Apprentissage profond pour l'audio branding

Dominique Fourer

15 mai 2019

IBISC - EA 4526, Univ. Évry/Paris-Saclay

Informatique, Bio-informatique et Systèmes Complexes

Labo STIC de l'Univ. d'Évry/Paris-Saclay créé en 2006 (Fusion du LAMI/LSC)

- 2 sites à Évry (IBGBI, Pelvoux)
- 4 équipes de recherche AROBAS, IRA2, COSMO, **SIAM**
- 19 PU ou equiv., 32 MCF ou eq., 36 doctorants et postdocs

Domaines d'application

STIC&SMART SYSTEM et STIC&VIVANT

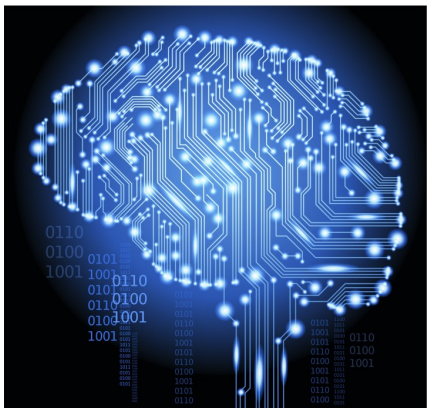
Plan

Objectifs de la présentation

- Une introduction à l'apprentissage profond
- Contributions dans l'*audio branding*

$\Rightarrow$ Perspectives d'application en logistique.

Naissance de l'IA [Minsky, McCarthy, 1956]



IA : une tentative de définition...

Ensemble des théories et techniques capables de simuler certains traits de l'intelligence humaine.

The Logic Theorist [Newell, Shaw, Simon 1956] => Résolution de problèmes et démonstration automatique de théorèmes.

Introduit 3 concepts :

- **Combinatoire** (i.e. exploration de graphe)
- **Heuristique** (solution calculable)
- **Liste de procédures** (e.g. λ-calcul, IPL, LISP [McCarthy, 58])

Algorithmes inspirés de la biologie

Algorithmes évolutionnaires

Inspirés de la théorie de l'évolution

- Stratégie d'évolution [Barricelli et al, 1954]
- Automates cellulaires [Neumann et al, 1966], (jeu de la vie [Conway, 1970])
- Algorithmes génétiques [Holland et al., 1973]

Intelligence distribuée "*swarm intelligence*"

Repose sur la stigmergie [Grassé, 1959]

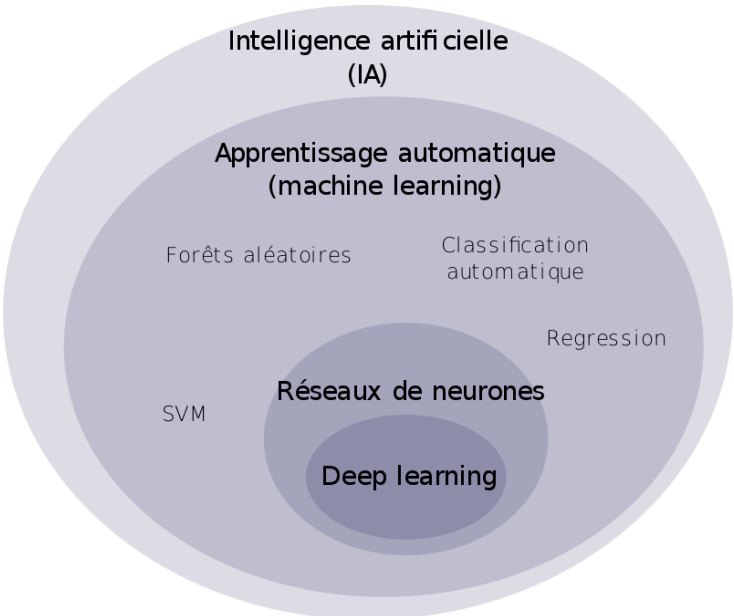
- Essaim particulaire [Kennedy, Eberhart, 1995]
- Algorithmes de colonies de fourmies [Ebling,Dorigo et al. 1989-2005]

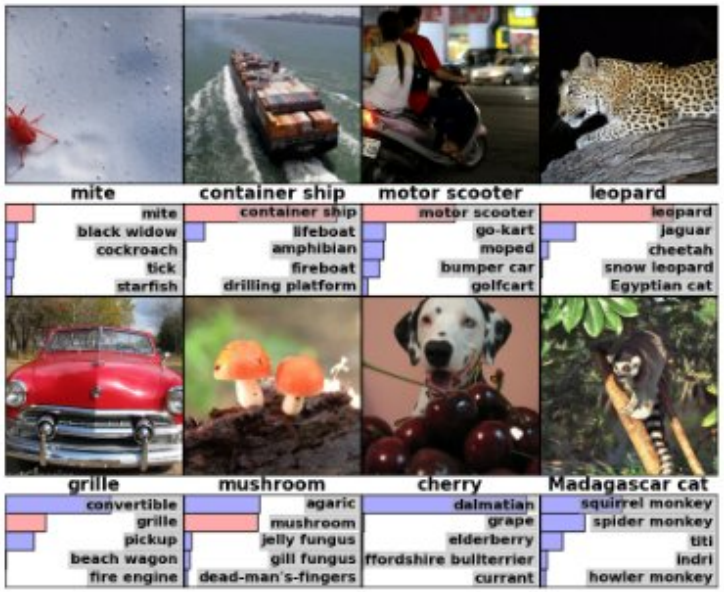
Réseaux de neurones artificiels

- Neurone formel [Hebb, 1949] [McCulloch, Pitts, 1959]
- Perceptron [Rosenblatt, 1957], récurrent [Hopfield, 1982], multi-couche [Rumelhart, LeCun et al. 1986]
- Rétropropagation du gradient [Werbos, 1975] [Parker, LeCun, 1985] [Hinton, Williams, 1986]
- Réseaux convolutifs (CNN) [LeCun, 1998]

Plan

Le succès du *deep learning* depuis 2012





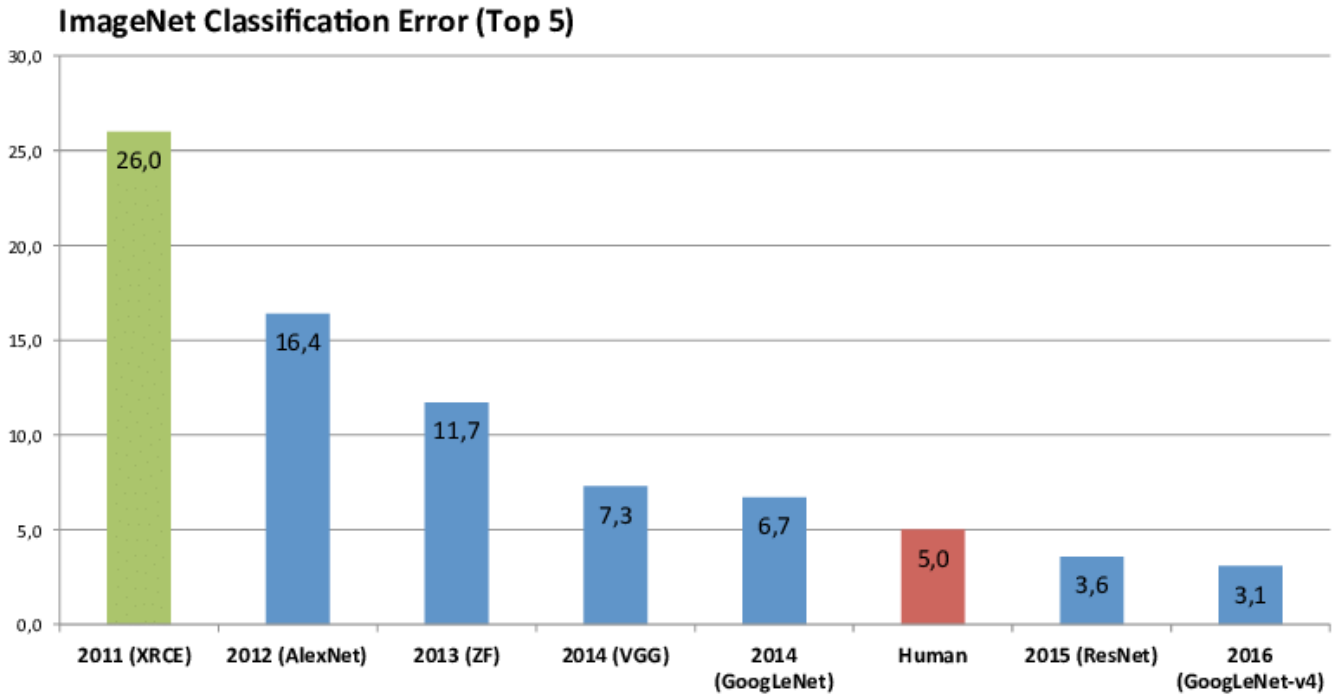
`http://image-net.org/challenges/LSVRC/`

