Apprentissage Minimax pour les Réseaux de Neurones

Cyprien GILET[1], Marie GUYOMARD[2], Susana BARBOSA[3], Lionel FILLATRE[2]

[1]Université de Technologie de Compiègne, Heudiasyc, France

[2]Université Côte d'Azur, I3S, France

[3]Université Côte d'Azur, IPMC, France

cyprien.gilet@hds.utc.fr, guyomard@i3s.unice.fr, sudocarmo@gmail.com, lionel.fillatre@i3s.unice.fr

Résumé – Cet article propose une nouvelle approche pour équilibrer les risques conditionnels d'un réseau de neurones convolutif avec un critère minimax. Les réseaux de neurones cherchent à minimiser l'erreur de classification, ce qui conduit généralement à des risques conditionnels par classe déséquilibrés. L'apprentissage minimax consiste à égaliser les risques conditionnels en remplaçant la dernière couche du réseau de neurones par un classifieur minimax approprié. Une telle égalisation est importante dans de nombreuses applications lorsque certaines classes sont mal reconnues. Des expériences numériques comparent notre approche à plusieurs algorithmes sur des images médicales et la base de données CIFAR-100. L'égalisation des risques conditionnels fonctionne bien, même lorsque le nombre de classes est élevé.

Abstract – This paper proposes a new approach for balancing the class-conditional risks of a convolutional neural network with a minimax criterion. Neural networks seek to minimize the classification error, which generally leads to imbalanced class-conditional risks. Minimax learning consists in equalizing the class-conditional risks by replacing the last layer of the neural network by an appropriate minimax classifier. Such an equalization is important in many real world applications when some classes are poorly recognized. Numerical experiments compare our approach to several state-of-the-art algorithms on medical images and CIFAR-100 database. They show the relevance of our approach when it is necessary to well classify the classes with the smallest number of images. Our approach works well when the number of classes is large.

1 Introduction

Les réseaux de neurones convolutifs (CNNs) deviennent incontournables pour la classification des images [1] et permettent généralement d'atteindre des performances de classification élevées. Cependant, les CNNs souffrent souvent lorsqu'ils traitent des ensembles de données déséquilibrés [2]. Ces problèmes surviennent couramment dans de nombreux domaines d'application tels que la médecine personnalisée et la biologie. Lorsque les classes sont inégalement représentées, la plupart des CNNs se concentrent essentiellement sur les classes dominantes qui contiennent le plus grand nombre d'images et sous-estiment les plus petites. En d'autres termes, une classe minoritaire avec juste un petit nombre d'images aura un grand risque conditionnel par classe. De plus, un classifieur déséquilibré peut devenir très sensible aux changements de probabilité a priori [3]. L'égalisation des risques conditionnels par classe est donc essentielle pour obtenir un classifieur robuste.

Dans [2], les auteurs fournissent un aperçu intéressant des approches pour résoudre le problème de données déséquilibrées dans l'apprentissage en profondeur. Une approche courante consiste à équilibrer les données en ré-échantillonnant l'ensemble d'entraînement [4] lorsque le nombre d'images est suffisamment grand. Une autre approche fréquente est l'apprentissage sensible aux coûts [5] qui vise à attribuer différents coûts d'erreurs de classification par classe afin de contrebalancer le nombre d'occurrences dans chaque classe. Cependant, ces coûts sont généralement difficiles à optimiser lorsqu'il s'agit d'un grand nombre de classes [3]. Les auteurs de [6] proposent de remplacer l'objectif standard d'entropie croisée lors de la procédure d'entraînement.

Une approche alternative pour rendre un classifieur robuste aux classes déséquilibrées et aux changements a priori est le critère minimax [3, 7]. Un classifieur minimax cherche à minimiser le maximum des risques conditionnels. Par conséquent, un classifieur minimax tend à égaliser ces risques par classe. Il est obtenu en maximisant l'erreur de classification de Bayes par rapport aux probabilités a priori. Le critère minimax appliqué aux réseaux de neurones a déjà été étudié dans [7] où les auteurs ont proposé un algorithme du point fixe qui nécessite de ré-échantillonner le jeu de données d'apprentissage à chaque itération. Nous proposons dans cet article un algorithme basé sur le gradient qui ne nécessite pas un tel ré-échantillonnage. Notre approche est plus appropriée lorsque certaines classes n'ont que quelques échantillons, que le nombre total d'échantillons d'apprentissage est limité ou que le nombre de classes est important. Notre approche remplace la couche de classifieur de sortie (c'est-à-dire la dernière couche du CNN) par une règle de décision minimax.

