Le manchot, la banane et la bibliothèque. . . (de la désambiguïsation d'une tâche de classification avec un exemple)

Yassir Bendou, Lucas Drumetz, Vincent Gripon, Giulia Lioi, Bastien Pasdeloup

▶ **To cite this version:**

Yassir Bendou, Lucas Drumetz, Vincent Gripon, Giulia Lioi, Bastien Pasdeloup. Le manchot, la banane et la bibliothèque. . . (de la désambiguïsation d'une tâche de classification avec un exemple). GRETSI 2022 : 28ème colloque du Groupement de Recherche en Traitement du Signal et des Images, Sep 2022, Nancy, France. hal-03694660

HAL Id: hal-03694660
https://hal.science/hal-03694660

Submitted on 13 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Le manchot, la banane et la bibliothèque…
(de la désambiguïsation d'une tâche de classification avec un exemple)

Yassir BENDOU, Lucas DRUMETZ, Vincent GRIPON, Giulia LIOI, Bastien PASDELOUP

IMT Atlantique
Lab-STICC, UMR CNRS 6285
Brest F-29238, France
prenom.nom@imt-atlantique.fr

Résumé – L'apprentissage à partir de peu d'exemples vise à exploiter les connaissances acquises par un modèle d'apprentissage profond pré-entraîné afin de reconnaître de nouvelles classes inconnues quand peu d'exemples sont étiquetés. Travailler avec un seul exemple est particulièrement difficile, même pour le modèle le mieux pré-entraîné, en raison de la possible présence d'autres objets dans une scène. Dans cet article, nous identifions d'abord le problème, puis proposons une méthodologie pour désambiguïser une image en y identifiant différents objets. Pour cela, nous avons recourt à l'augmentation de données, ainsi qu'à un processus d'optimisation. Nous combinons notre méthode avec une stratégie de classification simple qui permet d'améliorer l'état de l'art du domaine sur le jeu de données standard Mini-ImageNet.

Abstract – Few-shot learning aims at leveraring knowledge learned by a deep learning model in order to recognize unseen classes during training with limited labeled examples. Working with one example is particularly challenging as multiple objects in a scene can be ambiguous even for the best trained model. In this paper, we first identify the problem of having multiple objects in an image and propose a methodology to disambiguate them using data augmentation and an optimization process. We combine our method with a simple classification strategy which helps improve the state-of-the-art on the standard Mini-ImageNet dataset.

1 Introduction

L'apprentissage avec peu d'exemples (APE), est un domaine de recherche qui connaît une popularité croissante, en particulier en vision par ordinateur. Concilier les performances remarquables de l'apprentissage profond, généralement obtenues grâce à l'accès à d'immenses bases de données, avec la contrainte de disposer d'un très petit nombre d'exemples peut sembler paradoxal. Pourtant, la réponse réside dans la capacité de l'apprentissage profond à transférer les connaissances acquises lors de la résolution d'une tâche précédente vers une nouvelle tâche, potentiellement différente.

La disposition classique de l'apprentissage à partir de peu d'exemples se compose de deux parties :

— Un jeu de données *générique*, contenant de nombreuses images et classes, qui peut être utilisé pour entraîner efficacement les réseaux profonds ;
— Un jeu de données *nouveau*, contenant des classes distinctes du jeu de données générique. Chaque classe ne contient que peu d'exemples étiquetés, appelés *support*. Les autres exemples forment le jeu de données *requête*.

Un problème d'APE consiste alors à classifier des exemples de l'ensemble requête, étant donné un ensemble support de 1 ou 5 exemples par classe. Les méthodes d'APE sont comparées en moyennant les performances sur un grand nombre de problèmes générés aléatoirement à partir de jeux de données standardisés.

Afin de transférer les connaissances du jeu de données générique vers le nouveau, une approche classique consiste à entraîner un réseau profond sur le premier, et à l'utiliser ensuite comme extracteur de caractéristiques pour le second. Cette approche facilite souvent la classification des données, les caractéristiques ayant de bonnes propriétés de séparabilité.

