Mesure d'évaluation de la qualité des explications pour les réseaux neuronaux profonds

ADIL Nawfel, MOHAMMED Limame, ESKINAZI Etienne, NAJI Ramzi
Télécom Paris

Abstract
De nombreuses méthodes ont été développées pour expliquer le fonctionnement des réseaux de neurones profonds, mais peu d'efforts ont été consacrés à vérifier si ces explications sont objectivement pertinentes. Parmi ces méthodes, SHAP (SHapley Additive exPlanations) se distingue en s'appuyant sur la théorie des jeux pour attribuer à chaque caractéristique une contribution précise à la prédiction du modèle. Malgré ses atouts théoriques (additivité, cohérence), les outils d'évaluation de la qualité des explications restent limités. Dans ce contexte , nous avons recodé deux métriques introduites dans l'article *How Good is your Explanation? Algorithmic Stability Measures to Assess the Quality of Explanations for Deep Neural Networks* : la généralisabilité moyenne (MeGe) et la cohérence relative (ReCo). Le but de notre approche est de confirmer que ces deux métriques évaluent de manière cohérente les explications produites par la méthode SHAP.
Vous retrouverez la solution implémentée ici : https://github.com/ramzinaji/Explanation$_AI$

1 Introduction

Dans le cadre de la validation des métriques MeGe (Mean Generalizability) et ReCo (Relative Consistency), telles que proposées dans l'article, nous avons procédé à la réentraînement d'un modèle ResNet sur différents jeux de données. Une fois ces nouveaux modèles appris, nous avons appliqué la méthode d'explicabilité SHAP (SHapley Additive exPlanations) afin de générer des cartes d'explication pour les prédictions. Ces représentations SHAP ont ensuite servi de base au calcul des scores MeGe et ReCo, permettant ainsi d'évaluer la stabilité et la cohérence des explications produites par les différents modèles.

2 Méthode SHAP

2.1 Approche SHAP

Les valeurs SHAP (SHapley Additive exPlanations) sont une méthode utilisée dans l'apprentissage automatique pour expliquer la sortie d'un modèle en attribuant la contribution de chaque caractéristique à la prédiction finale. Dérivées de la théorie des jeux coopératifs, les valeurs SHAP fournissent une approche structurée pour comprendre l'importance des caractéristiques individuelles dans un modèle prédictif.

À la base, les valeurs SHAP visent à attribuer une valeur numérique à chaque caractéristique, en montrant à quel point cette caractéristique particulière a influencé la prédiction du modèle pour un point de données spécifique. En tenant compte de toutes les combinaisons possibles de caractéristiques et de leurs contributions, les valeurs SHAP offrent une compréhension complète de l'importance des caractéristiques dans le modèle.

2.2 Formulation mathématique de SHAP

Soit un modèle de prédiction entraîné F et une entrée $\mathbf{X} = (x_1, x_2, \ldots, x_p)$, avec p caractéristiques. L'objectif de SHAP (SHapley Additive exPlanations) est d'approximer la sortie du modèle $F(\mathbf{X})$ comme une somme des contributions individuelles de chaque caractéristique :

$$F(\mathbf{X}) \approx \phi_0 + \sum_{i=1}^{p} \phi_i$$

où :

- ϕ_0 est la valeur de base (par exemple, la prédiction moyenne du modèle),
- ϕ_i est la valeur de Shapley associée à la caractéristique x_i, représentant sa contribution à la prédiction.

Les valeurs de Shapley sont définies comme suit pour chaque caractéristique i :

$$\phi_i = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{i\}} \frac{|S|!(p - |S| - 1)!}{p!} \left[F_{S \cup \{i\}}(\mathbf{X}) - F_S(\mathbf{X}) \right]$$

avec :

- S un sous-ensemble de caractéristiques ne contenant pas i,
- $F_S(\mathbf{X})$ la prédiction du modèle lorsque seules les variables de S sont connues (les autres étant marginalisées ou fixées à une valeur de fond),
- Le coefficient est le poids de Shapley, assurant une équité sur toutes les permutations possibles.

Les valeurs de Shapley possèdent plusieurs propriétés fondamentales :

1. **Additivité** :

$$F(\mathbf{X}) = \phi_0 + \sum_{i=1}^{p} \phi_i$$

2. **Symétrie** : Si deux caractéristiques contribuent de façon identique à la prédiction dans tous les cas, leurs valeurs SHAP sont égales.

3. **Nullité** : Si une caractéristique n'influence jamais la prédiction, sa valeur SHAP est nulle.

4. **Efficacité** : La somme des contributions est égale à la différence entre la prédiction du modèle et la valeur de base.

2.3 SHAP pour classification

Dans le cadre de l'explicabilité par SHAP, pour une instance $X \in R^p$ et une sortie du modèle $F_k(X)$ correspondant à la classe k, le vecteur $\phi_k = (\phi_{k,1}, \phi_{k,2}, \ldots, \phi_{k,p}) \in R^p$ représente la contribution de chaque caractéristique à la prédiction de la classe k.

