Les défis du glissement de contexte géographique

T. Bayet[1,2,3], C. Denis[1,2], A. Bah[3,4], J-D. Zucker[2,5]

[1] Sorbonne Université, LIP6, 75005 Paris
[2] IRD, Sorbonne Université, UMMISCO, F-93143, Bondy, France
[3] UCAD, IRD, UMMISCO, Dakar, Sénégal
[4] Ecole Supérieure Polytechnique, UCAD, 15915 Dakar Fann, Sénégal
[5] Sorbonne Université, INSERM, NUTRIOMICS, F-75013, Paris, France

theophile.bayet@ird.fr

Résumé

Les modèles de vision par ordinateur sont majoritairement entraînés avec de larges bases de données publiques, dont l'exploration des biais permet l'identification de contextes dans lesquels ces modèles sont inadaptés. Le glissement d'un contexte adapté à un contexte inadapté est le témoin du manque de résilience d'un modèle à une variation dans les données. Nous tentons de caractériser le glissement de contexte géographique des données occidentales vers les données mondiales, et identifions les défis associés à cette caractérisation.

Mots-clés

Glissement de contexte, Chute de Performance, Vision par Ordinateur, Généralisation de Domaine

Abstract

Computer vision models are mostly trained with large public databases, of which exploring the biases allows the identification of contexts in which these models are unsuitable. The shift from a suitable context to an unsuitable context is a sign of the model's lack of resilience to a variation in the data. We attempt to characterize the shift in geographic context from Western to global data, and identify the challenges associated with this characterization.

Keywords

Performance Drop, Context Shift, Domain Generalization, Computer Vision

1 Introduction

Les progrès réalisés en intelligence artificielle (IA) ont provoqué un engouement pour son utilisation dans le cadre de la science soutenable [33, 47, 34, 40, 31]. Dans ce cadre, de nombreux travaux cherchent à lier IA et soutenabilité, notamment face à la crise climatique. On pourra distinguer deux approches majeures dans cette tendance : la soutenabilité de l'IA, qui promeut des pratiques plus éthiques et énergétiquement sobres pour les modèles développés [44, 36, 29], et l'IA pour la soutenabilité, qui met en avant des recherches alignant IA et réponses aux crises écologiques et humanitaires [50, 20, 24]. Cette seconde approche s'intéresse particulièrement au déploiement de modèles dans les pays en développement, puisque ces derniers sont plus touchés par les crises globales en cours [4]. Le domaine de la vision par ordinateur intègre ces nouvelles considérations, et pointe du doigt le manque de géodiversité des jeux de données communs [14, 26, 53] utilisés par un grand nombre de modèles comme un frein au déploiement de modèles soutenables [15, 43]. Ce déploiement est pourtant considéré nécessaire à la lutte contre les crises en cours et à venir [40], et donc à la résilience des systèmes humains.

Le manque de géodiversité dans les bases de données classiques a aussi des conséquences sur les performances des modèles, à travers le phénomène du glissement de contexte. Ce dernier survient quand un modèle est utilisé sur des données qui possèdent un contexte différent des données sur lesquelles le modèle a été entraîné. C'est un témoin des phénomènes de sur-apprentissage inhérents aux systèmes actuels d'apprentissage profond [17], ainsi que des biais inhérents à la conception des bases de données d'images [37]. Ce phénomène correspond aux baisses de performance souvent observées dans le déploiement sur le terrain de modèles entraînés en laboratoire [25]. La résilience d'un modèle peut pourtant se comprendre comme sa capacité à maintenir ses performances quelque soit le contexte de son déploiement, et donc de sa capacité à généraliser hors de son contexte d'entraînement.

Dans cet article, nous présentons une initiative pour documenter les perturbations liées au contexte géographique dans les modèles de vision par ordinateur, et nous montrons comment certains choix peuvent influencer ces perturbations. Nous faisons en section 2 un résumé de la littérature pour éclairer notre approche. Les expériences réalisés et leurs résultats sont décrits en section 3 et analysés en section 4. Enfin, nous concluons sur nos travaux en section 5.

2 Travaux connexes

Les jeux de données et les benchmarks sont les pierres angulaires de la communauté de l'apprentissage machine ; ils permettent la comparaison et le classement des modèles, in-

citant ainsi au développement de nouvelles architectures et techniques sur de vastes gammes de tâches. En vision par ordinateur, les collections de données sont devenues de plus en plus importantes et complexes [27, 21, 18, 14, 26, 22, 30] et couvrent de nombreux domaines. La corrélation positive entre volume de données et performance d'un modèle [45] souligne l'importance de ces avancées.

Pourtant, malgré des efforts soutenus [51, 45, 9, 52], ces bases de données sont soumises à de nombreux biais, desquels ne se déparent pas les modèles et applications utilisant ces dernières. Ces applications sont ensuite dénoncées comme nocives [13, 12, 10]. Dérivant de ces biais, le glissement de contexte cause aussi beaucoup de tort à ces modèles.

