Stratégie d'Apprentissage Guidée par la Focalisation de l'Attention pour la Généralisation de Domaine

Julien Miranda[1,2] Mathieu Giraud[2] Stanislas Larnier[2] Ariane Herbulot [1] Michel Devy [1]

[1] LAAS, CNRS, 7 avenue du colonel Roche, F-31400 Toulouse, France
Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France
[2] Donecle, IoT Valley, 201 Rue Pierre et Marie Curie, 31670 Labège, France

julien.miranda@laas.fr / julien.miranda@donecle.com

Résumé

Si les algorithmes de vision par ordinateur actuels atteignent d'excellentes performances en classification d'images, leur utilisation dans le monde réel en particulier lors de l'intégration à un système robotique dans un environnement ouvert reste un sujet difficile. En effet, les modèles les plus efficaces reposent sur des mécanismes d'apprentissage profond dont les performances, encore mal expliquées par la théorie mais constatées en pratique, sont associées au seul domaine couvert par les données d'entraînement. Dans cet article nous étudions la généralisation de domaine dans le cadre de l'inspection visuelle de fuselage d'avion par un drone autonome. Nous proposons une méthode permettant de mesurer la qualité de l'attention d'un réseau de classification pour expliquer le phénomène de transfert négatif. Dans nos recherches d'un modèle robuste aux changements de domaine, nous étudions l'impact positif de transformations préalables des images bien choisies sur différents indicateurs. Nous évaluons pour cela la complexité a priori du problème, la qualité de l'attention et le score de classification des réseaux après entraînement. Enfin nous mettons en perspective les propriétés de généralisation des modèles et leur complexité pour proposer une stratégie d'apprentissage guidée par la focalisation de l'attention. Cette méthode permet d'améliorer significativement les performances en généralisation de domaine sur nos jeux de données sans nécessiter de modification pour le système déployé.

Mots Clef

Classification d'images, généralisation de domaine, segmentation non supervisée, complexité de jeux de données, complexité de modèles, stratégie d'apprentissage

Abstract

Current computer vision algorithms achieve excellent performance in image classification. However their use in the real world when integrating into a robotic system in an open environment remains a difficult subject. Indeed, the most effective models are based on deep learning mechanisms whose performance is only associated with the domain covered by training data. Moreover those abilities are observed in practice but still poorly explained by theory. In this article we study domain generalization for visual inspection of aircraft fuselage by an autonomous drone. We propose a method to measure the quality of the attention of a classification network to explain the phenomenon of negative transfer. In our search for a robust model for domain changes, we study the positive impact of well-chosen prior transformations applied to images on the complexity of the problem and on the performance on test sets. Overall, we relate the generalization properties of models to those notions of complexity to propose a guided learning strategy for attention focusing. This method significantly improves performance in domain generalization on our datasets without requiring any modification for the deployed system.

Keywords

Image classification, domain generalization, unsupervised segmentation, dataset complexity, model complexity, learning strategy

1 Introduction

La classification automatique d'images est un problème très courant dans divers domaines industriels [1, 2]. Il n'est cependant pas rare que le nombre de données étiquetées dans le domaine d'utilisation (domaine cible) soit réduit voire inexistant alors qu'un plus grand nombre d'entre elles est disponible dans un ou plusieurs domaines voisins (domaines sources). En particulier, la robotisation des tâches de maintenance aéronautique, et plus spécifiquement l'automatisation des inspections visuelles basée sur des algorithmes de vision par ordinateur [3, 4, 5] se heurte à ce problème. Ces inspections peuvent aujourd'hui être effectuées sur la base d'images du fuselage acquises par un drone autonome [6].

Lors de telles analyses, tous les éléments présents sur le fuselage doivent être catégorisés afin de fournir un rapport d'inspection détaillé. Cette tâche est généralement résolue en entraînant un réseau de neurone profond sur

de grandes quantités de données, ou en effectuant un ré-apprentissage d'un réseau préalablement entraîné pour l'adapter à la tâche. Néanmoins la grande variabilité des modèles d'avions et des livrées (peinture) sur ces aéronefs rend le problème plus difficile à traiter selon ce paradigme. En effet, chaque appareil peut constituer un domaine spécifique, et parvenir à une généralisation à l'ensemble d'entre eux pourrait demander d'annoter de nombreux avions différents, avec des garanties limitées lorsque l'avion à inspecter présente une livrée jusqu'alors inconnue. Quelques exemples de différentes livrées sont présentés dans la figure 1. Cette diversité peut conduire à un problème important, appelé transfert négatif [7]. Ce phénomène survient en particulier lorsque des exemples d'un type d'objet disponibles dans un seul domaine sont ajoutés à l'ensemble d'entraînement qui ne contenait pas d'objets dans celui-ci. L'ajout de tels éléments vise à doter le système de la capacité à reconnaitre cette nouvelle classe.

FIGURE 1 – Différentes livrées d'avions.