Nos contributions sont les suivantes. Tout d'abord, la Section 2 montre comment coupler un CNN avec un classifieur

minimax. Nous discrétisons les descripteurs profonds calculés par le CNN pour obtenir un risque moyen sous forme analytique. Nous le maximisons avec un algorithme de sous-gradient projeté qui calcule le classifieur minimax. Ensuite, la Section 3 montre les avantages de notre approche sur les images médicales et la base de données CIFAR-100 pour s'assurer que les classes minoritaires sont prédites aussi bien que les autres. Enfin, la Section 4 conclut l'article.

2 Couplage CNN et Minimax

2.1 Couplage d'un CNN avec un classifieur

Soit $\mathcal{Y} = \{1, \dots, K\}$ l'ensemble des étiquettes des classes ($K \geq 2$). Soit $\Phi : \mathcal{X} \to \mathcal{Y}$ un CNN qui attribue une étiquette de classe à chaque image $X \in \mathcal{X}$. Fondamentalement, l'architecture d'un CNN Φ composé de s couches cachées $h_1, \dots, h_s$ peut être modélisée comme [8]

$$\Phi(X) \;=\; h_{s+1} \circ h_s \circ \cdots \circ h_1(X) \;=\; h_{s+1} \circ \varphi(X), \quad (1)$$

où $h_{s+1}(\cdot)$ désigne la couche de sortie, $\varphi(X)$ est la sortie de la dernière couche cachée et $f \circ g(X) = f(g(X))$ désigne la composition des fonctions. Dans le reste de l'article, $Z = \varphi(X) \in \mathbb{R}^d$ est appelé le descripteur profond et h_{s+1} est appelé le classifieur de sortie. Habituellement dans un CNN, la règle de décision de sortie h_{s+1} vise à approcher le classifieur de Bayes. Le classifieur softmax [8] est souvent utilisé pour effectuer cette approximation. Par conséquent, la règle de décision softmax et le classifieur de Bayes devraient atteindre des performances similaires sur les descripteurs profonds.

Cet article propose de remplacer le classifieur de sortie par une règle de décision minimax. Ainsi, nous étudions les réseaux de neurones profonds qui peuvent être exprimés comme

$$\Phi_\delta(X) \;=\; \delta \circ \varphi(X) \;=\; \delta(Z), \quad (2)$$

où $\delta : \mathbb{R}^d \to \mathcal{Y}$ est toute règle de décision jouant le rôle de classifieur de sortie. En d'autres termes, $\Phi_\delta(X)$ est un CNN qui prend une décision basée sur les descripteurs profonds Z. Nous ne voulons pas entraîner à nouveau les couches cachées du CNN mais juste coupler les descripteurs profonds avec un classifieur spécifique (seul ce classifieur sera entraîné). Ainsi, notre approche s'apparente à un réglage fin (qui a par ailleurs montré son efficacité pour l'analyse d'images médicales [9]).

2.2 Risque de Bayes du CNN couplé

Soit un multi-ensemble $\mathcal{S} = \{(Y_i, X_i), i \in \mathcal{I}\}$ soit l'ensemble de données d'entraînement contenant m entraînement étiqueté images, où $\mathcal{I}$ est un ensemble fini d'indices. Soit $L : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$ la fonction de perte telle que $L(k, l) := L_{k,l}$ correspond à la perte de prédiction de la classe l lorsque la classe réelle est k. Le risque empirique d'erreur de classification du CNN Φ_δ est

$$\hat{r}\left(\Phi_\delta\right) = \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \Phi_\delta(X_i)) = \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \delta(Z_i)), \quad (3)$$