$\approx$ 1,2 millions d'images, 1000 classes [ILSVRC2012] [Krizhevsky et al. NIPS'12]

- Excellentes performances
- Grande masse de données annotées
- Parallélisation des algorithmes
- Moyens de calculs accessibles (e.g. GPU, clusters, etc.)

	Affiliation	Taux d'erreur	Description
1	U. Toronto	15,31 % (-10,8%)	**Deep CNN**
2	U. Tokyo	26,172 %	desc. + classif.
3	U. Oxford	26,979 %	desc. + classif.
4	Xerox/INRIA	27,058 %	desc. + classif.

Avantages du deep learning

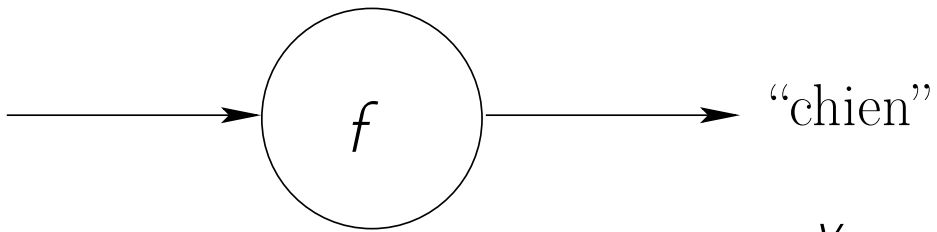


- État de l'art en vision par ordinateur [Azizpour et al., 2016]
- Forte capacité de généralisation (sans surapprentissage)
- Apprend automatiquement une représentation discriminante (*deep features*)
- Surpasse l'humain dans de plus en plus de domaines (e.g. reconnaissance d'images, jeu d'échec, Alphago, Alphastar, etc.)
- Une infinité de domaines d'application (e.g. biomédical, finance, astronomie, reconnaissance vocale, etc.)

Formulation du problème (classification)

Comment construire f qui fonctionne quelque soit x ?

x

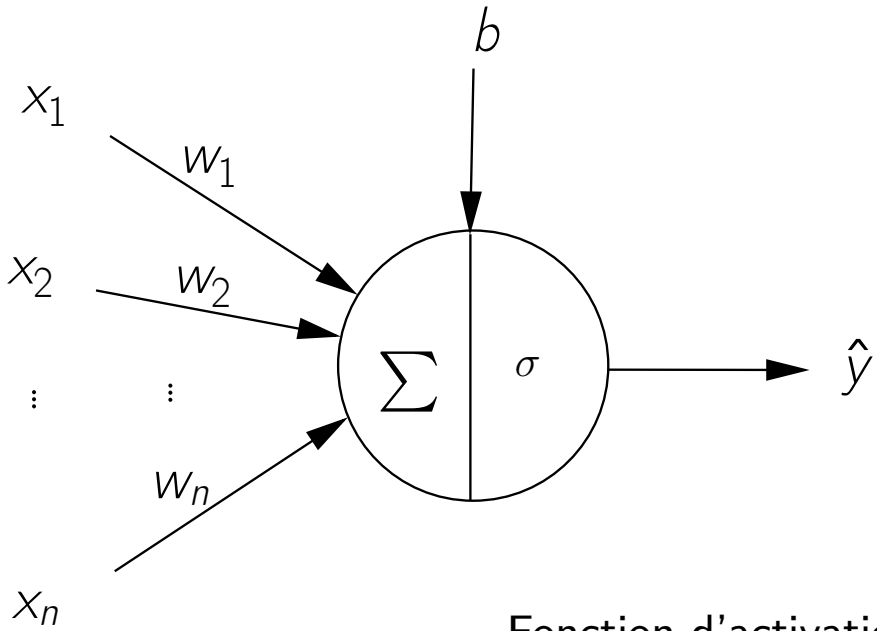


y

Difficultés

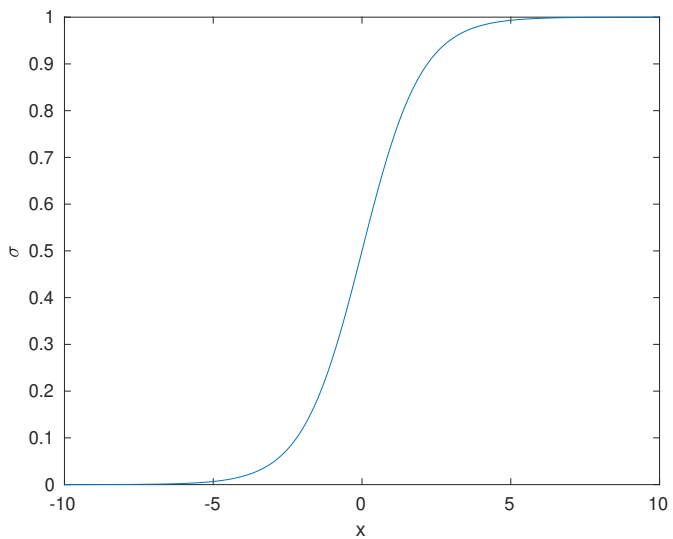
- $dim(x) \approx 10^6$ (*curse of dimensionality*)
- x et y n'ont pas la même grandeur physique :
  x : image (tenseur), y : label (scalaire)
- f n'a pas été entraîné sur x et doit **généraliser** le problème

Modèle neuronal (Perceptron) [Rosenblatt, 1957]

