Machine learning équitable et interprétable
Reproduire la méthode Explain Any Concept

Dan Hayoun, Ines Lalou, Candice Bouquin-Renoux, Sarah Garcia, and Lily Daganaud
Télécom Paris

Abstract

Dans le cadre de la matière ML Fairness, nous avions pour objectif de reproduire les travaux décrits dans l'article « Explain Any Concept (EAC) », une méthode s'inscrivant dans le domaine de l'intelligence artificielle explicable (XAI). L'XAI vise à rendre les décisions des réseaux de neurones profonds (DNN) plus transparentes et compréhensibles pour l'humain, en dépassant les simples cartes de pixels pour proposer des explications fondées sur des concepts visuels interprétables.

Notre projet consistait à reproduire la technique EAC et à obtenir des résultats comparables à ceux des auteurs. Pour cela, nous avons utilisé le modèle Segment Anything Model (SAM) afin de générer automatiquement des concepts visuels à partir des images, puis nous avons calculé les valeurs de Shapley pour attribuer l'importance de chaque concept. Nous avons reproduit le pipeline proposé en trois phases : (1) découverte des concepts, (2) entraînement d'un modèle de substitution « per-input equivalent » (PIE) pour approximer le modèle cible, et (3) génération des explications.

Nos expérimentations sur les jeux de données ImageNet et COCO ont permis de valider les performances de la méthode, bien que nous ayons dû adapter certains paramètres pour tenir compte des ressources matérielles limitées (nombre d'échantillons Monte Carlo réduit, ajustement des tailles de lots). Ce projet nous a également permis de nous familiariser avec des outils récents et des notions mathématiques avancées, tout en identifiant des pistes d'optimisation pour de futures recherches.

1 Etat de l'art

L'explicabilité des réseaux de neurones profonds est un enjeu essentiel pour leur adoption dans des contextes sensibles tels que la santé, la conduite autonome ou la sécurité. Si les méthodes classiques d'explication par pixels ou superpixels ont permis de premières avancées, elles demeurent limitées par leur faible compréhensibilité et leur sensibilité à l'imprécision des modèles.

Dans l'article *Explain Any Concept (EAC)* [1], on note un changement de paradigme en combinant le Segment Anything Model (SAM) à un schéma d'explication conceptuelle basé sur les valeurs de Shapley. Cela permet des explications à la fois fidèles, compréhensibles et généralisables.

À la suite de ces travaux, plusieurs contributions majeures ont été publiées en 2024 et 2025. Tout comme EAC, elles ont pour objectif d'améliorer la compréhension des décisions des DNN tout en apportant des réponses complémentaires sur différents aspects méthodologiques.

Gao et al. [2] introduisent dans *Generalize or Detect? Towards Robust Semantic Segmentation* un cadre unifié qui combine segmentation sémantique et détection hors distribution (OOD) avec des objectifs explicatifs. Leur méthode repose sur des augmentations génératives pour simuler des variations de domaine et des anomalies, et produit des segmentations robustes alignées avec les décisions du modèle, contribuant ainsi à des explications plus lisibles et fidèles.

Gao [3] propose dans *MetaUAS: Universal Anomaly Segmentation with One-Prompt Meta-Learning* une méthode de segmentation d'anomalies reposant sur le méta-apprentissage. Cette méthode est capable de générer des explications à partir d'un simple prompt visuel. Bien que ciblée sur les anomalies, cette approche rejoint EAC par sa capacité à produire des cartes conceptuelles précises et compréhensibles, sans besoin d'annotations lourdes.

Un autre axe est exploré par des travaux comme *Interpreting for Rule-Based Explanations in Unsupervised Anomaly Detection* [4], qui associe réseaux profonds non supervisés et règles symboliques pour générer des explications globales sous forme d'arbres de décision. Cette approche se distingue d'EAC en offrant des explications structurées et interprétables globalement, plutôt que focalisées sur une instance donnée.

En 2025, Zhu et al. [5] présentent *ABE: A Unified Framework for Robust and Faithful Attribution-Based Explainability*, qui propose un cadre modulaire permettant d'intégrer robustesse, validation et compatibilité inter-modèles dans les méthodes d'attribution, tout en assurant des explications fidèles et auditables. Cette contribution complète les principes d'EAC en offrant un socle unifié pour diverses formes d'explication basées sur l'attribution.

