

Axiomatic Attribution for Deep Networks

#PAPIER #ATTENTION

#EPITA

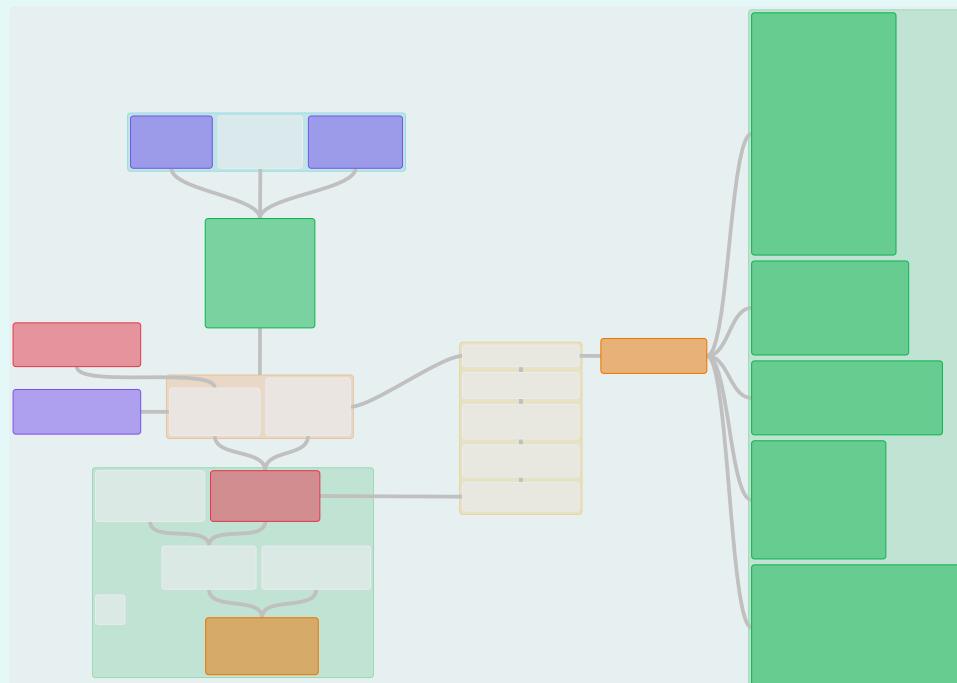
#DEEP-LEARNING

#ML

#EXPLICABILITE



- Lien : <https://arxiv.org/abs/1703.01365>
 - Canvas :
 - Axiomatic Attribution for Deep Networks



Abstract

Cet article s'intéresse à une méthode capable de se "*brancher*" à n'importe quel réseau de neurones pour expliquer les *features* à

l'origine de la prédiction du réseau. Ils ont appelé cette méthodes la [Méthode des Gradients Intégrés](#).

- Dans un premier temps, les auteurs ont identifié **deux axiomes** nécessaires à une bonne identification.
- Ils ont ensuite utilisé ces axiomes pour montrer les **faiblesses** des méthodes précédentes
- Ils ont finalement construit la [Méthode des Gradients Intégrés](#) à partir de ces axiomes.
- Ils l'ont évalué sur des images, du texte, des molécules chimiques etc.

Motivation & Résumé des résultats

✓ Définition mathématiques d'une méthode d'attribution

On peut définir une méthode d'attribution de la sorte:

Etant donné un **réseau de neurones** $F : (x_1, \dots, x_n) \rightarrow [0, 1]$ et une fonction d'attribution A_F associée au réseau F et à une entrée x étant donnée la **baseline** x' , on a

$$A_F(x, x') = (a_1, \dots, a_n)$$

avec (a_1, \dots, a_n) les importances de chaque feature dans la décision du réseau

Utilité de cette méthode

- Mieux comprendre le fonctionnement du réseau pour l'améliorer

- Indiquer à un spécialiste des points dignes d'intérêt
- Extraire de l'explication un ensemble de connaissances qui pourraient ensuite être utilisé dans un système de règles plus simple.

Difficulté de construction

Une méthode d'**attribution axiomatique** est difficile à découvrir car on ne peut jamais savoir si l'erreur provient de l'attribution ou du réseau.

L'article se concentre donc sur une méthode basé sur des **axiomes** solides démontrés mathématiquement.

Méthode des Gradients Intégrés

- Ne nécessite que quelques appels à la fonction de calcul des gradients
- Ne nécessite aucune modification du réseau
- Satisfait les **axiomes** décrit plus bas, *au contraire des autres méthodes de l'Etat de l'Art*.

⚠ Nécessité d'une Baseline

On définit une **Baseline** comme une entrée telle que toute prédiction du réseau est *équitablement correcte/incorrecte*.

Par exemple dans le cadre de la classification d'image, il pourrait s'agir d'une image intégralement noire.