FIGURE 1 – Exemple d'image étiquetée comme "bibliothèque".

Toutefois, même en disposant des meilleures caractéristiques possibles, les problèmes d'APE peuvent être intrinsèquement ambigus, par exemple si une image contient plusieurs objets d'intérêt. Dans cet article, nous cherchons à montrer qu'il est parfois possible de désambiguïser ces problèmes en étant attentif à la distribution des données dans l'espace des caractéristiques. Cette approche est motivée par l'image de la Figure 1, représentant une "bibliothèque" dans le jeu de données de Mini-ImageNet. La présence de cet exemple dans un problème d'APE fait typiquement chuter la performance de plusieurs %. En effet, le modèle n'est pas entraîné à reconnaître les objets saillants de la Figure 1. La "bibliothèque" étant en arrière-plan, le vecteur caractéristique de cette image décrit principalement le

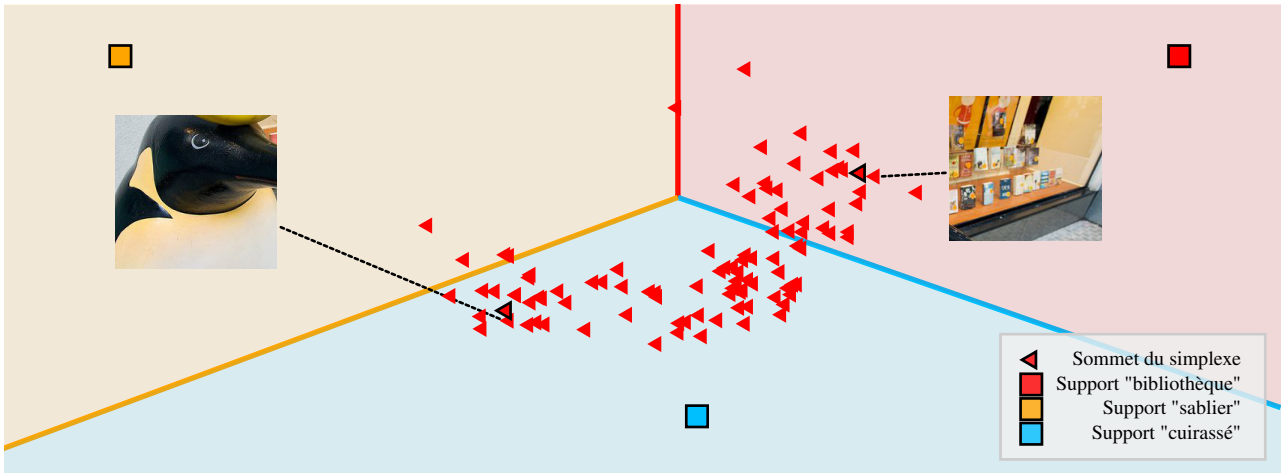

FIGURE 2 – Projection dans l'espace formé par les axes entre les centroïdes des classes "bibliothèque", "cuirassé" et "sablier". Les carrés et zones associées sont les centroïdes et cellules de Voronoï obtenus à partir des exemples du support. Chacun des 200 triangles représente le vecteur de caractéristiques d'un recadrage aléatoire de l'image en Figure 1. En utilisant l'optimisation décrite dans la partie 3.2, on obtient deux sommets, ici entourés de noir, dont nous visualisons le recadrage le plus proche.

"manchot" et la "banane" qui sont au premier-plan. Une conséquence est que cette image est éloignée des autres images de la classe "bibliothèque" dans l'espace des caractéristiques.

Dans cet article, nous proposons une méthode simple pour désambiguïser les objets présents dans les images en modélisant la distribution formée par les vecteurs caractéristiques de différentes régions de l'image sous forme de simplexe. Cette modélisation permet d'extraire différents sommets que l'on peut identifier à différents objets dans une image. Ensuite, nous exploitons ces sommets afin d'améliorer les performances sur des tâches de classification dans le cadre de l'APE.

