Un modèle statistique pour prédire la généralisation dans des tâches de classification avec peu d'exemples

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

IMT Atlantique
Lab-STICC, UMR CNRS 6285
Brest F-29238, France

Résumé – L'estimation de l'erreur de généralisation des classificateurs est difficile dans les scénarios d'apprentissage avec peu d'exemples du fait que l'ensemble de validation est souvent indisponible. Dans ce type de problèmes, il est typique d'utiliser des réseaux de neurones pré-entraînés avec des classificateurs basés sur la distance tels que le centre de la classe le plus proche. Dans cet article, nous proposons un modèle Gaussien pour estimer les paramètres de la distribution des vecteurs caractéristiques et prédire l'erreur de généralisation sur de nouvelles tâches de classification avec peu d'exemples étiquetés. Nous introduisons également un estimateur sans biais pour les distances entre les densités conditionnelles des classes et montrons l'importance d'une bonne estimation de ces distances. Nos expériences montrent que notre approche est plus performante que d'autres solutions, telles que la stratégie de validation croisée.

Abstract – The estimation of the generalization error of classifiers often relies on a validation set. Such a set is hardly available in few-shot learning scenarios, a highly disregarded shortcoming in the field. In these scenarios, it is common to rely on features extracted from pre-trained neural networks combined with distance-based classifiers such as nearest class mean. In this work, we introduce a Gaussian model of the feature distribution. By estimating the parameters of this model, we are able to predict the generalization error on new classification tasks with few samples. We observe that accurate distance estimates between class-conditional densities are the key to accurate estimates of the generalization performance. Therefore, we propose an unbiased estimator for these distances and integrate it in our numerical analysis. We empirically show that our approach outperforms alternatives such as the leave-one-out cross-validation strategy.

1 Introduction

Le problème d'apprentissage à partir de peu d'exemples (APE), où le nombre d'échantillons d'entraînement par classe est faible (généralement moins de dix [13]), a connu récemment un grand nombre de contributions [14]. La plupart des solutions à l'état de l'art consistent à utiliser un extracteur de caractéristiques à base de réseaux de neurones profonds pré-entraînés pour projeter les échantillons dans un espace latent où les classes sont censées être plus faciles à discriminer, suivi d'un classificateur à base de distances tel que le centre de la classe le plus proche [2]. Toutefois, il n'est pas simple de mesurer les performances d'un tel classificateur quand peu d'exemples sont disponibles. L'approche classique consiste à effectuer une validation croisée, où un échantillon est arbitrairement retiré de l'ensemble d'apprentissage pour être utilisé comme exemple de validation ; ce processus est répété plusieurs fois pour obtenir en moyenne la précision estimée.

Plusieurs alternatives à la validation croisée ont été proposées dans la littérature dans un contexte de classification standard où un grand nombre d'échantillons est disponible [10]. Dans ce travail, nous nous intéressons principalement à proposer une alternative à la validation croisée et aux méthodes de prédiction de la généralisation existantes dans le contexte de l'APE. La méthode proposée utilise un modèle statistique des densités conditionnelles des classes dans l'espace latent pour estimer la probabilité d'erreur.

Deux difficultés majeures se posent ici : 1) nous devons trouver un modèle qui soit suffisamment expressif pour capturer la distribution des données tout en dépendant de peu de paramètres pour les estimer avec précision dans un régime avec peu de données. 2) la probabilité d'erreur dépend des distances entre les centres de chaque classe. Nous observons que l'estimateur naïf des distances est biaisée, ce qui conduit à sous-estimer la probabilité d'erreur, notamment lorsqu'on travaille avec des espaces à haute dimension et avec peu d'échantillons.

Les principales contributions de notre travail sont les suivantes : 1) Nous introduisons un modèle statistique des densités conditionnelles des classes dans l'espace latent et proposons un estimateur sans biais pour les distances entre les centres des classes. 2) Nous démontrons que notre méthode est plus performante que d'autres alternatives sur des bancs d'essai d'APE standardisés.