Chaque valeur $\phi_{k,i}$ peut être interprétée comme suit :

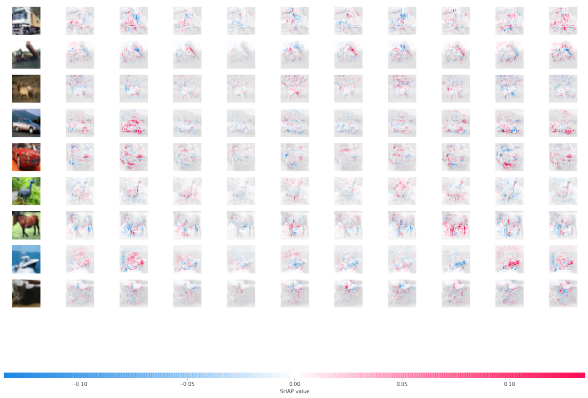
- $\phi_{k,i} > 0$: la $i^{ème}$ caractéristique augmente la prédiction pour la classe k,
- $\phi_{k,i} < 0$: la $i^{ème}$ caractéristique diminue la prédiction pour la classe k,
- $\phi_{k,i} \approx 0$: la $i^{ème}$ caractéristique a peu ou pas d'influence.

SHAP propose une décomposition additive :

$$F_k(X) = \phi_{k,0} + \sum_{i=1}^{p} \phi_{k,i}$$

où $\phi_{k,0}$ est la valeur de base (souvent la sortie moyenne du modèle), et chaque $\phi_{k,i}$ est la contribution marginale de la caractéristique i.

Dans les tâches de classification, on analyse généralement le vecteur $\phi_{\hat{k}}$, où $\hat{k} = \arg\max_k F_k(X)$, c'est-à-dire la classe prédite par le modèle.



Valeurs SHAP sur une instance CIFAR10

3 Implémentation de MeGe et ReCo

3.1 Construction du score MeGe

Le score **MeGe** (Mean Generalizability) permet d'évaluer la **généralisabilité** des explications d'un réseau de neurones profond. L'idée sous-jacente est que des explications fiables doivent être **stables**, c'est-à-dire qu'elles ne doivent pas varier de manière significative lorsque le modèle est entraîné sur des échantillons légèrement différents issus de la même distribution de données. Une explication est dite *généralisable* si elle se maintient d'un modèle à l'autre malgré des variations mineures dans les données d'entraînement.

Soit un modèle de prédiction entraîné $f : R^p \to R^C$ et une méthode d'explication additive, telle que SHAP, qui produit une explication $\phi_f(\mathbf{x}) \in R^p$ pour une entrée $\mathbf{x}$. Ce vecteur $\phi_f(\mathbf{x})$ attribue une importance à chaque caractéristique de $\mathbf{x}$ dans la décision du modèle f.

Pour construire le score MeGe, on applique une procédure de *k-fold cross-training* : on divise l'ensemble de données $\mathcal{D}$ en k sous-ensembles égaux $\{V_i\}_{i=0}^{k-1}$, puis on entraîne k modèles f_i, chacun sur $\mathcal{D} \setminus V_i$. Ainsi, chaque modèle est appris sur une portion différente de l'échantillon total. Cette méthode permet de simuler des perturbations naturelles dans l'apprentissage, afin d'analyser la stabilité des explications.

On compare ensuite les vecteurs d'explication $\phi_{f_i}(\mathbf{x})$ et $\phi_{f_j}(\mathbf{x})$ générés par deux modèles différents f_i et f_j, pour une même entrée $\mathbf{x} \in V_i$, donc non vue par f_i. On se limite au cas où *les deux modèles prédisent correctement* la sortie, c'est-à-dire $f_i(\mathbf{x}) = y$ et $f_j(\mathbf{x}) = y$, ce qui définit l'ensemble suivant :

$$\mathcal{S}^{=} = \left\{ \delta_{\mathbf{x}}^{(i,j)} : f_i(\mathbf{x}) = y \wedge f_j(\mathbf{x}) = y \right\}$$

où $\delta_{\mathbf{x}}^{(i,j)}$ désigne une mesure de distance (par exemple la norme L_2) entre les explications :

$$\delta_{\mathbf{x}}^{(i,j)} = \left\| \phi_{f_i}(\mathbf{x}) - \phi_{f_j}(\mathbf{x}) \right\|$$

On définit alors le score MeGe comme la moyenne des distances de l'ensemble $\mathcal{S}^{=}$:

$$MeGe = \frac{1}{|\mathcal{S}^{=}|} \sum_{\delta_{\mathbf{x}}^{(i,j)} \in \mathcal{S}^{=}} \delta_{\mathbf{x}}^{(i,j)}$$

Un score MeGe faible indique que les explications sont cohérentes d'un modèle à l'autre, ce qui témoigne d'une **stabilité algorithmique** et d'une **généralisabilité** des explications. À l'inverse, une forte variabilité entre les vecteurs $\phi_{f_i}(\mathbf{x})$ suggère une forte sensibilité aux données d'apprentissage, ce qui rend les explications moins fiables.