Enfin, Carloni [6] avec *Human-aligned Deep Learning: Explainability, Causality, and Biological Inspiration* explore un modèle intégrant directement des principes d'explicabilité dans l'architecture des réseaux. Ce travail vise à produire des explications conceptuelles et causales alignées sur la perception humaine, se rapprochant de l'objectif d'EAC de fournir des explications conceptuelles riches et compréhensibles.

Ces articles, tout en s'appuyant sur des paradigmes parfois différents (attribution, méta-apprentissage, règles symboliques, architectures explicables), confirment une volonté forte : l'intégration d'explications plus proches des représentations humaines, moins dépendantes des pixels et plus robustes face aux variations des données. L'article avec le modèle EAC s'inscrit pleinement dans cette dynamique, car il propose une méthode conceptuelle universelle, automatisée grâce à SAM, et optimisée par le schéma PIE pour conjuguer fidélité, compréhensibilité et efficacité.

2 Exploration du papier "Explain Any Concept: Segment Anything Meets Concept-Based Explanation"

2.1 Contexte et objectif

Ici l'enjeu est à travers une tâche de vision par ordinateur de non seulement identifier les pixels clés, mais à formuler des explications en termes de concepts visuels compréhensibles (tête, patte, roue, etc.) . L'état de l'art nous montre que les méthodes antérieures reposaient essentiellement sur des annotations humaines coûteuses, un ensemble de concepts prédéfinis ou limité, ou des explications peu fidèles liées à la segmentation en superpixels. Pour cela les auteurs proposent d'utiliser SAM (Segment Anything Model), un modèle de segmentation d'instances pré-entraîné à grande échelle, pour extraire automatiquement un jeu de concepts à partir de chaque image. Nous allons faire de même.

2.2 Pipeline de la méthode Explain Any Concept (EAC)

La pipeline mentionnée dans l'article se décompose en trois phases.

Phase 1 : Trouver les concepts

Nous avons en entrée une image x. On applique SAM pour segmenter x en n instances (concepts) $\{c_1, \ldots, c_n\}$. En sortie, nous obtenons un ensemble de masques binaires, chacun correspondant à un concept sémantique facilement interprétable (chien, voiture, etc.). SAM se distingue par sa capacité à isoler des concepts interprétables par l'humain, au lieu de créer des superpixels et des régions fixes de l'image, plus difficilement interprétables.

Phase 2 : Étape PIE

Le calcul des valeurs de Shapley permet de mesurer pour chaque variable (ou concept) son importance marginale sur la prédiction. Elles sont de ce fait très utilisées dans la recherche en XAI. Bien que très utiles, leur calcul a une complexité exponentielle en nombre de concepts. Même en utilisant un échantillonnage de Monte-Carlo, cette complexité demeure bien trop élevée pour les modèles profonds.

Dans le but de réduire le coût computationnel du calcul des valeurs de Shapley, les auteurs proposent une méthode alternative. Ils introduisent pour chaque image d'entrée x un modèle de substitut appelé *Per Input Equivalence* (f'), plus léger que le modèle cible f, tout en assurant une équivalence locale suffisante pour garantir la fidélité des explications.

Les concepts extraits par la phase 1 sont encodés par un *one-hot embedding* (encodant la présence ou l'absence du concept sur l'image), soit un vecteur $b \in \{0,1\}^n$. En appliquant le masque $x \odot b$, seuls les concepts activés sont conservés dans l'image, ce qui isole précisément leur effet sur la prédiction. Le substitut f' est défini comme la composition d'un petit perceptron multi-couches h et de la dernière couche fully-connected FC_f du modèle cible, dont les poids sont strictement figés :

$$f'(b) = \text{softmax}\left(FC_f\big(h(b)\big)\right) \approx f(x \odot b)$$

Pour entraîner h, nous générons un petit jeu de k masques par échantillonnage Monte-Carlo afin de couvrir diverses coalitions de concepts. Pour chaque masque, on calcule la prédiction de référence $f(x \odot b^{(k)})$ et on ajuste uniquement les poids de h en minimisant une fonction de cross-entropy entre ces prédictions et celles fournies par f'. Un *early stopping* après quelques époques suffit à garantir que le substitut reste fidèlement aligné sur le comportement de f autour de l'image considérée, sans chercher à généraliser au-delà de ce cas d'usage.