Le glissement de contexte est un phénomène qui survient quand le contexte des données d'entraînement et de test ou d'application d'un modèle diffèrent. Ces glissements arrivent de différentes manières [32, 37], et en particulier lors de déploiement de modèles sur le terrain, en dehors des conditions contrôlées de laboratoire [25, 3, 49, 5]. Les modèles ont du mal à généraliser sur des données hors-distribution (HD) [45] provenant de domaines inconnus, induisant une baisse de performance par rapport aux données en-distribution (ED) sur lesquelles le modèle a été entraîné. La généralisation sur de nouveaux domaines (GND) est un pan de recherche en vision par ordinateur [25] qui s'intéresse aux capacités des modèles à généraliser sur des données HD. A travers de nombreuses définitions, cette notion fait référence à la fois au problème qui survient quand un modèle est confronté à des données HD, et à la capacité de ce modèle à généraliser dans de telles situations [54, 25].

L'intérêt croissant pour la GND et les glissements de contexte ont permis l'émergence de nombreux benchmarks, dont le but est avant tout de promouvoir ces phénomènes et d'aider à les prévenir [6, 8, 48, 22]. Santurkar et al. [42] manipulent Imagenet pour introduire un glissement sur la répartition des sous-populations du jeu de données. Dans leur travail, le glissement de contexte est détecté grâce à la baisse de performance entre les données ED et HD. Koh et al. [25] introduisent 10 nouveaux jeux de données pour simuler le déploiement de modèles sur le terrain et stimuler la recherche scientifique dans ce domaine. Leur approche générale pour caractériser le glissement à l'oeuvre est de mesurer la performance entre données ED et HD. Ils font aussi état des soucis liés à cette approche, et décrivent cinq constructions différentes pour cette caractérisation. Enfin, Zhou et al. [54] soulignent l'existence de plus de 30 jeux de données qui sont communément utilisés en GND.

La connaissance du type de biais, de son origine (contrôlée, synthétique, artificielle,...), et de la tâche à réaliser sont nécessaires pour déterminer une stratégie pour le combattre. Cela conduit à des travaux centrés sur des types de biais spécifiques, comme l'utilisation d'une mesure d'équité pour surmonter des biais raciaux et de genre [5]. L'équité a de nombreuses métriques associées, conçues avec des objectifs qui peuvent différer de l'une à l'autre. "Equalized odds" [23] vise à faire des prédictions de manière à ce que les taux de vrai positifs et faux positifs soient distribués de manière égalitaire entre les groupes. "Demographic Parity" [16] a été conçue pour assurer des prédictions similaires pour des individus partageant des traits communs. Pour "Disparate Impact" [19], les auteurs utilisent l'hypothèse qu'entre deux groupes, les taux de classification positive devraient être à 80% similaires. Ces métriques utilisent des groupes définis par des attributs protégés que sont la race, la couleur de peau, l'ethnie, la religion, le genre, le handicap, ou le statut familial. Elles font néanmoins l'objet de critiques. Andrus et al. [5] soulèvent la problématique de la collecte de ces attributs protégés pour pouvoir créer des benchmarks utilisant ces derniers. Sambasivan et al. [41] soulignent que ces mesures ne sont pas exemptes de biais occidentaux, et Agarwal [1, 2] montre comment certaines de ces métriques s'opposent, et la nécessité de faire un compromis entre ces dernières.

Des efforts similaires sont réalisés pour améliorer la diversité géographique des bases de données [6, 39, 8]. Devries et al. [15] observent comment les systèmes de reconnaissance automatique publics se comportent sur la base de données Dollar Street Dataset [39]. Pour y caractériser le glissement de contexte géographique, ils utilisent deux métriques qui sont la précision par revenu et la précision par pays sur une carte mondiale. Shankar et al. [43] utilisent quant à eux des graphes de densité de log-vraisemblance sur des classes sélectionnées communes à Imagenet, Open Images, et une base de donnée crowd-sourcée. Ils montrent que les modèles usuels de vision par ordinateur sont plus performant dans un contexte occidental, compris ici comme européen et nord-américain, que dans le reste du monde. Nous interrogeons dans la suite de l'article la pertinence et la robustesse des métriques utilisées par ces travaux.

La littérature souligne que le manque de données dans certaines zones géographiques constitue un frein majeur au développement de modèles résilients pour la soutenabilité, et au déploiement de ces modèles. Dans un effort d'investigation de ce désagrément, nous proposons dans la section suivante une expérience pour évaluer ce dernier en fonction des contextes géographiques.

3 Expérience

De nombreux modèles de vision par ordinateur sont disponibles publiquement, par requête ou API. Chacun peut donc créer une application en envoyant des images en entrée à une API, et en récupérant les sorties d'un modèle pour une tâche particulière. C'est dans ce cadre que nous définissons notre tâche : la classification d'images par des modèles publics.