2 Etat de l'art

2.1 Classification avec peu d'exemples

Soit $\mathcal{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{\ell}$ un petit ensemble de données d'entraînement où $\forall j, (\mathbf{x}_j, y_j) \sim p_{\mathcal{X},\mathcal{Y}}$. L'objectif de la classification avec peu d'exemples est d'entraîner un classificateur C en utilisant $\mathcal{D}$. Étant donné qu'il y a peu d'échantillons d'entraînement dans les tâches d'APE, la plupart des méthodes consistent à pré-entraîner un extracteur de caractéristiques profond f_θ avec l'ensemble de paramètres θ sur un grand jeu de données générique. Cet extracteur de caractéristiques est ensuite adapté ou utilisé tel quel sur $\mathcal{D}$ pour produire des vecteurs caractéristiques $\mathbf{z} = f_\theta(\mathbf{x})$ dans un espace euclidien. Le

classificateur fonctionne ensuite avec la variante de l'ensemble de données $f_\theta(\mathcal{D}) = \{(\mathbf{z}_j, y_j)\}_{j=1}^{\ell}$.

Il existe de nombreuses stratégies d'entraînement de f_θ que l'on peut classer en deux catégories : les approches à base d'optimisation comme le méta-apprentissage [8] et les approches à base de transfer de connaissance qui visent à apprendre un bon extracteur de caractéristiques [14]. Cette dernière catégorie a connu un grand succès récemment en raison de ses performances compétitives tout en étant relativement simple à mettre en œuvre [2].

Différents classificateurs ont été proposés dans la littérature de l'APE, tels que l'utilisation d'un perceptron à multicouches [6] ou des classificateurs à base de distance tel que le centre de la classe le plus proche (CCP) [14]. Nous nous en tenons à l'approche à base de CCP en raison de ses performances bien reconnues et de sa simplicité.

Le classificateur à base de centre de classe le plus proche C_{CCP} est le classificateur optimal lorsque les densités conditionnelles des classes suivent une distribution gaussienne isotrope égale avec un à priori uniforme entre les classes.

$$p(\mathbf{z} \mid y = c) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_c, \sigma^2 \mathbf{I}) \ , \tag{1}$$

où $\boldsymbol{\mu}_c$ est le centre de la classe $c \in \mathcal{Y}$, σ est l'écart-type. La classification d'un nouvel exemple $\mathbf{z}$ se fait tel que :

$$C_{\text{NCM}}(\mathbf{z}) = \arg\min_{c \in \mathcal{Y}} \|\mathbf{z} - \boldsymbol{\mu}_c\|_2 \ . \tag{2}$$

En pratique les centres des classes sont les moyennes empiriques estimées à partir de données d'entraînement $\mathcal{D}$.

Une fois les centres de classe estimés, il y a au plus $(n-1)$ dimensions d'intérêt dans l'espace euclidien considéré, qui correspondent aux directions entre les centres de classes, où n est le nombre de classes. Les dimensions restantes peuvent être ignorées car elles produisent des contributions orthogonales aux axes entre les centres de classe. Ainsi, la projection sur un tel sous-espace de dimension $(n-1)$ n'a pas d'incidence sur les frontières de décision d'un classificateur CCP.

2.2 Prédiction de la géneralization

De nombreux travaux ont été proposés sur la généralisation des réseaux de neurones entraînés sur de grands jeux de données. Les méthodes proposées dans la littérature peuvent être résumées en quelques familles différentes. La première est celle des méthodes PAC-Bayes, où le comportement de généralisation d'un modèle est décrit par des bornes approximativement correctes[7]. Ces méthodes fournissent souvent une borne supérieure à l'erreur de généralisation et les résultats sont souvent limités à un petit ensemble de modèles. La deuxième famille est celle des méthodes basées sur les normes, qui analysent les poids des réseaux. Ces méthodes se sont révélées peu performantes [10]. La dernière famille de méthodes vise à analyser la représentation intermédiaire des données d'apprentissage dans l'espace latent, par exemple en utilisant l'indice de Davies-Bouldin [12] qui est une mesure de regroupement des données. Il convient de noter que l'objectif principal de ces méthodes est de prédire la généralisation d'un modèle entraîné avec un grand nombre de données. La prédiction de la généralisation dans un scénario avec peu d'exemples a été principalement abordée avec des méthodes de méta-apprentissage [7]. Le travail le plus proche du nôtre est [3], où certaines des stratégies mentionnées précédemment ont été testées pour l'APE dans un cadre de transfert de connaissance.