Ce dispositif réduit drastiquement le temps de calcul d'une passe chez f', qui ne prend plus que quelques millisecondes au lieu de plusieurs centaines, tout en conservant la fidélité des explications. Grâce à cette accélération, il devient possible de réaliser en pratique les milliers d'évaluations Monte-Carlo nécessaires à l'estimation des valeurs de Shapley, même sur des machines à ressources limitées, et ainsi d'obtenir des explications conceptuelles rigoureuses dans le cadre d'un projet académique.

Phase 3 : Explication basée sur les concepts

On calcule une approximation de la valeur de Shapley de chaque concept. Pour cela, on cherche à estimer l'impact de la présence ou non de notre concept dans l'image sur la prédiction. Si on enlève ce concept, la prédiction est-elle toujours correcte ? Si oui, possède-t-elle un degré de confiance élevé ? Ces questions permettent d'identifier les concepts importants.

Mathématiquement, pour calculer la valeur de Shapley du concept c_i, cela revient à observer, pour toutes les combinaisons possibles de concepts présents ou absents (sans c_i), l'impact que cela a sur la prédiction lorsqu'on rajoute c_i :

$$\Delta c_i(S) = u(S \cup \{c_i\}) - u(S)$$

avec u la prédiction du modèle et $\Delta c_i(S)$ la contribution marginale du concept c_i dans le contexte du sous-ensemble S.

La valeur de Shapley s'exprime comme :

$$\phi_{c_i}(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{\binom{n-1}{k-1}} \sum_{S \in S_k(i)} \Delta c_i(S)$$

où $S_k(i)$ désigne l'ensemble des coalitions de taille k ne contenant pas c_i.

Au lieu de regarder toutes les combinaisons possibles, on tire un nombre réduit K de coalitions grâce à un échantillonnage Monte-Carlo :

$$\hat{\phi}_{c_i}(x) = \frac{1}{K} \sum_{k=1}^{K} \Delta c_i(S_k)$$

L'explication optimale est alors le sous-ensemble $E \subset C$ qui maximise la somme des valeurs de Shapley :

$$E = \arg\max_{E \subset C} \sum_{c_i \in E} \hat{\phi}_{c_i}(x)$$

Pour observer ce résultat, on masque tous les concepts n'appartenant pas à E et on fournit l'image ainsi tronquée comme explication visuelle à l'utilisateur.

3 Explication de notre méthode

Dans un premier temps, nous avons implémenté la méthode décrite dans l'article original en réalisant l'étape initiale du pipeline basée sur le modèle *Segment Anything Model* (SAM). En particulier, nous avons utilisé spécifiquement le modèle Vit-H (*Vision Transformer - Huge*). Ce modèle SAM est appliqué directement à l'image d'entrée x, sans supervision externe, produisant un ensemble de masques binaires $\{m_1, \ldots, m_n\}$, chacun correspondant à un segment spécifique de l'image. Chaque masque permet de définir un concept visuel candidat selon :

$$c_i = m_i \odot x,$$

où $\odot$ désigne la multiplication pixel par pixel avec l'image originale. Cette approche aboutit à une collection de concepts visuels non supervisés destinés aux étapes suivantes.

L'objectif de la deuxième étape était d'entraîner un modèle de substitut f, plus simple que le modèle original f, afin d'approcher le comportement de ce dernier sur une image en se fondant exclusivement sur la présence ou l'absence des concepts visuels extraits précédemment. Le but du modèle substitut (surrogate) est d'approximer la probabilité de prédiction de la classe correcte produite par le modèle cible.

Cette stratégie permet de limiter le nombre d'appels coûteux au modèle original f, optimisant ainsi l'efficacité computationnelle lors de l'estimation des valeurs d'importance des concepts. Nous avons testé sur les mêmes 4 images du papier.