où $Z_i = \varphi(X_i)$. Chaque CNN de la forme $\Phi_\delta(X)$ peut être comparé en évaluant uniquement le risque $\hat{r}_\varphi(\delta)$ défini par

$$\hat{r}_\varphi(\delta) \;:=\; \hat{r}(\delta \circ \varphi) \;=\; \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \delta(Z_i)), \quad (4)$$

puisque $\varphi(\cdot)$ est commun à tous les CNN Φ_δ. Autrement dit, le risque empirique $\hat{r}(\Phi_\delta)$ d'un CNN Φ_δ est égal au risque empirique $\hat{r}_\varphi(\delta)$ de la règle de décision δ appliquée sur les entités profondes.

Notons $\pi := [\pi_1, \dots, \pi_K]$ les proportions de classes de l'ensemble d'apprentissage telles que π_k est la proportion d'images observées dans classe k. Comme expliqué dans [3, 10], le risque moyen $\hat{r}_\varphi(\delta)$ peut être écrit comme

$$\hat{r}_\varphi(\delta) = \sum_{k \in \mathcal{Y}} \pi_k \hat{R}_k(\delta), \quad (5)$$

$$\hat{R}_k(\delta) \;=\; \sum_{l \in \hat{\mathcal{Y}}} L_{kl} \hat{\mathbb{P}}(\delta \circ \varphi(X_i) = l \mid Y_i = k), \quad (6)$$

où $\hat{R}_k(\delta)$ est le risque conditionnel de classe empirique de δ associé à la classe k et $\hat{\mathbb{P}}(\cdot \mid \cdot)$ est la probabilité conditionnelle estimée à partir de l'ensemble de données d'apprentissage. Le classifieur de Bayes optimal qui minimise $\hat{r}_\varphi(\delta)$ est défini par

$$\delta_\pi^B \;:=\; \arg\min_{\delta \in \Delta} \hat{r}_\varphi(\delta), \quad \text{où } \Delta = \{\delta : \mathbb{R} \to \mathcal{Y}\}. \quad (7)$$

Le classifieur δ_π^B n'est optimal que pour l'a priori π.

2.3 Apprentissage Minimax

Le critère minimax vise à minimiser $M(\delta) = \max_k \hat{R}_k(\delta)$. Comme démontré dans [10], le classfieur minimax $\delta_{\bar{\pi}}^B$ est le classifieur de Bayes associé à $\bar{\pi}$ qui maximise $\hat{r}_\varphi\left(\delta_\pi^B\right)$ par rapport à π. Le classifieur minimax est souvent une règle d'égalisation telle que $\hat{R}_k\left(\delta_{\bar{\pi}}^B\right) = M(\delta_{\bar{\pi}}^B)$ pour tout k. Malheureusement, le calcul de $\hat{r}_\varphi\left(\delta_\pi^B\right)$ est insoluble à cause de la malédiction de la dimensionnalité : nous ne pouvons pas estimer précisément une distribution de probabilité empirique sur les images d'entrée. Cette difficulté persiste avec les descripteurs profonds même si leur taille est plus petite que celle des images d'entrée. Nous proposons donc de discrétiser les descripteurs profonds puis d'apprendre le classifieur minimax en utilisant une expression sous forme analytique. Les trois étapes pour calculer le classifieur minimax discret sont décrites ci-après.

1) K-means : Nous avons testé numériquement différentes méthodes de discrétisation mais l'algorithme des K-means est le meilleur compromis entre simplicité et efficacité. Par conséquent, l'espace des descripteurs profonds $\mathbb{R}^d$ est partitionné en T régions disjointes $\{\Omega_1, \dots, \Omega_T\}$ telles que $\cup_{t=1}^T \Omega_t = \mathbb{R}^d$. Ceci définit une fonction $\gamma : \mathbb{R}^d \mapsto \{1, \dots, T\}$ tel que $\gamma(Z) = t$ si et seulement si $Z \in \Omega_t$. La dimension réduite T est choisie pour obtenir un bon compromis entre précision et scalabilité de l'étape d'optimisation.