Définitions et notations. On appelle contexte d'une donnée ou d'un ensemble de données l'ensemble des circonstances dans lesquelles la donnée ou l'ensemble de données a été généré. On définit un ensemble de contextes $\mathcal{C} = \{1, ..., C\}$. A chaque contexte $c \in \mathcal{C}$, on associe une distribution de données P_c sur (x, y, c), où x est une entrée et y l'annotation associée. Un ensemble $\mathcal{C} = \{c_0, ..., c_n\}$ de n contextes aura pour distribution $P_{\mathcal{C}} = \sum_{c \in \mathcal{C}} P_c$.

Un ensemble d'images associé à un contexte $\mathcal{C}$ est donc

composé d'instances $(x, y, c) \sim \mathcal{C}$ avec $c \in \mathcal{C}$. Le glissement de contexte survient quand un modèle entraîné avec un ensemble d'images associé à un contexte $\mathcal{C}_1$ infère sur un ensemble d'images associé à un contexte $\mathcal{C}_2$, avec $\mathcal{C}_1 \neq \mathcal{C}_2$. Un jeu de données sera annoté grâce à un ensemble de k catégories $\mathcal{Y} = \{y_1, ..., y_k\}$, et une annotation pourra être composée de plusieurs catégories, soit $y = \{y_1, ..., y_l\}$ avec $\forall i, y_i \in \mathcal{Y}$ et $0 \leq t \leq k$

Dans la suite, on confondra le contexte $\mathcal{C}$ d'un ensemble d'images et celui du modèle entraîné sur cet ensemble d'images sous la même notation. On confondra de même l'ensemble de catégories $\mathcal{Y}$ de cet ensemble d'images et du modèle entraîné à l'aide de ce dernier.

3.1 Configuration de l'expérience

Afin de caractériser la baisse de performance par zone géographique, on souhaite reproduire dans cette expérience le modèle opératoire de Devries et al. [15], qui évalue les performances de modèles sur la tâche de classification d'images avec des objets communs. Dans cette partie, nous présentons les similarités et différences entre leur configuration et celle que nous proposons.

Données. Devries et al. estiment la baisse de performance par pays et revenus de plusieurs modèles de vision par ordinateur disponibles publiquement, en utilisant le Dollar Street Dataset (DSD), composé de 38 479 images de maisonnées dans 63 pays, annotées avec 289 catégories. Ces images sont réalisées par des photographes dans des maisonnées, et ciblent des éléments de la vie courante. Ces 63 pays couvrent tous les continents, mais délaissent une grande partie de l'Afrique. Dans un soucis de représentativité plus exhaustive, nous utilisons COCO World URLs [7], composé de 400 annotations pour 23 zones géographiques établies selon la norme M49 [46], et couvrant l'ensemble des pays du globe. Ce jeu de données est annoté avec les catégories de MS COCO, on notera donc son ensemble de catégories $\mathcal{Y}_{MS\ COCO}$.

Modèles. Les modèles utilisés par Devries et al. sont des modèles de détection publiquement disponibles des groupes Clarifai, Google Cloud Vision, Amazon Rekognition, IBM Watson, Microsoft Azure, ainsi qu'un modèle à l'état de l'art en reconnaissance d'objets, à priori indépendante du contexte géographique. Ils sont choisis pour leur performances, leur utilisation devenue commune, et la tâche pour laquelle ils ont été entraînés, qui est la classification d'images avec des objets et éléments de la vie courante. Au moment de l'expérience, IBM Watson ne proposait plus les services de détection en ligne, et le modèle à l'état de l'art pour la classification multilabels est CSRA décrit dans les travaux de Zhu and Wu [55]. Les données utilisées pour entraîner les modèles publics ne sont pas décrites par les différentes sociétés, on ne connaît donc pas le contexte $\mathcal{C}_{modele}$ qui leur est associé. Dans le cas de CSRA, différents poids correspondant à l'entraînement sur différents jeux de données sont disponibles. Nous sélectionnerons des poids correspondant à l'entraînement sur MS COCO, le contexte associé sera donc noté $\mathcal{C}_{MS\ COCO}$. Les ensembles de catégories $\mathcal{Y}_{modele}$ associés aux modèles publics ne sont pas forcément disponibles, tandis que celui associé à CSRA est $\mathcal{Y}_{MS\ COCO}$.