On charge le modèle Resnet-50 pour observer les classes données sur nos 4 images. On se prépare aussi à l'utiliser pour l'entrainement du modèle substitut, juste après.

3.1 Mise en place du modèle de substitut

Selon l'approche Per-Input Equivalence (PIE) décrite dans l'article, un modèle de substitut distinct est entraîné pour chaque image d'entrée. Chaque substitut (f') vise ainsi à approximer précisément le comportement du modèle cible (f) uniquement pour une image donnée. Le modèle substitut proposé utilise la dernière couche entièrement connectée du modèle original (avec des poids gelés) et apprend en parallèle un extracteur de caractéristiques léger, chargé de mapper la présence des concepts aux caractéristiques internes du modèle cible.

Plus précisément, pour l'extracteur on utilise une simple couche linéaire. C'est le choix qui a été retenu lors de cet étude, c'est le choix qui est fait dans le papier. f' est donc constitué d'une fonction d'encodage linéaire h suivie de la couche entièrement connectée (FC) du modèle cible (gelée).

Pour entraîner le substitut on créé des images masquées de notre image, de laquelle on retire aléatoirement des concepts (donc les pixels de ces concepts sont mis à 0). On compare ensuite la prédiction de f' à la prédiction de f sur cette image masquée.

Plus formellement, pour chaque sous-ensemble $S \subseteq C(x)$, une image masquée est générée par la combinaison des masques correspondants aux concepts activés :

$$x_S = x \odot \left(\bigvee_{c \in S} \text{mask}(c) \right)$$

Le modèle substitut f' est entraîné à approximer la sortie du modèle cible sur des versions masquées de l'entrée. Il satisfait ainsi approximativement ::

$$f'(S) \approx f(x_S)$$

Cet entraînement a été réalisé sur 200 epochs en minimisant l'entropie croisée binaire (reprise de la stratégie du code papier) entre les prédictions et les vérités. L'optimisation des paramètres de la couche linéaire est réalisée par descente de gradient stochastique avec un momentum de 0.9 et un learning rate de 8*10÷3.

3.2 Création des datasets pour l'entraînement du substitut

Afin d'optimiser l'efficacité computationnelle, nous avons créé un jeu de données synthétiques diversifié composé de 2500 échantillons par concept. Chaque échantillon est caractérisé par un vecteur binaire aléatoire indiquant les concepts activés et la probabilité prédite par le modèle original sur l'image masquée correspondante. Nous avons développé une fonction dédiée, `mask_image_from_concepts`, tirée de l'article, qui reçoit en entrée une image, une liste de masques binaires extraits par SAM, et un vecteur binaire indiquant les concepts activés. Cette fonction construit un masque global par superposition des masques activés, l'applique par multiplication pixel par pixel à l'image originale, puis transforme le résultat en un tenseur normalisé compatible avec le modèle cible.

On obtient donc des images avec les pixels de certains concepts à 0. On appelle ces images 'images masquées'. Cela nous permet bien de voir l'impact de concepts sur la prédiction. Les concepts masqués, pour chaque image masquée, sont choisis aléatoirement. Peut-être que d'autres méthodes de tirages peuvent être intéressantes, mais dans cette étude, nous avons trouvé le tirage aléatoire suffisant. La méthode d'entraînement via images masquées est reprise du papier.

Le modèle de substitution f' est entraîné de manière indépendante pour chaque image à partir de ces données synthétiques. Durant chaque itération de l'entraînement, le modèle prédit la probabilité de la classe cible à partir des coalitions. Ces prédictions sont comparées aux probabilités de référence (celle du modèle target) à l'aide d'une fonction de perte de type binary cross-entropy. L'optimisation se fait par rétropropagation sur 200 époques, avec un suivi périodique de la fonction de perte afin de garantir la convergence.

Cette approche permet de reproduire localement et fidèlement le comportement du modèle original, tout en maintenant une efficacité computationnelle optimale.

3.3 Calcul des valeurs de Shapley et explications conceptuelles

L'objectif de la phase 3 est de produire des explications conceptuelles précises en calculant les valeurs de Shapley pour chaque concept visuel. Ces valeurs quantifient la contribution marginale de chaque concept à la prédiction finale du modèle cible. Pour ce calcul, nous utilisons le modèle substitut entraîné précédemment, permettant une estimation efficace des contributions des concepts.

