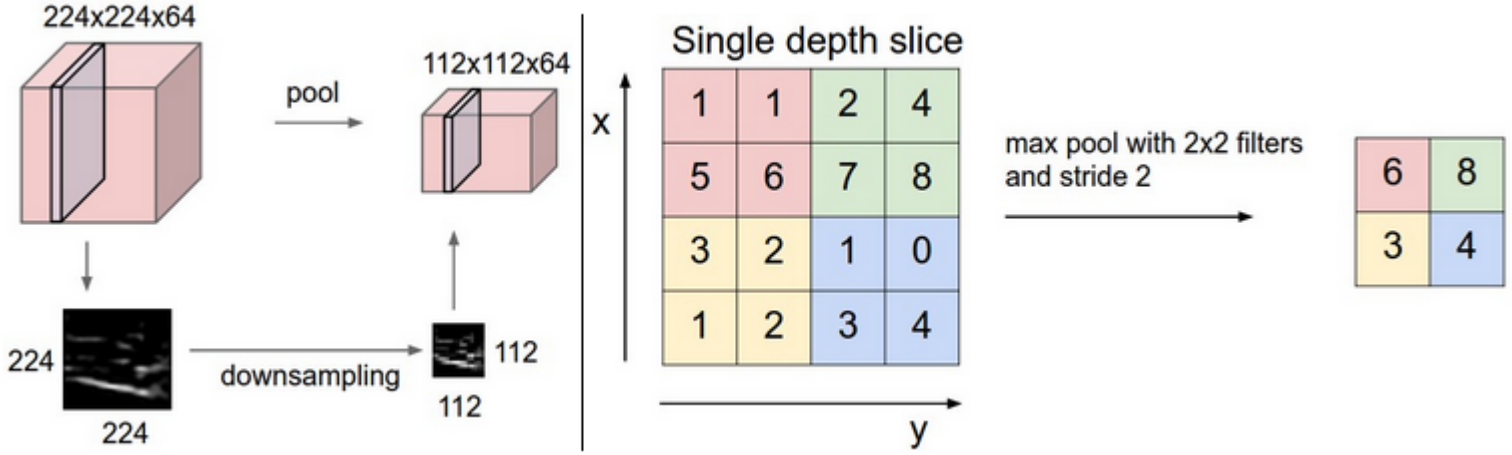
- Convolution layer : l'étape la plus importante dans le réseau.
  - On glisse le filtre tout au long de l'image input de façon à couvrir toute la surface ➔ on définit les paramètres de transition des filtres :
    - » Depth : égal au nombre de filtres
    - » Stride : le pas de transition
    - » Zero padding : ajouter des zéros sur la marge
  - A chaque fois, un output correspondant à la zone géographique est computé et une neurone multidimensionnelle est crée
  - Chaque zone géographique sera définie par un vecteur de neurones 3D
  - Contrairement au FC layer, les neurones ne dépendent que d'une seule zone géographique

CNNs/ ConvNets :



Pooling layer :

- Entre deux couches de convolution, effectuer du pooling pour :
  - Réduire la dimensionnalité : downsampling
  - Par conséquent réduire le nombre de paramètres et le risque du surapprentissage



100% DataScience

INSTITUT DES ACTUAIRES

1/ Présentation : par Imen et Nathalie

- Création du Club Algo, objectifs et présentation du sujet (Nathalie)
- Principes des Réseaux de Neurones et introduction au CNN (Imen)

2/ Application des CNN à la reconnaissance marque et modèle de voiture : par Victor et Arthur

- Présentation de la méthodologie retenue
- Résultats obtenus

3/ Questions

Base de données utilisée

- La base de données initiale est constituée de **100,000** photos de voitures, de la marque et du modèle associé
- Seuls les modèles pour lesquels plus de 100 photos étaient présentes dans la base ont été conservés ➔ conservation d'environ 150 modèles de voiture
- La base sur laquelle ont été lancés les réseaux de neurones est finalement composée de **60,000** observations
- Nous avons récupéré des listes de plusieurs millions de photos pour affiner les détections et classifications

