

INTERPRÉTABILITÉ DES RÉSEAUX DE NEURONES PROFONDS

Exemple du Convnet

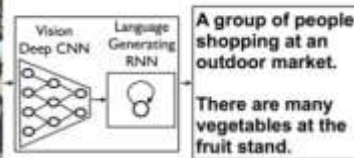
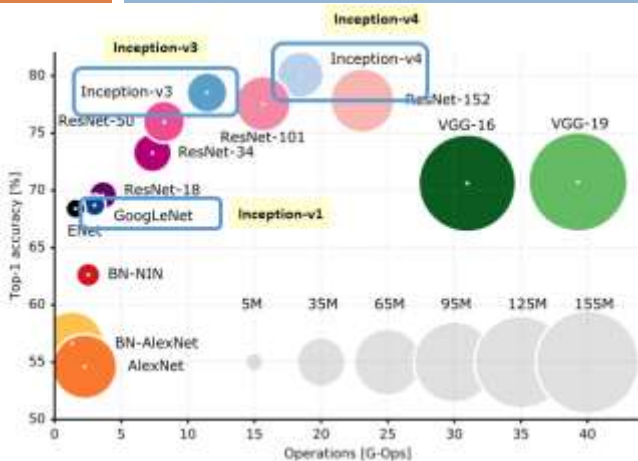
Hajji Hicham

- Associate Professor, Computer Sciences and Spatial Databases
- Research Interest:
 - ▣ Distributed Computing with : applications to Spatial Big Data
 - ▣ Deep Learning applications to Imageries and SpatioTemporal Time Series

Plan

- Pourquoi étudier l'interprétabilité en Deep Learning?
- Approches d'interprétabilité
- Au delà des approches classiques
- Code avec Keras-vis, LIME, Lucid, DeepExplain
- Conclusion

Malgré les Succès stories en Deep Learning



<http://googleresearch.blogspot.com.es/2014/11/a-picture-is-worth-thousand-words.html>

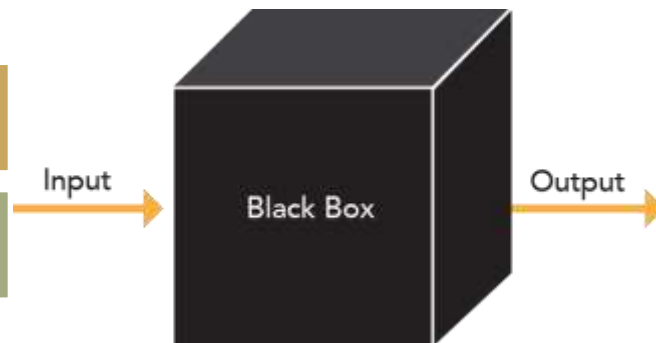
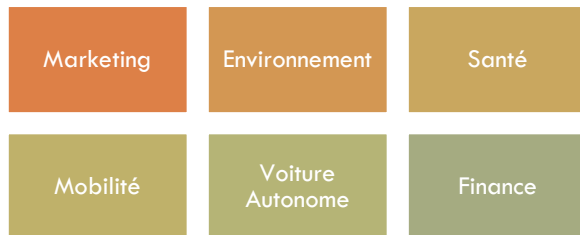
CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar*, Jeremy Irvin*, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

<https://stanfordmlgroup.github.io/projects/chexnet/>

Applications dans

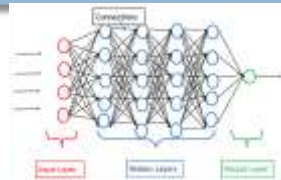


Interprétabilité en DL

- Dans la bibliographie, il existe deux termes:
 - Interprétabilité/Interpretability:
 - 📖 Comprendre ce que le modèle a prédit
 - Explicabilité /Explainability
 - 📖 Résumer les raisons derrière le comportement d'un réseau de neuronne



Mon Convnet pense que vous avez besoin d'une opération. Je ne sais pas pourquoi, mais il a été toujours à 99% précis dans le passé



Je vous propose une politique économique parce que mon IA m'as dis qu'elle va améliorer votre vie de citoyens. Allez vous voter pour lui/elle?

Besoins en interprétabilité

Comprendre la décision



Input
Chest X-Ray Image

CheXNet
121-layer CNN

Output
Pneumonia Positive (85%)



source: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning [Pranav Rajpurkar](#) et, , al

Et L'erreur aussi



Volcano misclassified as Barn spider



Visualizing CNN's decision shows why the Volcano was misclassified as Barn spider

Source: http://mxnet.incubator.apache.org/versions/master/tutorials/vision/cnn_visualization.html

Besoins en interprétabilité

Intriguing properties of neural networks

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus

(Submitted on 21 Dec 2013 (v1), last revised 19 Feb 2014 (this version, v4))

Explaining and Harnessing Adversarial Examples

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy

(Submitted on 20 Dec 2014 (v1), last revised 20 Mar 2015 (this version, v3))

Identifier les exemples contradictoires

- Récemment, des exemples contradictoires (adversarial examples) trompent des réseaux bien robustes dans leurs tâches de classification
- Ils ne sont spécifiques aux réseaux de neurones de convolutions (presque tous les modèles DL)



x
"panda"
57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$
"nematode"
8.2% confidence

=



$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
"gibbon"
99.3 % confidence

Besoins en interprétabilité



Vers “un droit à l’explication” dans la directive européenne GDPR (25 Mai 2018):

« La loi numérique exige depuis 2017 une transparence entière et sans concession de tous les algorithmes décisionnels à caractère individuel, utilisés par la fonction publique.
Écrit autrement, les algorithmes qui fournissent à l'administration une décision sur un citoyen doivent être explicités publiquement ... ces algorithmes ... doivent être disponibles à tout moment par tous les individus de notre pays».

 **Pedro Domingos**
@pmddomingos

Following

Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal.

7:59 PM · 28 Jan 2018

Le Point

Par Aurélie Jean

Publié le 20/04/2019 à 16:29

Is Deep Learning Going To Be Illegal In Europe?



RICHA BHATIA · JAN 30, 2018

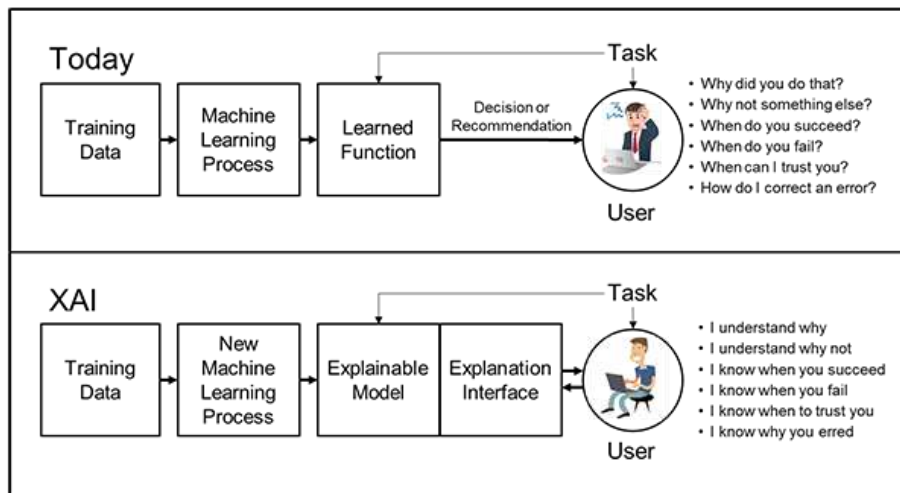
Besoins en interprétabilité



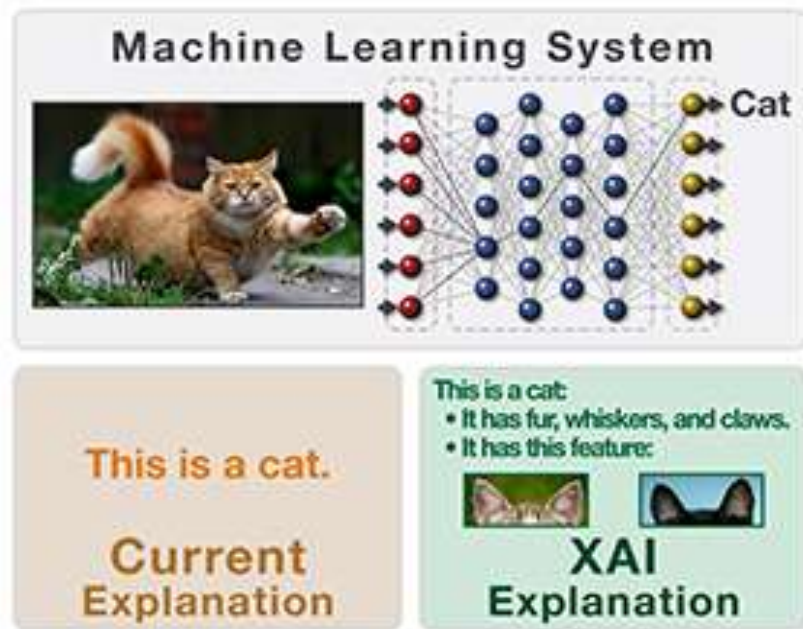
Defense Advanced Research Projects Agency > Program Information

Remettre l'utilisateur au centre

Explainable Artificial Intelligence (XAI)



source: <https://www.darpa.mil/program/explainable-artificial-intelligence>



Besoins en interprétabilité

Rendre le modèle responsable de ses prédictions

(Accountability to model predictions)

Tesla car that crashed and killed driver was running on Autopilot, firm says

- Company says driver took no action despite system's warnings
- Uber settles with family of woman killed by self-driving car



Besoins en interprétabilité

Comprendre les diagnostics médicaux:


Dans le domaine médical, plusieurs modèles voient le jour et qui surpassent les pratique médicales actuelles pour le diagnostic de maladies

Des décisions importantes découlent de ces diagnostics, d'où le besoin de comprendre (dans les deux cas: **classification et mis-classification**)

nature
biomedical engineering

Editorial | Published: 10 October 2018

Towards trustable machine learning

Nature Biomedical Engineering **2**, 709–710 (2018) | [Download Citation](#) 