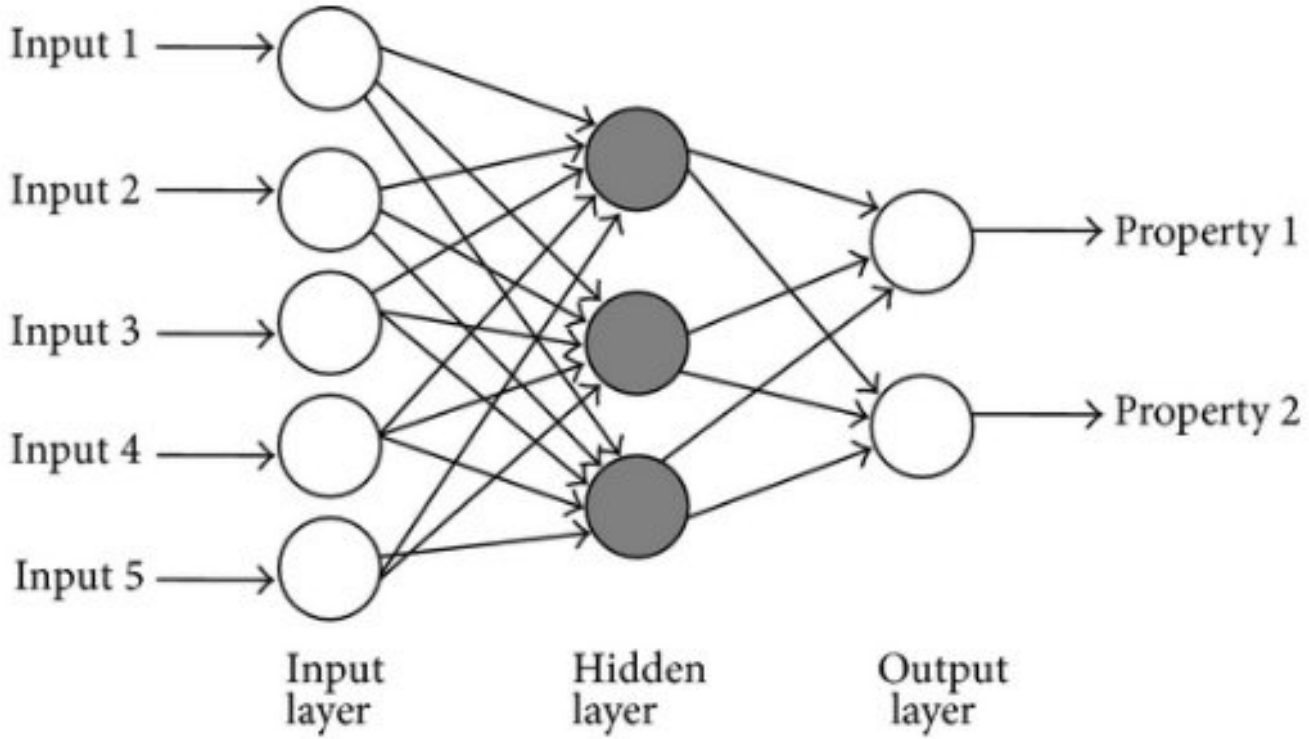


- entrées : $\mathbf{x} = (x_1, x_2, ..., x_n)$
- poids : $w_1, w_2, ..., w_n$
- biais : $b = x_0 w_0$ avec $x_0 = 1$
- fonction d'agrégation : Σ
- fonction d'activation : $\sigma(x) = \frac{1}{1+e^{-x}}$
- sortie : $\hat{y} = \sigma\left(\sum_{i=1}^{n} x_i w_i + b\right)$

Fonction d'activation sigmoïde.

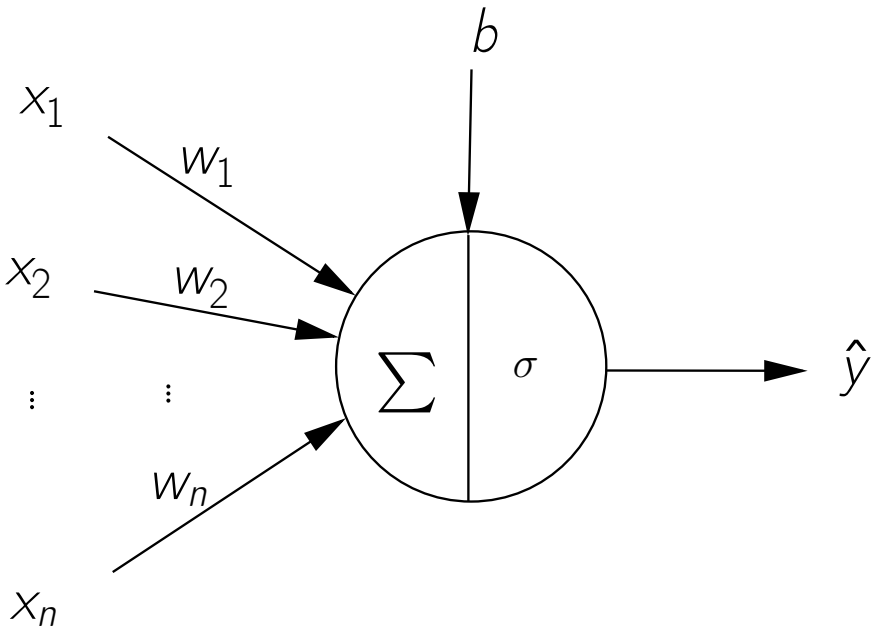


Réseau de neurones (MLP)



- N'est plus soumis aux limitations théoriques de [Minsky, Papert, 1969]
  $\Rightarrow$ Peut approximer toute fonction booléenne avec au moins une couche cachée
- Les neurones de la couche l servent d'entrée à la couche $l+1$
- Toutes les connexions sont pondérées par des poids $w_{l,i}$
- Les couches cachées permettent d'apprendre l'espace latent (*features*)
- Plus il y a de couches cachées et plus le réseau est **profond**

Entraînement (apprentissage supervisé)



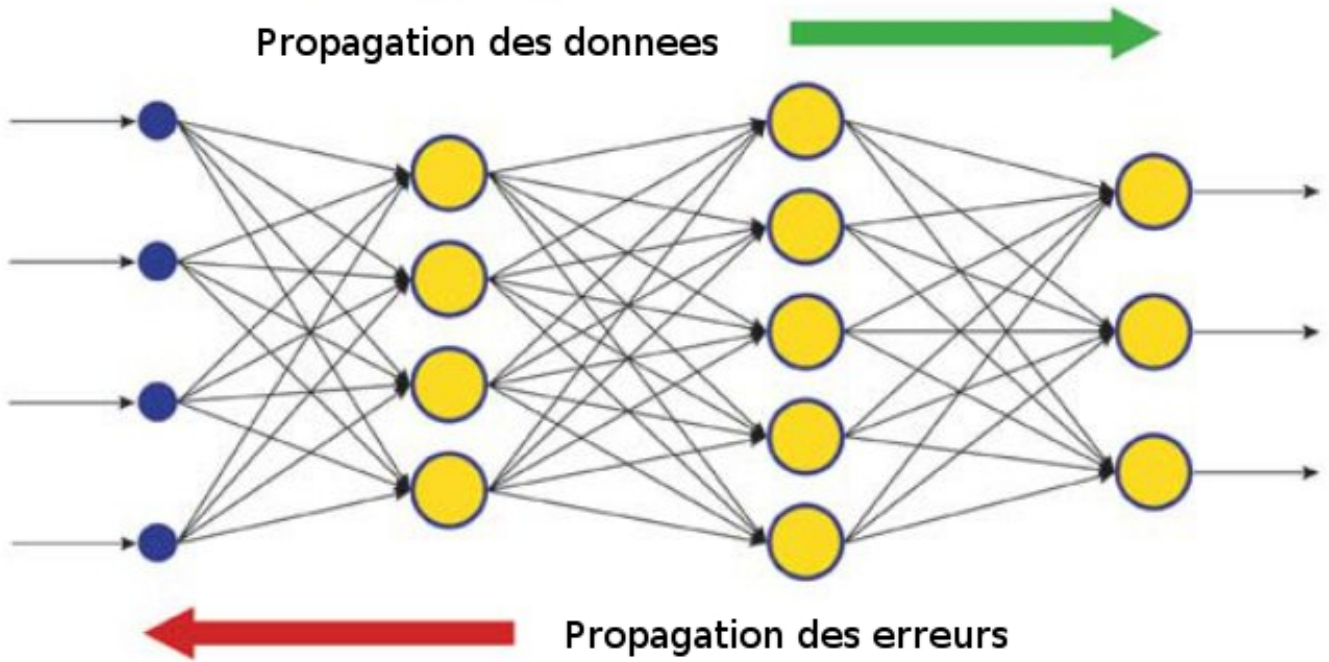
Apprentissage $\Leftrightarrow$ Optimisation