2 État de l'art

Plusieurs familles de méthodes existent afin de résoudre des problèmes d'APE. On distingue notamment celles basées sur les distances [1, 2], le meta-apprentissage [3], ou le transfert de connaissance à partir d'un réseau pré-entraîné [4]. Cette dernière approche est généralement préférée, car apportant des meilleures performances tout en étant relativement simple.

L'entraînement de l'extracteur de caractéristiques exploite généralement de l'augmentation de données, en créant différentes versions des images du jeu de données générique par des rotations ou recadrages aléatoires [5]. L'augmentation du nouveau jeu de données, bien que moins fréquente, a aussi été explorée. On peut par exemple calculer le transport optimal entre différents recadrages des images de l'ensemble du support et de l'image requête afin de ré-estimer les distances entre images [6], utiliser différents recadrages des images afin de distinguer les objets d'intérêt des arrières-plans [7], ou moyenner les caractéristiques obtenus à partir de recadrages de l'image afin de réduire l'influence d'objets parasites [8].

3 Méthodologie

Afin de désambiguïser un problème d'APE, il est important de détecter la présence de plusieurs objets d'intérêt dans une image. Pour ce faire, nous proposons de générer plusieurs recadrages aléatoires et de générer pour chacun d'entre eux un vecteur caractéristique. Notre hypothèse est que ces vecteurs vont s'inscrire naturellement dans un simplexe, dont les sommets sont des objets d'intérêt.

3.1 L'exemple de la bibliothèque

Nous représentons en Figure 2 la projection, dans l'espace formé par les axes entre les centroïdes des classes "bibliothèque", "cuirassé" et "sablier", de vecteurs caractéristiques obtenus à partir de recadrages aléatoires de l'exemple en Figure 1. On observe que ces vecteurs semblent se distribuer le long d'une courbe tracée dans le plan des trois axes, dont les extrémités correspondent à différents objets dans la scène : le manchot et la banane d'un côté et la bibliothèque de l'autre.

3.2 Extraction des sommets

Nous modélisons la distribution de chaque image dans l'espace des caractéristiques sous forme d'un simplexe à K sommets, avec K supposé fixé pour le moment. Nous verrons dans la partie 3.3 comment identifier le bon nombre de sommets pour chaque image. Nous notons N le nombre de recadrages de chaque image. Soit D la dimension d'un vecteur de caractéristiques (typiquement quelques centaines [8]). On définit $\mathbf{X} \in \mathcal{M}_{N\times D}(\mathbb{R})$ la matrice contenant tous les vecteurs caractéristiques de l'image. Soit $\mathbf{D} \in \mathcal{M}_{K\times D}(\mathbb{R})$ la matrice contenant les sommets du simplexe et $\mathbf{W} \in \mathcal{M}_{N\times K}(\mathbb{R})$ la matrice qui contient la pondération de chaque sommet à chaque vecteur. On