En pratique, on observe une association systématiquement entre ce nouveau domaine et la nouvelle classe, peu importe l'objet présent dans l'image. Cela conduit à des performances désastreuses lors de l'utilisation.
Dans cet article nous proposons d'étudier la généralisation de domaine pour la classification d'images dans le cas où le domaine peut être identifié en considérant l'arrière-plan de l'image (l'objet d'intérêt est au premier plan). Nos travaux se placent dans le cadre d'inspection automatisée de fuselage d'avions de ligne et se basent sur des acquisitions réalisées à l'aide des drones Donecle. La section 2 est un état de l'art des différents champs constitutifs de notre étude, à savoir la généralisation de domaine, l'étude de la complexité des modèles et des jeux de données pour la généralisation et les stratégies d'apprentissage prenant en compte la complexité des échantillons. Dans la section 3 nous montrons que les faibles propriétés de généralisation du réseau de neurone utilisé peuvent être attribuées à une mauvaise focalisation de l'attention, portée sur le domaine aux dépends des objets. Nous introduisons pour cela une mesure de la qualité de l'attention du modèle sur les objets. Sur ces bases, nous faisons l'hypothèse en section 4 qu'un pré-traitement des images favorisant la prise en compte du premier plan est une réponse possible au problème de généralisation de domaine et nous proposons une transformation adaptée. Nous explorons alors la possibilité d'évaluer la difficulté du problème à partir du jeu de données disponibles pour l'apprentissage. Pour cela nous utilisons une métrique basée sur la théorie du regroupement spectral issue de la littérature, évaluée pour les différentes transformations. Cette dernière vient compléter les analyses basées sur la qualité de l'attention et les performances en classification menées en section 5. Nous utilisons ces analyses pour étendre notre approche et parvenir à une stratégie d'apprentissage par cursus (*curriculum learning*) en section 6. Nous mettons enfin en perspective en section 7 la difficulté estimée des jeux de données et les performances en classification avec une mesure théorique de la complexité du classifieur et la borne de généralisation dont elle dérive.

2 État de l'art

2.1 Généralisation de domaine

Différentes méthodes de transfert (ou adaptation) de domaine pour l'apprentissage profond sont disponibles dans la littérature. On différencie la tâche d'adaptation de domaine pour laquelle des données sont disponibles dans le domaine source et le domaine cible [8, 9], de la tâche de généralisation de domaine, pour laquelle des données sont exploitables uniquement dans le domaine source [10, 11]. On considère ici le cas de la généralisation de domaine. De nombreux travaux mettent en évidence le lien entre une mauvaise localisation de l'attention du réseau et une erreur importante lors de la généralisation de domaine [12, 13]. Il a en outre été montré que les mécanismes d'attention sont fortement corrélés avec les performances des réseaux [14]. Des études proposent d'utiliser de nombreux domaines, ou d'ajouter des modules tels qu'un auto-encodeur variationnel afin d'apprendre un espace latent robuste au domaine [15, 16]. L'utilisation d'un nombre conséquent de domaines nécessaires à la mise en place d'une stratégie de méta-apprentissage implique un effort d'annotation supplémentaire et ne permet pas de s'affranchir avec certitude des effets de transfert négatif. D'une part l'ajout de modules rend le système plus complexe, ce qui diminue la généralisation des capacités observées lors de l'apprentissage, mais ce point n'est en général pas abordé. D'autre part, le phénomène de mauvaise focalisation de l'attention n'y est observé que qualitativement. Enfin, on trouve l'idée de l'introduction d'une simplification du problème d'apprentissage via une réduction de la dimension de l'espace de représentation [17], mais l'application d'une transformation directement sur les images n'est pas traitée dans ces travaux.
Dans cet article, nous proposons d'utiliser un indicateur simple pour quantifier le phénomène de mauvaise focalisation, nous introduisons également un pré-traitement adapté à la généralisation de domaine et mettons en perspective les effets de cette transformation avec les propriétés théoriques connues de généralisation des réseaux de neurones.

2.2 Complexité du modèle et des données

Plusieurs approches théoriques visent à décrire les propriétés de généralisation des modèles issus de l'apprentissage automatique, et en particulier aujourd'hui des réseaux de neurones profonds. En effet, les capacités de généralisation

observées pour ces réseaux ne sont pas bien expliquées par les bornes de généralisations, issues de la théorie de l'apprentissage statistique telles que la dimension VC (pour dimension de Vapnik-Tchervonenkis) ou le moyenne de Rademacher [18, 19]. Ces bornes permettent toutefois de mieux comprendre le rôle de la complexité de la classe de modèles dans la capacité de généralisation de ses éléments. La prise en compte des propriétés des données du problème est soit inexistante (bornes basées uniquement sur la dimension VC), soit indirecte (Rademacher). En particulier, la complexité spectrale [20] est basée sur les propriétés des matrices de poids obtenues après entraînement du réseau.
L'évocation de ce résultat nécessite l'introduction de quelques notations. Soient $(\sigma_1, ..., \sigma_L)$ les L non-linéarités du réseau, avec pour $i \in (1, ..., L)$, $\sigma_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ ρ_i-Lipschitzienne. Soient $\mathcal{A} = (A_1, ..., A_L)$ les matrices de poids du réseau. On note $\mathcal{F}_\mathcal{A}$ la fonction appliquée par ce réseau, telle que :

$$\mathcal{F}_\mathcal{A}(x) = \sigma_L(A_L \sigma_{L-1}(...))$$

On note alors complexité spectrale $\mathcal{R}_\mathcal{A}$ le produit :

$$\mathcal{R}_\mathcal{A} := \left(\prod_{i=1}^{L} \rho_i \left\|A_i\right\|_\sigma\right) \left(\sum_{i=1}^{L} \frac{\left\|A_i^T - M_i^T\right\|_{2,1}^{2/3}}{\left\|A_i\right\|_\sigma^{2/3}}\right)^{3/2} \quad (1)$$

où $M = (M_1, ..., M_L)$ est une collection de *matrices de référence* fixée. Nous nous focalisons ici sur la généralisation de domaine et ne pouvons donc pas nous appuyer sur ces travaux qui étudient les garanties théoriques dérivées de bornes de généralisation existantes en prenant en compte des échantillons disponibles dans le domaine cible [21, 22].
Dans cet article nous tentons d'évaluer la capacité des bornes de généralisation à traduire l'effet des traitements effectués sur les données, en questionnant la limite d'une prise en compte seulement implicite des caractéristiques liées au jeu de données. Nous explorons pour ce faire des mesures plus directes afin de prévoir a priori la difficulté du problème. La littérature propose des approches [23] avec pour application notamment le nettoyage de jeux de données par sous-échantillonnage. Plus récemment une mesure spectrale adaptée à la classification par un réseau convolutif (*Cumulative Spectral Gradient*, ou CSG) [24] a été proposée. Elle tire sa pertinence de la corrélation observée entre les performances en classification obtenues après entraînement et la complexité estimée du jeu de données.