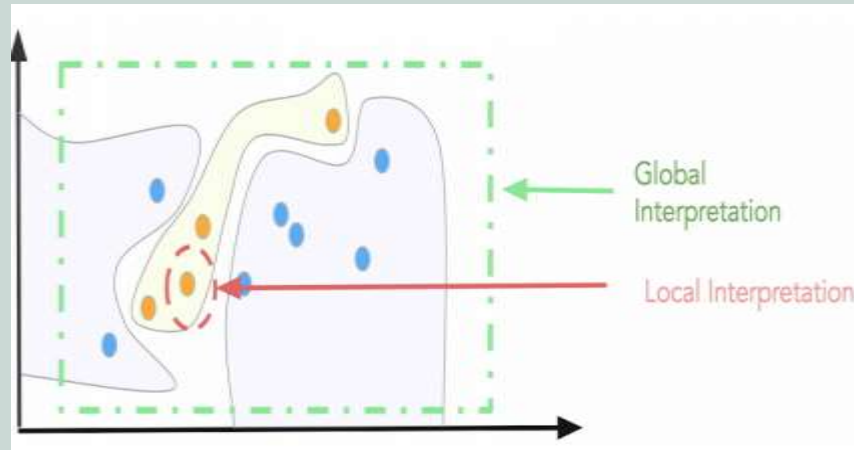
Real-time detection of polyps during colonoscopy, via machine learning.

source: <https://www.nature.com/articles/s41551-018-0315-x>

Autres concepts autour de l'interprétabilité



Scope: Global vs Local



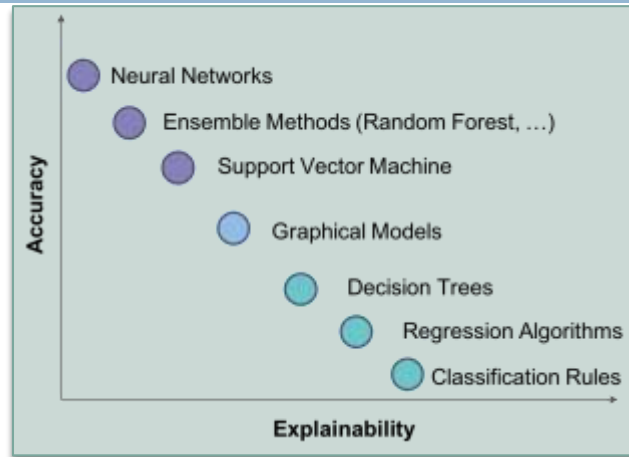
Type de la technique: Model agnostic vs Model dependent

Autres concepts autour de l'interprétabilité

- **Domaines d'interprétabilité**

- **Avant la construction du modèle DL**
 - comprendre les données
- **Pendant la construction du modèle DL**
 - comprendre l'entraînement de données
- **Après la construction**
 - Attribution
 - Feature visualization

Nous nous concentrerons dans cette présentation sur la 3ème partie

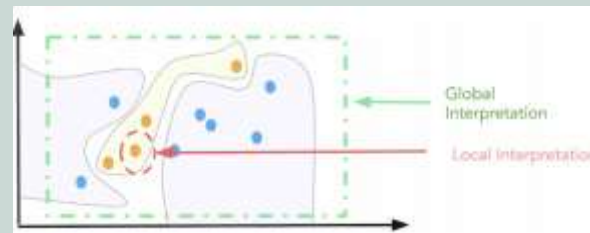


Autres Besoins

- Comprendre et élucider le processus d'apprentissage
- Construire une confiance (Trust) dans le modèle d'apprentissage
- Comprendre pourquoi le réseau a échoué dans la classification
- Elle aider à proposer de meilleurs architectures
- Besoin de "**Déboguer**" un réseau de neurones comme un programme traditionnel
- ...



Scope: Global vs Local



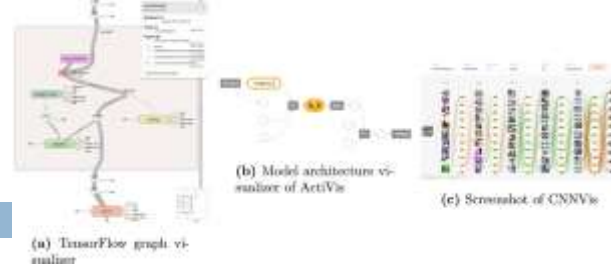
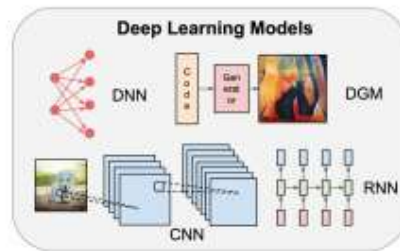
Type of technique: Model agnostic vs Model dependent

Outils de visualisations



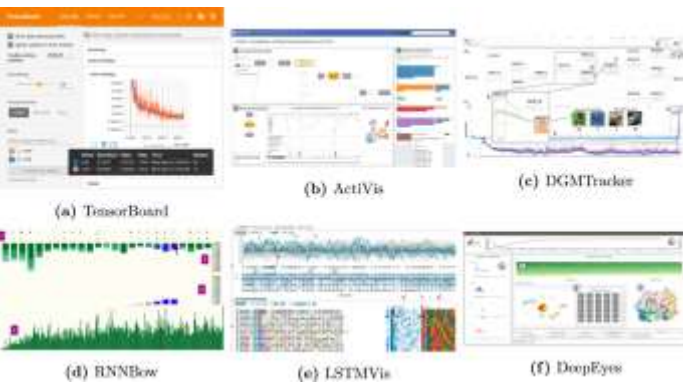
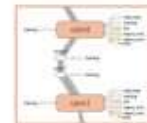
Beginner

Tools for Teaching Concepts



Practitioner

Architecture Assessment



Tools for Debugging and Improving Models

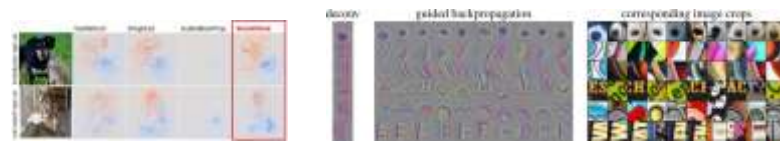


Expert

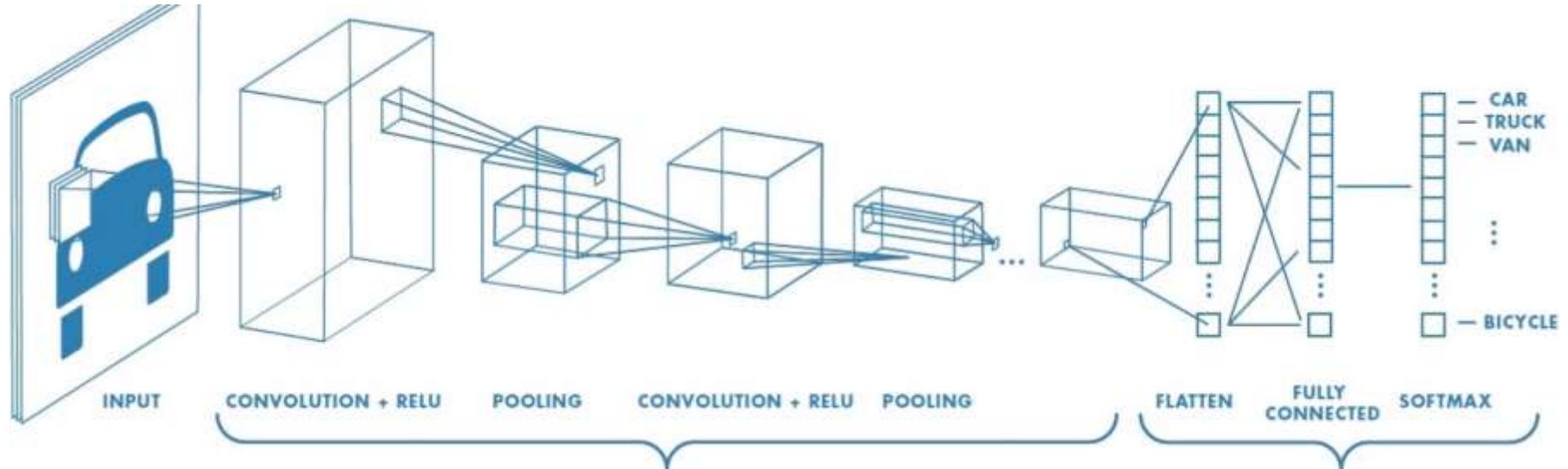
Visual Explanation



Source: A user-based taxonomy for deep learning visualization [RuleiYu, LeiShi, Sept 2018](#)



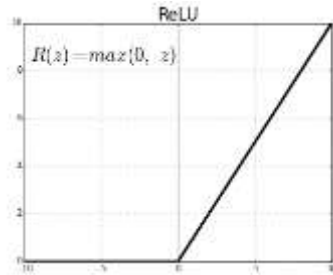
Rappel : Réseau de Neurones de Convolutions



- 1- Une succession de couches de convolution suivies
- 2- fonctions d'activations (ajouter de la non linéarité au réseau)
- 3- couche de pooling pour réduire la dimensionnalité des données
- 4- on obtient à la fin un vecteur de représentation que l'on passe à une couche de classification

Rappel : Passe avant et arrière

Utilisant la fonction d'activation RELU



ReLU Layer

Filter 1 Feature Map

9	3	5	-8
-6	2	-3	1
1	3	4	1
3	-4	5	1



9	3	5	0
0	2	0	1
1	3	4	1
3	0	5	1

Passe avant

Function

ReLU activation

Formula

$$R = \max(0, Z)$$

Derivative

$$R'(Z) = \begin{cases} 0 & Z < 0 \\ 1 & Z > 0 \end{cases}$$

La fonction d'activation de ReLU $g(z) = \max\{0, z\}$
n'est pas différentiable à $z = 0$.

Une fonction est différentiable à un point particulier s'il existe des dérivées de gauche et des dérivées de droite et que les deux dérivées sont égales à ce point.