Annotations et évaluation. Les métriques automatiques d'évaluation de performance d'un modèle sur un jeu de données nécessitent $\mathcal{Y}_{modele} = \mathcal{Y}_{jeu\ de\ donnees}$. Ce n'est le cas ici que pour le modèle CSRA. Devries et al. contournent ce problème à l'aide d'une solution d'évaluation nécessitant une intervention humaine, où un évaluateur estime si l'une des cinq sorties les plus probables d'un modèle correspond à l'annotation de l'image. Nous souhaitons conserver une évaluation automatique pour faciliter le passage à l'échelle. Nous procédons à une récupération des ensembles de catégories $\mathcal{Y}_{modele}$ quand c'est possible, et à un croisement entre ces derniers. $\mathcal{Y}_{Google\ Cloud\ Vision}$ n'est pas disponible et $\mathcal{Y}_{Microsoft\ Azure}$ est globalement incompatible avec les autres ensembles ; nous utilisons donc comme ensemble d'annotation $\mathcal{Y}_{experience} = \mathcal{Y}_{Clarifai} \cap \mathcal{Y}_{Amazon\ Rekognition} \cap \mathcal{Y}_{MS\ COCO}$, et par conséquent, uniquement les modèles associés. Les annotations y d'une image seront alors réduites à cet ensemble, soit $y_{experience} = y \cap \mathcal{Y}_{experience}$, de même que les prédictions des modèles. Cette opération pourra générer des données images sans annotations, où $y_{experience} = \emptyset$, appelées "données vides" dans la suite. La métrique utilisée sera la hamming loss puisque cette dernière semble la plus adaptée aux configurations multilabels pour la classification [11].

Métriques pour l'observation du glissement. Le glissement de contexte est observé dans les travaux de Devries et al. par deux métriques : une carte de la précision obtenue pour chaque pays testé, et un graphe traçant la précision sur une image par revenu associé à cette image. Ne connaissant pas le revenu associé à chaque image, nous témoignerons du glissement uniquement grâce à une carte de performance par zone géographique dans cette configuration. L'intérêt n'est pas porté aux performances brutes du modèle, mais aux variations de ces performances sur les différentes zones géographiques.

3.2 Résultats et Controverses

Selon les précédentes observations sur les répartitions de données dans les jeux de données d'images communs [15, 43], chacun des modèles utilisés dans l'expérience aura un contexte particulièrement occidental. On pourra découper les contextes des modèles de la manière suivante : $\mathcal{C}_{modele} = \mathcal{C}_{modele}^{occident} + \mathcal{C}_{modele}^{non\ occident}$, et théoriser que pour les modèles utilisés ici, $\mathcal{C}_{modele}^{non\ occident} \approx \emptyset$. De fait, on attend des modèles sélectionnés qu'ils soient plus performants dans les pays présentant un contexte occidental. La figure 1 présente la carte du monde générée par l'expérience. On peut y observer que les pays occidentaux ont des scores homogènes, de même pour les pays en développement. La variation de performance entre ces derniers est pourtant l'inverse du résultat attendu. Les pays présentant un contexte occidental semblent obtenir des scores plus élevés, et donc moins bons, que leurs homologues non occidentaux.

Ces résultats surprenants sont partagés pour chacun des trois modèles sélectionnés pour l'expérience, éliminant la

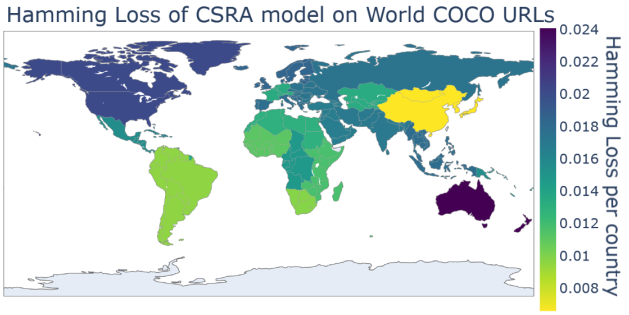

FIGURE 1 – Carte de performance du modèle CSRA sur les données de COCO World URLs [7]. La métrique utilisée est la Hamming Loss, qui est meilleure plus elle est basse.

remise en cause de ces derniers. Il est donc nécessaire de questionner le reste du processus expérimental afin d'identifier la cause de l'inversion du phénomène attendu.

Changement des données. On modifie en premier lieu dans la configuration de l'expérience le jeu de données sur lequel les modèles sont utilisés. On utilisera le jeu témoin DSD, ce dernier possédant une granularité fine d'information liant image et information géographique. L'ensemble de catégories de DSD possédant 9 catégories en commun avec celui de MS COCO, on utilisera leur intersection $\mathcal{Y}_{MS\ COCO} \cap \mathcal{Y}_{DSD}$ pour l'expérience. Seul le modèle CSRA sera utilisé, ce dernier utilisant aussi l'ensemble de catégories $\mathcal{Y}_{MS\ COCO}$.
La figure 2 présente la carte du monde générée par l'expérience.