2.3 Stratégies d'apprentissage

L'une des raisons pouvant potentiellement expliquer la focalisation de l'attention sur le domaine observé au cours de nos expériences réside dans la stratégie d'apprentissage naïve adoptée par défaut pour ce type de problèmes. En effet, lors de la phase d'entraînement, les données sont présentées aléatoirement, donc sans structure ou ordre. Or plusieurs travaux montrent l'effet positif d'une structuration des échantillons lors de cette étape, généralement par complexité croissante pour former un apprentissage par cursus [25], via un apprentissage "à son propre rythme" (*self-paced learning*) [26] ou encore utilisant un planificateur d'échantillons [27]. La mise en place d'une telle stratégie suppose de pouvoir définir à l'avance quel échantillon est le plus à-même d'être pris en compte en premier lors de l'apprentissage. Dans cet article nous nous basons sur les différentes approches de la complexité évoquées ci-dessus pour proposer un ordonnancement des transformations et donc des échantillons. Cette stratégie d'apprentissage guidé permet de restreindre les étapes de pré-traitement à la seule phase d'apprentissage.

3 Analyse quantitative de la mauvaise focalisation de l'attention

Dans le cadre des travaux présentés ici, on considère un domaine source D_1 correspondant à des éléments présents sur une zone du fuselage sans peinture spécifique. Le domaine cible D_2 correspond à des éléments présents sur une zone de fuselage spécifique à la livrée.

FIGURE 2 – Domaines D_1 (gauche) et D_2 (droite)

La figure 2 présente des acquisitions du drone sur deux domaines. La tâche de classification consiste ici à reconnaitre sur le fuselage les différents types d'éléments à prendre en compte lors d'une inspection visuelle. On se restreint pour notre étude à certaines classes d'objet, présentées dans la table 1. La base de données d'entraînement est constituée de 1500 éléments pour chacune des 7 classes, la base de test est constituée de 128 éléments pour chaque classe. Les échantillons proviennent d'une base de donnée industrielle propriétaire.

Domaine	Vis	Rivet	Étiquette	Marquage
D_1				
D_2				

TABLE 1 – Quelques éléments dans différents domaines.

Pour la suite des travaux, on utilise un réseau de neurones profond classique (VGG-11 [28]). Nos expériences portent sur la classification de différents éléments appris ou testés soit uniquement sur le domaine D_1, soit uniquement sur

le domaine D_2, soit sur les deux domaines $D_{1,2}$. On observera les performances en classification selon les situations (généralisation d'un domaine source vers un domaine cible, ou fusion des domaines). Les performances du réseau de référence dans ces situation sont présentées dans la table 2.

Source	Cible	Précision	Rappel	F1-score
D_1	D_2	0.86	0.71	0.78
$D_{1,2}$	$D_{1,2}$	0.94	0.87	0.91
D_2	D_1	0.78	0.96	0.86

TABLE 2 – Performances du réseau VGG-11 selon la composition des domaines source et cible.

On remarque que les transferts de domaine induisent soit un score de rappel faible (D_1 vers D_2) soit un score de précision faible (D_2 vers D_1). Nous faisons l'hypothèse que l'erreur observée est due à une focalisation de l'attention du réseau de classification sur le domaine plutôt que sur l'objet contenu dans l'image. Afin de valider cette hypothèse, nous observons la carte d'activation spatiale du réseau pour la classe retenue grâce une localisation basée sur le gradient (GradCam [12]). La superposition à l'image d'entrée permet de visualiser les zones utilisées pour la décision.

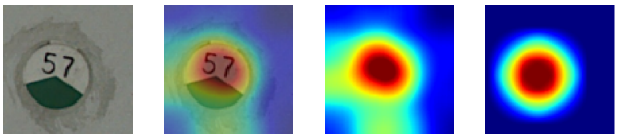


FIGURE 3 – De gauche à droite : image, visualisation de la décision, $\hat{G}$, G.

Une étude quantitative est menée à partir de masques objets binaires issus d'une segmentation manuelle qui forme la référence pour l'attention du réseau.
On construit G le masque d'attention et $\hat{G}$ la matrice d'attention estimée par l'algorithme GradCam tels que $G_{i,j} \in [-1;1]$ et $\hat{G}_{i,j} \in [0;1]$. Ces cartes sont illustrées dans la figure 3. Les valeurs négatives du masque de référence permettent de pénaliser l'attention portée dans des zones de l'image qui ne contiennent pas l'objet. Plus précisément, le masque d'attention est construit à partir d'une image binaire $(-1, 1)$ issue de la segmentation manuelle des objets, à laquelle on applique un flou gaussien paramétré sur la base de la mesure de la variance inter-humains pour cette tâche.
Le score s_a est alors défini par :

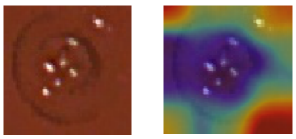
$$s_a = \frac{1}{A} \sum_i \sum_j G_{i,j} \hat{G}_{i,j}$$

Où $A = \sum_i \sum_j \max(G_{i,j}, 0)$ est la surface de l'objet. Par construction, un score plus élevé témoigne d'une meilleure prise en compte des zones appartenant à l'objet, alors qu'un score faible indique une décision basée sur des éléments n'appartenant pas à celui-ci. En particulier les scores négatifs sont symptomatiques d'une décision prise sur des zones hors-objet. L'observation comparée du score moyen S pour les prédictions correctes ($f(x) = y$) et incorrectes ($f(x) \neq y$) présentée en table 3 met en évidence plusieurs phénomènes.

