Transparence des résultats de l'apprentissage profond

Sara Bouteggui[1], Lucile Bruel[1], Florent Gimenez[1], Léo Saulières[1], Josiane Mothe[2]

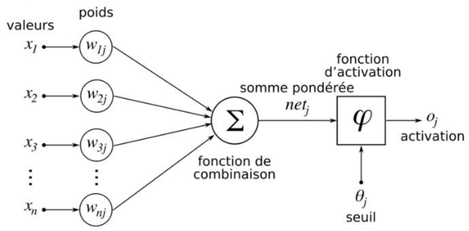
1.Etudiants en Master 1 IARF, Université Paul Sabatier, Toulouse
2.IRIT UMR5505 CNRSLab, Université de Toulouse, INSPEE, France

Réseaux de neurones

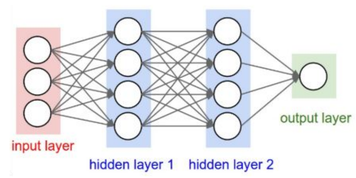
L'**apprentissage profond** (Deep Learning) utilise des **neurones artificiels** s'inspirant du cerveau humain.
Les algorithmes de deep learning s'améliorent de façon autonome grâce aux **réseaux de neurones**, en utilisant un grand nombre de données

"Un réseau de neurones artificiels est un système informatique s'inspirant du fonctionnement du cerveau humain pour apprendre."

Fonctionnement d'un neurone :
1) Calcul de la somme pondérée de ses entrées multipliées par leurs poids respectifs
2) Ajout du biais
3) Cette valeur est transmise à une fonction d'activation qui va décider de la valeur en sortie du neurone.



Structure générale d'un réseau de neurones :

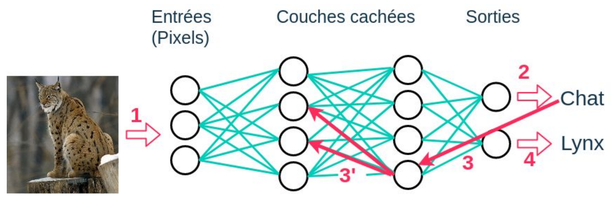


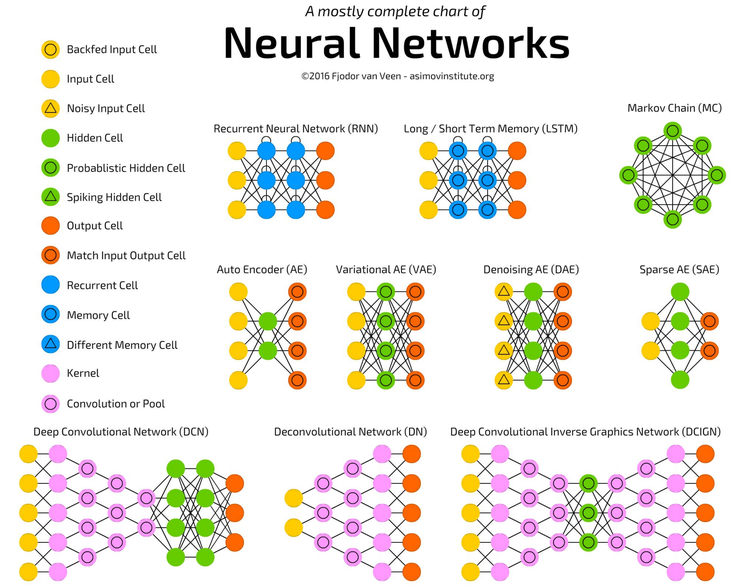
Différents types d'apprentissage:

- **Apprentissage supervisé :**
  Le réseau apprend à partir de données labellisées par des utilisateurs.
  Exemple : pour classifier des images, on doit indiquer quel est le type de l'image pour l'ensemble du jeu de données.
- **Apprentissage non supervisé :**
  Le réseau apprend par lui-même à partir de données non labellisées.

Entraînement du réseau :

- Si un résultat est mauvais :
  → **Rétropropagation** : Re-parcourir le réseau pour corriger les **poids** et les **biais**.
  On répète cela de manière itérative jusqu'à ce que l'**erreur** du réseau soit **minimisée**.



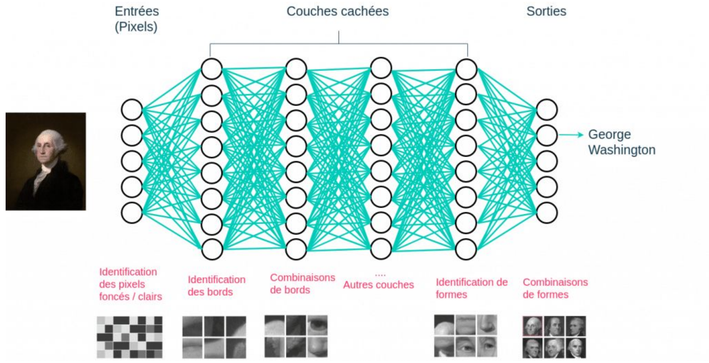


Principales structures de réseaux de neurones

Réseaux convolutifs

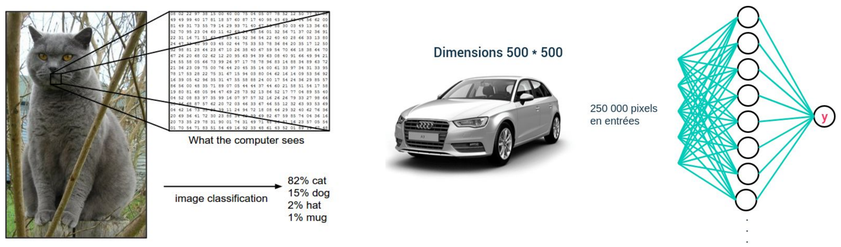
- **Filtre** : opération mathématique qui modifie la valeur d'un ou plusieurs pixels localement selon sa taille (fixée) sur une zone d'une image.
- **Convolution** : Appliquer un filtre sur la totalité d'une image en le faisant 'glisser'.