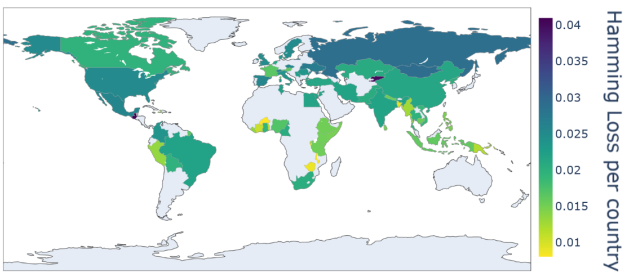

FIGURE 2 – Carte de performance du modèle CSRA sur les données de Dollar Street Dataset [39]. La métrique utilisée est la Hamming Loss, qui est meilleure plus elle est basse.

Ces résultats sont similaires à l'expérience précédente, puisqu'il y a une variation de performance entre pays occidentaux et non occidentaux, mais que cette variation se fait encore à l'inverse du résultat attendu.

Changement de l'évaluation. On modifie en second lieu la métrique de performance du modèle, en passant de la Hamming Loss à la précision@5, afin de se rapprocher des travaux de Devries et al. L'expérience est d'abord faite sur COCO World URLs et avec les trois modèles sélectionnés précédemment, puis sur DSD avec le modèle CSRA uniquement.
La figure 3 présente la carte du monde générée par l'inférence avec le modèle CSRA sur COCO World URLs, les cartes générées par les modèles Clarifai et Amazon Rekognition présentant des caractéristiques similaires.

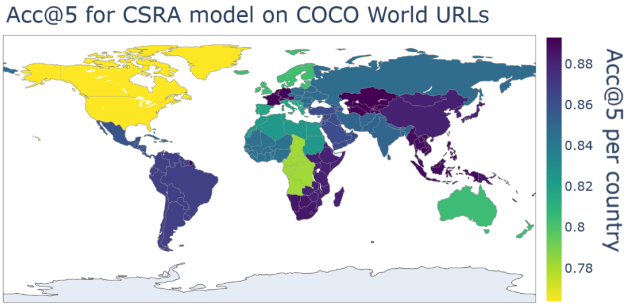

FIGURE 3 – Carte de performance du modèle CSRA sur les données de COCO World URLs [7]. La métrique utilisée est la précision @5, qui est meilleure plus elle est haute.

La lecture de cette carte est moins évidente ; les scores des pays occidentaux ne sont plus homogènes, avec l'Amérique du Nord présentant le score le plus bas, et l'Europe du Nord présentant le deuxième meilleur score, derrière l'Asie centrale. De même pour les pays non occidentaux, l'Afrique centrale présente un score faible tandis que l'Afrique de l'Est et du Sud présentent des scores élevés. Cette expérience relève deux points : l'influence de la métrique de performance du modèle sur le témoignage du glissement de contexte, et les faiblesses d'une carte comme outil pour témoigner de ce dernier.

Changement des données et de l'évaluation. Dans un soucis d'exhaustivité, nous reproduisons l'expérience en cumulant les deux changements précédant. La figure 4 présente la carte de résultats pour le modèle CSRA sur DSD avec la précision@5 comme métrique de performance pour le modèle.

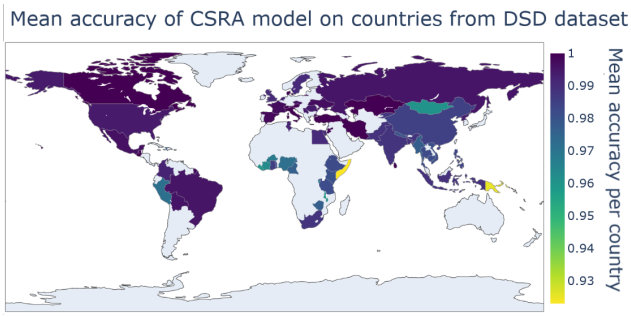

FIGURE 4 – Carte de performance du modèle CSRA sur les données de Dollar Street Dataset [39]. La métrique utilisée est la précision@5, qui est meilleure plus elle est haute.

C'est le premier résultat pour lequel nous arrivons à ce qui était attendu : une claire variation de performance entre pays occidentaux et non occidentaux, avec un avantage net pour les premiers. Ces résultats semblent souligner l'importance des données et de la métrique de performance dans le témoignage du phénomène de glissement de contexte.

4 Analyse et discussions

Nous analysons dans cette section les multiples résultats obtenus dans la partie précédente et tentons d'expliquer les paramètres influençant ces résultats.

Influence des données vides. La répartition des données vides (les images où $y_{experience} = \emptyset$) n'étant pas uniforme dans les différentes zones géographiques ou pays (voir la figure 5), et les données vides ayant un score uniforme, cela perturbe l'estimation des variations de performance des modèles en fonction des localités. Ces deux facteurs étant indépendants des modèles, ils portent atteinte à leur résilience.

Il convient de se poser la question de la pertinence de ces données vides et de la configuration de l'expérience. Retirer les données vides de l'expérience revient à amputer les jeux de données d'une partie de leur volume. Cela revient à biaiser les résultats et à manquer d'analyser une réponse du modèle face à une donnée non pertinente pour ce dernier. Ces données vides sont quant à elles issues d'autres biais, qui sont les choix humains ayant mené à la construction des ensembles de catégories différents pour les ensembles d'images.