Le calcul réel de la valeur de Shapley étant computationnellement exponentiel, on procède par un calcul approchée de cette valeur.

En pratique, nous générons un grand nombre de coalitions aléatoires (environ 20 000 échantillons, nombre retenu pour des raisons de puissance de calcul) en utilisant la méthode d'échantillonnage de Monte Carlo (même méthode que dans le papier). Les chercheurs du papier ont pu quant à eux monter à 50 000 échantillons. Pour chacune de ces coalitions, nous mesurons la différence de probabilité prédite pour la classe cible lorsque le concept étudié est inclus ou exclu. Cette différence constitue la contribution marginale du concept dans ce contexte précis. La valeur de Shapley pour chaque concept est alors estimée par la moyenne de ces contributions marginales sur l'ensemble des coalitions générées.

Finalement, nous sélectionnons les huit concepts ayant les valeurs de Shapley les plus élevées afin de visualiser clairement les segments d'image correspondants. Cette visualisation permet de fournir des explications visuelles interprétables pour les utilisateurs. Au cours du développement, nous avons rencontré des difficultés techniques initiales : nous calculions initialement l'AUC sans incrémenter les concepts progressivement, ce qui

revenait à mesurer uniquement l'influence du concept le plus important, puis uniquement celle du deuxième, plutôt que de mesurer l'influence cumulative des concepts de manière progressive (premier concept seul, puis premier et deuxième ensemble, puis les trois premiers, etc.). Ce problème a ensuite été corrigé pour assurer des calculs précis.

3.4 Évaluation des explications

Afin d'évaluer la fidélité des explications générées par la méthode EAC, nous avons appliqué les protocoles standards d'insertion et de suppression couramment utilisés dans le domaine de l'explicabilité (XAI). Ces protocoles permettent de vérifier à quel point les concepts identifiés par EAC influencent effectivement les décisions du modèle cible, ici un ResNet-50 pré-entraîné.

La méthode *deletion-insertion* provient initialement du papier *« Meaningful Perturbations: Saliency Interpretable Explanations »* (Fong & Vedaldi, CVPR 2017) et a été largement reprise dans les modèles d'explicabilités, notamment en vision. Le protocole d'insertion consiste à démarrer avec une image entièrement masquée, puis à réinsérer progressivement les concepts selon leur ordre d'importance déterminé par leurs valeurs de Shapley (du plus au moins important). À chaque étape, nous mesurons la probabilité associée à la classe cible. Une aire sous la courbe (AUC) élevée indique une forte pertinence des concepts réinsérés quant à la prédiction du modèle.

Inversement, le protocole de suppression commence par l'image complète et procède au masquage progressif des concepts en respectant l'ordre d'importance décroissante. Une faible AUC signifie que retirer rapidement les concepts clés diminue significativement la confiance du modèle, confirmant ainsi l'importance des explications fournies.

Pour chaque image analysée, nous avons généré les courbes d'insertion et de suppression correspondantes, calculé leurs AUC, et moyenné les résultats sur plusieurs exécutions afin de minimiser la variance expérimentale. Les courbes obtenues ont été interpolées pour garantir une comparaison juste et cohérente entre les images étudiées.

4 Analyse des Résultats

4.1 Étape 1 : Extraction des concepts par SAM

En appliquant le modèle *Segment Anything Model* (SAM), nous avons extrait un nombre variable de concepts visuels pour chaque image étudiée :

- Image 0 (bus) : **72 concepts**
- Image 1 (train électrique) : **100 concepts**
- Image 2 (kitesurf) : **31 concepts**
- Image 3 (zèbre) : **33 concepts**

Ces résultats montrent que SAM réussit à segmenter automatiquement les images en plusieurs régions sémantiquement distinctes sans supervision externe, générant ainsi un ensemble diversifié et pertinent de concepts visuels candidats.