Source	Cible	$s_a \vert f(x) = y$	$s_a \vert f(x) \neq y$
D_1	D_2	0.63	0.58
$D_{1,2}$	$D_{1,2}$	0.90	0.74
D_2	D_1	0.23	0.11

TABLE 3 – Scores d'attention lors d'une classification correcte et incorrecte.

Tout d'abord, le score d'attention est en moyenne plus faible lorsque le résultat de la classification est faux. Cette observation conforte l'hypothèse selon laquelle les zones de l'image n'appartenant pas aux objets contribuent à l'erreur observée. La figure 4 illustre l'un de ces cas : dans cet exemple la classe 'étiquette' est prédite en lieu et place de 'vis', et on peut observer que l'attention est portée dans les coins de l'image plutôt qu'au centre, où se trouve l'objet.



$$\begin{cases} y &= \text{vis} \\ f(x) &= \text{étiquette} \end{cases}$$

FIGURE 4 – Exemple de mauvaise focalisation.

D'autre part, pour la généralisation de D_2 vers D_1, on observe un score très faible, qui indique que l'apprentissage s'est porté sur le domaine plus que sur l'objet, et explique le faible taux de précision dans ce cas. Après avoir mis en évidence le lien entre les zones de l'image participant à la classification et l'erreur en généralisation de domaine, nous proposons de trouver un traitement correctif des images permettant de limiter la contribution négative du domaine pour la classification. Dans la section suivante nous cherchons une transformation des images permettant de réduire l'influence de l'arrière plan lors de la phase d'apprentissage.

4 Transformations des images

4.1 Transformation en niveaux de gris

Le traitement le plus simple à appliquer pour des domaines liés à la livrée semble être d'utiliser les images en niveaux de gris plutôt qu'en couleur. La table 4 présente les résultats obtenus, en comparaison avec le réseau de référence sur les données initiales.
Ces résultats permettent de noter des effets positifs pour les cas de généralisation de domaine mais montrent une régression par rapport au réseau de référence lors d'une

Source	Cible	Précision	Rappel	F1-score
D_1	D_2	0.88	0.74	0.80 (+2%)
$D_{1,2}$	$D_{1,2}$	0.88	0.89	0.89 (-2%)
D_2	D_1	0.83	0.93	0.88 (+2%)

TABLE 4 – Performances du réseau VGG-11 selon la composition des domaines sources et cibles, sur des images traitées en niveaux de gris.

fusion de domaines. Ne pas prendre en compte l'information de couleur pour les zones appartenant aux objets est en soi dommageable car cette information peut être considérée comme utile, en particulier pour dissocier certains éléments. Quelques exemples sont visibles en figure 5, où on peut constater que les deux première images (étiquettes rouge et verte) se distinguent principalement par leur couleur.

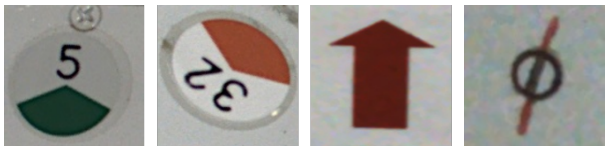


FIGURE 5 – Exemples d'objets dont la couleur est une caractéristique discriminante.

Un traitement préalable des images peut donc améliorer la capacité de généralisation du système à d'autres domaines, mais une simple conversion en niveaux de gris ne permet pas un progrès significatif sur cet aspect sans régression dans le cas de fusion de domaines. Ainsi, nous proposons une transformation des images qui permet de réduire l'influence négative des variations de domaine sans perte d'information utile. Nous avons vu que le risque majeur est la focalisation sur l'arrière-plan, ce qui nous conduit à une méthode basée sur une segmentation de l'objet (premier plan) vis à vis du domaine (arrière-plan).

4.2 Segmentation du premier plan

Des méthodes permettant une segmentation non supervisée du premier plan ont été proposées, et nombre d'entre elles [29] sont construites autour d'un algorithme de segmentation semi-automatique très répandu, GrabCut [30]. Cependant, la définition de la notion d'arrière plan est très souvent spécifique à l'application envisagée. Ici, une partie de notre jeu de données est composé d'objets possédant des couleurs proches de l'arrière-plan. Un exemple est visible sur la figure 5, à gauche. Ce cas de figure n'est pas le cas nominal des méthodes présentes dans littérature. Ainsi nous avons cherché une solution plus adaptée également basée sur l'algorithme GrabCut. Cette dernière utilise un modèle de mélange de Gaussiennes pour séparer l'arrière-plan et le premier plan à partir d'un masque d'initialisation. Ce masque associe à chaque pixel de l'image l'un des quatre états suivants : probablement à l'arrière-plan (PAP), certainement à l'arrière-plan (CAP), probablement au premier plan (PPP) ou certainement au premier plan (CPP). Dans la version semi-automatique, une opération humaine préalable est nécessaire pour afin d'indiquer quelles zones appartiennent à l'une au l'autre de ces catégories. Nous cherchons à rendre cet algorithme complètement automatique par la création d'un masque d'initialisation sans intervention humaine. Les méthodes existantes ne permettent pas de dissocier certaines zones ambiguës des objets de l'arrière plan. Afin d'améliorer cet aspect, de telles zones peuvent être initialement marquées comme PPP. Pour ce faire, différentes approches sont comparées en observant l'indice de Jaccard (ou *Intersection Over Union*, IoU) calculé à partir de la vérité terrain annotée manuellement.
Soient S_c les pixels en bordure d'image et m_c la médiane du canal valeur de l'image en HSV de S_c. L'hypothèse faite ici est que la majorité des pixels de S_c sont à l'arrière-plan. Cette hypothèse est basée sur le fait que les régions d'intérêt qui contiennent les objets ont été extraites dans l'image d'origine avec une marge adaptée à cette considération. La méthode usuelle (qui initialise les pixels n'appartenant pas à S_c à PPP et S_c à PAP), admet une variante pour laquelle les pixels centraux sont initialisés à CAP. Deux autres méthodes d'initialisation par diffusion de l'arrière plan et par contour actif on été testée. On ne détaille ici que cette dernière, qui a été retenue