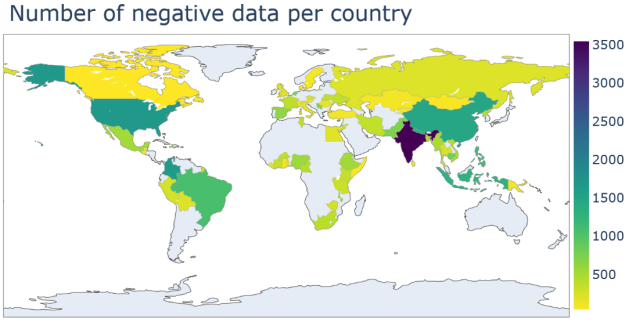

FIGURE 5 – Répartition des données vides dans DSD avec $\mathcal{Y}_{experience} = \mathcal{Y}_{MS\ COCO} \cap \mathcal{Y}_{DSD}$

Il existe des approches permettant d'adapter les modèles aux tâches désirées, comme le Transfer Learning [35], le One-shot Learning [28] ou encore le Zero-shot Learning [38]. Ces approches sont très utilisées dans la littérature, mais ne reflètent pas le cadre d'utilisation des modèles via API que nous reproduisons ici. Elles requièrent des capacités de calcul conséquentes pour entraîner ou accorder des modèles, et l'accès aux couches et poids de ces derniers - des facteurs qui constituent un frein à un déploiement rapide dans les pays non occidentaux.

Le rôle des données vides pourrait être réduit en utilisant d'autres correspondances entre les ensembles de catégories. Des approches telles que le "fuzzy matching" ont été utilisées dans des travaux antérieurs [39], mais à ce jour, la plupart des appariements sont effectués à la main et dépendent donc de l'effort et des choix humains. Les méthodes de transformation des problèmes telles que la "binary relevance", le "label powerset" ou les méthodes d'ensemblage [11] devraient être envisagées pour pallier le problème des données vides. Avec le développement des embeddings de mots, des recherches ultérieures pourraient également envisager d'automatiser le processus de "fuzzy matching" en utilisant la proximité des mots dans les espaces d'embeddings et ainsi permettre l'automatisation de la généralisation de domaines pour les applications de vision par ordinateur.

Influence des jeux de données. Les analyses des figures 3 et 4 soulignent l'influence des données pour le témoignage du phénomène de glissement de contexte. Cette différence s'explique en partie par le rôle du volume de données, bien plus important dans DSD que dans COCO World URLs, ce qui permet une meilleure confiance dans les résultats obtenus. Ces deux jeux de données diffèrent aussi par la précision des méta-données : elles sont vérifiées pour DSD tandis que COCO World URLs a été généré grâce à des requêtes en langue anglaise sur la plateforme Flickr. Ce second jeu de données aura donc des biais non maîtrisés, qui peuvent perturber l'expérience réalisée, puisqu'ils se retrouvent dans le contexte associé au jeu de données.

La configuration choisie diffère des problèmes classiques de glissement de contexte [25], car les données ED et HD ne sont pas clairement définies. Ceci est du en partie à une multiplicité des glissements dans notre expérimentation : glissement d'un jeu de données à un autre, d'un ensemble de catégories à un autre, d'un contexte occidental à un contexte non occidental. Isoler le dernier glissement en utilisant un modèle entraîné sur les données occidentales de DSD et en le testant sur l'ensemble de DSD permettrait de caractériser ce glissement, mais ne permet pas de le caractériser pour l'ensemble des modèles disponibles publiquement.

Influence des métriques. Deux types de métriques sont à considérer : les métriques de performance des modèles et les métriques de témoignage du glissement. Pour les premières, deux métriques différentes sont utilisées ici : la hamming loss et la précision@5. Les constats opposés entre les figures 2 et 4 soulignent l'importance du choix de la métrique de performance des modèles. Ces différences de constat s'expliquent par la configuration de l'expérience (ensemble de catégories, influence des données vides) et par la sensibilité des métriques à différents aspects entre prédictions et annotations.

La métrique utilisée pour caractériser le glissement de contexte a deux inconvénients majeurs. Premièrement, elle est sujette à l'interprétation visuelle, et ne définit pas clairement quand une baisse de performance est significative ou non. Cette absence d'interprétation claire est illustrée en figure 3. Deuxièmement, elle repose sur une intervention humaine, et n'est donc pas adaptée à un passage à l'échelle, en benchmark par exemple. Cette métrique manque aussi de recul sur sa robustesse par rapport à de possibles préjudices, comme un déséquilibre dans les catégories ou les populations dans le jeu de données, ou des données irrégulières. Ces critiques sont également valables pour les autres métriques utilisées dans la littérature, comme les graphes de densité de log-vraissemblance pour une classe particulière ou de performance par revenu associé aux données.