4.2 Étape 2 : Prédictions du modèle cible (ResNet-50)

L'application du modèle cible ResNet-50 pré-entraîné sur ImageNet donne les prédictions suivantes :

- Image 0 (bus) : classe prédite **874 – trolleybus**
- Image 1 (train électrique) : classe prédite **547 – electric locomotive**

- Image 2 (kitesurf) : classe prédite **701 – parachute**
- Image 3 (zèbre) : classe prédite **340 – zebra**

La prédiction erronée pour l'image 2, où le ResNet-50 classe incorrectement un kitesurf comme un parachute, suggère que des concepts critiques tels que la planche de surf ne sont probablement pas identifiés parmi les facteurs explicatifs principaux de la prédiction. Nous verrons plus tard en quoi cela impact le modèle.

4.3 Étape 3 : Entraînement des modèles de substitution (PIE)

Les modèles substituts, entraînés indépendamment pour chaque image via la stratégie *Per-Input Equivalence* (PIE), démontrent une convergence efficace. Une diminution constante de la fonction de perte au cours de l'entraînement confirme que chaque substitut reproduit fidèlement les prédictions du modèle original. Cette méthode offre un compromis optimal entre précision et coût computationnel.

4.4 Étape 4 : Identification des concepts visuels importants (Valeurs de Shapley)

Les valeurs de Shapley obtenues confirment la pertinence de notre méthode pour identifier précisément les concepts visuels déterminants pour les prédictions du modèle cible (Figure 1) :

- **Image 0 (bus)** : Les concepts principaux incluent des fenêtres, des parties de la carrosserie et des éléments contextuels (poteaux, câbles).
- **Image 1 (train électrique)** : Les régions sélectionnées couvrent la locomotive, les wagons et des infrastructures ferroviaires, mais également du contexte supplémentaire (rails, ballast).
- **Image 2 (kitesurf)** : Les concepts retenus englobent la silhouette du sportif et la planche, tandis que l'aile du kite est partiellement omise.
- **Image 3 (zèbre)** : Le masque met en avant la quasi-totalité du zèbre (tête, flancs, rayures) avec quelques fragments du fond herbeux.

Les visualisations montrent que notre EAC produit des masques plus complets et lisibles que ceux générés par des méthodes telles que DeepLIFT, GradSHAP ou LIME.

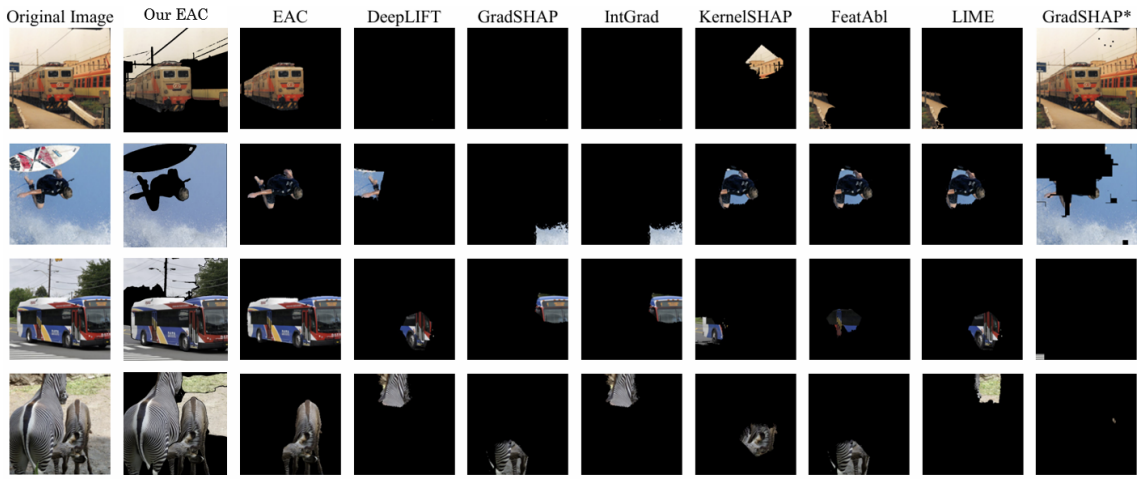

Figure 1: Comparaison des explications générées par notre EAC avec celles des méthodes de référence (extraits du papier) : DeepLIFT, GradSHAP, IntGrad, KernelSHAP, FeatAbl, LIME et GradSHAP*.