Apprentissage des w_i et b à partir d'exemples annotés

Principe : on minimise une *loss function* $\mathcal{L}(y_t, \hat{y})$, $\forall(\mathbf{x}_t, y_t) \in D$

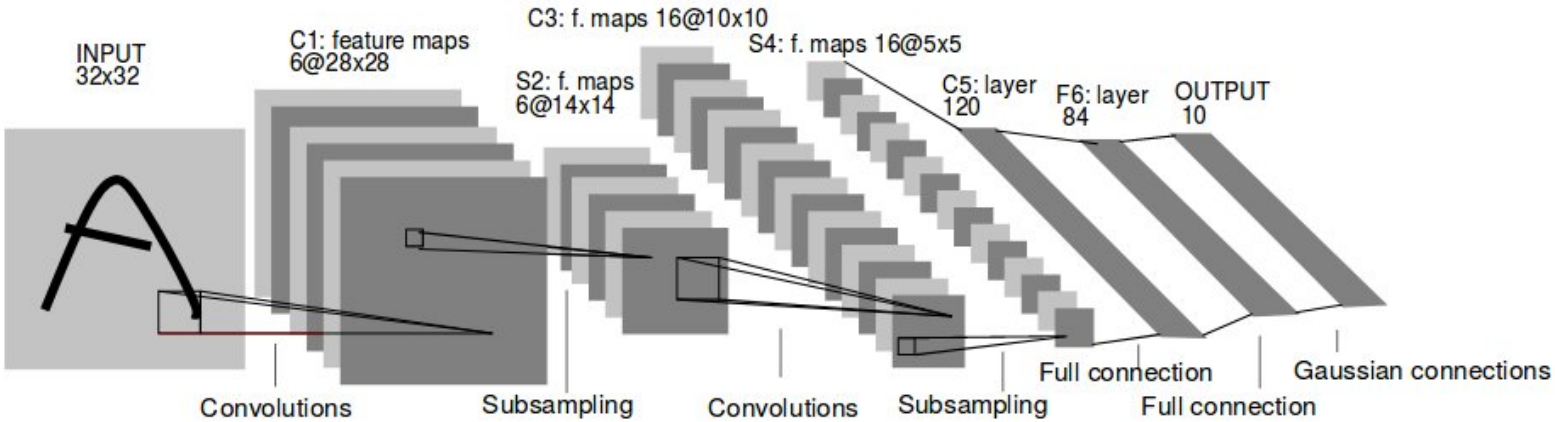
- On calcule la sortie du réseau $\hat{y}$
- On met à jour les $w_i \leftarrow w_i + \alpha \frac{\partial \mathcal{L}(y_t, \sigma(\sum_{i=0}^{n} x_i w_i))}{\partial w_i}$
  avec $\alpha \in [0; 1]$ le pas d'entraînement.

Entraînement par rétropropagation [Hinton, Williams, 1986]



- Permet de mettre à jour tous les neurones et toutes les synapses
- Prise en compte de tous les exemples disponibles pour l'apprentissage
- Minimise une fonctionnelle (*loss function*) $\mathcal{L}(y, \hat{y})$ arbitraire.

Lenet-5 [LeCun 1998]



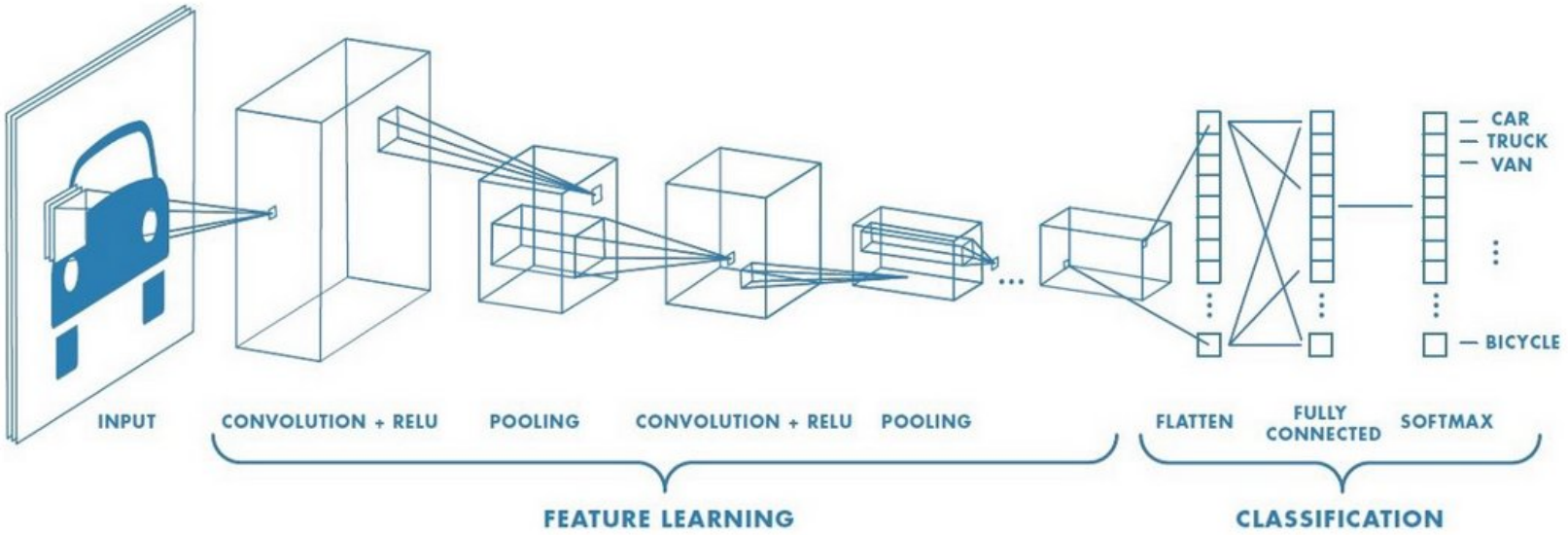
Principe

- Chaque couche est composée d'un ensemble de noyaux (K_s)
- Les features y_s sont obtenues en filtrant (convolution) l'entrée x par chaque noyau K_s

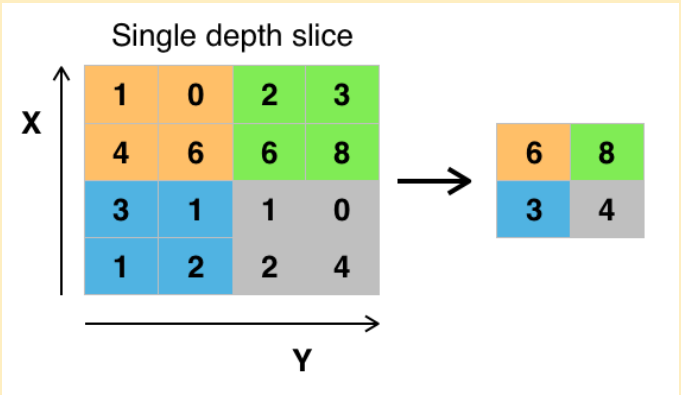
$$y_s[n,m] = x * K_s = \sum_i \sum_j x[n-i, m-j] K_s[i,j] \quad (1)$$