Ces facteurs d'influence sont une fois de plus indépendants

des modèles, et constituent donc une possible atteinte à la résilience de ces derniers. Des recherches supplémentaires sont nécessaires pour élaborer une caractérisation pertinente de la chute de performance due au contexte géographique. Étant donné que le pays d'origine et le revenu sont des attributs démographiques, comme le sexe, la race, la religion et d'autres attributs qui sont abondamment traités dans les travaux relatifs à l'équité, les futures recherches sur ce sujet ont tout intérêt à se rapprocher des métriques d'équité.

5 Conclusion

La caractérisation du glissement de contexte des modèles publics dans le cadre de leur déploiement dans des pays non occidentaux est une tâche qui soulève de nombreux défis, dont l'identification de jeux de données adaptés, le passage d'un ensemble de catégories à un autre, et le choix de métriques de performance appropriées.
Cette caractérisation souffre aussi du manque d'une métrique pouvant clairement témoigner du phénomène de glissement de contexte. Les métriques visuelles proposées dans les travaux antérieurs, si elles ont l'avantage d'apporter de l'explicabilité en comparaison à leurs homologues numériques, ne permettent pas de s'affranchir de l'interprétation humaine.
Les défis identifiés sont autant d'obstacles à la résilience des modèles d'IA pour la soutenabilité. L'IA est engagée dans la lutte contre le changement climatique, mais ces obstacles sont des freins à son utilisation dans les contextes appropriés.
Nous soulignons dans ces travaux les conséquences des biais de construction des jeux de données d'images. Bien que nous encourageons l'amélioration de la visibilité des communautés les moins représentées dans ces bases de données d'images, nous souhaitons également souligner les risques liés à la collecte de ces données. De la même manière que les attributs de genre ou d'ethnicité peuvent causer des dommages lorsqu'ils sont collectés ou affichés dans une base de données, la diffusion d'attributs démographiques pourrait nuire aux communautés sous-représentées déjà désavantagés. Compte tenu à la fois de l'incitation à produire des données plus inclusives et du risque associé aux métadonnées démographiques, le partage de ces métadonnées devrait toujours se faire avec une grande prudence.

Remerciements

Ce travail est financé par Sorbonne Université via la bourse du Programme Doctoral International de Modélisation pour les Systèmes Complexes.

Références

[1] Sushant Agarwal. Trade-Offs between Fairness and Interpretability in Machine Learning. In *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[2] Sushant Agarwal. Trade-Offs between Fairness and Privacy in Machine Learning. In *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[3] Kendra Albert, Maggie Delano, Jonathon Penney, Afsaneh Rigot, and Ram Shankar Siva Kumar. Ethical Testing in the Real World : Evaluating Physical Testing of Adversarial Machine Learning. *SSRN Electronic Journal*, 2020.

[4] Glenn Althor, James E. M. Watson, and Richard A. Fuller. Global mismatch between greenhouse gas emissions and the burden of climate change. *Scientific Reports*, 6(1) :20281, February 2016.

[5] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What We Can't Measure, We Can't Understand : Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 249–260, Virtual Event Canada, March 2021. ACM.

[6] James Atwood, Yoni Halpern, Pallavi Baljekar, Eric Breck, D. Sculley, Pavel Ostyakov, Sergey I. Nikolenko, Igor Ivanov, Roman Solovyev, Weimin Wang, and Miha Skalic. The Inclusive Images Competition. In Sergio Escalera and Ralf Herbrich, editors, *The NeurIPS '18 Competition*, pages 155–186. Springer International Publishing, Cham, 2020. Series Title : The Springer Series on Challenges in Machine Learning.

[7] Theophile Bayet. COCO-style geographically unbiased image dataset for computer vision applications, 2023. Type : dataset.

[8] Theophile Bayet, Christophe Denis, Alassane Bah, and Jean-Daniel Zucker. Distribution Shift nested in Web Scraping : Adapting MS COCO for Inclusive Data. In *ICML Workshop on Principles of Distribution Shift 2022*, Baltimore, United States, July 2022.

[9] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet ? *arXiv :2006.07159 [cs]*, June 2020. arXiv : 2006.07159.

[10] Abeba Birhane and Vinay Uday Prabhu. Large image datasets : A pyrrhic win for computer vision ? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, Waikoloa, HI, USA, January 2021. IEEE.

[11] Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203 :117215, October 2022.

[12] Joy Buolamwini and Timnit Gebru. Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, February 2018.

[13] Kate Crawford and Trevor Paglen. Excavating AI.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet : A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, June 2009. IEEE.

[15] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does Object Recognition Work for Everyone ? In *CVPR Workshops*, June 2019. arXiv : 1906.02659.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, Cambridge Massachusetts, January 2012. ACM.