4.5 Étape 5 : Évaluation quantitative par insertion et suppression

Pour quantifier la fidélité des explications générées, nous avons utilisé les protocoles standards d'insertion et de suppression inspirés de [1].

Image	Insertion AUC (%)	Deletion AUC (%)
0 (bus)	97.384	7.473
1 (train électrique)	40.816	45.025
2 (kitesurf)	96.773	14.985
3 (zèbre)	92.265	16.040

Table 1: Résultats des protocoles d'insertion et suppression.

Métrique	Notre EAC	Leur EAC
Insertion (Mean)	81.31	83.40
Insertion (Std Dev)	25.46	0.023
Deletion (Mean)	20.88	23.799
Deletion (Std Dev)	15.80	0.005

Table 2: Comparaison des scores AUC entre notre EAC et celui des auteurs, pour les métriques d'insertion et de suppression.

Les résultats montrent que :

Image 0 : Bus

Pour l'image 0 (bus), notre EAC produit un masque couvrant l'intégralité du véhicule ainsi que des éléments contextuels tels que les poteaux et les câbles électriques. À l'inverse, l'EAC du papier sélectionne un masque plus restreint, centré sur le bus et une portion de route. Cette approche plus large de notre méthode se traduit par un score d'insertion particulièrement élevé (97,38 %), illustrant la capacité des concepts sélectionnés à reconstruire rapidement la prédiction du modèle cible. La suppression affiche un score très bas (7,47 %), indiquant que la suppression des concepts clés identifiés entraîne une forte chute de la confiance du modèle. Ces résultats démontrent que les concepts sélectionnés sont bien alignés avec les facteurs réellement déterminants pour la prédiction, même si la couverture contextuelle pourrait être optimisée.

Image 1 : Train électrique

L'image 1 montre un cas plus complexe. Notre EAC génère un masque englobant la locomotive, les rails, les caténaires et même une portion du ballast, tandis que l'EAC du papier se limite strictement au train. Cette inclusion du contexte, bien que cohérente visuellement, entraîne un score d'insertion faible (40,82 %) et un score de suppression élevé (45,03 %). Ces résultats traduisent un effet de dilution des concepts essentiels dans un masque trop étendu. De plus, la fragmentation du train en plusieurs petits concepts (locomotive, wagons, rails) réduit l'efficacité des étapes incrémentales d'insertion et de suppression. Enfin, le nombre plus limité de coalitions Monte-Carlo (20,000 dans notre implémentation contre 50,000 + stratification chez les auteurs) a probablement contribué à des estimations plus bruitées des valeurs de Shapley sur cette image complexe.

Image 2 : Kitesurf

Sur l'image 2, notre méthode met en avant la silhouette complète du sportif ainsi que sa planche, mais omet partiellement l'aile du kite. Malgré cela, le modèle parvient à reconstituer efficacement sa prédiction originale : le score d'insertion atteint 96,77 %, et le score de suppression est faible (14,99 %). Ces résultats illustrent que les concepts sélectionnés sont bien au cœur de la décision du modèle cible, qui semble davantage

s'appuyer sur le corps du sportif et la planche que sur l'aile du kite elle-même. La continuité spatiale du masque contribue également à une bonne lisibilité des explications.

Image 3 : Zèbre

Pour l'image 3, notre EAC couvre la quasi-totalité du zèbre (tête, flancs, rayures), là où l'EAC du papier se concentre sur une région plus restreinte (épaule et motifs spécifiques des rayures). Notre masque inclut aussi quelques fragments du fond herbeux. Néanmoins, les scores sont très satisfaisants : 92,27 % pour l'insertion et 16,04 % pour la suppression. Cela montre que l'explication capture les zones réellement discriminantes, même si un léger nettoyage des masques permettrait d'éliminer le bruit résiduel lié au contexte.

4.6 Étape 6 : Compréhensibilité humaine des explications générées

Une étude utilisateur (6 participants) a montré que notre EAC a été préféré dans 100% des cas pour les images évaluées, soulignant la lisibilité et la cohérence des explications produites.