2) Maximisation du risque : À partir de la discrétisation des K-means, nous obtenons les instances d'apprentissage étiquetées $\mathcal{S}_T = \{(Y_i, t_i), i \in \mathcal{I}\}$ où $t_i = \gamma(\varphi(X_i))$ est un descripteur profond discrétisé. Nous pouvons calculer les probabilités

$\hat{p}_{k,t}$ d'observer le profil des descripteurs $\gamma(Z) = t$ étant donné que l'étiquette de classe est k

$$\hat{p}_{k,t} := \frac{|\,(Y_i, t_i) \in \mathcal{S}_T : t_i = t, Y_i = k|}{|\,(Y_i, t_i) \in \mathcal{S}_T : t_i = t|}, \tag{8}$$

où $|A|$ est le nombre d'éléments de l'ensemble A. Puisque les descripteurs profonds sont maintenant discrétisés, un bref calcul montre que le risque moyen est donné par

$$\hat{r}_\varphi\left(\delta_\pi^B\right) = \sum_{t=1}^{T} \min_{1 \le q \le K} \lambda_{q,t} \text{ avec } \lambda_{q,t} = \sum_{k=1}^{K} L_{k,q}\, \pi_k\, \hat{p}_{k,t}. \tag{9}$$

La valeur $\lambda_{q,t}$ représente la perte moyenne pour décider de la classe q lorsqu'on observe le profil discret t. La fonction $\hat{r}_\varphi\left(\delta_\pi^B\right)$ est concave mais non dérivable par rapport à π à cause de la fonction min. Un algorithme de sous-gradient projeté peut être utilisé pour trouver le maximum de cette fonction (voir par exemple [3]). Un tel algorithme d'optimisation permet de calculer très précisément la valeur $\bar{\pi} = \text{argmax}_{\pi \in \mathbb{S}}\, \hat{r}_\varphi\left(\delta_\pi^B\right)$ où $\mathbb{S}$ est le simplexe probabiliste.

3) Classifieur Minimax : La sortie finale de notre algorithme est le classifieur minimax $\delta_{\bar{\pi}}^B$ donné par

$$\delta_{\bar{\pi}}^B(X_i) = \arg\min_{1 \le q \le K} \lambda_{q,t_i} \text{ avec } t_i = \gamma(\varphi(X_i)). \tag{10}$$

Ce classifieur est exprimé sous une forme analytique et est facile à utiliser en pratique. De plus, notre approche est évolutive puisque i) le clustering K-means nous permet de contrôler la dimension T et ii) l'algorithme de sous-gradient fonctionne bien en haute dimension.

3 Expériences Numériques

Données médicales : on considère trois bases de données médicales réelles (*BreastMNIST, OCTMNIST, DermaMNIST*) [11] qui diffèrent selon le nombre d'images, le nombre de classes et les proportions de classes. Chaque base de données contient un ensemble d'apprentissage, un ensemble de validation et un ensemble de test d'images de dimensions 28×28.

Pour illustrer que notre approche peut être couplée à tout type de CNN, nous avons considéré deux CNNs : *ResNet-18* [12] et *EfficientNet-B7* [13]. Nous avons calibré chaque CNN sur l'ensemble d'apprentissage avec 100 époques en utilisant la fonction de perte d'entropie croisée et un optimiseur SGD. Nous avons comparé six classifieurs de sortie : le Discrete Bayes Classifier (DBC), les K-Nearest Neighbors (KNN), le Support Vector Machine (SVM), les Weighted Random Forests (WRF), le classifieur softmax qui était considéré dans les CNNs initiaux, et notre Discrete Minimax Classifier (DMC). Chaque classifieur de sortie (du CNN ResNet-18 et du CNN EfficientNet-B7) a été ajusté sur les descripteurs profonds associés à la base de validation afin d'éviter un surentraînement éventuellement dû aux descripteurs profonds provenant de l'ensemble d'apprentissage. Les performances de généralisation ont été évaluées sur l'ensemble de test.