La **Baseline** permet d'identifier les éléments présents dans l'image réelle et absents dans la **Baseline** qui ont contribué à la prédiction.

Deux Axiomes Fondamentaux

Sensitivité

Sensitivité

- ✓ Une méthode d'attribution satisfait la **Sensitivité** ssi

Une différence d'une seule **feature** entre image & Baseline suffisant à modifier la prédiction du réseau **implique** une attribution non nulle pour cette feature.

On note que la méthode des **Gradients** ne respecte pas la **Sensitivité** :

≡ Prenons le modèle : $f(x) = 1 - \text{ReLU}(1 - x)$

Avec une **Baseline** valant 0 & un **Input** valant 2, la sortie du réseau n'est pas la même, mais l'attribution reste nulle car la fonction est constante entre 0 & 1.

Invariance de l'implémentation

Invariance de l'implémentation

- ✓ Une méthode d'attribution satisfait l'**Invariance de l'implémentation** ssi

Deux réseaux **fonctionnellement équivalents** (ie effectuant toujours des prédictions identiques pour toute entrée x) ont

toujours les mêmes attributions, *même lorsque leurs architectures ne sont pas les mêmes.*

Méthode des Gradients Intégrés

La méthode des gradients intégrés suit les étapes suivantes :

1. Générer n *images interpolées* entre la **Baseline** et l'**Input** :

$$x' + \frac{k}{n} \times (x - x')$$

2. Calculer les *Gradients* entre les prédictions de sortie du modèle F par rapport aux caractéristiques d'entrée :

$$\frac{\partial F(\text{Image Interpolée})}{\partial x_i}$$

3. Cumuler les *Gradients* ainsi calculés (On approxime l'intégrale au moyen d'une Somme de Riemann) :

$$\sum_{k=1}^n \text{Gradients} \times \frac{1}{n}$$

4. Mettre les *Gradients Intégrés* à l'échelle par rapport à l'image d'origine :

$$(x_i - x_{i'}) \times \text{Gradients Intégrés}$$

Toutes ces étapes se résument par l'équation suivante :

$$\text{Gradients Intégrés}_{i(x)} = (x_i - x_{i'}) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i}$$

Il s'avère que la méthode des **Gradients Intégrés** est aussi

Complète :

Complétude des méthodes d'attribution

✓ Une méthode d'attribution est dite Complètessi

La somme de ses attributions vaut la différence entre la **Baseline** et l'**Input** :

$$\sum_{i=1}^n \text{Méthode d'attribution}_{i(x)} = F(x) - F(x')$$

Ainsi, la Méthode des Gradients Intégrés respecte :

- La **Sensibilité** car *Complétude* \implies *Sensibilité*
- L'**Invariance de l'Implémentation** car elle est calculée uniquement en fonction des Gradients de la fonction décrite par le modèle

Unicité des Gradients Intégrés

Méthode de Parcours

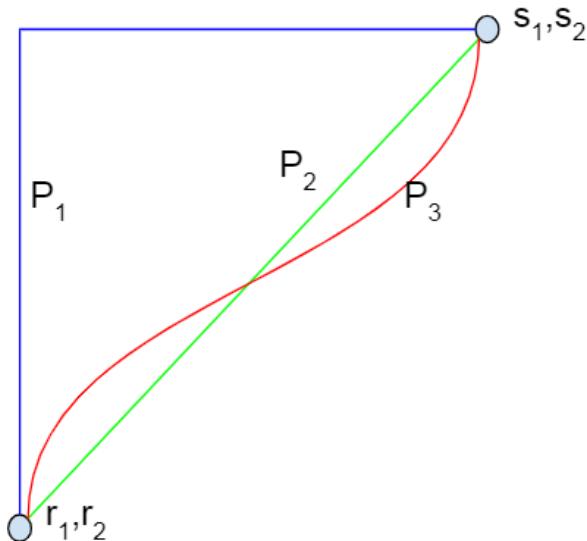


Figure 1. Three paths between a baseline (r_1, r_2) and an input (s_1, s_2) . Each path corresponds to a different attribution method. The path P_2 corresponds to the path used by integrated gradients.

Il existe différents trajets permettant de générer des images interpolées entre la **Baseline & l'Input**. La Méthode des Gradients Intégrés corresponds au Barycentre pondéré à coefficients positifs entre les deux points.

Toutes ces méthodes de parcours respectent les deux axiomes *désirables suivants* :

Sensitivité (b)

- ✓ Une méthode d'attribution respecte la **Sensitivité (b)**ssi

Toute feature dont le réseau ne dépend pas mathématiquement a une attribution **nulle**.

Linéarité

- ✓ Une méthode d'attribution respecte la **Linéarité**ssi

Toute composition linéaire de deux réseaux F_1 & F_2

$$a \times F_1 + b \times F_2$$

voit ses attributions également composées linéairement en fonction de a & b .