- On entraîne ce réseau en apprenant les coefficients des filtres K_s

Architecture type



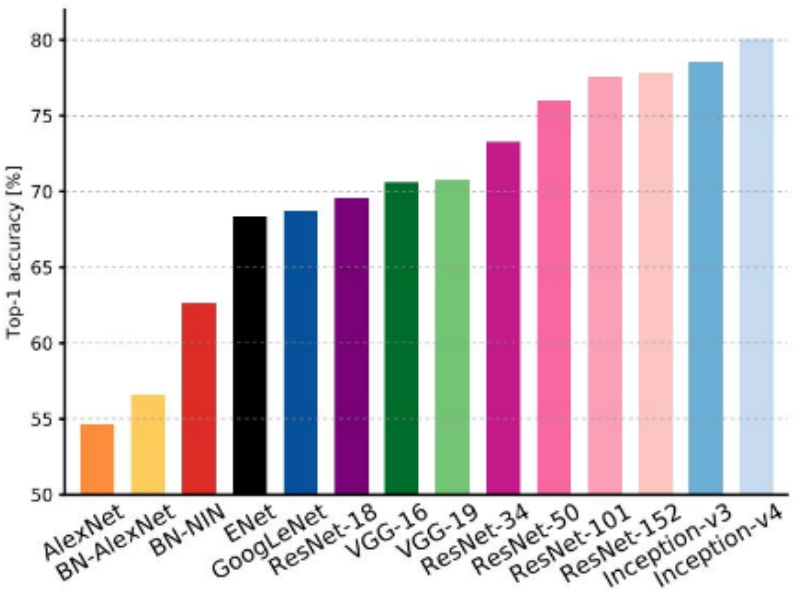
Couche de *pooling* (e.g. max-pooling)

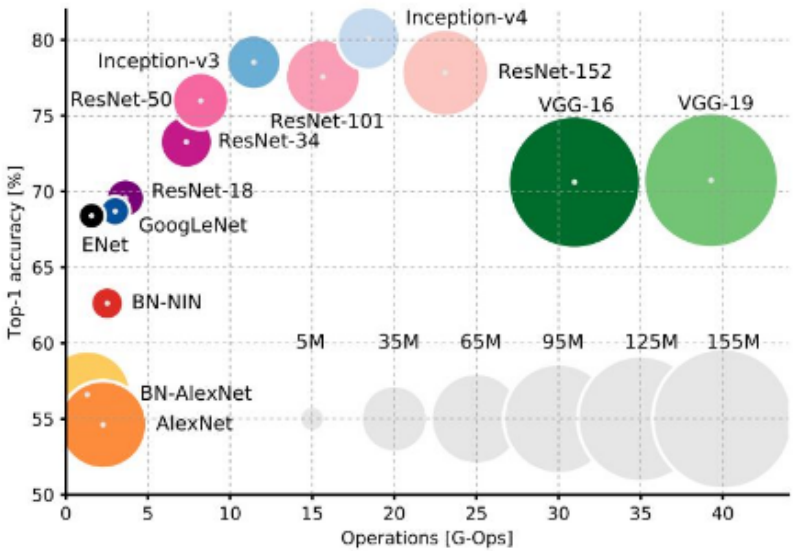


Permet de réduire la dimension de l'entrée

Comparatif des architectures actuelles

Beaucoup de progrès depuis 2012 !

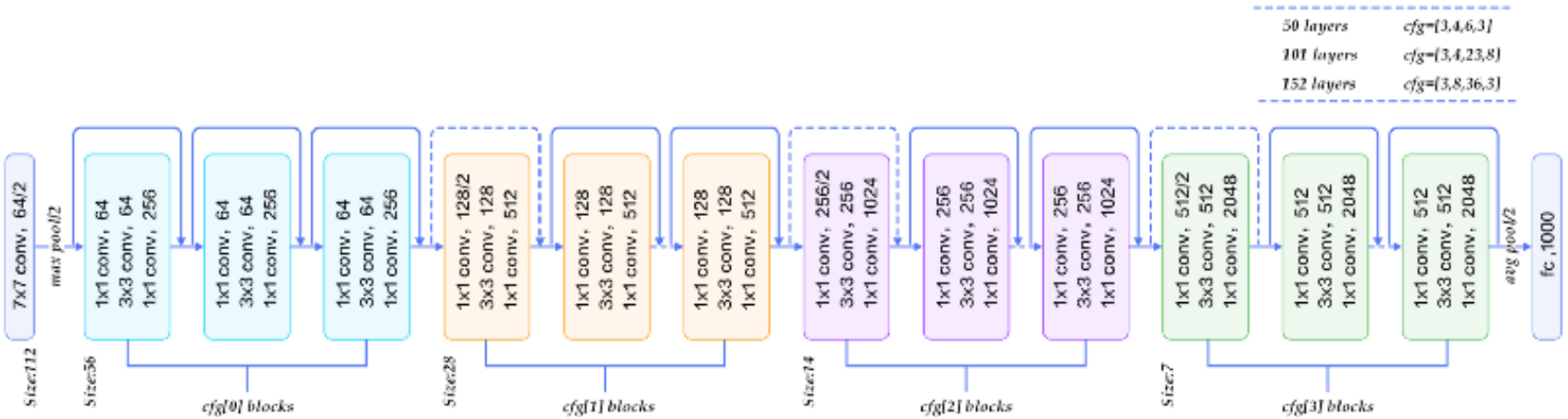




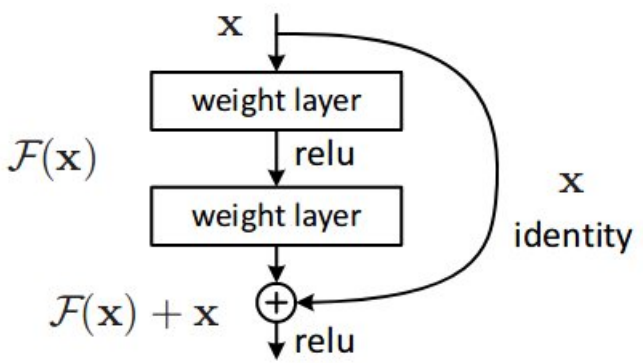
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

- Plus de paramètres entraînables n'améliore pas forcément la précision
- Meilleurs résultats (en 2016) obtenus par Inception-v4 (google inc.) [Szegedy et al, 2017]