Construction du score ReCo

Le score **ReCo** (Relative Consistency) permet d'évaluer la **cohérence relative** des explications produites par des modèles entraînés sur des sous-ensembles de données différents. Ce score repose sur l'idée que deux explications issues de prédicteurs différents doivent être

proches (i.e., avoir une faible distance) lorsque les prédictions sont identiques, et éloignées lorsque les prédictions divergent. Cela permet d'évaluer si les explications sont bien corrélées à la sortie du modèle, ce qui est un critère de confiance important.

Soit $\phi_{f_i}(\mathbf{x})$ et $\phi_{f_j}(\mathbf{x})$ les vecteurs d'explication générés respectivement par les modèles f_i et f_j pour une même entrée $\mathbf{x}$. On définit une distance $\delta_{\mathbf{x}}^{(i,j)} = \left\|\phi_{f_i}(\mathbf{x}) - \phi_{f_j}(\mathbf{x})\right\|$, qui mesure l'écart entre les explications. On regroupe ces distances dans deux ensembles :

- $\mathcal{S}^{=}$: l'ensemble des distances où $f_i(\mathbf{x}) = f_j(\mathbf{x}) = y$, i.e., les deux modèles donnent la bonne prédiction,
- $\mathcal{S}^{\neq}$: l'ensemble des distances où $f_i(\mathbf{x}) = y$ mais $f_j(\mathbf{x}) \neq y$, i.e., un seul des modèles prédit correctement.

L'idée est que les distances de $\mathcal{S}^{=}$ devraient être plus faibles que celles de $\mathcal{S}^{\neq}$, car deux prédictions identiques devraient être expliquées de manière similaire. Pour quantifier cette séparation, on introduit un seuil $\gamma \in \mathcal{S} = \mathcal{S}^{=} \cup \mathcal{S}^{\neq}$, puis on calcule :

$$\text{TPR}(\gamma) = \frac{|\{\delta \in \mathcal{S}^{=} : \delta \leq \gamma\}|}{|\{\delta \in \mathcal{S} : \delta \leq \gamma\}|}$$
$$et \quad TNR(\gamma) = \frac{|\{\delta \in \mathcal{S}^{\neq} : \delta > \gamma\}|}{|\{\delta \in \mathcal{S} : \delta > \gamma\}|}$$

Le **TPR** (True Positive Rate) mesure la proportion de distances faibles provenant de prédictions cohérentes, tandis que le **TNR** (True Negative Rate) mesure la proportion de grandes distances provenant de prédictions incohérentes.

Enfin, on définit le score ReCo comme la meilleure séparation possible entre les deux distributions, en maximisant la précision équilibrée :

$$ReCo = \max_{\gamma \in \mathcal{S}} \left(TPR(\gamma) + TNR(\gamma) - 1\right)$$

Un score ReCo proche de 1 indique que les explications sont fortement cohérentes avec les prédictions, tandis qu'un score proche de 0 traduit une incohérence importante entre explications similaires et différentes.

4 Conclusions

Limites des métriques MeGe et ReCo. Bien que les métriques MeGe et ReCo permettent d'évaluer respectivement la généralisabilité et la cohérence des explications, elles présentent plusieurs limites lorsqu'elles sont appliquées à des architectures complexes comme ResNet sur des jeux de données visuels comme CIFAR-10. Par exemple, dans nos expériences avec SHAP appliqué à un ResNet entraîné sur CIFAR-10, nous avons observé que MeGe pouvait donner un score élevé même lorsque les explications étaient peu informatives, simplement parce que les modèles entraînés sur des folds différents produisaient des cartes SHAP similaires mais peu discriminantes. Cela montre que MeGe peut parfois surévaluer la qualité des explications si celles-ci sont uniformément pauvres. De son côté, ReCo dépend fortement du choix du seuil γ. Dans le cas de CIFAR-10, certaines classes très similaires visuellement (comme *chat* et *chien*) peuvent générer des distances d'explication ambiguës, rendant difficile la séparation nette entre les prédictions correctes et incorrectes. Par ailleurs, ni MeGe ni ReCo ne tiennent compte de la pertinence sémantique des cartes de chaleur générées : une carte SHAP centrée sur le fond de l'image peut être évaluée positivement par ReCo si elle est stable, même si elle ne met pas en évidence les objets pertinents. Cela limite l'utilité de ces scores dans des scénarios où l'interprétation humaine est cruciale.

References

[1] Dillon Bowen and Lyle Ungar. Generalized shap: Generating multiple types of explanations in machine learning. *arXiv preprint arXiv:2107.07883*, 2021. Operations, Information, and Decisions Department, The Wharton School of Business, University of Pennsylvania; Department of Computer and Information Science, University of Pennsylvania.

[2] Thomas Fel, David Vigouroux, Rémi Cadène, and Thomas Serre. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. *arXiv preprint arXiv:2305.XXXX*, 2023. Carney Institute for Brain Science, Brown University; Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France; IRT Saint-Exupéry.