ReLU est différentiable à tout point sauf le pt 0:
la dérivée de gauche à $z = 0$ est 0 et la dérivée de droite est 1

Passe arrière

Approches d'interprétabilité

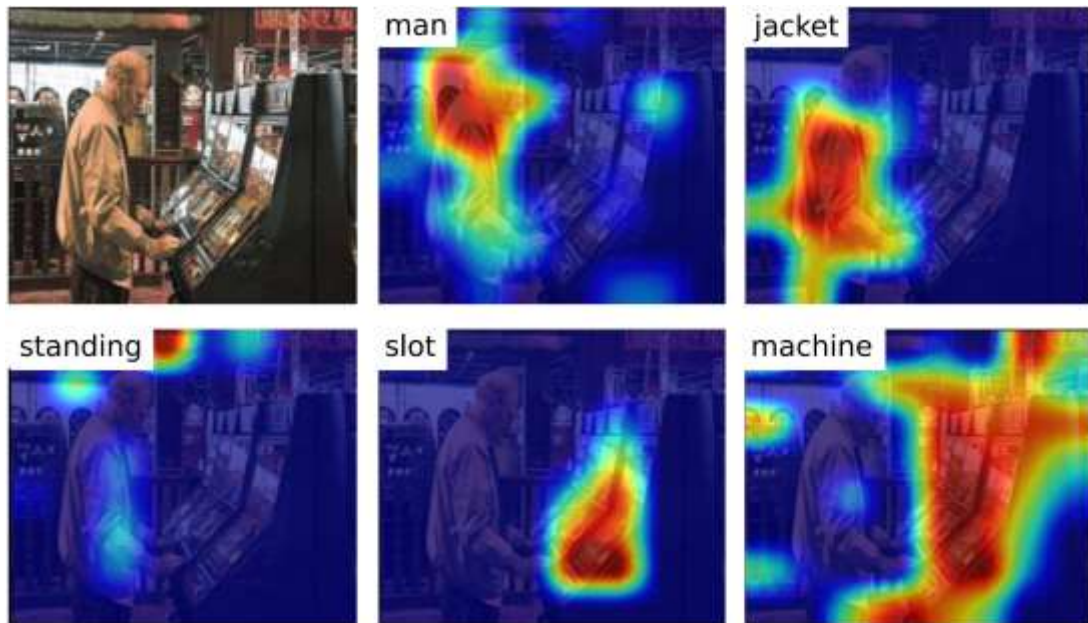
**Il y a deux
principales
approches:**

Attributions

**Visualisation des
patterns appris par le
réseau (Feature
Visualization)**

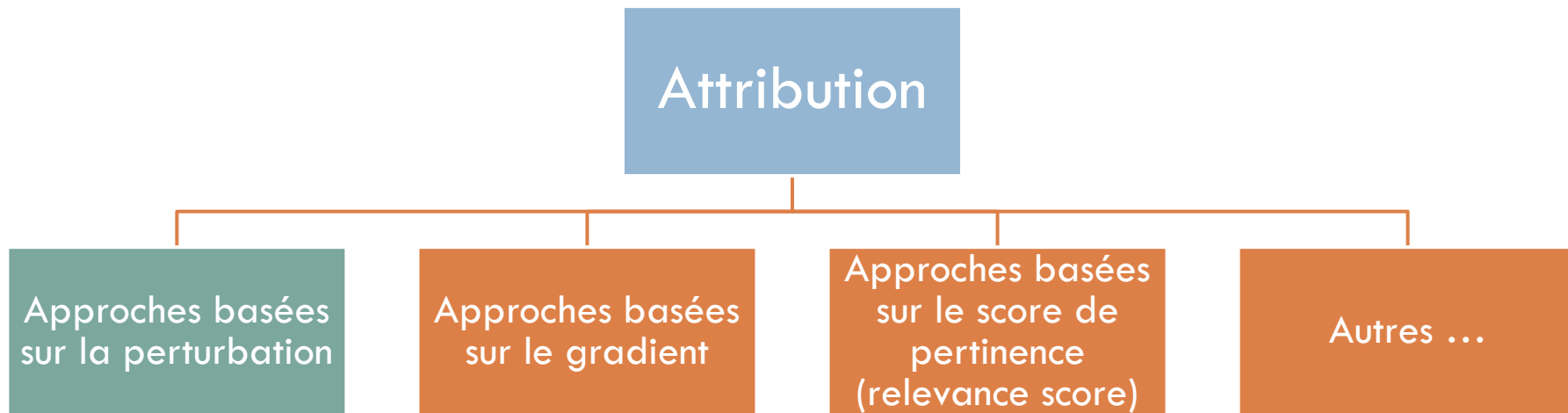
Attribution

Spécifier le rôle et l'importance de chaque attribut INPUT par rapport à la décision de classification



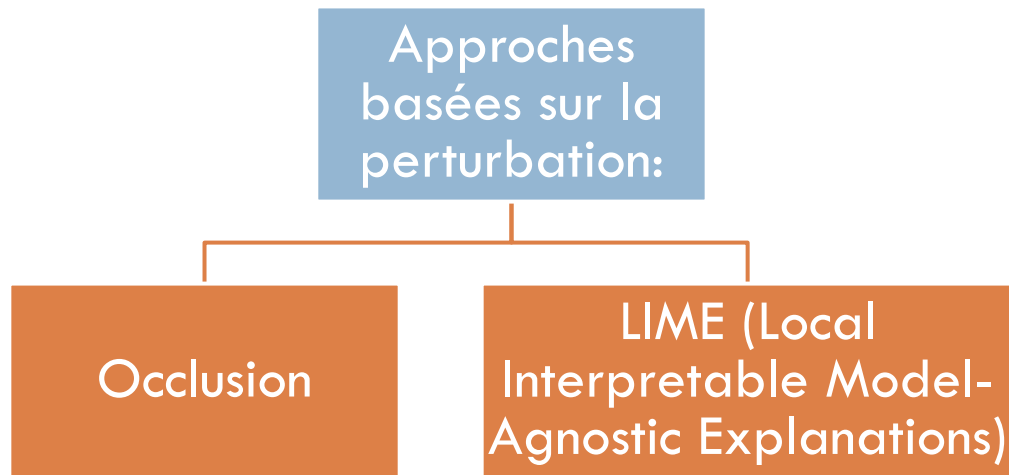
Source: <http://ai.bu.edu/caption-guided-saliency/index.html>

Attribution



Approches basées sur la perturbation

- Nous cherchons à comprendre quelles parties de l'image sont responsables pour maximiser la probabilité d'une certaine classe
- Nous pourrions masquer (griser) différents parties dans l'image et voir les effet sur la probabilité prédite de la bonne classe



Approches basées sur la perturbation

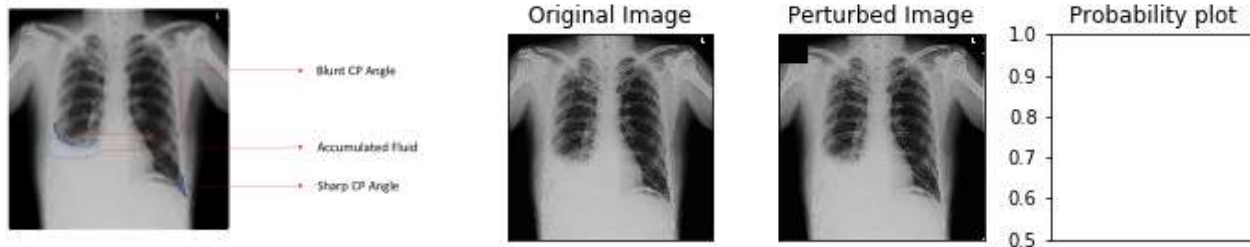
Occlusion

Visualizing and Understanding Convolutional Networks

Matthew D Zeiler, Rob Fergus

(Submitted on 12 Nov 2013 (v1), last revised 28 Nov 2013 (this version, v3))

Occlusion Visualization



source:<http://blog.que.ai>

Problème

- Sensibles suite à l'occlusion partielle de l'image
- L'objet à identifier est souvent bouché par endroits, ce qui entraîne une décision inappropriée de la part du réseau:

Solution proposée: **Super-Pixel Perturbation**

Approches basées sur la perturbation

LIME

Local Interpretable Model-Agnostic Explanations

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

(Submitted on 16 Feb 2016 (v1), last revised 9 Aug 2016 (this version, v3))



Original Image
 $P(\text{tree frog}) = 0.54$






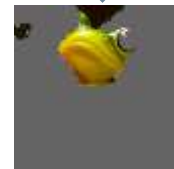
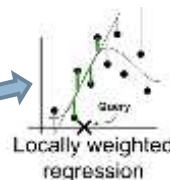
Segmentation avec **Quick Shift**
en SuperPixels contigus



Interpretable
Components



Perturbed Instances	$P(\text{tree frog})$
	<div><div></div></div> 0.85
	<div><div></div></div> 0.00001
	<div><div></div></div> 0.52



- **Approche agnostique** : approche utilisable avec n'importe quel modèle de machine learning: Arbre de décision, Random forest, Réseau de neurones...
- La technique tente de comprendre le modèle **en perturbant des échantillons de données** et en comprenant comment les prévisions changent. Le résultat de LIME est une liste d'explications reflétant la contribution de chaque caractéristique à la prédiction d'un échantillon de données.