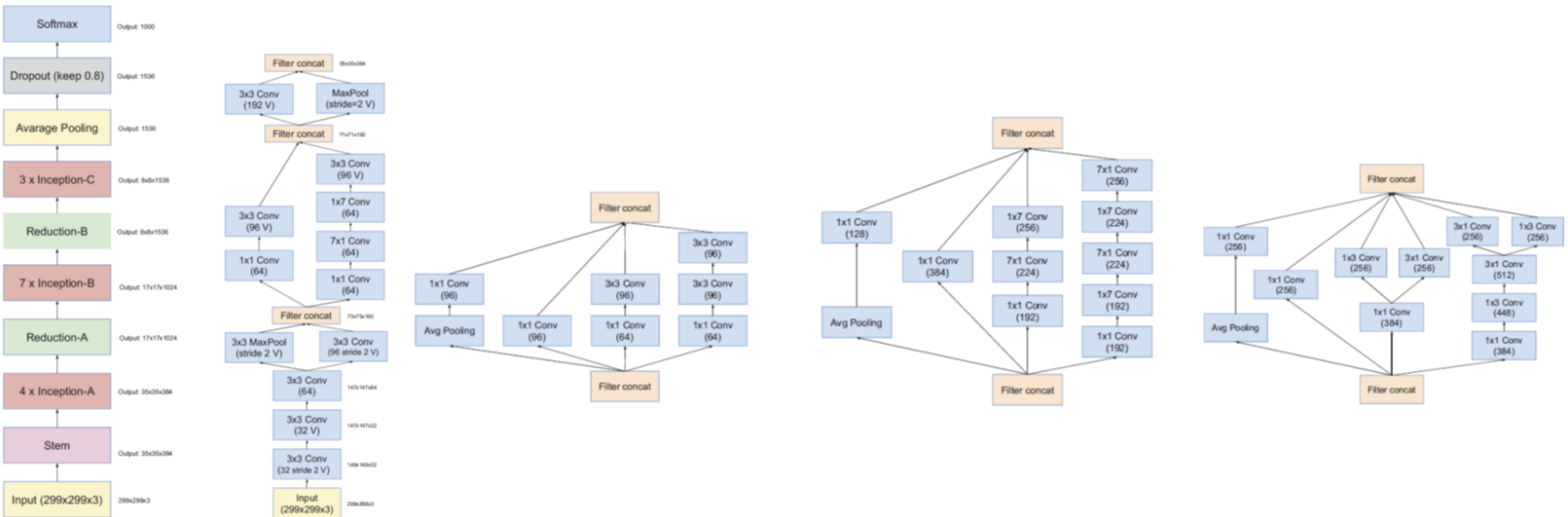
Resnet 2015 (Microsoft) [He et al. 2016]



- Version modifiée de GoogleNet (Inception-v1)
- Utilise des connexions résiduelles :



Googlenet-v4 2016

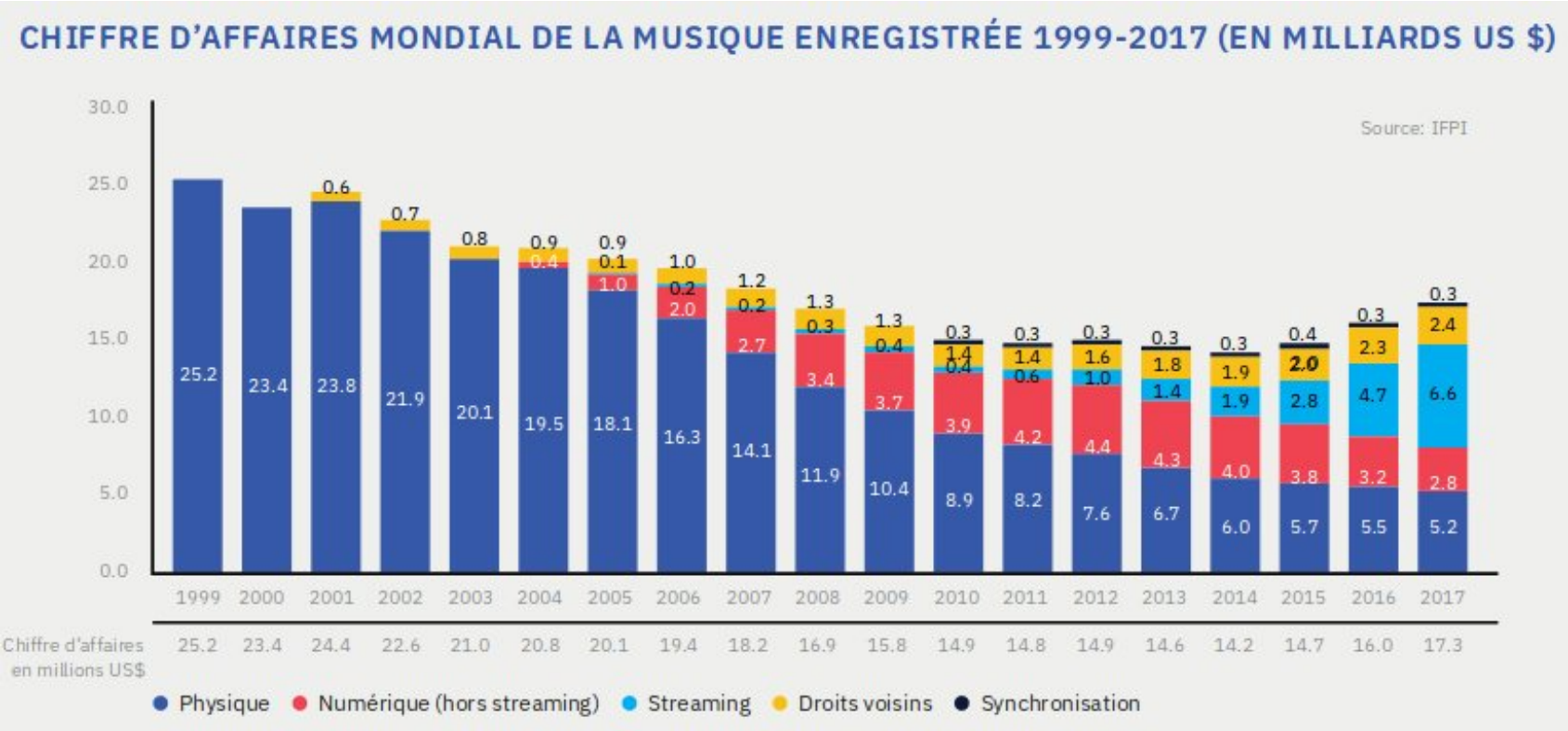


- Combine 3 architectures en cascade (Inception-A, B et C)
- Un peu moins de paramètres entraînables que ses concurrents (i.e. plus efficace)

$\Rightarrow$ En 2018, plusieurs variantes des réseaux précédents (Resnet, Inception, etc.) utilisant des extra-datas font moins de 3% d'erreur.

Plan

État de l'industrie de la musique



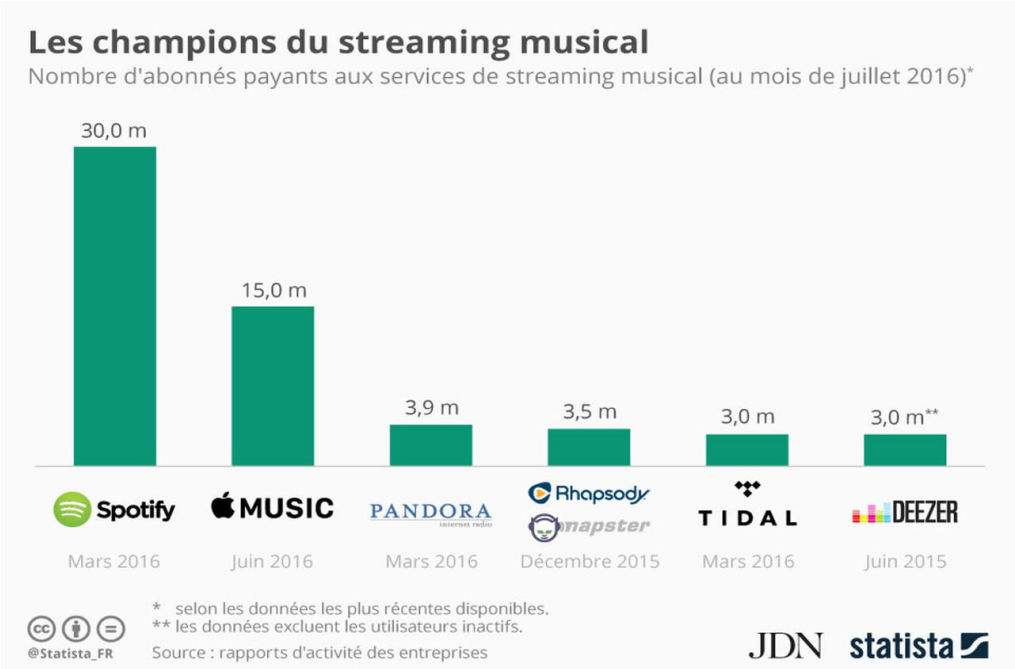
- Forte croissance pour la consommation de la musique en *streaming* et des droits voisins (e.g. copie privée)
- Croissance globale depuis 2014 après plusieurs années de baisse successives

Qui consomme le plus de musique ?