Sourcing des photos

- Récupération de photos publiques provenant de sites de ventes en ligne
  - La centrale
  - Le bon coin
  - watchmycar.fr (site fiabilisant les transactions en demandant des photos sous tous les angles)

- Récupération de données faiblement labellisées :
  - Utilisation de la technologie Velours pour corriger certains labels
  - Et pour avoir une bonne orientation des véhicules

Technologie utilisée

- Les calculs ont été réalisés en utilisant le package Caffe
- Des tests ont également été effectués en utilisant TensorFlow et Keras
- Utilisation d'une carte graphique Titan X de 2015 avec 2880 cœurs et 12 GB de mémoire vive sur un serveur assez basique (4 coeurs et 8GB de mémoire vive)
- Les apprentissages ont généralement pris quelques heures

Méthodologie retenue pour la prédiction des modèles de voiture

Prise en compte la qualité des photos dans la prédiction du modèle grâce à deux étapes :

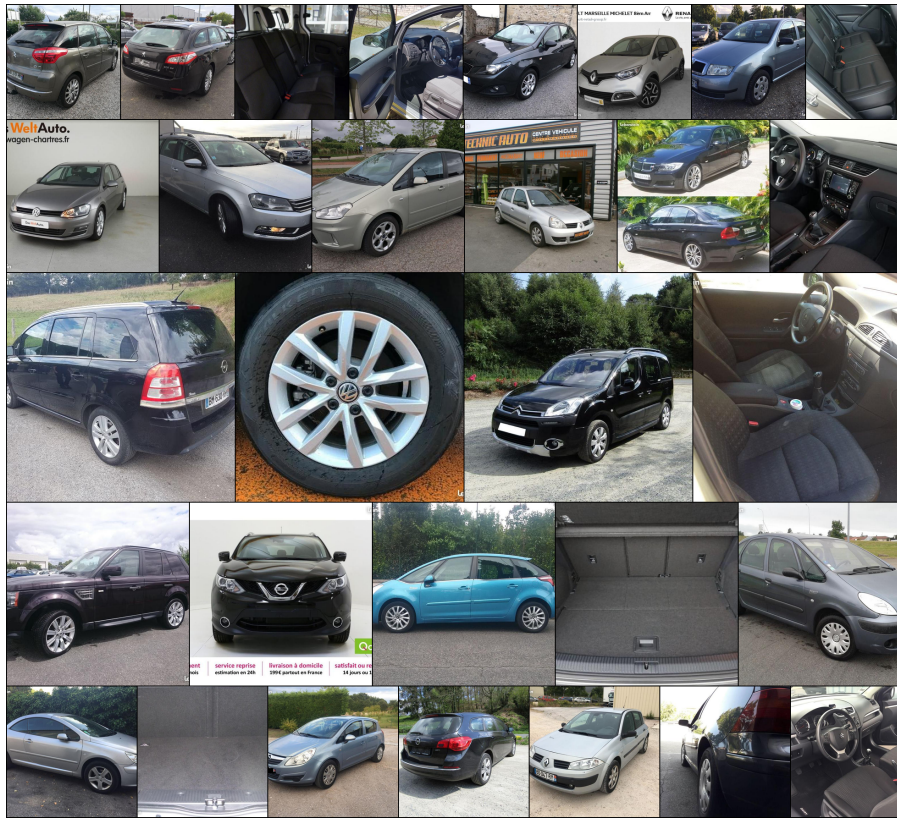
- **Etape 1** : Classification des photos pour distinguer celles sur lesquelles la voiture est bien visible de l'avant des autres
  - 1300 photos labellisées à la main : 1000 bien visibles de l'avant et 300 autres
  - Utilisation de la technologie Velours de Fotonower pour étendre la classification aux 60,000 photos
  - Cette technologie utilise des descripteurs d'AlexNet (couche pool5) et une classification SVM linéraire. Elle a été développé par Fotonower en 2015 grâce à l'apport de l'expertise de 3 chercheurs de l'ENPC : Matthieu Aubry, Guillaume Obonzinski et Nikos Komodakis