Contrairement aux travaux mentionnés précédemment, nous visons dans cet article à dériver un modèle statistique des densités conditionnelles des classes dans l'espace latent et à utiliser ce modèle pour estimer l'erreur de généralisation. Comme nous le démontrerons dans les expériences, la méthodologie proposée peut être plus performante que celles mentionnées précédemment dans des contextes de l'APE.

3 Méthodologie

3.1 Modèle statistique

La première étape de notre méthode[1] consiste à proposer un modèle statistique pour les densités conditionnelles des classes dans l'espace latent. Supposons que chaque classe suive une distribution gaussienne avec un à priori uniforme sur les classes c. -à-d. $\frac{1}{n}$, où n est le nombre de classes, ce qui est une hypothèse raisonnable dans un cadre avec peu d'exemples [9]. Les densités conditionnelles sont définies comme suit :

$$p(\mathbf{z} \mid y = c) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \ , \tag{3}$$

où $\boldsymbol{\Sigma}_c$ est la matrice de covariance de la classe c.

Notre hypothèse repose sur le fait que, compte tenu d'un extracteur de caractéristiques bien entraîné, chaque classe dans l'espace des caractéristiques devrait suivre une distribution gaussienne multivariée centrée autour du centre de la classe. Cette hypothèse a été largement adoptée dans la littérature de l'APE [5, 9, 15, 4]. En outre, les performances que nous obtenons dans nos résultats expérimentaux dans la section 4 montrent que ce modèle est adapté à nos données.

La prédiction de la généralisation peut être définie comme la prédiction de la probabilité d'erreur du classificateur C. Soit $R_c = \{\mathbf{z} \mid C(\mathbf{z}) = c\}$ la région de décision du classificateur C pour la classe c et $R = \cup_{c \in \mathcal{Y}} R_c$, l'erreur théorique est :

$$P_e = \sum_{c \in \mathcal{Y}} \int_{R \setminus R_c} p(\mathbf{z} \mid y = c) p(y = c) \, d\mathbf{z} \ . \tag{4}$$

3.2 Analyse

Il est souvent difficile d'obtenir une forme close de la solution de l'équation 4 lorsque $n > 2$. Dans cette section, nous nous concentrons sur le cas de la classification binaire et dérivons une expression analytique pour P_e. Dans le cas d'un classificateur binaire de données gaussiennes isotropes avec un écart-type σ égal et des centres de classe μ_a et μ_b, P_e a une forme close qui ne dépend que de la distance entre les centres de classe $r = \|\mu_a - \mu_b\|_2$ et σ : $P_e = 1 - \phi\left(\frac{r}{2\sigma}\right)$.

Pour estimer P_e, nous estimons généralement r et σ à l'aide d'échantillons i.i.d. tels que $\hat{r} = \|\hat{\mu}_a - \hat{\mu}_b\|_2$, où $\hat{\mu}$ est la moyenne empirique. Sous cette forme analytique, nous pouvons dériver une borne supérieure pour $\hat{P}_e = 1 - \phi\left(\frac{\hat{r}}{2\hat{\sigma}}\right)$ et prouver que l'évolution de cette borne est de $\mathcal{O}(\frac{1}{\sqrt{k}})$ où k est le nombre d'exemples disponibles. Plus de détails sur cette borne sont disponibles dans la version longue de cette article [1].

[1] Le code est disponible au lien suivant : `https://github.com/ybendou/fs-generalization`.

L'estimation de l'erreur de probabilité dépend de l'estimation de la distance entre les centres de classe. L'approche naïve consiste à estimer les moyennes de chaque distribution à l'aide des moyennes empiriques et en calculant $\hat{r} = \|\hat{\mu}_a - \hat{\mu}_b\|_2$, ce que nous appelons l'estimateur naïf. Cependant, cette estimation de la distance entre les centres de classe est biaisée.