Isolement du premier plan par contour actif. Un contour actif [32] est initialisé comme étant un cercle centré et entièrement inclus dans l'image. Dans notre application, les paramètres de cet algorithme ont été déterminés empiriquement. Les pixels restant dans le contour seront initialisés en PPP et le reste en PAP. La table 5 présente les résultats obtenus en comptant la proportion d'échantillons pour lesquels l'indice de Jaccard est supérieur à 0.5. Les histogrammes cumulés des résultats obtenus pour les méthodes principales peuvent être observés sur la figure 6.

Initialisation	IoU > 0.5 (%)
Référence	39.5
Référence initialisé en CAP	61.4
Diffusion	62.4
Contour actif	72.8

TABLE 5 – Comparaison des différentes méthodes de segmentation du premier plan.

Sur la base de ces résultats, on considérera pour la suite un masque de l'objet obtenu par GrabCut après initialisation à l'aide d'un modèle de contour actif. Cette solution permet d'obtenir le plus grand nombre de segmentations correctes (pour le critère IoU> 0.5), et le plus petit nombre de segmentations très mauvaises (aucun échantillon avec IoU< 0.20).

4.3 Désaturation de l'arrière-plan

Une fois le masque de segmentation obtenu, l'utilisation d'une carte de distance(qui associe à chaque pixel la distance au pixel de valeur différente le plus proche) comme

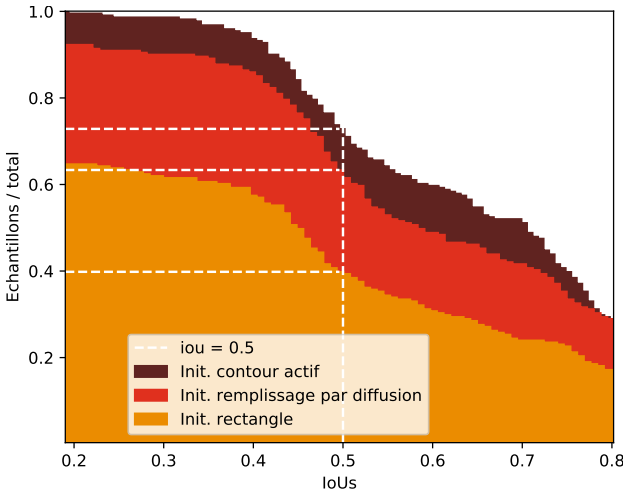


FIGURE 6 – Comparaison de méthodes d'initialisation du masque du GrabCut

pondération permet d'obtenir un masque de pondération. Ce dernier permet une transformation progressive de la composante de saturation des pixels dans l'espace HSV de manière à obtenir une image en niveaux de gris sur les zones d'arrière-plan, et une image colorée pour les zones appartenant à l'objet. La figure 7 illustre ce procédé et quelques exemples d'images transformées sont montrés sur la figure 10.

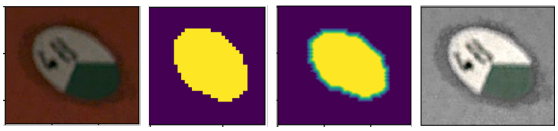


FIGURE 7 – Transformation, de gauche à droite : image d'origine, segmentation de l'arrière-plan, transformée de distance, image obtenue desaturation progressive de l'arrière-plan.

Afin d'évaluer quantitativement l'influence du prétraitement proposé, nous utilisons dans la section suivante un mesure de la complexité (ou difficulté) a priori des jeux de données pour analyser la méthode proposée.

5 Évaluation des transformations

5.1 Mesure de complexité du problème

Des cadres empiriques pour l'évaluation de la difficulté a priori d'un jeu de données sont accessibles dans la littérature. On parle de c-mesure. Il s'agit souvent de mesures faisant intervenir des systèmes de classification naïfs [23]. Nous utilisons ici une mesure de complexité proposée dans [24], basée sur la théorie du regroupement spectral.

5.2 Analyses quantitatives

Nous comparons dans un premier temps les effets des transformations effectuées sur les mesures de complexité a priori. Nous utilisons également un jeu de données sans transformation pour lequel les classes sont associées aux échantillons de manière aléatoire. La figure 8 montre les scores calculés par la méthode CSG.