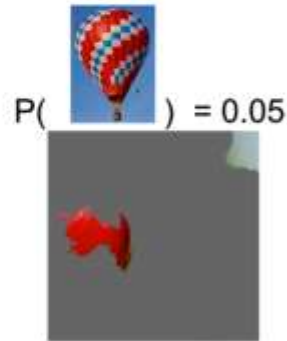
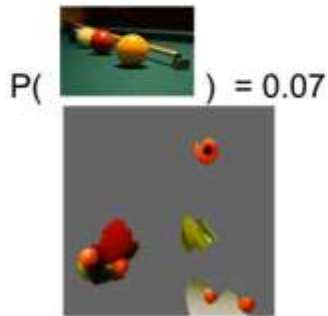
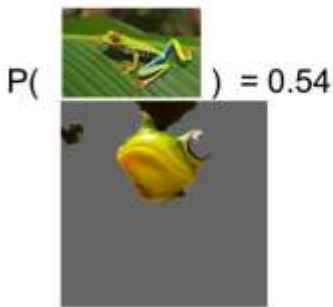
Approches basées sur la perturbation

LIME

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

(Submitted on 16 Feb 2016 (v1), last revised 9 Aug 2016 (this version, v3))

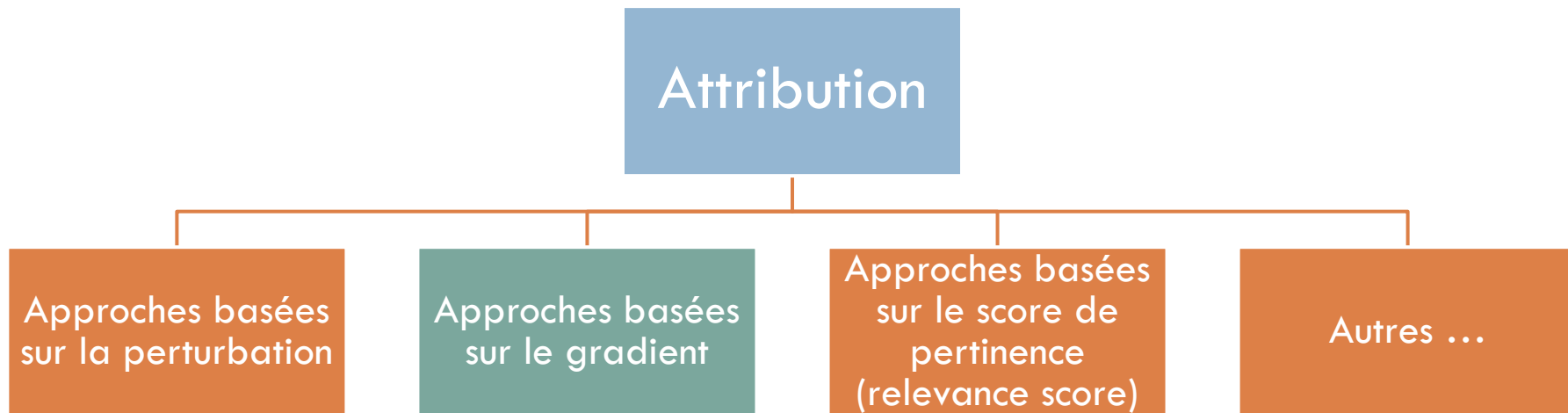


Explication d'une prédiction avec le modèle Inception. Les trois principales classes prédites sont "rainette", "table de billard" et "ballon". Sources: Marco Tulio Ribeiro, Pixabay (grenouille, billard, montgolfière).

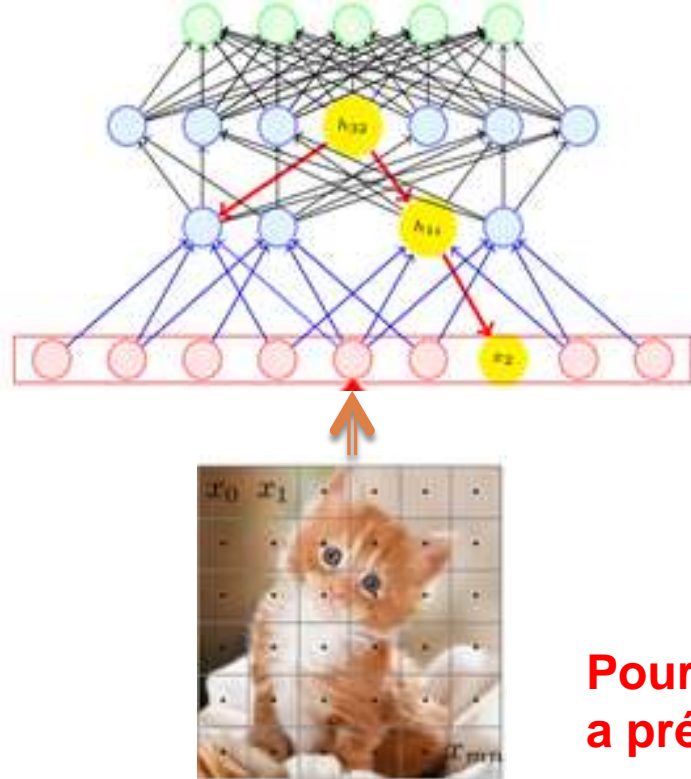
Problèmes des approches basées Perturbation

- Sensibles suite à l'occlusion partielle de l'image
- Elles affectent drastiquement les prédictions des Convnets
- Critiques de LIME??

Attribution



Approches basées sur le gradient



- Nous avons une image $m \times n$ pixels ($x_1, x_2 \dots x_m \times n$)
- Nous sommes intéressés par trouver l'influence de pixels (x_i) sur chaque neurone (h_j)
- Si un changement petit de x_i cause un grand changement dans h_j , nous disons que x_i a beaucoup d'influence sur h_j
- Cela revient à calculer le gradient $\frac{\partial h_j}{\partial x_i}$

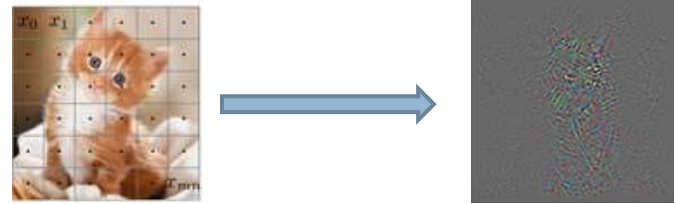
$$\frac{\partial h_j}{\partial x_i} = 0 \quad \rightarrow \text{Aucune Influence}$$

$$\frac{\partial h_j}{\partial x_i} = \text{large} \quad \rightarrow \text{Grande Influence}$$

$$\frac{\partial h_j}{\partial x_i} = \text{small} \quad \rightarrow \text{Faible Influence}$$

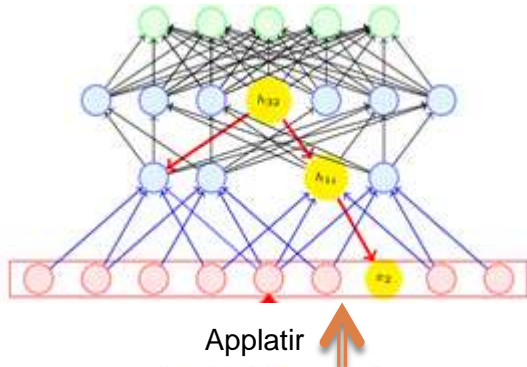
- Et visualiser cette matrice de gradient dans une image

**Pourquoi le modèle
a prédit un chat**



Approches basées sur le gradient

- Trouver l'influence des pixels de l'image en utilisant la **rétropropagation** (En calculant le gradient du score de la classe par rapport aux pixels de l'entrée)



Principales approches

Saliency Maps

Deconvolution

Guided
Backpropagation

Elles se distinguent par la manière avec laquelle **le gradient est calculé**

Saliency Maps

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

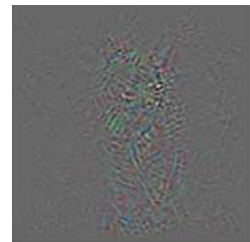
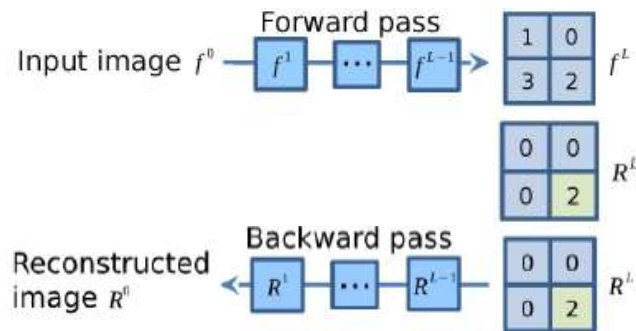
Karen Simonyan, Andrea Vedaldi, Andrew Zisserman

(Submitted on 20 Dec 2013 (v1), last revised 19 Apr 2014 (this version, v2))

Saliency Maps

$$\frac{\partial S_c}{\partial I} \Big|_{I_0} \quad \text{Ou} \quad \frac{\partial S_c}{\partial I} \Big|_{I_0} * I_0$$

Papier DeepLIFT: gradient * input



Ne montre pas une influence assez claire des pixels sur la totalité de l'image
(**Trop de bruits**)

Pourquoi le modèle a prédit un chat

Etant une image I_0 , une classe c , et une classification convnet avec fonction score de la classe $S_c(I)$.

La carte heatmap est calculée comme **la valeur absolue du gradient** de S_c par rapport à I à I_0

Deconvolution



Visualizing and Understanding Convolutional Networks

Matthew D Zeiler, Rob Fergus

[Submitted on 12 Nov 2013 (v1), last revised 28 Nov 2013 (this version, v3)]

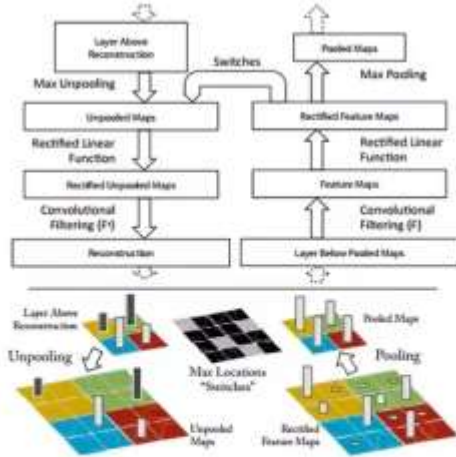
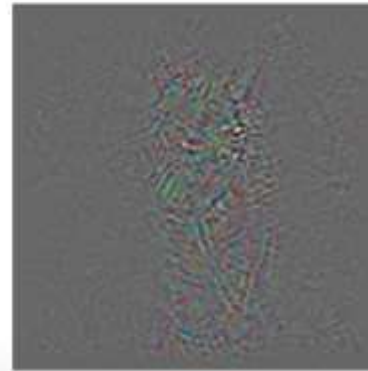


Fig. 1. Top: A deconvnet layer (left) attached to a convnet layer (right). The deconvnet will reconstruct an approximate version of the convnet features from the layer beneath. Bottom: An illustration of the unpooling operation in the deconvnet, using *switches* which record the location of the local max in each pooling region (colored zones) during pooling in the convnet. The black/white bars are negative/positive activations within the feature map.