	Classifieur	DERMA		BREAST		OCT	
	δ	Val	Test	Val	Test	Val	Test
ResNet-18							
$\hat{r}_\varphi(\delta)$	CNN	0.29	0.30	0.17	0.16	0.06	0.28
	DBC	0.26	0.32	0.14	0.18	0.07	0.20
	DMC	0.48	0.54	0.17	0.19	0.13	0.21
	KNN	0.22	0.29	0.09	0.14	0.06	0.26
	WRF	0.32	0.41	0.08	0.17	0.08	0.20
	SVM	0.00	0.33	0.00	0.20	0.02	0.28
$\max_{k \in \mathcal{Y}} \hat{R}(\delta)$	CNN	1.00	1.00	0.43	0.50	0.35	0.76
	DBC	1.00	1.00	0.43	0.57	0.47	0.41
	DMC	0.54	0.83	0.19	0.19	0.13	0.32
	KNN	1.00	1.00	0.19	0.21	0.41	0.73
	WRF	0.43	0.91	0.12	0.24	0.27	0.52
	SVM	0.00	1.00	0.00	0.47	0.46	0.76
$\psi(\delta)$	CNN	0.83	0.84	0.36	0.46	0.33	0.69
	DBC	0.90	0.87	0.39	0.54	0.46	0.33
	DMC	0.21	0.37	0.03	0.00	0.01	0.24
	KNN	0.93	0.90	0.14	0.09	0.38	0.70
	WRF	0.43	0.65	0.12	0.07	0.20	0.43
	SVM	0.00	0.83	0.00	0.20	0.44	0.73
EfficientNet-B7							
$\hat{r}_\varphi(\delta)$	CNN	0.27	0.27	0.22	0.23	0.07	0.27
	DBC	0.24	0.28	0.19	0.24	0.08	0.25
	DMC	0.48	0.49	0.23	0.29	0.14	0.22
	KNN	0.25	0.28	0.21	0.22	0.07	0.28
	WRF	0.30	0.36	0.13	0.25	0.09	0.25
	SVM	0.25	0.27	0.22	0.22	0.07	0.28
$\max_{k \in \mathcal{Y}} \hat{R}(\delta)$	CNN	1.00	0.87	0.48	0.59	0.41	0.72
	DBC	0.96	0.91	0.48	0.69	0.47	0.45
	DMC	0.43	0.83	0.24	0.52	0.15	0.29
	KNN	1.00	1.00	0.38	0.52	0.41	0.73
	WRF	0.83	0.84	0.14	0.43	0.27	0.52
	SVM	1.00	1.00	0.67	0.79	0.46	0.76
$\psi(\delta)$	CNN	0.90	0.77	0.35	0.49	0.38	0.69
	DBC	0.82	0.87	0.39	0.61	0.44	0.36
	DMC	0.11	0.41	0.01	0.31	0.01	0.17
	KNN	0.95	0.94	0.24	0.41	0.38	0.70
	WRF	0.65	0.72	0.05	0.24	0.20	0.43
	SVM	0.96	0.96	0.61	0.78	0.44	0.73

TABLE 1 – Résultats sur les ensembles de validation et de test pour chaque classifieur sur les descripteurs profonds.

Le tableau 1 compare les résultats sur les bases de validation et de test de chaque classifieur de sortie par rapport aux trois critères suivants : le risque d'erreur moyen $\hat{r}_\varphi(\delta)$ défini à l'équation (4), le maximum des risques conditionnels de classe (6), et la différence entre les risques conditionnels de classe maximum et minimum, définie par

$$\psi(\delta) := \max_{1 \le k \le K} \hat{R}_k(\delta) - \min_{1 \le k \le K} \hat{R}_k(\delta). \tag{11}$$

Les conclusions de ces expériences sont les suivantes, à la fois pour ResNet-18 et EfficientNet-B7. Premièrement, les classifieurs de sortie DBC, KNN et SVM donnent généralement des résultats similaires à ceux du CNN initial en utilisant la couche de sortie softmax. Ceci illustre que toutes ces règles de décision tendent à converger vers le classifieur de Bayes. Cependant, les risques conditionnels maximaux associés à ces classifieurs de sortie apparaissent toujours trop élevés. En d'autres termes, ces classifieurs ne fournissent pas de prédictions efficaces pour les classes avec le plus petit nombre d'images, même si ces classes correspondent à des maladies.

