Modèle d'ensemble d'apprentissage profond basé sur la représentation au second ordre de descripteurs multi-couches d'un CNN

Sara Akodad, Lionel Bombrun, Yannick Berthoumieu, Christian Germain

Modèle d'ensemble d'apprentissage profond basé sur la représentation au second ordre de descripteurs multi-couches d'un CNN

Sara AKODAD, Lionel BOMBRUN, Yannick BERTHOUMIEU, Christian GERMAIN

Université de Bordeaux, CNRS, IMS, UMR 5218, Groupe Signal et Image, F-33405 Talence, France
prenom.nom@ims-bordeaux.fr

Résumé – L'architecture d'un réseau de neurones convolutif (CNN) est formée par une succession de blocs de traitement, appelés aussi couches, qui permettent d'extraire et d'apprendre des caractéristiques discriminantes sur les images d'entrée. En particulier, les réseaux de neurones convolutifs sont constitués d'une couche dite de *pooling*. C'est une opération de sous-échantillonnage typiquement appliquée après une couche convolutive. Par ailleurs, les réseaux de neurones convolutifs de second ordre sont caractérisés par un opérateur de *pooling* qui permet de calculer la matrice de covariance globale. Il a été démontré dans la littérature que ce type d'opération permet de capter des statistiques plus riches, améliorant ainsi la représentation et la capacité de généralisation des CNN. Cependant, cette matrice est calculée uniquement sur les cartes de caractéristiques les plus profondes. Pour bénéficier de différents niveaux d'abstraction, nous proposons d'étendre les modèles existants en utilisant une approche multi-couches. Par ailleurs, pour obtenir de meilleures performances prédictives, une architecture d'ensemble est proposée. Des expérimentations sont menées sur quatre jeux de données afin de valider le potentiel du modèle proposé pour diverses applications de traitement d'image telles que la classification de scène de télédétection, la reconnaissance de scènes d'intérieur et la classification de textures.

Abstract – The architecture of a convolutional neural network (CNN) is formed by a succession of processing blocks, also called layers, which allow the extraction and learning of discriminating features on the input images. In particular, convolutional neural networks are made up of a layer called pooling. This is a sub-sampling operation typically applied after a convolutional layer. Moreover, second order convolutional neural networks are characterized by a pooling operator that allows to compute the global covariance matrix. It has been shown in the literature that this type of operation allows to capture richer statistics, thus improving the representation and the generalization capacity of CNNs. However, this matrix is computed only on the deepest feature maps. To benefit from different levels of abstraction, we propose to extend the existing models using a multi-layer approach. Moreover, to obtain better predictive performances, an ensemble architecture is proposed. Experiments are conducted on four datasets to validate the potential of the proposed model for various image processing applications such as remote sensing scene classification, indoor scene recognition and texture classification.

1 Introduction

Ces dernières années, l'apprentissage profond a gagné une grande popularité et de nombreux modèles ont été proposés dans la littérature [1]. En effet, il a été démontré que les algorithmes d'apprentissage automatique améliorent nettement les performances dans une large variété de contextes. En particulier, les réseaux de neurones convolutifs (CNN) ont été utilisés avec succès dans des applications de classification d'images. Ils sont constitués d'une succession de couches, dont chacune est en charge de l'extraction et l'apprentissage de caractéristiques spécifiques sur les images d'entrée. La représentation au premier ordre, telle que le calcul de moyenne ou de la valeur maximale, sont des opérations courantes dans les modèles CNN, aussi appelées *pooling*. Toutefois, de nombreux auteurs ont exprimé l'intérêt à utiliser une représentation d'ordre supérieur, comme la représentation au second ordre. Cela consiste essentiellement à calculer la matrice de covariance des cartes de caractéristiques issues des CNN [2]. Par ailleurs, les matrices de covariance sont des matrices symétriques définies positives qui vivent dans un espace non Euclidien. Il est donc nécessaire de réadapter les outils classiques de la géométrie Euclidienne pour manipuler ce type de données. Pour cela, des outils de la géométrie de l'information sont alors exploités pour les traiter. Dans ce contexte, plusieurs auteurs ont proposé diverses architectures de réseaux neuronaux de second ordre pour bénéficier à la fois de statistiques de second ordre et d'architectures d'apprentissage profond [3–6]. Cependant, dans ces modèles, la représentation de second ordre n'est utilisée que pour les couches les plus profondes de ces modèles. Pour bénéficier de différents niveaux d'abstraction, nous proposons d'étendre ces modèles en utilisant une approche multi-couches, ce qui représente la principale contribution de ce travail.