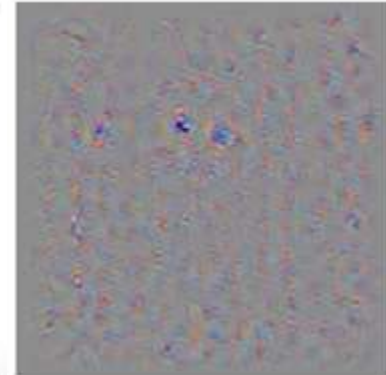
Input image



Backpropagation



Deconvolution



**Pourquoi le modèle
a prédit un chat**

Toujours Ne montre pas une influence assez claire des pixels sur la totalité de l'image (**Trop de bruits**)

Guided Backpropagation



Striving for Simplicity: The All Convolutional Net

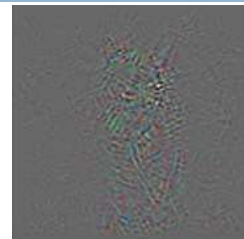
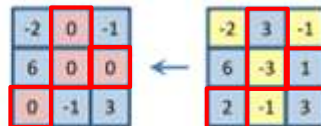
Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller

(Submitted on 21 Dec 2014 (v1), last revised 13 Apr 2015 (this version, v3))



backpropagation: $R_i^l = (f_i^l > 0) \cdot R_i^{l+1}$, where $R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}$

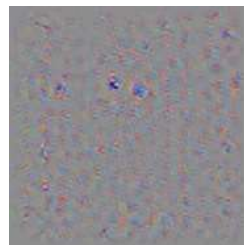
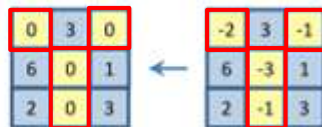
Backward pass:
backpropagation



le gradient ne passera que
par si les activations sont
POSITIVES

backward
'deconvnet': $R_i^l = (R_i^{l+1} > 0) \cdot R_i^{l+1}$

Backward pass:
"deconvnet"



Le gradient ne passera que
s'il est POSITIF

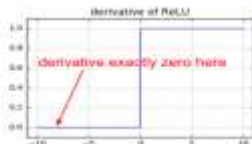
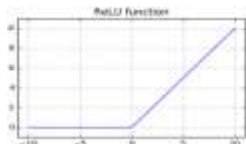
b)

Forward pass



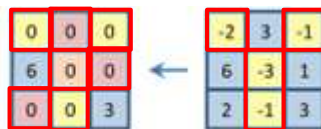
activation:

$$f_i^{l+1} = \text{relu}(f_i^l) = \max(f_i^l, 0)$$



guided
backpropagation: $R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1}$

Backward pass:
guided
backpropagation



le gradient ne passera que
s'il est POSITIF
et si les activations sont
POSITIVES

Inconvénients du Guided Backprop



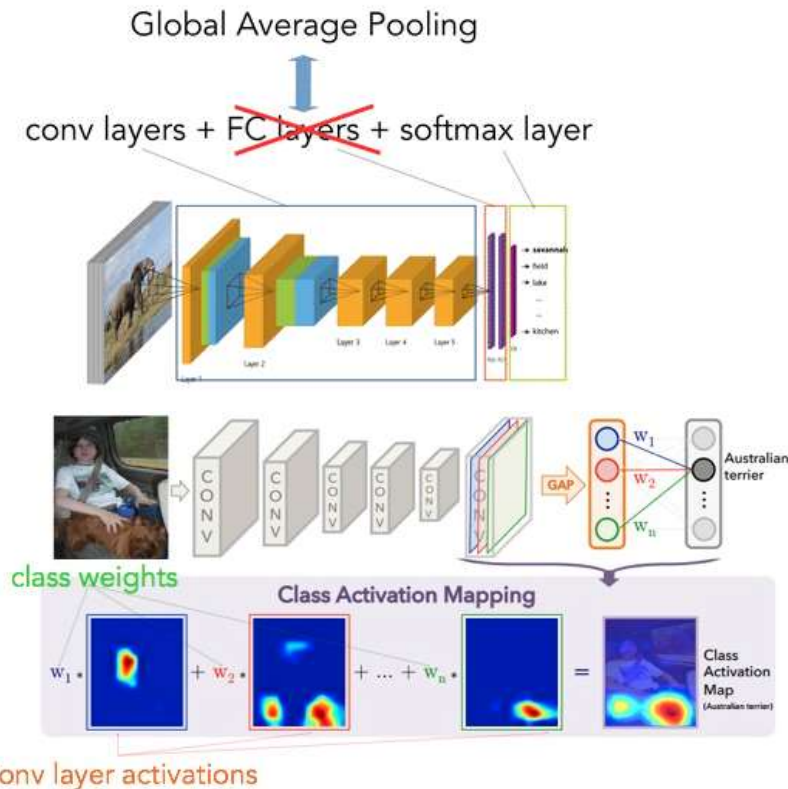
Pourquoi le modèle
a prédit un Chat

Pourquoi le modèle
a prédit un Chien

- La visualisation est **incapable de distinguer** les pixels du chat et du chien.
- Non class-discriminative : Ne discrimine pas entre les classes

Class Activation Mapping

- Les couches profondes dans le réseau captent les concepts visuels de haut niveau, mais elle sont perdues dans les FC (vecteur scalaire et non une image).
- hypothèse: pourquoi ne pas se concentrer sur les dernières couches de convolutions qui ont:
 - 1- le max de sémantique haut niveau (parties d'objets)
 - et 2- la localisation des concepts plus détaillée
- Solution CAM:
 - Modifier la structure du réseau en utilisant GAV
 - et le réentraîner
- Les poids appris dans la dernière couche correspondent aux poids d'importance des neurones, c'est-à-dire l'importance des cartes de caractéristiques pour chaque classe d'intérêt.



Grad-Class Activation Mapping

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

(Submitted on 7 Oct 2016 (v1), last revised 21 Mar 2017 (this version, v3))

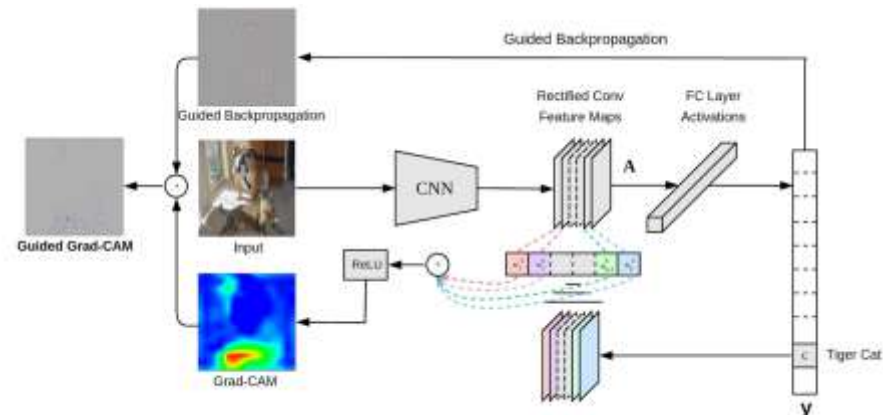
Problème avec CAM:

- Peut on avoir ces visualisations sans à avoir à changer l'architecture du réseau et sans réentraînement??? ==> Grad-CAM

Comment?????? Grad-CAM fait appel au gradient :-)

- Grad-CAM calcule le gradient du score pour la classe c par rapport à la feature maps de la couche conv A : $\frac{\partial y^c}{\partial A_{ij}^k}$
- Ces gradients retro propagés sont mis en commun pour obtenir l'importance des poids:

$$\alpha_k^c = \overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$



Grad-CAM for "Cat"



Grad-CAM for "Dog"



Grad-Class Activation Mapping

Advantages:

- Class-discriminative
- Donner des “explications” sur des prédictions apparemment déraisonnables,
- Applicable à une large variété de modèles qui utilisent l’architecture (<http://gradcam.cloudcv.org/>):
 - (1) CNNs avec fully-connected layers (exp. VGG),
 - (2) CNNs pour des sorties structurées (Image captioning),
 - (3) CNNs utilisés dans des tâches avec des inputs multi-modals(e.g. VQA)

Visual Question Answering

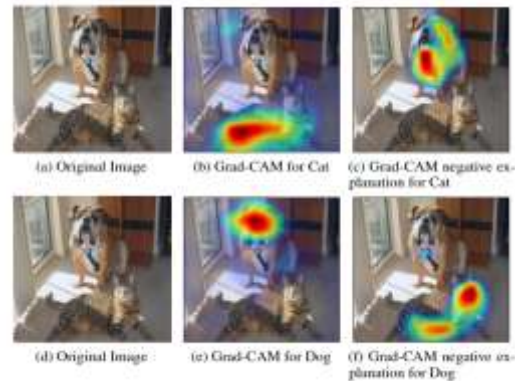
[Try Visual Question Answering Demo](#)

Classification

[Try Classification Demo](#)

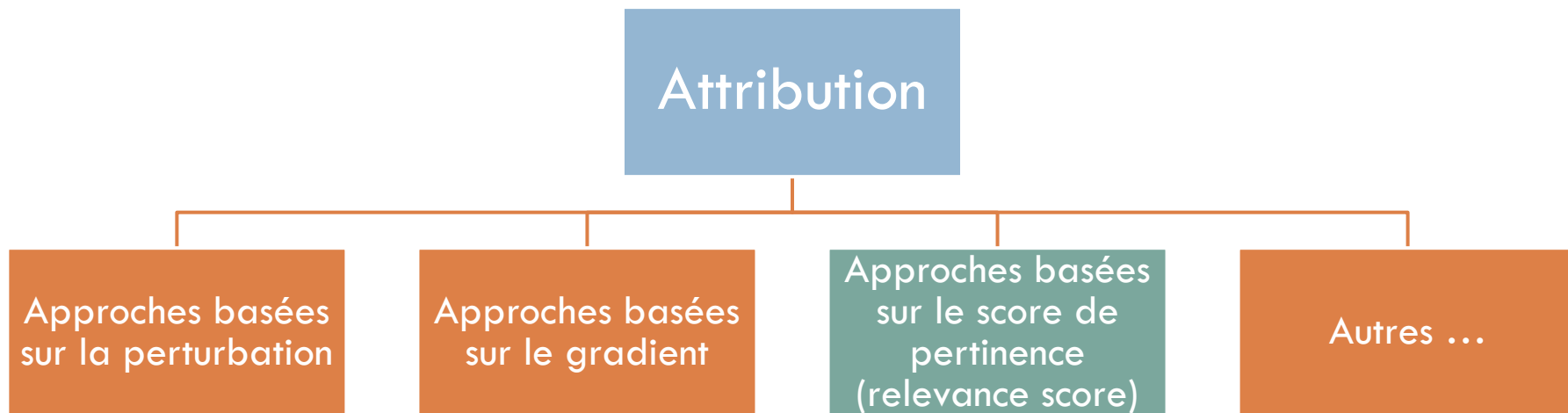
Captioning

[Try Captioning Demo](#)



- Robuste aux images “adversarial”, (selon les auteurs)

Attribution



Les approches basées sur le score de pertinence

- les approches basées sur le gradients souffrent de problèmes:
 - Gradients discontinus** pour certaines fonctions d'activations non linéaires:
 - les discontinuités dans les gradients provoquent des artefacts indésirables.
 - De plus, l'attribution ne se répand pas correctement en raison de telles non-linéarités, ce qui entraîne une distorsion des scores d'attribution.
 - Saturation des gradients:**
 - comme expliqué dans ce réseau simpliste, les gradients lorsque i_1 ou i_2 est supérieur à 1, le gradient de la sortie par rapport à chacun d'eux ne change pas tant que $i_1 + i_2 > 1$.