- **Etape 2** : Apprentissage sur le modèle de la voiture en distinguant les voitures bien visibles de l'avant des autres
  - 150 classes de modèle de voitures bien visibles de l'avant
  - 150 classes des mêmes modèles dans d'autres situations

INSTITUT DES ACTUAIRES

Amélioration de l'étape 1 : utilisation de technologie semi-supervisée pour organiser des photos de voitures

- Utilisation : Tri automatique de quelques milliers a quelques centaines de milliers de photos en quelques minutes ou heures
- Basé sur une implémentation en CUDA (GPU) d'algorithme de Kmean avec différentes variantes
- Travail réalisé en collaboration avec Gilles Pagès du LPMA (Jussieu) et soutenu par l'AMIES durant le CEMRACS 2017



Calibration sur les modèles de voiture (étape 2)

- Utilisation des poids du VGG16 pour les premières couches du réseau convolutif
- Calibration en cascade des dernières couches :
  - 1ère calibration grossière avec peu de classes et des classes grossières comme les marques de voitures
  - Calibration plus précise, sur les modèles

100% DataScience

INSTITUT DES ACTUAIRES

1/ Présentation : par Imen et Nathalie

- Création du Club Algo, objectifs et présentation du sujet (Nathalie)
- Principes des Réseaux de Neurones et introduction au CNN (Imen)

2/ Application des CNN à la reconnaissance marque et modèle de voiture : par Victor et Arthur

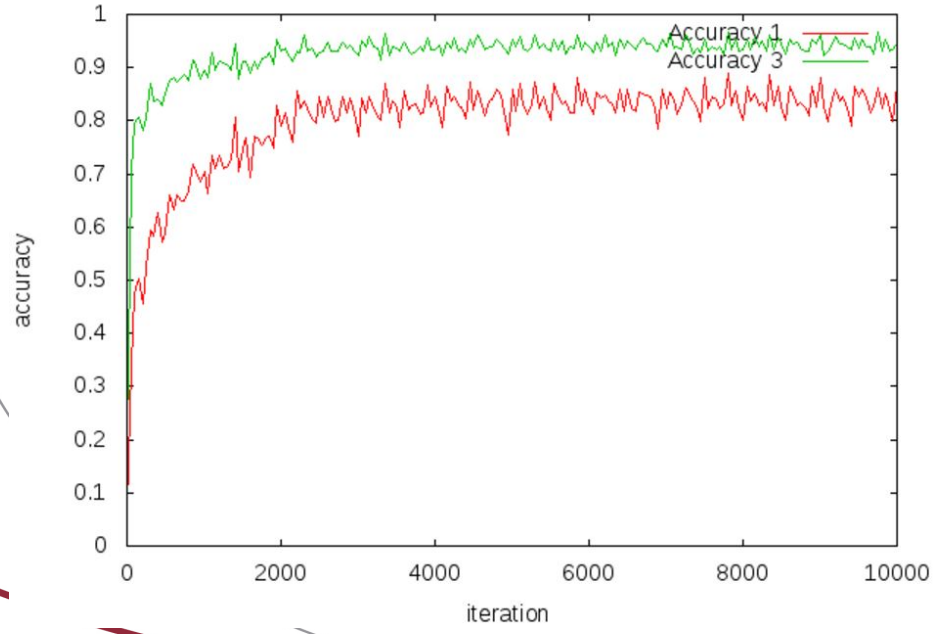
- Présentation de la méthodologie retenue
- Résultats obtenus

3/ Questions

Résultats obtenus sur l'ensemble de la base

- La précision est globalement de 75% pour le 1er choix et de 90% si on tolère de considérer les 3 premiers choix proposés par le modèle