Ce travail est structuré comme suit. La section 2 présente le modèle proposé en détaillant chaque étape de l'architecture. Ensuite, la section 3 expose quelques expérimentations sur quatre ensembles de données incluant la reconnaissance de scène et la classification de textures. Une étude d'ablation et une cartographie d'activation de classe à l'aide de l'approche Grad-CAM sont réalisées afin d'évaluer la valeur ajoutée de chaque partie du réseau. Enfin, la section 4 donne quelques conclusions.

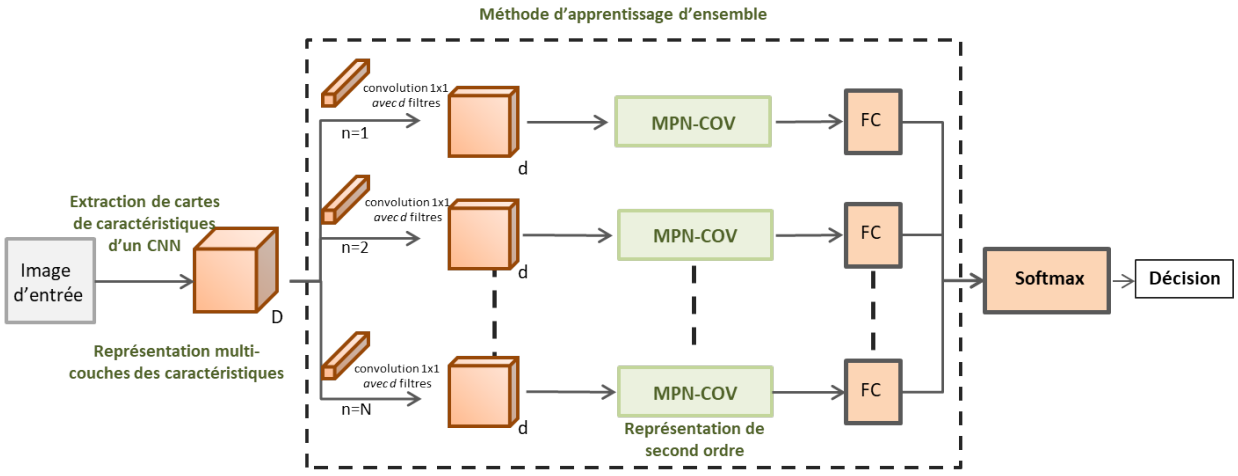

FIGURE 1 – Architecture du modèle d'ensemble proposé.

2 Architecture proposée

La Figure 1 montre une illustration du modèle d'apprentissage d'ensemble proposé basé sur la représentation du second ordre (*i.e.* matrices de covariances) des cartes de caractéristiques d'un CNN multi-couches. Il est essentiellement composé d'un réseau de neurones principal qui joue le rôle d'un extracteur de caractéristiques. Pour cela, un réseau de neurones convolutif standard peut être utilisé comme le réseau VGG-16. Ensuite, trois étapes composent l'architecture du modèle proposé : une extraction de descripteurs multi-couches, une représentation au second ordre de ces descripteurs, et une stratégie d'apprentissage d'ensemble. Les sous-sections suivantes décrivent ces différentes étapes.