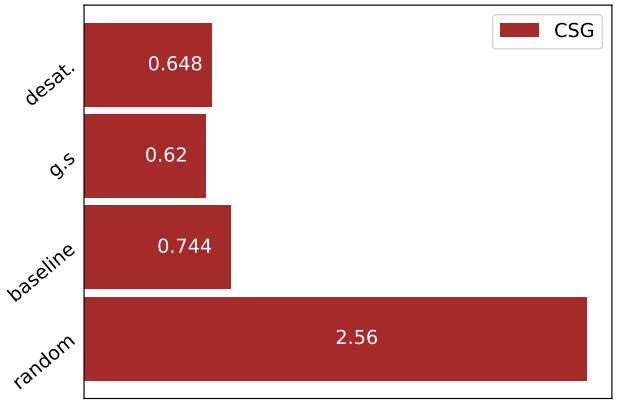


FIGURE 8 – Complexité a priori des jeux de données pour les différentes transformations selon la mesure CSG .

Le score utilisé donne une estimation de la complexité plus grande pour le cas d'étiquetage aléatoire (*random*) que pour les autres, ce qui était attendu. Parmi les autres scénarios, le plus complexe est alors le jeu de données non traité (*baseline*). Ce résultat était également attendu. Enfin l'écart pour les cas d'images en niveaux de gris (g.s pour *grayscale*) et traitées par désaturation progressive (*desat.*) est faible, mais les images en niveaux de gris sont considérées plus faciles.
Ces résultats indiquent un effet clair de la transformation sur la difficulté du problème. Nous pouvons donc nous attendre à une amélioration des performances après entraînement sur les données selon les différents traitements. Les scores présentés dans la table 6 montrent un impact positif significatif des transformations sur la qualité de l'attention (scores d'attention moyens sur les trois situations de généralisation).

Transformation	s_a
Aucune	0.57
Niveaux de gris	0.66
De-saturation progressive	0.74

TABLE 6 – Score d'attention s_a pour les différentes transformations.

Comme attendu, l'utilisation de la transformation proposée améliore les résultats en classification par rapport aux deux références, les résultats sont visibles dans la table 7. On note une amélioration des résultats en utilisant notre transformation, ce qui n'était pas prédit par la mesure de difficulté a priori.

5.3 Discussion

La méthode proposée apporte une réponse au problème de généralisation de domaine, mais présente plusieurs incon-

Source	Cible	Précision	Rappel	F1-score
D_1	D_2	0.97	0.86	0.91 (+13%)
$D_{1,2}$	$D_{1,2}$	0.89	0.97	0.93 (+ 2%)
D_2	D_1	0.96	0.96	0.96 (+ 11%)

TABLE 7 – Performances du réseau VGG-11 selon la composition des domaines source et cible, sur des images traitées par désaturation progressive.

vénients. En introduisant un traitement systématique impliquant une segmentation du premier plan, une source d'erreur potentielle est ajoutée, et les hyper paramètres doivent être réglés minutieusement. Ce prétraitement nécessite un temps de calcul supplémentaire qui peut s'avérer préjudiciable pour une application en temps réel. Ces limitations ont conduit à approfondir notre étude dans le but de remplacer la transformation systématique des images par un guidage de l'attention prenant place uniquement lors de la phase d'entraînement.

6 Apprentissage par cursus

L'apprentissage par cursus consiste à structurer le processus d'apprentissage en ordonnant les exemples par ordre de complexité croissante [25]. Le travail précédent permet d'entrevoir les gains potentiels d'une telle structuration dans notre cas. En particulier, nous avons observé que l'introduction de la couleur rend la tâche d'apprentissage plus difficile au regard de la vitesse de convergence des modèles. Ces observations sont en adéquation avec les scores de complexité a priori calculés dans les sections précédentes. Nous mettons en place une stratégie consistant à guider l'apprentissage en créant un cursus : les images sont d'abord montrées en niveaux de gris, puis transformées par désaturation progressive, et enfin sans traitement. La figure 9 présente l'évolution des performances sur la base de test (moyenne des trois situations D_1 vers D_2, D_2 vers D_1 et $D_{1,2}$ ver $D_{1,2}$). L'apprentissage par cursus est représenté par la courbe noire dont les marqueurs hexagonaux indiquent la transformation utilisée durant chacune des trois phases. L'effet bénéfique du cursus proposé est visible sur cette figure en observant la progression constante des performances sur l'ensemble test lorsque les autres méthodes atteignent un palier de stagnation. Les performances obtenues en fin d'apprentissage traduisent les effets des approches comparées dans cet article par rapport à la situation de référence (courbe orange) en particulier le gain apporté par notre stratégie d'apprentissage.

Source	Cible	Précision	Rappel	F1-score
D_1	D_2	0.96	0.87	0.90 (+12%)
$D_{1,2}$	$D_{1,2}$	0.89	0.97	0.94 (+ 3%)
D_2	D_1	0.96	0.97	0.95 (+ 10%)

TABLE 8 – Performances du réseau VGG-11 selon la composition des domaines source et cible, sur des images d'origine après entraînement par cursus.

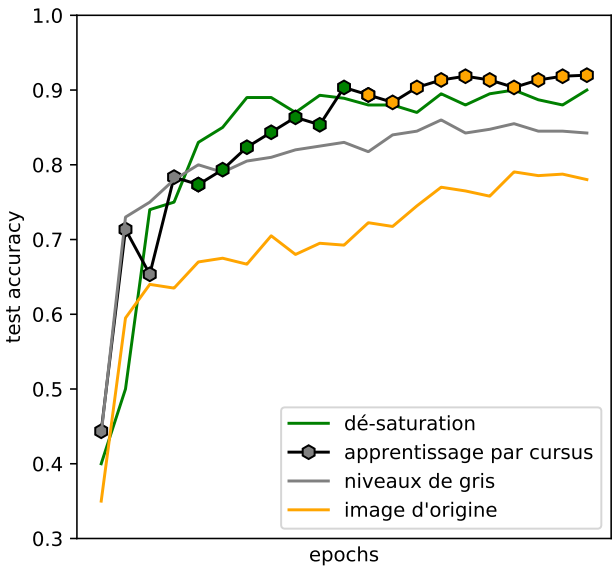


FIGURE 9 – Évaluation des performances en apprentissage au cours du temps pour les différentes méthodes.