[17] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. Identifying Statistical Bias in Dataset Replication. *arXiv :2005.09619 [cs, stat]*, September 2020. arXiv : 2005.09619.

[18] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2) :303–338, June 2010.

[19] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, Sydney NSW Australia, August 2015. ACM.

[20] Luciano Floridi. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6) :261–262, June 2019.

[21] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech 256, April 2022. Version Number : 1.0 Type : dataset.

[22] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS : A Dataset for Large Vocabulary Instance Segmentation. *arXiv :1908.03195 [cs]*, September 2019. arXiv : 1908.03195.

[23] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[24] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9) :389–399, September 2019.

[25] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS : A Benchmark of in-the-Wild Distribution Shifts. *arXiv :2012.07421 [cs]*, July 2021. arXiv : 2012.07421.

[26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4 : Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7) :1956–1981, July 2020. arXiv : 1811.00982.

[27] Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine*, 29(6) :141–142, November 2012.

[28] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4) :594–611, April 2006.

[29] Anne-Laure Ligozat and Sasha Luccioni. A Practical Guide to Quantifying Carbon Emissions for Machine Learning researchers and practitioners. 2021.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO : Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing, Cham, 2014. Series Title : Lecture Notes in Computer Science.

[31] Lynn Miller, Christoph Rüdiger, and Geoffrey I. Webb. Using AI and Satellite Earth Observation to Monitor UN Sustainable Development Indicators. In *AI for Social Good Workshop*, 2020.

[32] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1) :521–530, January 2012.

[33] Rohit Nishant, Mike Kennedy, and Jacqueline Corbett. Artificial intelligence for sustainability : Challenges, opportunities, and a research agenda. *International Journal of Information Management*, 53 :102104, August 2020.

[34] Chimango Nyasulu, Awa Diattara, Assitan Traore, Abdoulaye Deme, and Cheikh Ba. Towards Resilient Agriculture to Hostile Climate Change in the Sahel Region : A Case Study of Machine Learning-Based Weather Prediction in Senegal. *Agriculture*, 12(9) :1473, September 2022.

[35] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10) :1345–1359, October 2010.

[36] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon Emissions and Large Neural Network Training. *arXiv :2104.10350 [cs]*, April 2021. arXiv : 2104.10350.

[37] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet ? *arXiv :1902.10811 [cs, stat]*, June 2019. arXiv : 1902.10811.

[38] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. ML-Decoder : Scalable and Versatile Classification Head. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 32–41, Waikoloa, HI, USA, January 2023. IEEE.

[39] William A. Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The Dollar Street Dataset : Images Representing the Geographic and Socioeconomic Diversity of the World. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[40] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, 55(2) :1–96, February 2023.

[41] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, and Vinodkumar Prabhakaran. Non-portability of Algorithmic Fairness in India, December 2020. arXiv :2012.03659 [cs].

[42] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS : Benchmarks for Subpopulation Shift. In *International Conference on Learning Representations*. arXiv, August 2020. arXiv :2008.04859 [cs, stat].

[43] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No Classification without Representation : Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv :1711.08536 [stat]*, November 2017.

[44] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019. Association for Computational Linguistics.

[45] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, Colorado Springs, CO, USA, June 2011. IEEE.

[46] United-Nations. UNSD — M49 methodology - https ://unstats.un.org/unsd/methodology/m49/.

[47] Aimee van Wynsberghe. Sustainable AI : AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3) :213–218, August 2021.

[48] Mei Wang and Weihong Deng. Deep visual domain adaptation : A survey. *Neurocomputing*, 312 :135–153, October 2018.

[49] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive Inequity in Object Detection, February 2019. arXiv :1902.11097 [cs, stat].

[50] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. Sustainable AI : Environmental Implications, Challenges and Opportunities. *arXiv :2111.00364 [cs]*, October 2021. arXiv : 2111.00364.

[51] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets : filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, Barcelona Spain, January 2020. ACM.

[52] Kaiyu Yang, Jacqueline H. Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A Study of Face Obfuscation in ImageNet. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25313–25330. PMLR, July 2022.

[53] Daniele Zanaga, Ruben Van De Kerchove, Wanda De Keersmaecker, Niels Souverijns, Carsten Brockmann, Ralf Quast, Jan Wevers, Alex Grosu, Audrey Paccini, Sylvain Vergnaud, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Ivelina Georgieva, Myroslava Lesiv, Sarah Carter, Martin Herold, Linlin Li, Nandin-Erdene Tsendbazar, Fabrizio Ramoino, and Olivier Arino. ESA WorldCover 10 m 2020 v100, October 2021. Version Number : v100 Type : dataset.

[54] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain Generalization : A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2022.

[55] Ke Zhu and Jianxin Wu. Residual Attention : A Simple but Effective Method for Multi-Label Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 184–193, October 2021. ISSN : 2380-7504.