2.1 Extraction de descripteurs multi-couches

Dans un CNN, chaque couche fournit un niveau plus avancé d'abstraction conceptuelle que la couche précédente. Afin de bénéficier de ces différents niveaux d'abstraction, une approche d'extraction de descripteurs multi-couches est envisagée. Elle consiste en la combinaison de cartes d'activation obtenues en sortie de diférentes couches convolutives d'un CNN. En pratique, les cartes de caractéristiques M_1, M_2 et M_3 produites par trois couches convolutives profondes ($conv_{3-3}$, $conv_{4-3}$ et $conv_{5-3}$) du réseau VGG-16 sont considérées. Cependant, ces couches ont des dimensions spatiales variées. Les dimensions du réseau VGG-16 sont $M_1 \in \mathbb{R}^{56\times56\times256}$, $M_2 \in \mathbb{R}^{28\times28\times512}$ et $M_3 \in \mathbb{R}^{14\times14\times512}$. Pour fusionner les cartes de caractéristiques de ces couches, un sous-échantillonnage à la plus petite dimension spatiale est effectué par le biais de l'interpolation bilinéaire.

2.2 Représentation au second-ordre

Les statistiques du second ordre ont démontré d'excellents résultats dans des tâches de traitement du signal et des images telle que la classification des scènes de télédétection ou l'identification des textures [4, 7]. Motivée par ces travaux et le succès des réseaux neuronaux profonds, l'architecture proposée intègre un opérateur de *covariance pooling*. Pour cela, la structure de covariance normalisée (MPN-COV) introduite dans [8] est utilisée. Cette approche est constituée des opérations détaillées ci-après.

Considérons que le descripteur multi-couches soit un tenseur de dimension $h \times w \times d$ avec h la hauteur spatiale, w la largeur d la profondeur. Après avoir re-dimensionné le tenseur en une matrice de descripteurs $\mathbf{X}$ de taille $d \times n$ constituée de $n = wh$ caractéristiques de dimension d, la représentation au second ordre est effectuée en calculant la matrice de covariance de l'échantillon tel que $\mathbf{\Sigma} = \mathbf{X}\mathbf{J}\mathbf{X}^T$ où $\mathbf{J} = \frac{1}{n}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)$. $\mathbf{I}$ est la matrice d'identité de taille $n \times n$, $\mathbf{1}$ est le vecteur unitaire de dimension n, et T désigne la transposition. Ensuite, une étape de pré-normalisation est effectuée en divisant la matrice de covariance par sa trace de sorte que $\mathbf{A} = \frac{\Sigma}{tr(\Sigma)}$. L'étape suivante consiste à utiliser la métrique *power-Euclidean* pour comparer les matrices de covariance [9]. Pour cela, la normalisation de la matrice par la racine carrée est calculée. Comme $\mathbf{A}$ est une matrice symétrique définie positive (SPD), sa racine carrée est unique et peut être calculée par la décomposition en valeurs propres telle que $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\lambda_i)\mathbf{U}^T$, avec $\mathbf{U}$ une matrice orthogonale et $\mathrm{diag}(\lambda_i)$ la matrice diagonale des valeurs propres λ_i de $\mathbf{A}$. Ensuite, la racine carrée de $\mathbf{A}$ est $\mathbf{Y} = \mathbf{U}\,\mathrm{diag}(\lambda_i^{1/2})\mathbf{U}^T$, avec $\mathbf{Y}^2 = \mathbf{A}$. Cependant, l'implémentation rapide de la décomposition sur un GPU reste assez difficile. Pour s'affranchir de cette complexité, une solution itérative approximant la racine carrée de la matrice est employée. En effet, le calcul de la racine carrée $\mathbf{Y}$ de $\mathbf{A}$ revient à itérer (cinq fois en pratique) les équations de Newton-Schulz suivantes :

$$\mathbf{Y}_k = \frac{1}{2}\mathbf{Y}_{k-1}(3\mathbf{I} - \mathbf{Z}_{k-1}\mathbf{Y}_{k-1}), \quad (1)$$

$$\mathbf{Z}_k = \frac{1}{2}(3\mathbf{I} - \mathbf{Z}_{k-1}\mathbf{Y}_{k-1})\mathbf{Z}_{k-1}), \quad (2)$$

où $\mathbf{Y}_0 = \mathbf{A}$ et $\mathbf{Z}_0 = \mathbf{I}$ la matrice d'identité. Une étape de post-