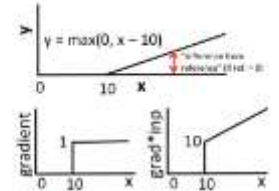


Figure 2. Discontinuous gradients can produce misleading importance scores. Response of a single rectified linear unit with a bias of -10 . Both gradient and gradient \times input have a discontinuity at $x = 10$; at $x = 10 + \epsilon$, gradient \times input assigns a contribution of $10 + \epsilon$ to x and -10 to the bias term (ϵ is a small positive number). When $x < 10$, contributions to x and the bias term are both 0. By contrast, the difference-from-reference (red arrow, top figure) gives a continuous increase in the contribution score.

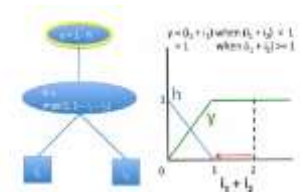
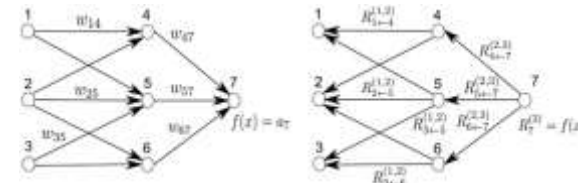
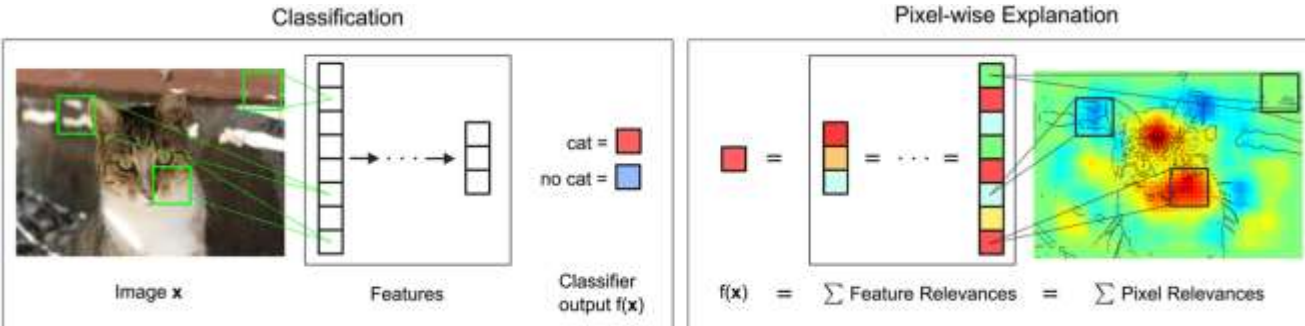
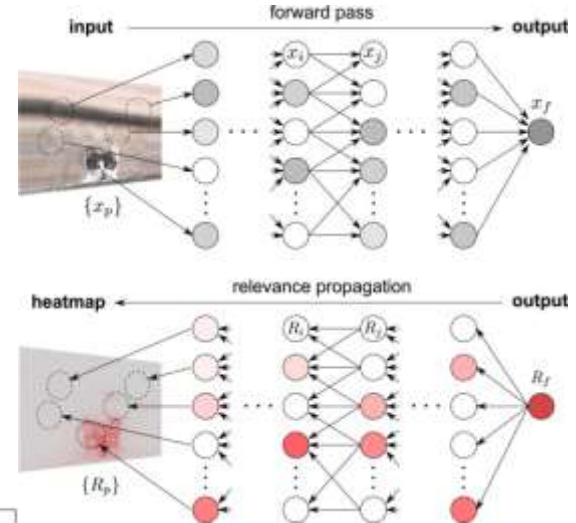


Figure 1. Perturbation-based approaches and gradient-based approaches fail to model saturation. Illustrated is a simple network exhibiting saturation in the signal from its inputs. At the point where $i_1 = 1$ and $i_2 = 1$, perturbing either i_1 or i_2 to 0 will not produce a change in the output. Note that the gradient of the output w.r.t the inputs is also zero when $i_1 + i_2 > 1$.

Layerwise Relevance Propagation

- Layer-wise Relevance Propagation (LRP) est une méthode qui identifie les pixels importants en effectuant une rétropropagation dans le réseau de neurones: **demo: <http://heatmapping.org/>**
 - La passe en arrière est une procédure de redistribution de pertinence conservatrice, où les neurones qui contribuent le plus à la couche supérieure en reçoivent la plus grande pertinence.
- À chaque couche, la valeur de pertinence totale est conservée** et la propagation finale mappe la pertinence sur le vecteur d'entrée. Ce processus est semblable dans l'esprit à la rétropropagation.



Demo

<http://www.heatmapping.org/>



www.heatmapping.org

DeepLIFT

Learning Important Features Through Propagating Activation Differences

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje

(Submitted on 10 Apr 2017)

- Au lieu d'expliquer directement la prédiction de sortie dans les modèles précédents, les auteurs expliquent **la différence entre la prédiction de sortie et la prédiction sur une image de référence de base**
- Les auteurs soulèvent une préoccupation valable avec les méthodes basées sur les gradients décrites ci-dessus - les gradients n'utilisent pas de référence qui limite l'inférence.
 - En effet, les méthodes basées sur les gradients décrivent uniquement le comportement local de la sortie à une valeur d'entrée spécifique, **sans tenir compte du comportement de la sortie sur une plage d'entrées.**

Comment choisir une approche d'attributions?

Comment choisir une approche?

Liées uniquement au
Convnet ou autres
architectures DL(RNN,
LSTM,...)

Ré-entraîner le réseau
ou pas

Quantité de calcul à
réaliser

Type d'approches:
Perturbation,
Gradient,
Pertinence...

Est ce qu'on veut des
explications Fine-
Grained ou Coarse-
Grained (Guider la
rétropropagation)

Agnostique ou
dépend d'un modèle

Approches d'interprétabilités

**Il y a deux
principales
approches:**

Attributions

**Visualisation des patterns
appris par le réseau
(Feature Visualization)**

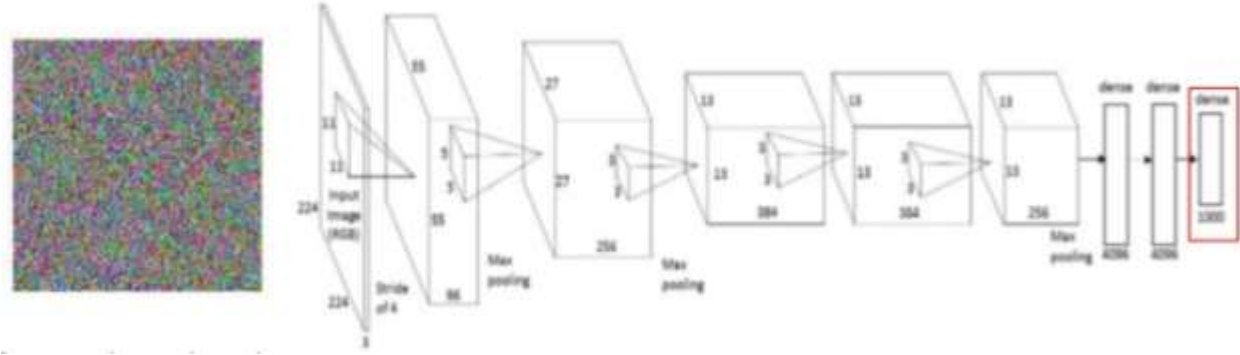
On s'intéresse non pas à l'explication d'une
prédiction du réseau, mais à ce qu'a appris le
réseau lors de la phase d'entraînement

Activation Maximization

L'idée simple:

Générer une image d'entrée qui maximise les activations de sortie du filtre

Il exagère les features liées à la classe



Dans le cas d'un score de classe:

Envoyer dans le réseau une image de bruit (aléatoire)

Fixer le gradient du vecteur des score à 1, 0,0,0,0,0,0,0

Rétropropager

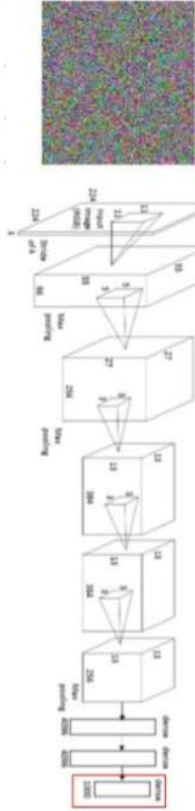
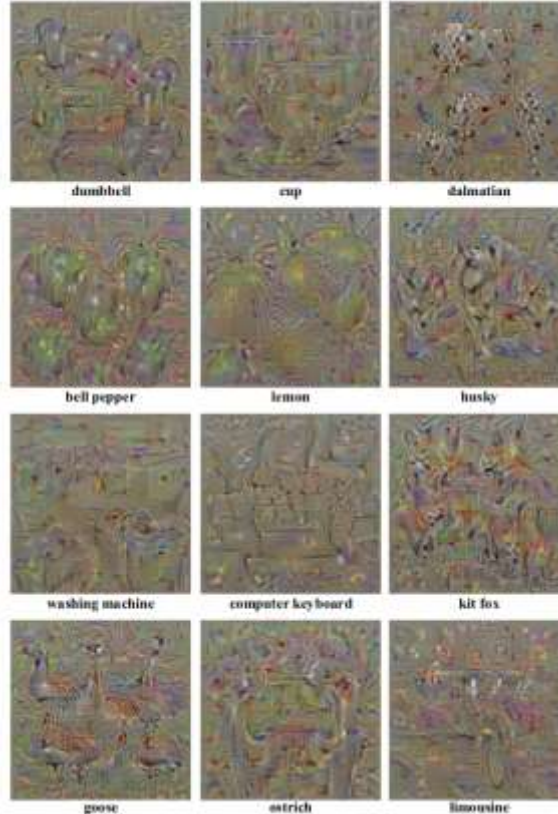
mettre à jour l'image initiale

Répéter

Activation Maximization

L'optimisation isole les causes du comportement de simples corrélations.