- L'Europe consomme le plus de musique après les USA et le Japon
- Forte croissance dans les autres pays (e.g. Chine, Brésil)

Intérêt pour les industriels

Une musique qui correspond à votre marque (*brand fit*) [McInnis, Park, 1991].



Comment faire ?

- Extraire des informations (*hard/soft features*) à partir de l'audio.
- Associer les morceaux de musique d'une base de données dont les informations correspondent à une cible.

Objectif : Système de recommandation musicale à la carte.

Projet ABC-DJ

Artist-to-Business to Business-to-Consumer Audio Branding System

TU BERLIN
TECHNISCHE UNIVERSITÄT BERLIN, AUDIO COMMUNICATION GROUP

HEARDIS!
HEARDIS! GMBH

IRCAM
INSTITUT DE RECHERCHE ET COORDINATION ACOUSTIQUE/MUSIQUE

FINCONS
FINCONS SPA

INTEGRAL
INTEGRAL MARKT- UND MEINUNGS-FORSCHUNGSGES.M.B.H.

LOVEMONK
LOVEMONK SL

PIACENZA
FRATELLI PIACENZA SPA

Données chiffrées

Horizon 2020
European Union funding
for Research & Innovation

Durée :	janvier 2016 - décembre 2018
Budget :	3 488 650 €
Coordinateur :	Richard Wages / TU Berlin
Site internet :	https ://www.abcdj.eu

Tâches d'indexation considérées dans le projet ABC-DJ (soft-features)

Tag	Classes considérées	train set
Voix (1)	**voice / instrument**	**413**
Voix (2)	male / female / other	633
Pop-Appeal	5 niveaux	983
Intensité	5 niveaux	812
Instrumentation	13 instruments (drums,piano,guitar,...)	1983
Timbre	(6) hard/dark/cold/bright/warm/soft	786
Genre	(10) world/pop/soul/folk/dance/rock/...	1987
Style	61 tags	9417

Comment traiter des bases de données de plusieurs millions de titres ?

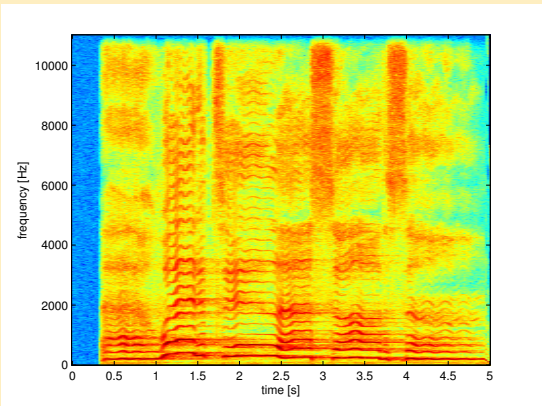
- Automatiser l'indexation de la musique pour alimenter les services de musique à la demande
- Développement de méthodes innovantes d'audition par ordinateur (traitement du signal + apprentissage machine)

Alimenter le lecteur de la société HearDis ! GmbH (`https://www.heardis.com`)

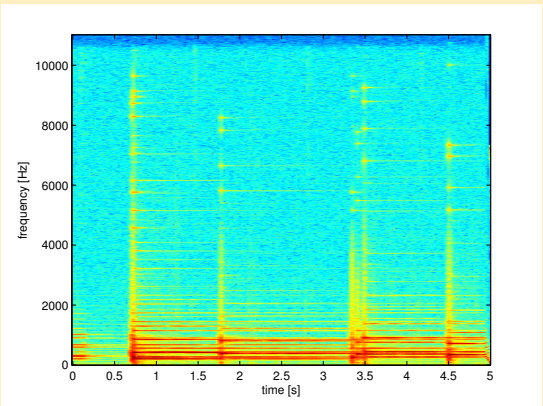
Spectrogramme d'une source sonore

Observation

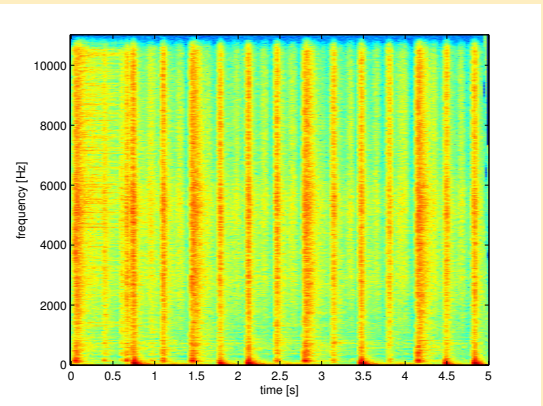
Chaque source sonore a une structure temps-fréquence spécifique.



voix chantée



piano



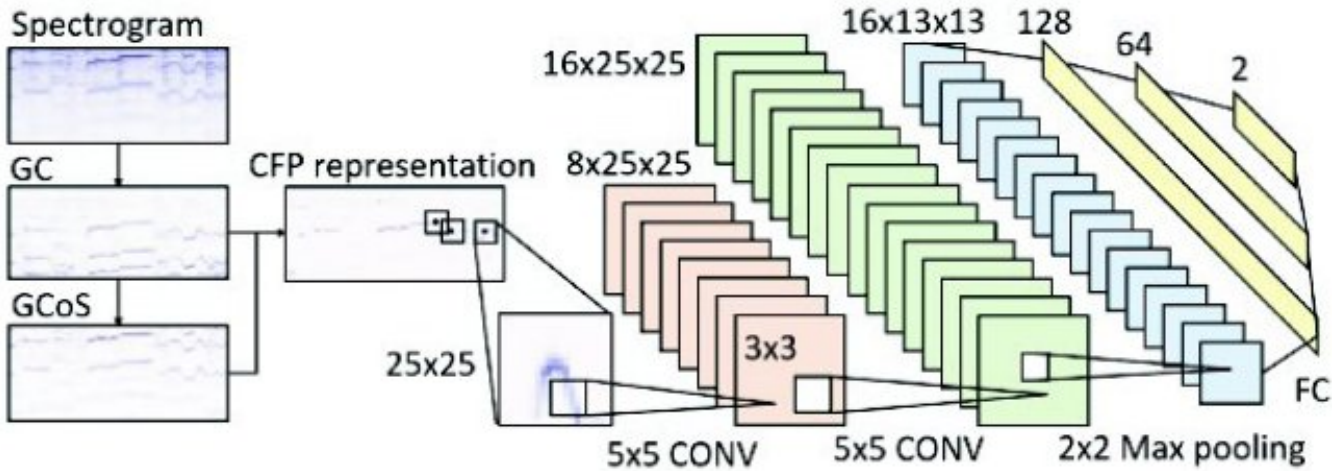
batterie

Intuition

Chaque instrument peut être détecté dans un mélange à partir de filtres spécifiques (e.g. deep features produites par un CNN) :