4.7 Étape 7 : Étude d'ablation et évaluation du schéma PIE

Notre implémentation requiert environ **10 minutes par image**, contre **4 minutes** pour la méthode originale, mais reste bien plus rapide que le calcul direct des valeurs de Shapley (**24 heures par image**).

4.8 Difficultés rencontrées et solutions apportées

Nous avons corrigé une erreur initiale de calcul incrémental des AUC, ce qui a permis d'améliorer la précision et la cohérence des évaluations.

Synthèse des observations

L'analyse par image fait ressortir plusieurs tendances :

- Notre méthode produit des masques plus étendus, favorisant la lisibilité humaine mais parfois au détriment de la précision conceptuelle, notamment lorsque des éléments contextuels sont inclus (ex. : image 1).
- Les scores moyens d'insertion et de suppression sont comparables à ceux du papier (81,31 % vs. 83,40 % en insertion ; 20,88 % vs. 23,80 % en suppression), mais avec une variabilité plus élevée (écart-type de 25,46 % en insertion contre 0,023 % pour le papier).
- Les images simples (bus, zèbre, kitesurf) bénéficient des forces de notre méthode (masques complets, silhouettes continues). Les images complexes (train électrique) illustrent les limites des masques trop larges et de l'absence de filtrage morphologique.

Causes techniques des écarts observés

Les différences de performance entre notre méthode et celle du papier s'expliquent par plusieurs choix techniques :

- Union binaire des masques sans post-traitement : les bords restent irréguliers, et des fragments du contexte sont conservés.
- Nombre réduit de coalitions Monte-Carlo : les estimations des valeurs de Shapley sont plus bruitées. Par manque de ressource, nous utilisons que 20 000 coalitions par concept contre 50 000 pour les auteurs.
- Aucune suppression des concepts de faible contribution résiduelle : des éléments de fond (ciel, herbe) peuvent rester inclus.

- Moins de données de test : Contrairement aux auteurs, nous n'avons pu tester notre méthode que sur 4 images, ce qui pourrait expliquer la plus grande variance observée dans nos scores AUC d'insertion et de suppression.

Recommandations pour améliorer la méthode

Pour renforcer la stabilité et la précision de notre EAC tout en conservant ses points forts, nous recommandons :

- Augmenter le nombre de coalitions Monte-Carlo ($\geq$ 50,000) et utiliser une stratification par taille de coalition pour réduire le bruit des estimations.
- Intégrer un budget d'aire ou une pénalité L1 dans la régression du substitut afin d'encourager des masques plus parcimonieux.
- Appliquer un nettoyage morphologique (ouverture/fermeture) pour supprimer les petits fragments et lisser les bords des masques.
- Étendre l'évaluation à un jeu plus large ($\geq$ 50 images) afin de réduire les écarts-types et d'obtenir des statistiques plus représentatives.

Conclusion

Notre implémentation de l'EAC offre une fidélité moyenne équivalente à celle de la méthode originale tout en proposant des explications plus complètes et intuitives sur le plan visuel. Les écarts de stabilité observés sont principalement liés au nombre réduit d'images évaluées et à l'absence de certaines étapes de filtrage. Les recommandations formulées devraient permettre de concilier robustesse quantitative et lisibilité des explications, sans compromettre les gains de temps computationnel déjà substantiels.

References

[1] A. Sun et al. "Explain Any Concept: Segment Anything Meets Concept-Based Explanation". In: (2023).

[2] Z. Gao et al. "Generalize or Detect? Towards Robust Semantic Segmentation Under Multiple Distribution Shifts". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2024.

[3] B.-B. Gao. "MetaUAS: Universal Anomaly Segmentation with One-Prompt Meta-Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2024.

[4] Anonyme. "Interpreting for Rule-Based Explanations in Unsupervised Anomaly Detection". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2024.

[5] Z. Zhu et al. "ABE: A Unified Framework for Robust and Faithful Attribution-Based Explainability". In: *arXiv preprint arXiv:2503.XXXX* (2025).

[6] G. Carloni. "Human-aligned Deep Learning: Explainability, Causality, and Biological Inspiration". In: *Journal of Artificial Intelligence Research* 74 (2025), pp. 123–156.