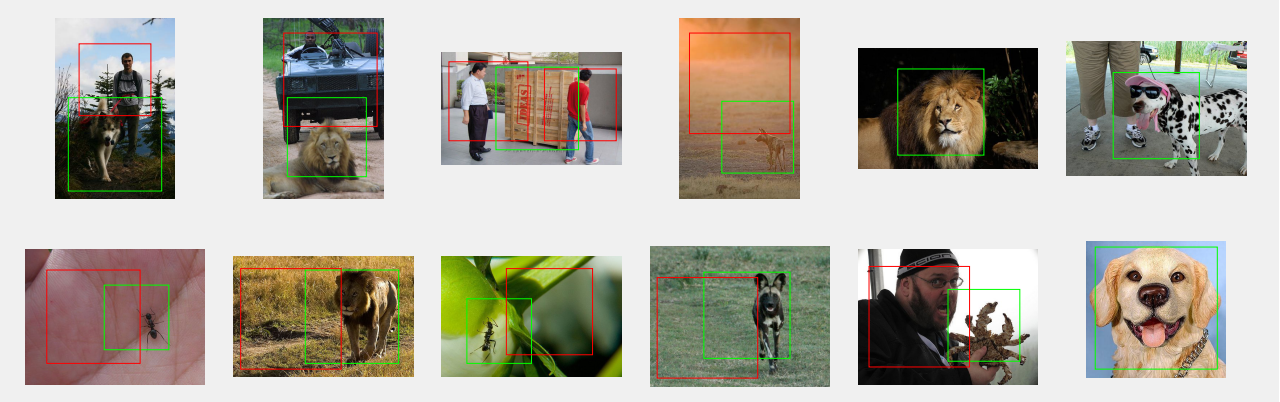

FIGURE 3 – Résultats de la méthode du simplexe sur différentes images de Mini-ImageNet. Les cadres représentent les recadrages les plus proches des sommets des simplexes obtenus. Nous utilisons la couleur verte pour les recadrages les plus adaptés aux étiquettes des classes correspondantes.

note $\mathbf{1}_N$ le vecteur de dimension N rempli de 1.

La première partie de notre problème d'optimisation consiste à reconstruire nos points à partir des K sommets non connus. Ceci revient à résoudre l'équation suivante :

$$\mathbf{D}^*, \mathbf{W}^* = \underset{\mathbf{D},\mathbf{W}}{\arg\min} \quad \|\mathbf{WD} - \mathbf{X}\|_2^2 \,. \qquad (1)$$
$$\text{s.t.} \begin{cases} 0 \leq \mathbf{W} \\ \mathbf{W}\mathbf{1}_K = \mathbf{1}_N \end{cases}$$

Cette forme est très courante en séparation de sources [9, 10], et peut être résolue en alternant entre l'estimation de $\mathbf{D}$ et de $\mathbf{W}$. Un tel problème a une infinité de solutions (il suffit de prendre des sommets assez éloignés les uns des autres pour englober parfaitement les données), ce qui le rend mal posé. Parmi les solutions potentielles, nous nous intéressons à celles dont les sommets sont relativement proches des données. En effet, ceux-ci devraient correspondre approximativement à des recadrages possibles des images d'entrée. Il convient donc d'ajouter un terme de régularisation visant à limiter l'éparpillement des sommets dans l'espace des caractéristiques. Parmi les régularisations possibles [10], certaines méthodes cherchent à minimiser le volume du simplexe. Un tel objectif étant complexe à atteindre, une bonne méthode approchée revient à minimiser le périmètre du simplexe, en minimisant la somme des distances entre ses sommets, comme suit :

$$\mathbf{D}^*, \mathbf{W}^* = \underset{\mathbf{D},\mathbf{W}}{\arg\min} \quad \left(\begin{array}{l} \lambda \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \|\mathbf{D}_k - \mathbf{D}_{k'}\|_2^2 \\ +(1-\lambda)\|\mathbf{WD} - \mathbf{X}\|_2^2 \end{array} \right) , \quad (2)$$
$$\text{s.t.} \begin{cases} 0 \leq \mathbf{W} \\ \mathbf{W}\mathbf{1}_K = \mathbf{1}_N \end{cases}$$