Préservation de la Symétrie

↳ Les Gradients Intégrés sont la seule Méthode de Parcours à préserver la symétrie

Une méthode d'attribution préserve la symétrie ssi pour deux features symétriques (ie telles que $F(x, y) = F(y, x)$), les attributions sont identiques.

Applications

Pour rendre le calcul possible pour un ordinateur, on approxime le calcul de l'intégrale au moyen d'une Somme de Riemann :

$$\text{Gradients Intégrés}_{i(x)} \simeq (x_i - x'_i) \times \sum_{k=1}^n \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \times \frac{1}{n}$$

Un réseau de reconnaissance d'objets

Original image



Top label and score

Top label: reflex camera
Score: 0.993755

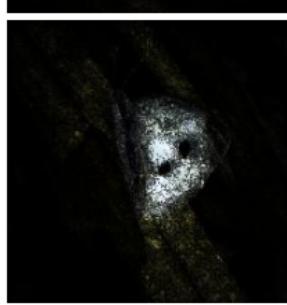
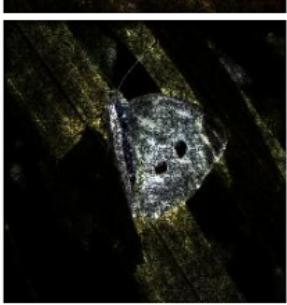
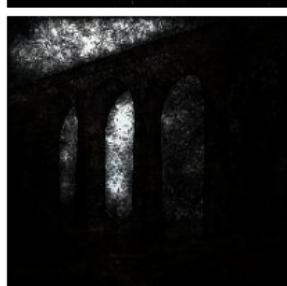
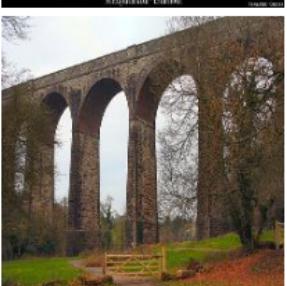
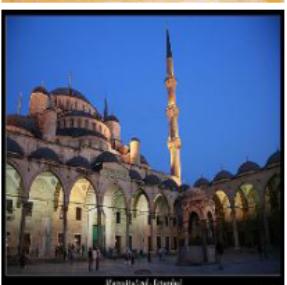
Top label: fireboat
Score: 0.999961