Un neurone peut ne pas détecter ce que vous pensiez au départ.



Deep Dream

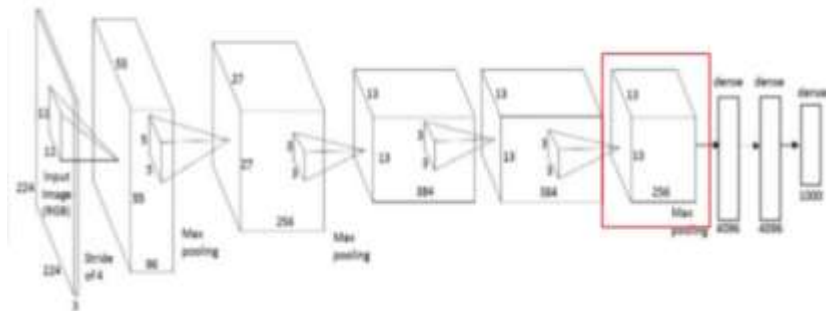


Au lieu de démarrer avec une image bruit, on démarre avec une image réelle, et on exagère les caractéristiques:

Demo : <https://deepdreamgenerator.com/>.

Etapes:

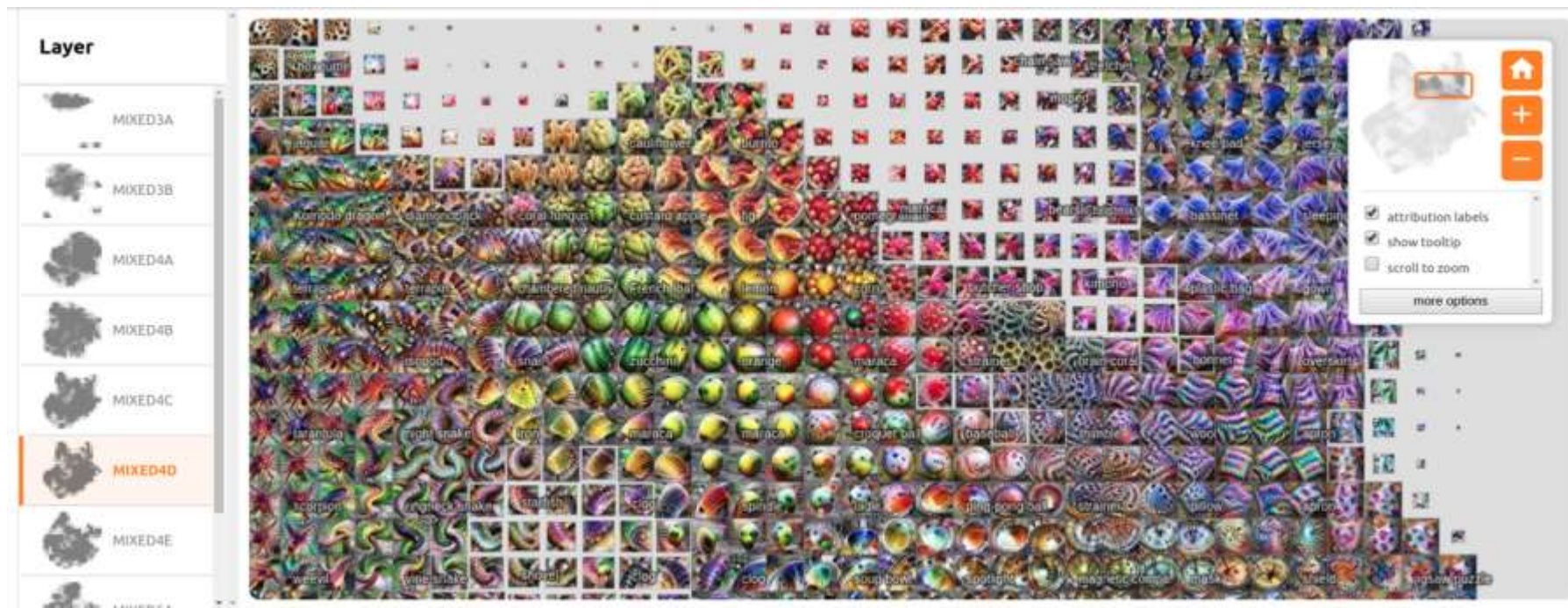
- Transférez l'image jusqu'à une couche (par exemple, conv5)
- Définissez les ~~dégradés~~ de manière à ce qu'ils soient activés sur ce calque
- Backprop (avec régularisation)
- Mettre à jour l'image
- Répétez



À chaque itération, l'image est mise à jour pour **Booster** toutes les caractéristiques (features) activées dans cette couche lors du forward pass.

Atlas d'activation

<https://distill.pub/2019/activation-atlas/>



Network Dissection

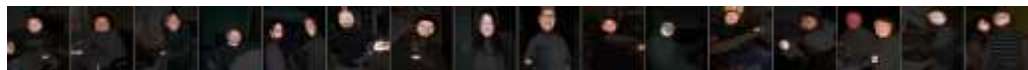
Network Dissection: Quantifying Interpretability of Deep Visual Representations

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba

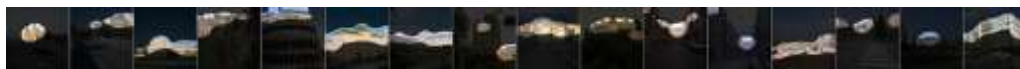
(Submitted on 19 Apr 2017)

Vers des Unités interprétables

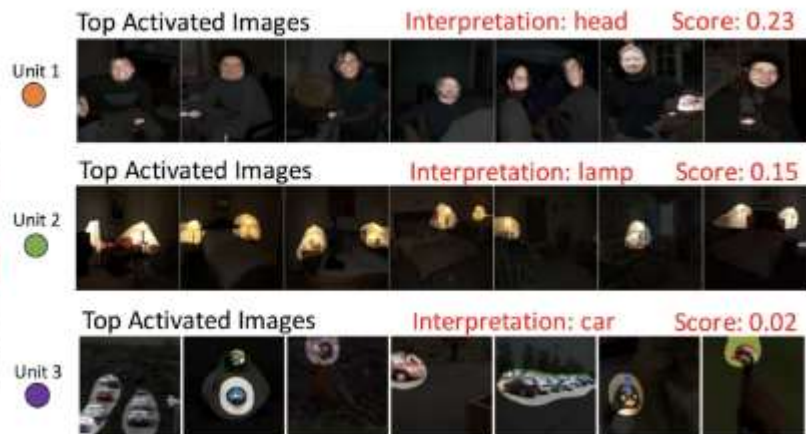
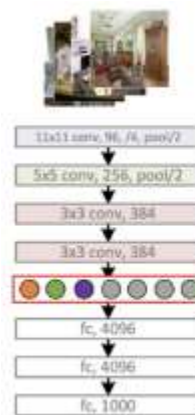
- La dissection de réseau est une méthode pour quantifier l'interprétabilité des unités individuelles dans un CNN profond.
- Il fonctionne en mesurant l'alignement entre la réponse unitaire et un ensemble de concepts tirés d'un ensemble de données de segmentation large et dense appelé Broden.



[AlexNet-Places205 conv5](#) unit 138: heads



[AlexNet-Places205 conv5](#) unit 53: stairways



MIEUX QUE DES MOTS,
UNE DÉMO

AUTRES

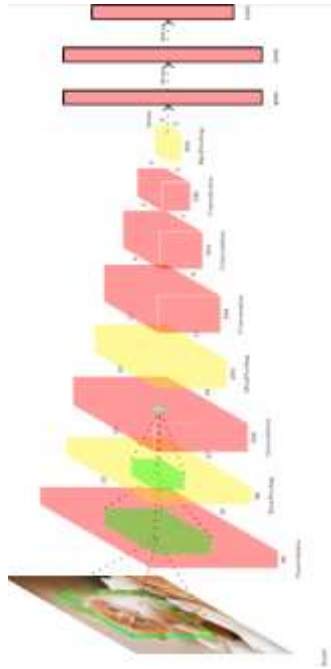
Autres visualisations

les Images qui activent au maximum un neurone

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

(Submitted on 11 Nov 2013 (v1), last revised 22 Oct 2014 (this version, v5))



- Repérer les images qui activent au maximum un neurone
- Prendre un grand ensemble de données d'images, les alimenter à travers le réseau
- et garder une trace des images qui activent au maximum un neurone

Activation maximale des images pour certains neurones **POOL5 (5ème couche de pooling)** d'un AlexNet



Les valeurs d'activation et le champ récepteur du neurone particulier sont indiqués en blanc

Autres visualisations

les Images qui activent au maximum un neurone

Intriguing properties of neural networks

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus

(Submitted on 21 Dec 2013 (v1), last revised 19 Feb 2014 (this version, v4))

Problème:

- Il faut passer un volume important d'images dans le réseau
- Les neurones ReLU n'ont pas nécessairement de sens sémantique par eux-mêmes

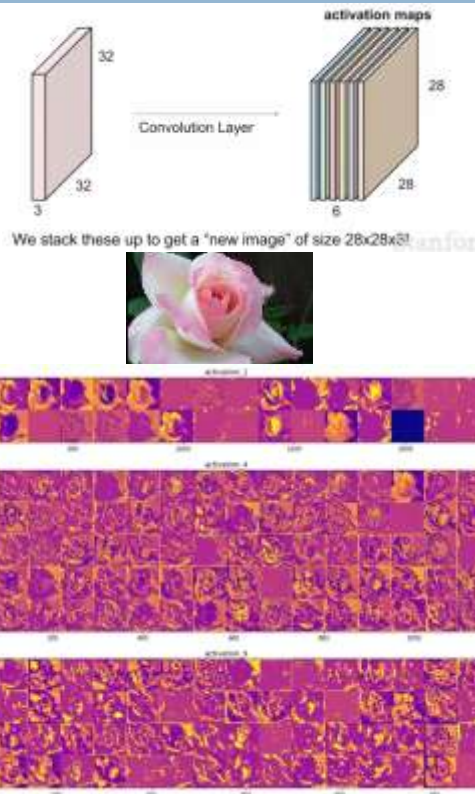
Autres visualisations

- **Deux utilisations:**

- **Visualiser les activations de couches durant les forward pass:** Prendre une image et la passer dans le réseau
 - les couches peu profondes visualisent les motifs généraux tels que les bords, la texture, l'arrière-plan, etc.
 - les couches les plus profondes visualisent les caractéristiques spécifiques aux données d'entraînement (plus liées à la classe)
- **Visualiser les activations de couches durant la phase d'entraînement:**
 - les activations commencent généralement par être relativement denses,
 - A mesure que l'entraînement avance, les activations deviennent généralement plus clairsemées et localisées.