Réseaux convolutifs : utilisés pour la classification d'images (exemple ci-dessus).
Ces réseaux appliquent une succession de convolutions à différentes régions d'une image.



Exemples d'application des réseaux de neurones

- Reconnaissance vocale (assistants vocaux, …)
- Classification d'images (reconnaissance faciale, chiffres digitaux, …)



- Traitement du langage naturel (Bert, Elmo, …)
- Algorithmes de recommandation (Netflix, Youtube ,..).

Transparence dans les réseaux de neurones

Transparence d'un système :

Le fait, pour un utilisateur humain, de pouvoir comprendre tous les résultats de ce système et la manière dont ils ont été obtenus.[Iyer,2018]

Différentes méthodes :

1) Par visualisation :
- **Grad-CAM** s'appuie sur une classification de pixels d'une image
- **Carte de saillance** des objets : classification plus générale des objets

⊕
- Facilement compréhensible

⊖
- Peut produire de faux positifs
- Reste opaque en grande partie

2) Neural Interaction Transparency (NIT) :
Environnement qui met en lumière les liens cachés des neurones en les classant sous forme de groupes de neurones avec des interactions.[Tsang,2018]

⊕
- Étudie les couches cachées du réseau
- Modèle statistique **GAM** (Generalized Additive Models) rapide

⊖
- Interactions apprises parfois fausses
- Phase de démêlage longue

3) Méta-prédicteurs :
Déduction de certaines règles de prédiction du réseau grâce à des perturbations faites sur les entrées. [Fong,2017]

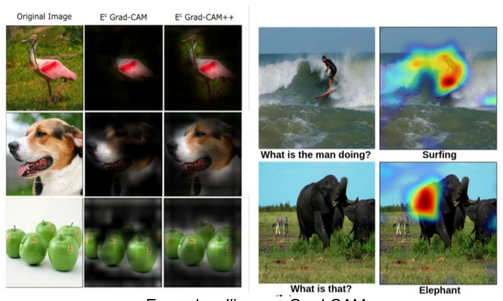
⊕
- Met en évidence les points faibles du réseau
- Pour tout modèle d'apprentissage
- Aucune modification du réseau

⊖
- Prend beaucoup de temps

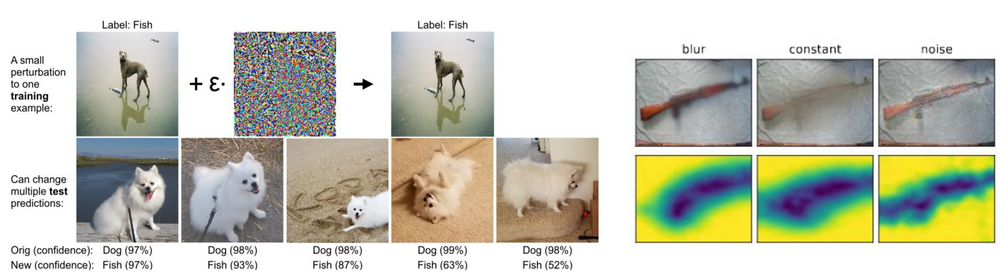
4) Fonction d'influences :
"La fonction d'influence est une mesure de l'importance de la dépendance des paramètres du modèle ou des prédictions par rapport à une instance de formation" [Molnar,2019]

⊕
- Différenciation de modèles en apparence identiques
- Met en évidence la vulnérabilité du modèle
- Fiable : méthodes statistiques robustes

⊖
- Ne s'applique pas à tous les modèles d'apprentissage
- Moins facile à comprendre que la visualisation



Exemples d'images Grad-CAM



Exemples de perturbations pour ces deux dernières techniques et des conséquences qu'elles peuvent avoir.

Références

- Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Why did you say that? (January 2017)
- Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized Gradient-BasedVisual Explanations for Deep Convolutional Networks (March 2018) https://cs231n.github.io/classification
- https://culturegeek.ninja/2019/10/18/le-deep-learning-definition-et-explication
- Fong, R.C., Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation.(October 2017)
- Molnar, C.: Interpretable Machine Learning. (2019) https://christophm.github.io/interpretable-ml-book/
- Tsang, M., Liu, H., Purushotham, S., Murali, P., Liu, Y.: Neural Interaction Transparency (NIT): Disentangling Learned Interactions for Improved Interpretability (2018)
- Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Proceedings of the34th International Conference on Machine Learning ICML'17, Sydney, NSW, Australia, JMLR.org(August 2017)

IRIT Institut de Recherche en Informatique de Toulouse CNRS - INP - UT3 - UT1 - UT2J
UNIVERSITÉ TOULOUSE III PAUL SABATIER