Lemma 1 *Soit $(a_i)_{i\in[\![1..k]\!]}$ et $(b_i)_{i\in[\![1..k]\!]}$ deux séquences de variables aléatoires i.i.d tirées de leurs distributions de probabilités multivariées respectives p_a and p_b supposées indépendantes avec des espérances finies μ_a and μ_b et des moments de second ordre finis et matrices de covariance $\mathbf{\Sigma}_a$ et $\mathbf{\Sigma}_b$. Soit $\hat{r}$ l'estimateur naïf des distances utilisant $\hat{\mu}_a$ et $\hat{\mu}_b$ les moyennes empiriques respectives de chacune des deux séquences, alors :*

$$\mathop{\mathbb{E}}_{\substack{a\sim p_a\\ b\sim p_b}}\left(\hat{r}^2\right) - r^2 = \frac{\mathrm{Tr}(\mathbf{\Sigma}_a + \mathbf{\Sigma}_b)}{k} \,. \tag{5}$$

Le biais est une fonction du bruit et du nombre d'échantillons k. Notre approche numérique comprend une étape de réduction de ce biais. La correction est effectuée dans l'espace original à haute dimension. La preuve du lemme 1 ainsi que des expériences montrant l'étendue de ce biais sont incluses dans la version longue de cet article [1]. Nous fournissons également des expériences avec et sans la correction du biais pour démontrer l'importance de ce biais.

3.3 Aspect numérique

Le calcul analytique de P_e dans l'équation 4 pour $n > 2$ est difficile. En pratique, nous pouvons approximer P_e à l'aide d'une méthode de Monte Carlo. Pour chaque classe, nous tirons un grand nombre de points de données à partir de distributions gaussiennes ajustées à l'ensemble de données du problème d'APE pour l'enrichir artificiellement et calculer les décisions du classificateur.

Nous prenons en compte les distances positivement biaisées entre les centres de classe qui conduisent à une sous-estimation de P_e (comme démontré dans la version longue de cet article [1]). La correction de ce biais est essentielle à l'estimation précise de P_e. En fait, pour un classificateur basé sur la distance, le positionnement absolu des centres de classe n'affecte pas ses décisions. Afin d'effectuer l'échantillonnage, nous générons un ensemble de points qui respectent les nouvelles distances estimées à l'aide de l'algorithme de Positionnement multidimensionnel (MDS) [11].

L'estimation de P_e par échantillonnage signifie qu'il n'y a aucune restriction quant au choix des matrices de covariance des données. Dans la section 4, nous comparons les performances pour différentes matrices de covariance et réalisons une expérience pour valider notre choix, c'est-à-dire : 1) la matrice identité, 2) une matrice de covariance isotrope partagée entre les classes, 3) une matrice de covariance isotrope par classe, 4) la matrice de covariance complète par classe.

4 Expériences

4.1 Données

Nous utilisons deux jeux de données standards pour les problèmes de classification à partir de peu d'exemples : Tiered-ImageNet et Meta-dataset (qui inclut ImageNet et VGG-Flower). Nous échantillonnons 10^3 problèmes de chaque ensemble de données qui contient au moins 500 échantillons par classe. Nous pré-entraînons un réseau de neurones convolutif nommé ResNet-18 avec un espace latent de dimension 512 en utilisant la procédure standard de [2]. Les détails peuvent être trouvés dans la version longue de cet article [1].

4.2 Estimation des moments

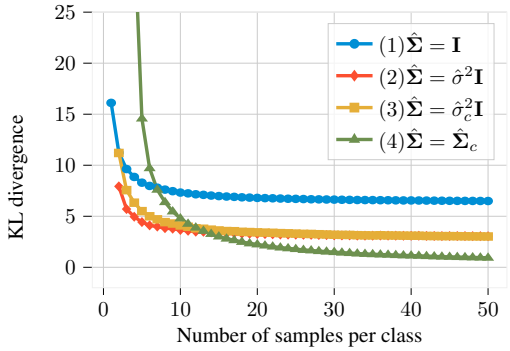
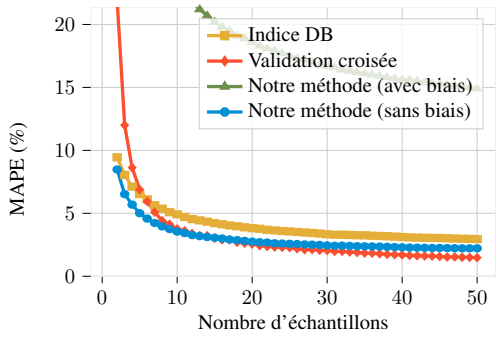


FIGURE 1 : Divergence KL entre une distribution gaussienne ajustée à partir d'un nombre limité d'échantillons et l'approximation gaussienne à l'aide d'un grand nombre d'échantillons.