Lorsqu'il s'agit de bien classer les classes les plus petites et d'équilibrer les risques par classe, un compromis s'impose : laisser le risque global $\hat{r}_\varphi(\delta)$ augmenter pour mieux classer les plus petites classes. Ceci est confirmé par WRF (qui est généralement connu pour être pertinent lorsqu'il s'agit de données déséquilibrées) et notre DMC. Notre DMC atteint généralement les risques par classe maximaux les plus bas et les équilibre mieux que les autres méthodes. Le classifieur discret de Bayes et le CNN initial ne parviennent pas à équilibrer les risques.

Données CIFAR100 : Nous considérons également la base de données CIFAR-100 [14] qui contient 60 000 images avec des classes $K = 100$. Nous avons considéré un ensemble d'apprentissage, respectivement un ensemble de test, composé de 40 000 images, resp. 20 000 images. Les ensembles d'apprentissage et de test satisfaisaient les proportions par classe équilibrées $\pi = [1/100, \ldots, 1/100]$. Nous avons considéré les descripteurs extraits de la dernière couche cachée du CNN EfficientNet-B0 [13] et nous avons comparé le DMC avec la régression logistique pondérée (WLR) appliqués tous deux sur les descripteurs profonds. Puisque les proportions par classe sont ici parfaitement équilibrées, il en résulte que les poids de WLR n'aident pas : la WLR équivaut à considérer le classifieur softmax initial de la dernière couche du CNN. Comme illustré sur la Fig. 1, ce classifieur n'est pas capable d'équilibrer les risques conditionnels par classe. Malgré ces difficultés, nous pouvons observer que notre approche DMC est performante pour minimiser au maximum les risques conditionnels sur cette base de données à grande échelle.

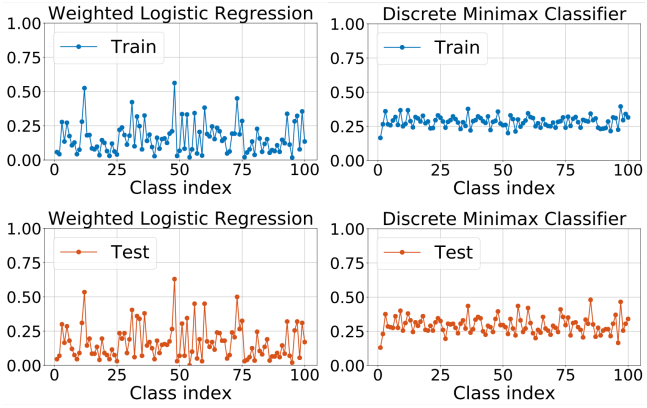

FIGURE 1 – Risques conditionnels par classe sur la base de données CIFAR-100.

4 Conclusion

Cet article présente une nouvelle approche pour équilibrer les risques conditionnels d'un CNN pour traiter des données déséquilibrées. Notre approche couple un CNN avec un classifieur minimax. Le calcul de ce classifieur est simple et efficace, même si le nombre de classes est élevé. De futurs travaux seront consacrés à l'étude de l'erreur de généralisation.

Références

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, 2018.

[3] C. Gilet, S. Barbosa, and L. Fillatre, "Discrete box-constrained minimax classifier for uncertain and imbalanced class proportions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[4] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1263–1284, 2009.

[5] M. Kukar and I. Kononenko, "Cost-sensitive learning with neural networks," *European Conference on Artificial Intelligence*, 1998.

[6] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems*, vol. 32, pp. 1567–1578, 2019.

[7] A. Guerrero-Curieses, R. Alaíz-Rodríguez, and J. Cid-Sueiro, "A fixed-point algorithm to minimax learning with neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, pp. 383–392, 2004.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[9] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis : Full training or fine tuning ?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[10] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer-Verlag New York, 2nd ed., 1994.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "*MedMNIST* databases." `https://zenodo.org/record/4269852#.X_mdsulKiHE`.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016.

[13] M. Tan and Q. Le, "Efficientnet : Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, 2019.

[14] A. Krizhevsky, "Learning multiple layers of features from tiny images." https ://www.cs.toronto.edu/ kriz/cifar.html, 2009.