La table 8 montre que les performances d'un réseau ayant été entraîné selon le cursus proposé sont équivalents à ceux obtenus en effectuant une transformation systématique des images par désaturation progressive. Cependant cette méthode d'apprentissage ne nécessite plus de prétraitement pour l'utilisation en conditions réelles. Restreindre ces opérations à une phase d'apprentissage est un gage de robustesse car l'erreur de généralisation des traitements eux-même à d'autres domaines n'intervient plus. Ainsi la segmentation du premier plan devient un simple outil d'automatisation de la création des masques d'attention qui sont utilisés lors de la transformation des images (deuxième phase du cursus d'apprentissage indiquée par des marqueurs verts sur la figure 9).

7 Complexité et généralisation

Nous observons dans cette section la complexité des réseaux entrainés lors des différentes expériences. Nous avons également considéré un réseau entraîné sur les images d'origine après avoir aléatoirement attribué une classe aux échantillons. Nous utilisons l'expression de la complexité spectrale définie dans l'équation (1). Pour les expériences, nous avons utilisé un réseau VGG-11, dont les non-linéarités sont uniquement des fonctions *ReLU* et *MaxPool* 1-Lipschitziennes. En prenant $M = 0$, cette équation se simplifie ainsi en :

$$\mathcal{R}_{\mathcal{A}} := \left(\prod_{i=1} \|A_i\|_\sigma \right) \left(\sum_{i=1}^{L} \frac{\|A_i^T\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}} \right)^{3/2} \qquad (2)$$

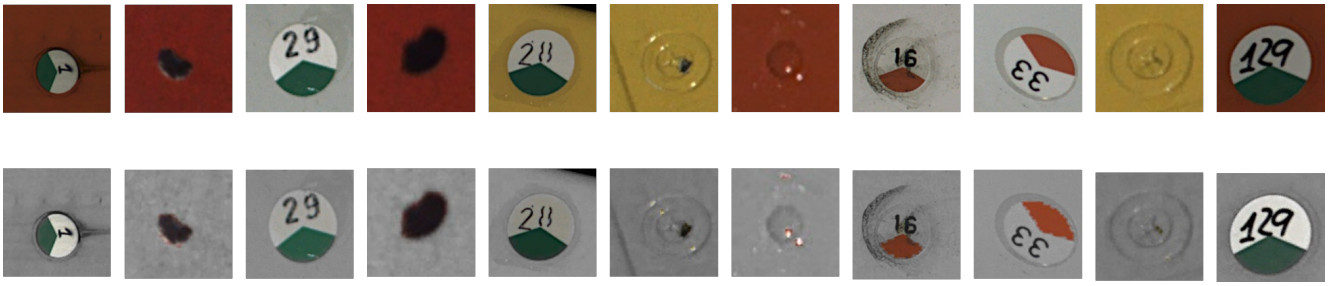


FIGURE 10 – Effet de la désaturation progressive pour un panel d'échantillons : sur la première ligne des images acquises dans le domaine D_1 ou D_2, sur la seconde les mêmes images après transformation.

On observe alors les fonctions de répartition des marges normalisées, données par :

$$(x, y) \rightarrow \frac{\mathcal{F}_{\mathcal{A}}(x)_y - \mathcal{F}_{\mathcal{A}}(x)_i i \neq y}{\mathcal{R}_{\mathcal{A}} \|X\|_2 / n}$$

où x est l'entrée (image) et y le label attendu. Le résultat obtenu est exposé sur la figure 11.

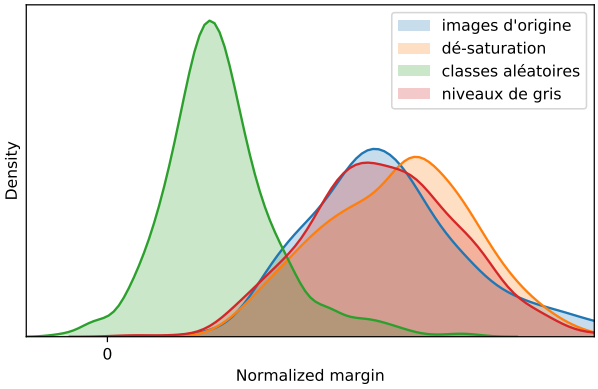


FIGURE 11 – Distributions des marges après normalisation spectrale.

La répartition des marges permet bien de séparer le cas le plus difficile (classes aléatoires), des autres. Le cas des images traitées par notre méthode apparaît comme plus simple, ce qui est cohérent avec les résultats en classification. Néanmoins les distributions pour les images originales et en niveaux de gris n'apparaissent pas clairement dissociables alors que les observations sur la difficulté a priori des jeux de données, sur l'attention et sur la généralisation des réseaux correspondants indiquent une différence importante. Ainsi, la notion de complexité spectrale censée capturer implicitement la difficulté intrinsèque du jeu de données ne permet pas d'expliquer complètement certains phénomènes plus visibles en utilisant des estimations directes.