Integrated gradients

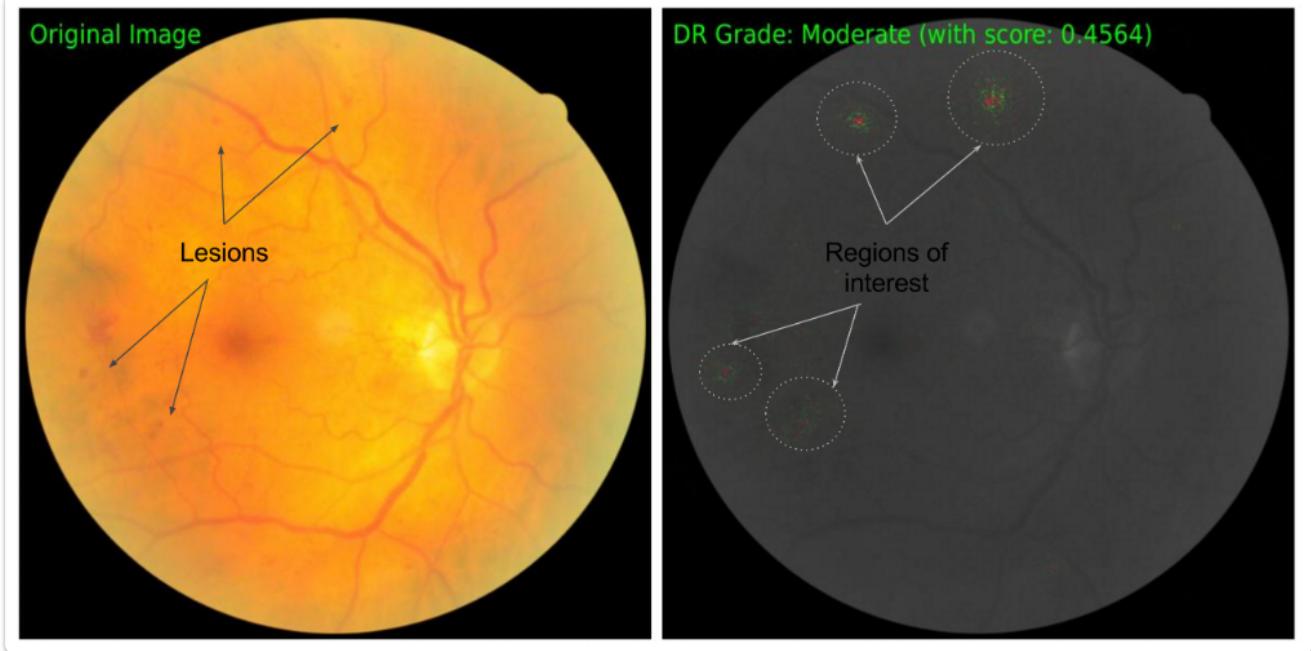


Gradients at image





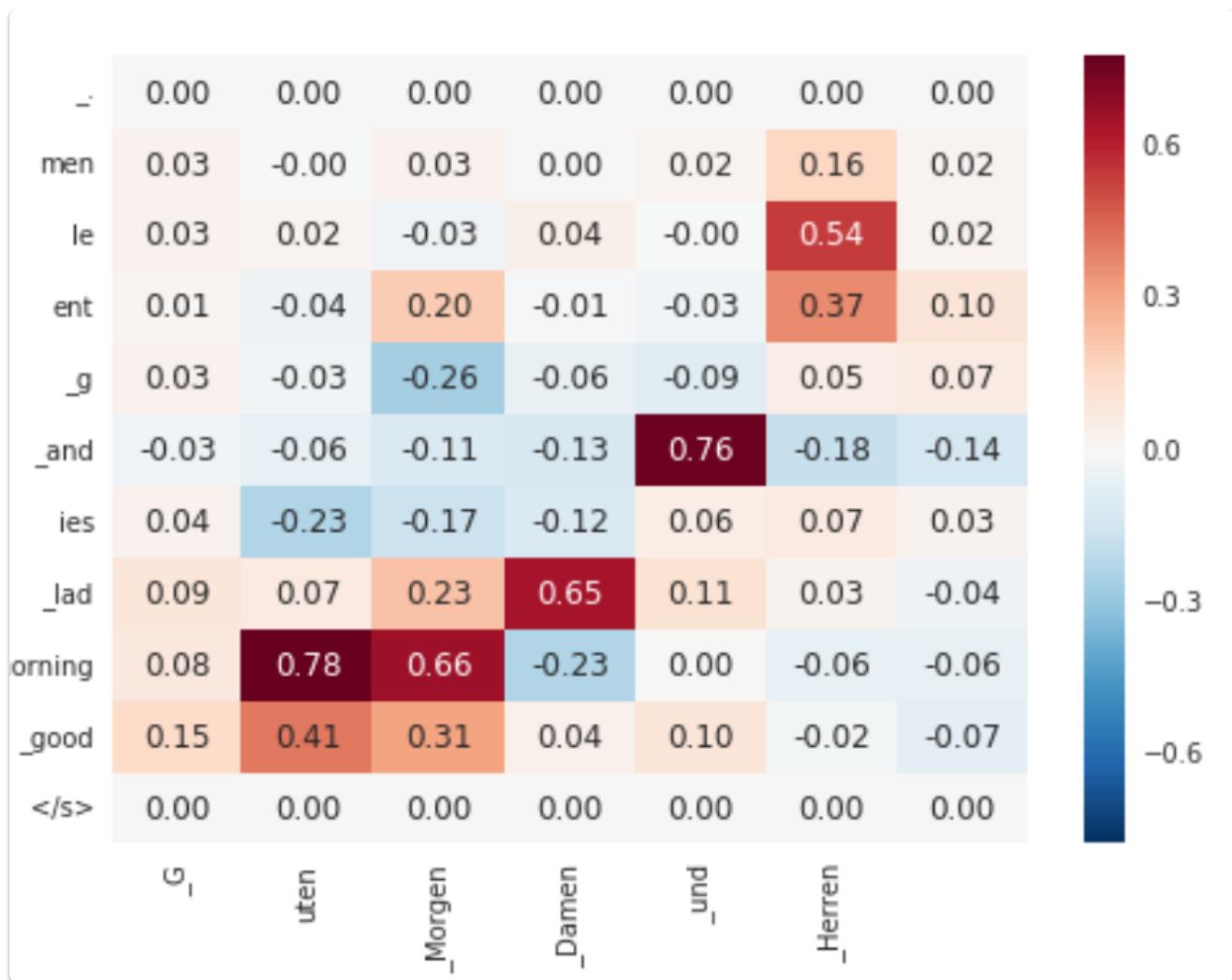
Prédiction d'une complication diabétique



Classification de questions

- how many townships have a population above 50 ? [prediction: NUMERIC]
- what is the difference in population between flora and masilo [prediction: NUMERIC]
- how many athletes are not ranked ? [prediction: NUMERIC]
- what is the total number of points scored ? [prediction: NUMERIC]
- which film was before the audacity of democracy ? [prediction: STRING]
- which year did she work on the most films ? [prediction: DATETIME]
- what year was the last school established ? [prediction: DATETIME]
- when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
- did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Traduction



Ligand-Based Virtual Screening