- La voix chantée a une large bande de fréquence et contient du vibrato
- Les instruments harmoniques sont stables temporellement (lignes horizontales)
- Les instruments percussifs sont brefs avec une large bande de fréquences (lignes verticales)

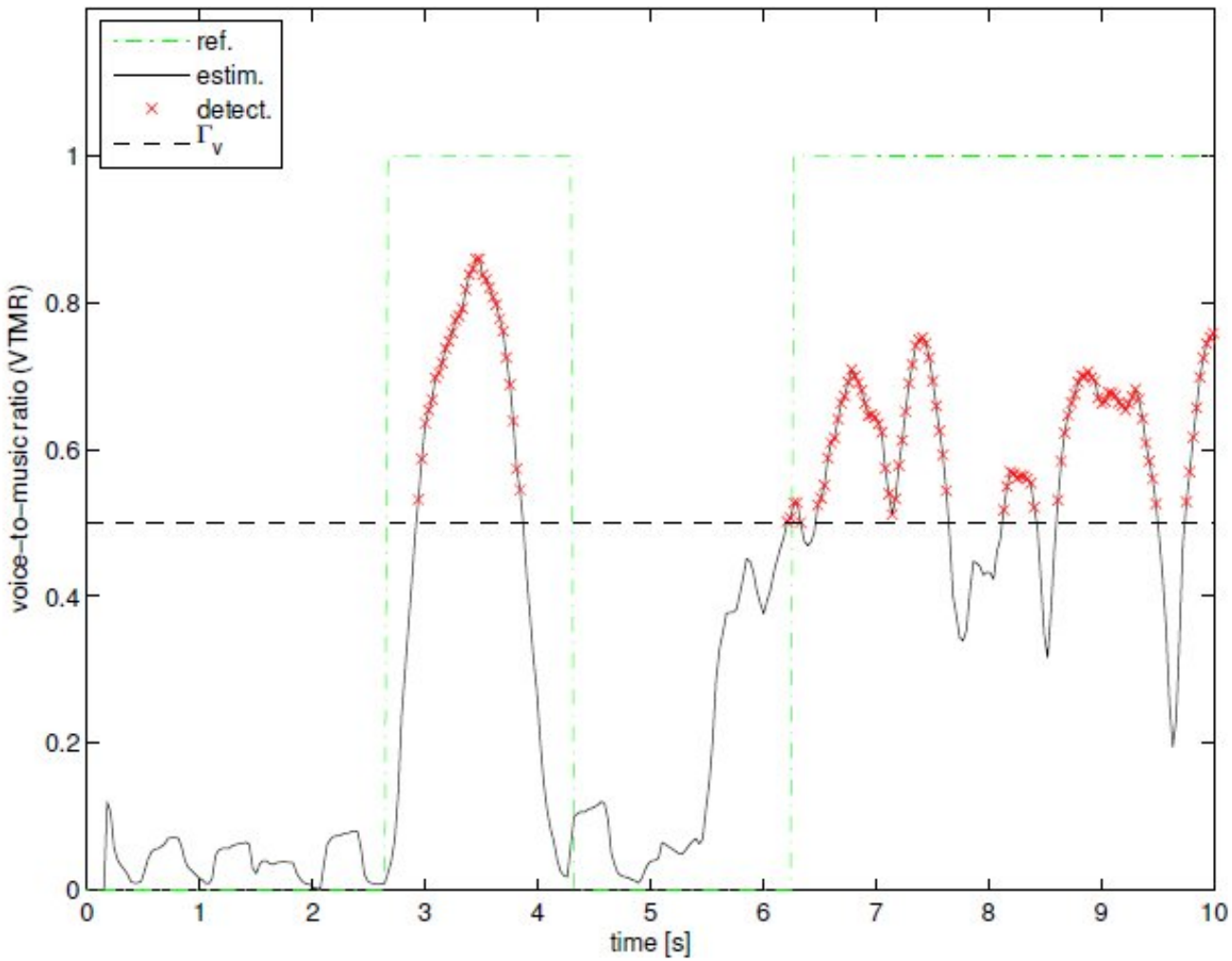
Deep CNN pour la détection supervisée de la voix [Schlüter, 16]



- Prend en entrée une représentation temps-fréquence du mélange (déduite du spectrogramme).
- Apprend automatiquement les filtres permettant de dissocier la voix du reste du mélange.
- Fournit une fonction de saillance (i.e. probabilité que le mélange contienne de la voix à l'instant t)

Exemple

Détection sur une pièce musicale *MusicDelta Punk* de MedleyDB :



Résultats comparatifs sur plusieurs *datasets* publics

Table : Rappel moyen pour la détection de la voix chantée

(a) évaluation sur chaque dataset

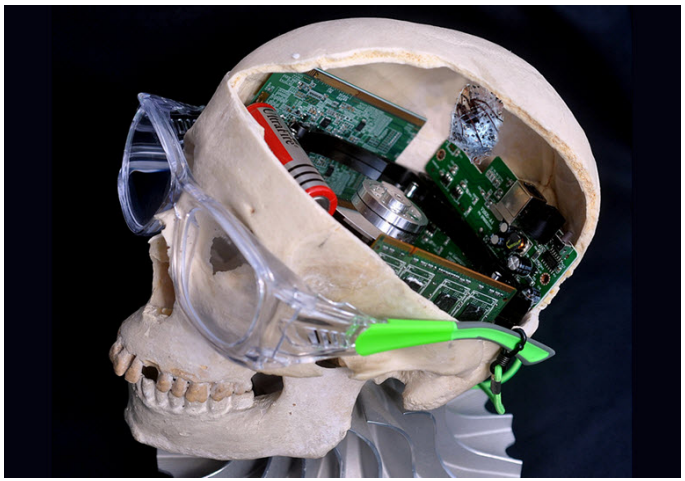
Dataset	Best unsupervised	SVM (MFCC+SCT)	CNN
Jamendo	0.58	0.81	**0.86**
MIR1K	0.68	0.77	**0.9**
MedleyDB	0.59	0.79	**0.86**

(b) évaluation en mélangeant les datasets

Training datasets	SVM (MFCC+SCT)		CNN	
	self-DB	cross-DB	self-DB	cross-DB
Jamendo + MIR1K	0.81	0.73	**0.89**	**0.75**
Jamendo + MedleyDB	0.80	0.59	**0.86**	0.65
MedleyDB + MIR1K	0.80	0.76	**0.84**	**0.77**

[Fourer, D., & Peeters, G. (2018). Single-Channel Blind Source Separation for Singing Voice Detection : A Comparative Study. arXiv preprint arXiv :1805.01201.]

Conclusion



Contributions

- Les raisons du succès du deep learning
- Une introduction au fonctionnement du deep learning
- Exemple d'application dans le domaine audio

Questions ouvertes sur le deep learning

- Interprétabilité des modèles $\Rightarrow$ théorie
- L'entraînement automatique (pas encore d'IA universelle)
- Problème qualitatif et quantitatif des données annotées *(weak-labeled / semi-supervised / unsupervised)*

Perspectives dans un cadre logistique

Perception

- Capteurs sonores (cf. détection d'événements [DCASE challenge])
- Reconnaissance vocale
- Capteurs visuels

Modélisation et optimisation de systèmes complexes

- Transport optimal
- Aide à la décision
- Véhicules autonomes / drones
- Robotique
- ...