où $\mathbf{D}_k$ correspond à la k-ième colonne de $\mathbf{D}$, et $\lambda \in [0, 1]$ est un hyperparamètre de régularisation fixé à 0.05. Le problème est convexe en $\mathbf{D}$ et en $\mathbf{W}$, mais pas simultanément en les deux variables. Afin de résoudre ce problème d'optimisation, nous procédons par alternance de minimisations. Nous utilisons la méthode proposée par [11] qui est une forme close pour la résolution de $\mathbf{D}$. Afin de garantir les contraintes du problème, la résolution de $\mathbf{W}$ se fait par descente de gradient sur une matrice $\mathbf{V} \in \mathcal{M}_{N\times K}(\mathbb{R})$ telle que $\forall k : \mathbf{W}_k = \text{softmax}(\mathbf{V}_k) = \frac{\exp(\mathbf{V}_k)}{\sum_{k'=1}^{K} \exp(\mathbf{V}_{k'})}$. Par ailleurs, l'initialisation des sommets de la matrice $\mathbf{D}$ se fait en choisissant K vecteurs aléatoires de la matrice $\mathbf{X}$. Les entrées de $\mathbf{V}$ sont initialisées uniformément aléatoirement selon une loi $\mathcal{N}(0, 1)$. La solution est un point stationnaire de la fonction de coût sans garantie d'optimalité.

3.3 Choix du nombre de sommets

Jusqu'à présent, nous avons considéré K connu et fixé. Toutefois, le nombre d'objets d'intérêt dans une image n'est pas une information connue en pratique. Dans cette partie, nous proposons une méthode automatique afin d'identifier ce nombre. Afin de simplifier l'analyse, nous exploitons le fait qu'une image naturelle contient souvent peu d'objets d'intérêt ($K \leq 3$). Ceci se confirme empiriquement en visualisant la répartition de recadrages aléatoires de différentes images dans l'espace des caractéristiques. Ensuite, nous exploitons l'erreur de reconstrution de l'équation 1 afin d'établir un critère de sélection du nombre de sommets, similaire à celui utilisé dans la méthode du coude permettant la sélection du nombre de centroïdes dans l'algorithme des K-moyennes. L'idée est que si l'erreur de reconstruction ne diminue pas significativement (seuil de 1.5) quand on augmente le nombre de sommets, alors le nombre précédent est retenu. Cette méthode est appliquée sur l'ensemble des données afin d'identifier le nombre d'objets présents dans chaque image.

3.4 Classification

Une fois les sommets identifiés pour toutes nos images, nous régularisons ceux-ci afin de réduire leur dispersion dans l'espace. Pour cela, nous interpolons linéairement les sommets (0.75) avec le vecteur caractéristique moyen de tous les recadrages de cette image (0.25). Ces coefficients sont fixés par recherche d'hyperparamètres. La classification d'une image requête se fait ensuite en identifiant son sommet le plus proche (métrique euclidienne) de l'un des sommets des images du support.

4 Résultats

Afin d'évaluer notre méthode, nous utilisons le jeu de données standard Mini-ImageNet. Nous étudions le cadre usuel en APE qui considère un nouveau jeu de données contenant 5 classes, avec 1 exemple étiqueté et 15 requêtes par classe. Nous reportons nos résultats pour 10^5 problèmes d'APE générés aléatoirement à partir de Mini-ImageNet. La Table 1 compare les résultats de notre méthode [1] avec ceux de l'état de l'art.

TABLE 1 – État de l'art des méthodes proposées sur **Mini-ImageNet** avec un seul exemple étiqueté par classe.

Method	Performance (%)
Deep EMD v2 [6]	68.77 ± 0.29
PAL [12]	69.37 ± 0.64
invariance-equivariance [13]	67.28 ± 0.80
CSEI [14]	68.94 ± 0.28
COSOC [7]	69.28 ± 0.49
ASY [8]	70.77 ± 0.06
Notre méthode	$\mathbf{70.90 \pm 0.06}$

En complément de la Figure 2, nous visualisons en Figure 3 le plus proche recadrage des sommets obtenus pour quelques images du jeu de données. On observe que les sommets sont proches d'objets d'intérêt, dont souvent un correspond à l'objet associé à l'étiquette de l'image. Les autres sommets correspondent quant à eux à des objets non présents dans les étiquettes du jeu de données. De fait, lorsque une image requête est comparée à une image support, il est possible de désambiguïser le problème posé en identifiant les sommets les plus proches.