8 Conclusions et perspectives

Nous avons mis en évidence le fait qu'une transformation bien choisie des images peut être bénéfique sur la complexité a priori de la tâche de classification, sur l'attention et sur les performances des systèmes d'apprentissage automatique. Nous avons pour cela proposé une mesure de la qualité de l'attention. Nous avons également observé une mesure de la complexité du problème issue de la littérature. Cette mesure explique en partie l'effet positif des prétraitements appliqués systématiquement et a servi à l'élaboration d'une stratégie d'apprentissage par échantillons de difficulté croissante. En effet, afin de s'affranchir de pré-traitements systématiques nous les avons incorporés à la phase d'apprentissage qui se fait alors par cursus. Les bornes de généralisation existantes prenant en compte la complexité du jeu de donnée de manière indirecte n'expliquent que partiellement les résultats obtenus lors de nos tests.

La démarche conduite ici est aisément transférable à d'autres domaines d'applications : le choix de la méthode de segmentation est indépendant du reste des traitements, et la désaturation progressive introduite peut être remplacée par tout autre transformation qui permet une meilleure focalisation de l'attention.

Plusieurs perspectives sont ouvertes par ces travaux et pourront faire l'objet d'approfondissements. Le choix des classes, qui ne sont en général pas mutuellement exclusives pour les applications dans le monde réel, est un facteur important dans les mesures de complexité des jeux de données. La pratique consistant à considérer un ensemble étiqueté aléatoirement comme référence d'un problème plus complexe est d'ailleurs courante. Une meilleure répartition des éléments entre les classes pour un problème de classification supervisé est une piste d'approfondissement envisagée. Un autre axe majeur pourra s'articuler autour du lien entre mesure de complexité du jeu de données et la complexité du modèle obtenu sur cet ensemble, dans l'optique d'approfondir la compréhension de la généralisation des modèles. Enfin, la mesure de complexité (CSG) utilisée n'explique pas totalement les résultats obtenus, en particulier le meilleur taux de réussite lors de l'utilisation de la désaturation progressive alors que la complexité du jeu de données est estimée plus grande que celle des images en niveaux de gris. Les travaux exploratoires sur une nouvelle mesure de complexité faisant intervenir la variance expliquée par une analyse en composantes principales sont prometteurs.

Références

[1] D. Oppenheim and G. Shani, "Potato disease classification using convolution neural networks," *Advances in Animal Biosciences*, vol. 8, no. 2, pp. 244–249, 2017.

[2] A. Luckow, M. Cook, N. Ashcraft, E. Weill, E. Djerekarov, and B. Vorster, "Deep learning in the automotive industry : Applications and tools," in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3759–3768, IEEE, 2016.

[3] M. Siegel, P. Gunatilake, and G. Podnar, "Robotic assistants for aircraft inspectors," *IEEE Instrumentation Measurement Magazine*, vol. 1, no. 1, pp. 16–30, 1998.

[4] I. Jovančević, S. Larnier, J.-J. Orteu, and T. Sentenac, "Automated exterior inspection of an aircraft with a pan-tilt-zoom camera mounted on a mobile robot," *Journal of Electronic Imaging*, vol. 24, no. 6, p. 061110, 2015.

[5] M. Rice, L. Li, G. Ying, M. Wan, E. T. Lim, G. Feng, J. Ng, M. Nicole, T. Jin-Li, and V. S. Babu, "Automating the visual inspection of aircraft," in *Aerospace Technology and Engineering Conference*, 2018.

[6] J. Miranda, S. Larnier, and M. Claybrough, "Caractérisation d'objets sur des images acquises par drone," in *Conférence Reconnaissance des Formes, Image, Apprentissage et Perception*, 2018.

[7] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[8] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion : Maximizing for domain invariance," *arXiv preprint arXiv :1412.3474*, 2014.

[9] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Advances in neural information processing systems*, pp. 136–144, 2016.

[10] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1446–1455, 2019.

[11] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*, pp. 10–18, 2013.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam : Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[13] X. Wang, L. Li, W. Ye, M. Long, and J. Wang, "Transferable attention for domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5345–5352, 2019.

[14] S. Zagoruyko and N. Komodakis, "Paying more attention to attention : Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv :1612.03928*, 2016.

[15] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together : A generalized framework for domain adaptation," *IEEE Transactions on Cybernetics*, vol. 49, no. 6, pp. 2144–2155, 2018.

[18] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks.," *Journal of Machine Learning Research*, vol. 20, no. 63, pp. 1–17, 2019.

[19] S. Sun, W. Chen, L. Wang, X. Liu, and T.-Y. Liu, "On the depth of deep neural networks : A theoretical view," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[20] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

[21] Z. Wang, "Theoretical guarantees of transfer learning," *arXiv preprint arXiv :1810.05986*, 2018.

[22] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning bounds for domain adaptation," in *Advances in neural information processing systems*, pp. 129–136, 2008.

[23] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 289–300, 2002.

[24] F. Branchaud-Charron, A. Achkar, and P.-M. Jodoin, "Spectral metric for dataset complexity assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3215–3224, 2019.

[25] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

[26] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems 23* (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds.), pp. 1189–1197, Curran Associates, Inc., 2010.

[27] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 1171–1179, Curran Associates, Inc., 2015.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv :1409.1556*, 2014.

[29] S. Prakash, R. Abhilash, and S. Das, "Snakecut : An integrated approach based on active contour and grabcut for automatic foreground object segmentation," *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 6, p. 13, 07 2007.

[30] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut" : interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, pp. 309–314, 2004.

[31] S. Dikshit, J. Raghav, G. Shrivastava, and K. Sharma, "Graphic system based on flood fill algorithm with images," 11 2012.

[32] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes : Active contour models," *IEEE Proc, on Computer Vision and Pattern Recognition*, vol. 1, pp. 321–331, 01 1988.