- **Demo :**

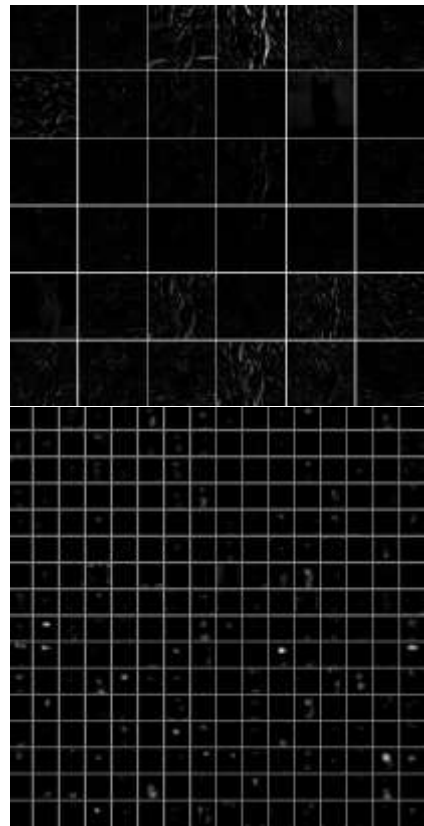
<https://immortal3.github.io/OnlineActivationMap/>



Autres visualisations

Problème:

Avec cette visualisation, certaines cartes d'activation (activation maps) peuvent être toutes nulles pour de nombreuses entrées différentes, ce qui peut indiquer des filtres morts et peut être le symptôme de taux d'apprentissage élevés.



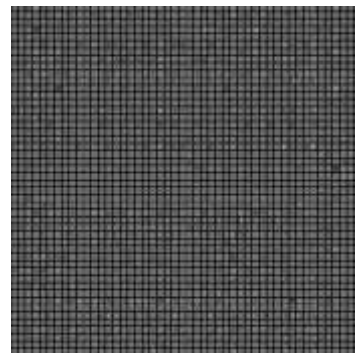
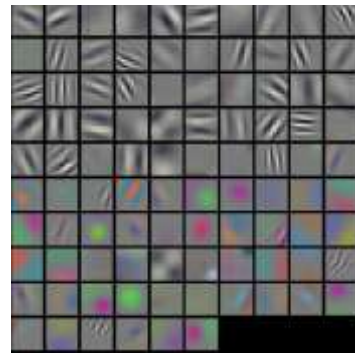
Autres visualisations

Visualiser les filtres Conv/FC durant les forward pass

- Observer les filtres construits lors de la phase d'entraînement
- Les poids sont utiles à visualiser car des réseaux bien formés affichent généralement des filtres lisses et agréables sans motifs bruyants.
- Les schémas avec bruits peuvent indiquer :
 - un réseau qui n'a pas été entraîné suffisamment
 - ou peut-être une régularisation très faible qui peut avoir conduit à un surajustement

Problème:

Uniquement les filtres des la premières couches qui sont interprétable



Autres visualisations

Visualisation de l'espace de représentation

On passe des images dans le réseau

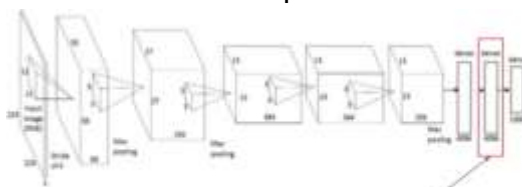
On récupère le vecteur composant la couche FC7 : un vecteur de 4096 neurones

On applique l'algorithme t-sne sur cet ensemble de vecteurs

t-SNE arrange les images qui ont des vecteurs FC7 similaires voisin dans l'espace 2Ds

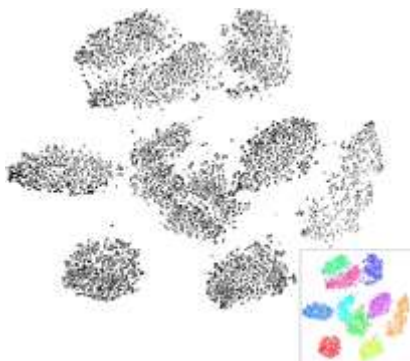
On Affiche ces projections sous forme de carte 2D

Approche non linéaire



Extraire le vecteur FC7 de dimensions 4096.

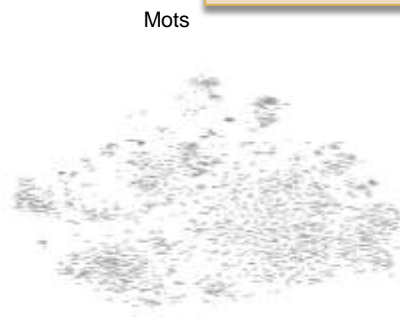
données MNIST



Olivetti
faces



Mots



Visualizing Data using t-SNE

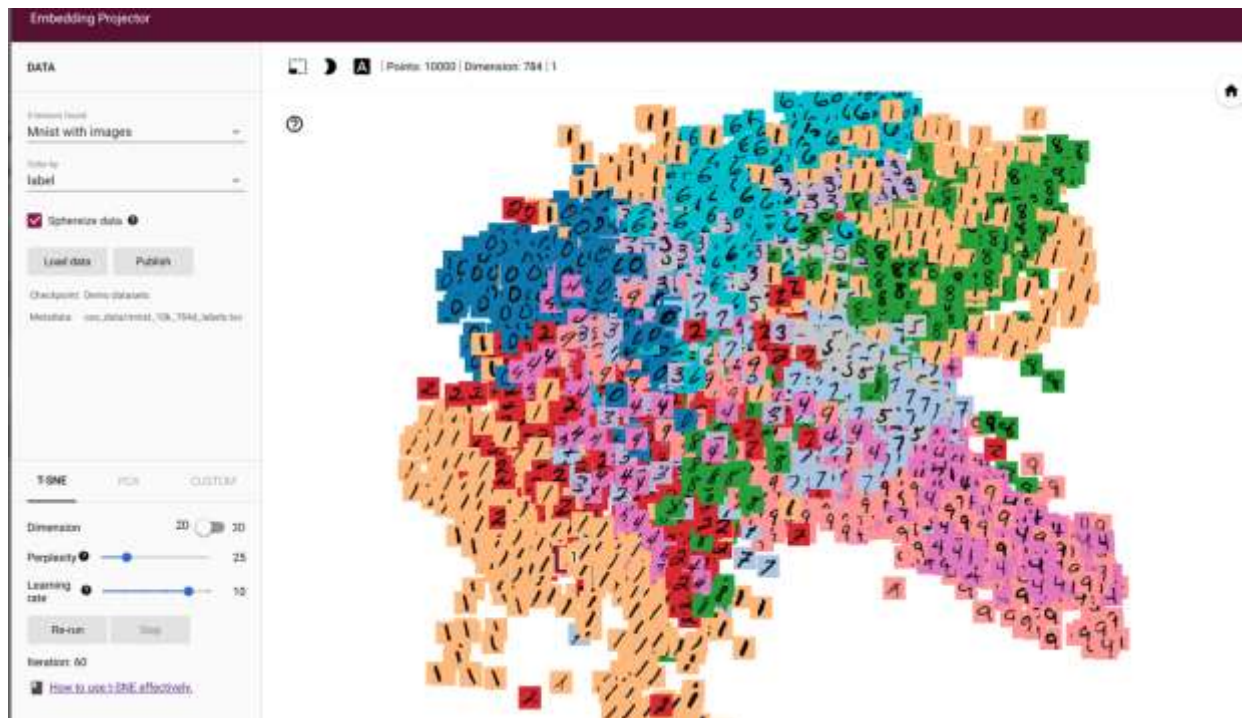
Laurens van der Maaten, Geoffrey E. Hinton • Published 2008

L'algorithme t-SNE (t-distributed stochastic neighbor embedding) est une technique de réduction de dimension pour la visualisation de données développée par Geoffrey Hinton et Laurens van der Maaten

Les distances entre les points à plusieurs dimensions sont conservées dans l'espace à 2D de t-sne: HAND MANIFOLD

t-sne avec Embedding projector

<https://projector.tensorflow.org/>



Précautions avec t-sne

- C'est vrai qu'elle permet de créer des “cartes ” bi-dimensionnelles à partir de données avec des centaines voir des milliers de dimensions, mais **t-sne peut donner des résultats contradictoires**
- De bonnes pratiques sont présentées ici: <https://distill.pub/2016/misread-tsne/>
 - 1: les hyperparamètres à utiliser avec précaution: **perplexité, itération...**



- 2: les tailles des clusters dans t-sne ne signifient rien: pas le nombre de points mais le BBOX

Conclusions

- L'interprétabilité est devenue un Hot Topic dans la recherche AI/DL
- Pour mettre en services de nouveaux produits basés sur l'IA, il est obligatoire de produire des explications aux décisions données
- En connaissant le dessous des modèles DL, on peut mieux les déboguer, voire les améliorer
- On peut les corriger dans les situations de décisions erronées