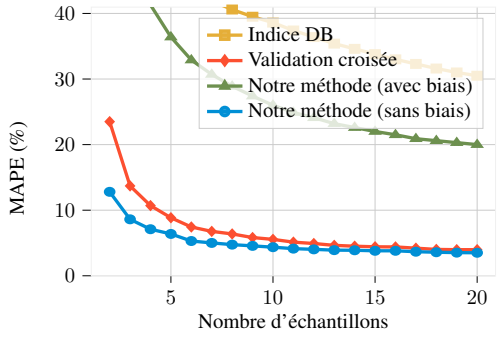
Nous avons testé différents types de matrice de covariance pour notre modèle gaussien. Nous avons mesuré la divergence moyenne de Kullback-Leibler (KL) entre les matrices de covariance obtenues avec peu d'exemples et celle obtenue avec un grand nombre d'échantillons étiquetés sur 10^3 problèmes à 5 classes. La figure 1 montre que le modèle libre avec plus de paramètres est plus performant pour un grand nombre d'échantillons, tandis que le modèle isotrope partagé est préférable pour un nombre réduit d'échantillons. Nous utilisons une matrice de covariance isotrope partagée pour $k \leq (n-1)^2$ (intersection entre les modèles 2 et 4 dans la Figure 1), et un modèle complètement libre dans le cas contraire.

4.3 Performance de la prédiction de la généralisation

Nous comparons notre méthode à la validation croisée et à l'indice de Davies-Bouldin (DB) qui mesure le regroupement intra-classe et inter-classe. Nous utilisons un jeu de données de validation pour entraîner une régression linéaire pour la méthode DB et l'appliquons aux tâches d'APE. Pour VGG-Flower, nous utilisons l'ensemble de validation d'ImageNet. Nous prédisons la précision $(1 - \hat{P}_e)$ pour chaque problème et la comparons à la précision réelle en utilisant l'erreur absolue moyenne en pourcentage (MAPE) comme métrique. La Figure 2 montre que notre méthode est plus performante que la validation croisée lorsque peu d'échantillons sont disponibles. Tandis que notre méthode performe mieux que l'indice de DB Index dans tous les scénarios. Celle-ci est également plus efficace pour prédire la généralisation et ses prédictions sont plus alignées avec la vérité de terrain, comme le montre la figure 3 pour des tâches d'APE d'ImageNet. Par ailleurs, notre méthode sans correction de biais ne donne pas de bons résultats, d'où l'importance de l'étape de correction de biais.



(a) Tiered-ImageNet



(b) VGG-Flower

FIGURE 2 : Pourcentage d'erreur absolu moyen (MAPE) de différents prédicteurs de généralisation en fonction du nombre d'échantillons sur 10^3 problèmes d'APE à 5 classes. La figure (a) est intra-domaine et la figure (b) est inter-domaine.

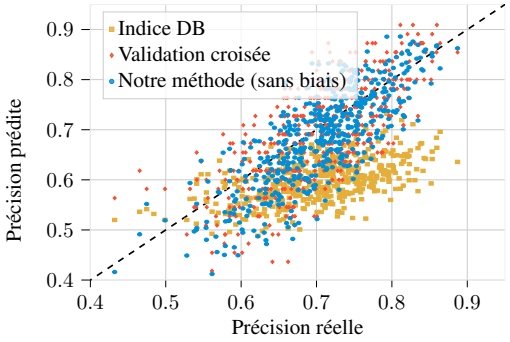


FIGURE 3 : Diagramme de dispersion de problèmes d'APE d'ImageNet. Chaque point représente un problème différent avec une précision réelle contre la précision prédite.

5 Discussion

Tout d'abord, l'indice DB est peu performant sur les ensembles de données inter-domaines en raison de l'écart important entre les jeux de données, ce qui entraîne une basse performance de la régression linéaire apprise. Les prédictions de l'indice DB sont mal alignées avec les précisions de la vérité terrain, comme illustré dans la Figure 3. Au contraire, notre méthode et la validation croisée ne souffrent pas de ce comportement. Pour prédire la généralisation avec des échantillons limités, la méthode proposée est similaire à l'indice DB avec un noyau gaussien pour la classification binaire, mais pour la classification multi-classes, elle est avantageuse car elle estime les densités conditionnelles des classes pour calculer le chevauchement entre les classes. En outre, l'erreur de prédiction dépend de la précision du problème d'APE. Les ensembles de données inter-domaines ont un écart de performance avec les ensembles de données intra-domaine en raison d'une meilleur séparation entre les classes dans l'espace latent et un meilleur rapport signal à bruit, ce qui conduit à de meilleures résultats.