Une autre façon d'exploiter les sommets extraits par notre méthode consiste à nettoyer notre jeu données. En effet, les images contenant plus d'un objet vont agir en facteur de confusion et impactent la performance de classification. La Figure 4 présente la performance obtenue, quand on considère uniquement les exemples pour lesquels $K = 1$ est optimal, en fonction du paramètre de régularisation λ. Lorsque ce dernier est haut, on retombe sur le cas classique avec toutes les données.

5 Conclusion

Dans cet article, nous avons observé que désambiguïser les données, en identifiant les objets d'intérêt, permet une augmentation des performances de classification en APE. Ces résultats sont encourageants, et ouvrent la porte à une amélioration supplémentaire quand on considérera des problèmes à plusieurs données étiquetées par classe, car une approche similaire devrait permettre de mieux identifier l'objet commun entre plusieurs exemples d'une même classe.

1. Le code permettant de reproduire les résultats de nos expériences est disponible dans le lien suivant : `https://github.com/ybendou/banana-penguin`.

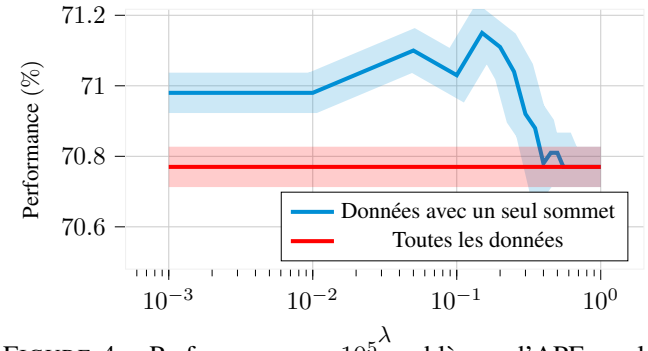

FIGURE 4 – Performance sur 10^5 problèmes d'APE sur les images de Mini-ImageNet ayant un seul sommet, en fonction de la régularisation du simplexe.

Références

[1] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," 2017. [Online]. Available : http://arxiv.org/abs/1703.05175

[2] O. Vinyals, C. Blundell, T. Lillicrap, koray Kavukcuoglu, and D. Wierstra, in *Advances in Neural Information Processing Systems*, 2016.

[3] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," 2017.

[4] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, "Simpleshot : Revisiting nearest-neighbor classification for few-shot learning," 2019. [Online]. Available : http://arxiv.org/abs/1911.04623

[5] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, and V. N. Balasubramanian, "Charting the right manifold : Manifold mixup for few-shot learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

[6] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd : Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[7] X. Luo, L. Wei, L. Wen, J. Yang, L. Xie, Z. Xu, and Q. Tian, "Rectifying the shortcut learning of background for few-shot learning," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[8] Y. Bendou, Y. Hu, R. Lafargue, G. Lioi, B. Pasdeloup, S. Pateux, and V. Gripon, "Easy : Ensemble augmented-shot y-shaped learning : State-of-the-art few-shot classification with simple ingredients," 2022. [Online]. Available : http://arxiv.org/abs/2201.09699

[9] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview : Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2012.

[10] L. Drumetz, J. Chanussot, C. Jutten, W.-K. Ma, and A. Iwasaki, "Spectral variability aware blind hyperspectral image unmixing based on convex geometry," *IEEE Transactions on Image Processing*, 2020.

[11] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J. Huntington, "Ice : A statistical approach to identifying endmembers in hyperspectral images," in *IEEE Transactions on Geoscience and Remote Sensing*, 2004.

[12] J. Ma, H. Xie, G. Han, S.-F. Chang, A. Galstyan, and W. Abd-Almageed, "Partner-assisted learning for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[13] M. N. Rizve, S. Khan, F. S. Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[14] J. Li, Z. Wang, and X. Hu, "Learning intact features by erasing-inpainting for few-shot classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.