Graphique d'apprentissage pour 22 modèles de voiture



Exemples de résultats

Xsara : 0,40
307 : 0,14
C4 : 0,13
206 : 0,08
Laguna : 0,04

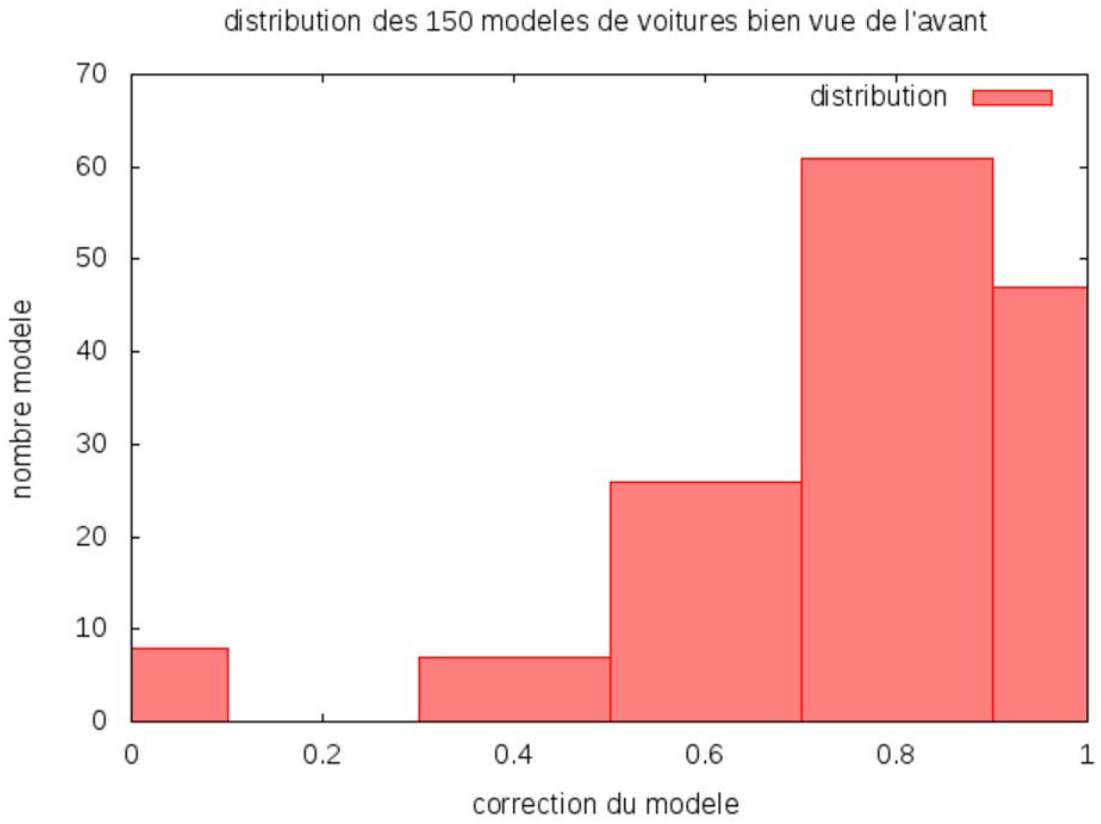
⇒ **Mauvaise prédiction**

Laguna : 0,96
Megane : 0,03
Clio : 0,01
Scenic : 0,001
Autres : 0,0001

⇒ **Bonne prédiction**

Résultats inégaux entre les deux classes de véhicules

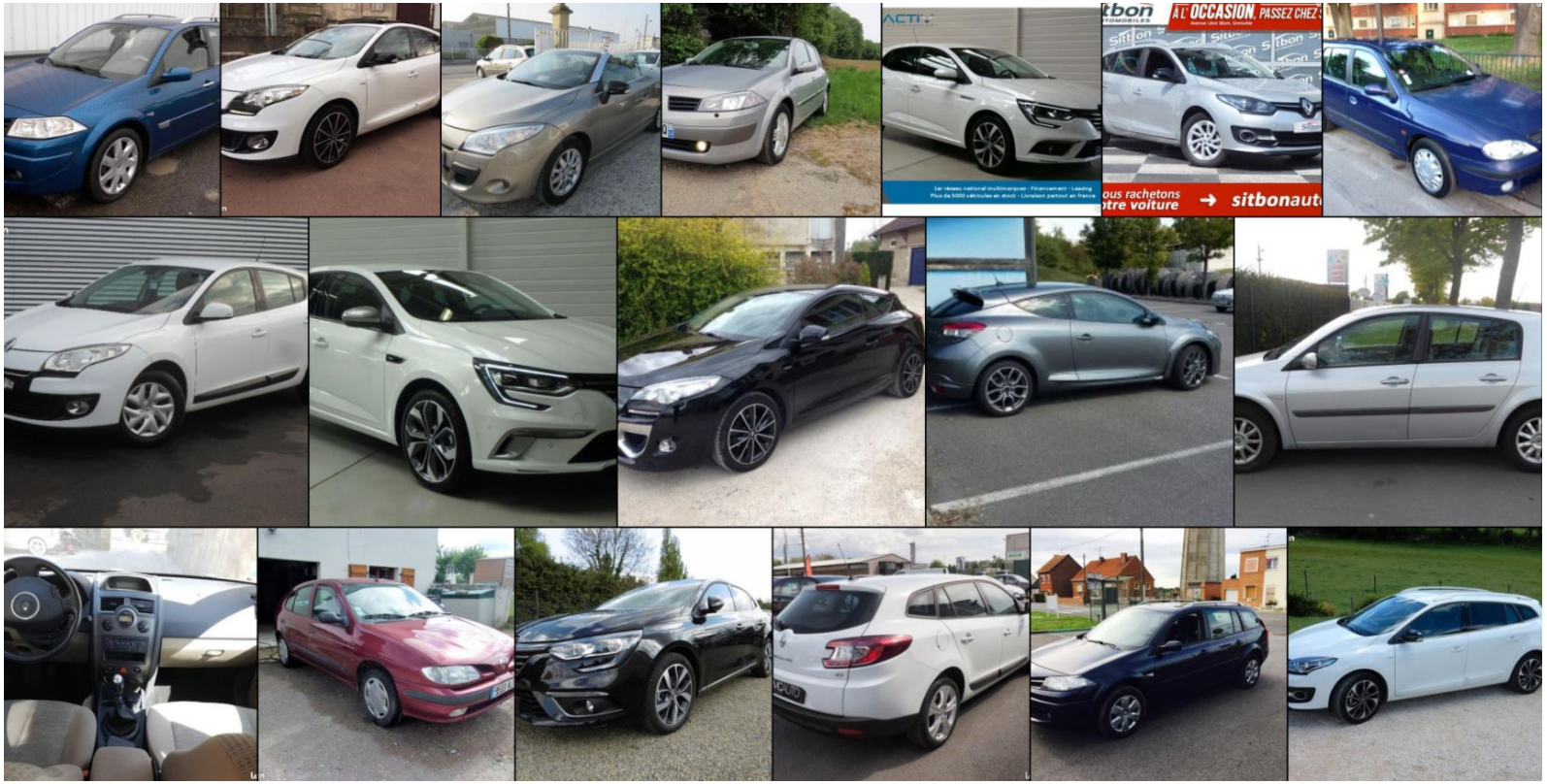
- La précision sur le 1er choix est de :
  - **86%** pour les véhicules bien visibles de face
  - **46%** pour les autres véhicules



INSTITUT DES ACTUAIRES

Exemples de résultats : 1er choix de détection correct

Exemples de résultats : bon choix dans les 3 premières propositions



INSTITUT DES ACTUAIRES

Exemples de résultats : aucune des 3 premières propositions n'est correcte

2/ Application des CNN à la reconnaissance marque et modèle de voiture

INSTITUT DES ACTUAIRES

Exemple : autonower.site

QUESTIONS

INSTITUT DES ACTUAIRES

Merci de votre attention !

Pour nous contacter :

club-algo@institutdesactuaires.com

LinkedIn