6 Conclusion

Cet article propose un modèle pour estimer la capacité de généralisation des classificateurs entraînés avec peu d'échantillons. Notre méthode est plus performante que la validation croisée et que l'estimateur basé sur le score de Davis-Bouldin pour différentes tâches. Notre méthode s'appuie fortement sur des estimations non biaisées des distances inter-classes, une contribution essentielle de cet article. Notre méthode peut être généralisée à d'autres classificateurs basés sur le transfert et sur les distances. Bien que nous améliorions les méthodes existantes, nous pensons que cela ouvre de nouvelles directions de recherche intéressantes.

Références

[1] Yassir BENDOU, Vincent GRIPON, Bastien PASDELOUP, Lukas MAUCH, Stefan UHLICH, Fabien CARDINAUX, Ghouthi Boukli HACENE et Javier Alonso GARCIA : A statistical model for predicting generalization in few-shot classification. 2022.

[2] Yassir BENDOU, Yuqing HU, Raphael LAFARGUE, Giulia LIOI, Bastien PASDELOUP, Stéphane PATEUX et Vincent GRIPON : Easy : Ensemble augmented-shot y-shaped learning : State-of-the-art few-shot classification with simple ingredients. 2022.

[3] Myriam BONTONOU, Louis BÉTHUNE et Vincent GRIPON : Predicting the accuracy of a few-shot classifier. *arXiv preprint arXiv :2007.04238*, 2020.

[4] Tianshi CAO, Marc LAW et Sanja FIDLER : A theoretical analysis of the number of shots in few-shot learning. *arXiv preprint arXiv :1909.11722*, 2019.

[5] Tomáš CHOBOLA, Daniel VAŠATA et Pavel KORDÍK : Transfer learning based few-shot classification using optimal transport mapping from preprocessed latent space of backbone neural network. *In AAAI Workshop on Meta-Learning and MetaDL Challenge*, pages 29–37. PMLR, 2021.

[6] Guneet S DHILLON, Pratik CHAUDHARI, Avinash RAVICHANDRAN et Stefano SOATTO : A baseline for few-shot image classification. *arXiv preprint arXiv :1909.02729*, 2019.

[7] Alec FARID et Anirudha MAJUMDAR : Generalization bounds for meta-learning via pac-bayes and uniform stability. *Advances in Neural Information Processing Systems*, 34:2173–2186, 2021.

[8] Chelsea FINN, Pieter ABBEEL et Sergey LEVINE : Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, pages 1126–1135, 2017.

[9] Yuqing HU, Stéphane PATEUX et Vincent GRIPON : Adaptive dimension reduction and variational inference for transductive few-shot classification. *arXiv preprint arXiv :2209.08527*, 2022.

[10] Yiding JIANG, Behnam NEYSHABUR, Hossein MOBAHI, Dilip KRISHNAN et Samy BENGIO : Fantastic generalization measures and where to find them. *arXiv preprint arXiv :1912.02178*, 2019.

[11] Joseph B KRUSKAL : Nonmetric multidimensional scaling : a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[12] Parth NATEKAR et Manik SHARMA : Representation based complexity measures for predicting generalization in deep learning. *arXiv preprint arXiv :2012.02775*, 2020.

[13] Mengye REN, Eleni TRIANTAFILLOU, Sachin RAVI, Jake SNELL, Kevin SWERSKY, Joshua B TENENBAUM, Hugo LAROCHELLE et Richard S ZEMEL : Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv :1803.00676*, 2018.

[14] Yan WANG, Wei-Lun CHAO, Kilian Q. WEINBERGER et Laurens van der MAATEN : Simpleshot : Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv :1911.04623*, 2019.

[15] Jingyi XU et Hieu LE : Generating representative samples for few-shot classification. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9003–9013, 2022.