Approche multi-couches	Représentation au second ordre	Stratégie d'ensemble	UC Merced	AID	DTD	Indoor
✗	✗	✗	91.96	81.26	65.54	62.49
✓	✗	✗	95.31	87.56	66.15	63.68
✗	✓	✗	91.07	80.39	65.36	66.00
✓	✓	✗	96.21	88.00	69.27	72.43
✓	✓	✓	**98.44**	**88.95**	**71.27**	**73.79**

TABLE 1 – Étude d'ablation du modèle d'apprentissage d'ensemble basé sur la représentation au second ordre des descripteurs d'un CNN. Application pour les quatre bases de données d'images considérées.

compensation est ensuite utilisée pour contrer l'effet de la prénormalisation, de sorte que $\mathbf{C} = \sqrt{tr(\Sigma)}\mathbf{Y}_5$. Enfin, un opérateur de vectorisation est appliqué pour considérer les entrées triangulaires supérieures de la matrice symétrique résultante, afin d'obtenir un vecteur de dimension $d(d+1)/2$. Toutes ces étapes peuvent être apprises de bout en bout dans le modèle. Pour plus d'informations sur la rétro-propagation du gradient, le lecteur intéressé est référé à [8].

2.3 Méthode d'ensemble

Les méthodes d'ensemble reposent sur la combinaison de plusieurs classifieurs faibles afin de construire un classifieur plus performant [10]. A titre d'exemple, la méthode de forêts aléatoires est une technique d'apprentissage d'ensembles qui exploite des arbres de décision pour entraîner chaque modèle sur un échantillon distinct du même ensemble de données d'apprentissage. Les prédictions des membres de l'ensemble sont ensuite regroupées pour élire la décision finale par le biais d'opérations simples, tel que le vote majoritaire dans le cas de la classification. La diversité dans l'ensemble, qui est en fait assurée par les différences au sein des données sur lesquelles chaque classifieur est appris, est la raison fondamentale du succès des techniques d'apprentissage d'ensemble. Inspirés par ce principe, nous proposons un modèle d'apprentissage d'ensemble profond comme illustré dans la Figure 1. L'ensemble des cartes de caractéristiques multi-couches est d'abord séparé en N sous-ensembles de d descripteurs. Toutefois, au lieu de sélectionner de façon aléatoire les d descripteurs sur l'ensemble de départ D, cette opération de sélection est apprise dans le modèle. Pour cela, une couche convolutive 1×1 est utilisée pour former chaque sous-ensemble. Notez que l'apprentissage de la couche convolutive 1×1 pour le deuxième sous-ensemble dépend de l'apprentissage de la couche convolutive 1×1 du premier sous-ensemble. Ensuite, pour chaque sous-ensemble, la représentation de second ordre des descripteurs multi-couches est introduite dans une couche entièrement connectée (FC) de 4096 neurones. Leurs sorties sont ensuite concaténées et passées à une dernière couche entièrement connectée utilisant la fonction d'activation *softmax* pour obtenir la décision finale.

3 Expérimentations

Dans cette partie, nous illustrons le potentiel du modèle d'apprentissage d'ensemble profond proposé dans trois problèmes de classification d'images. Pour cela, quatre bases de données sont considérées. Les deux premières sont des images de télédétection de référence utilisées pour la classification des scènes aériennes, à savoir les bases de données UC Merced et AID. Elles se composent respectivement de 2 100 et 10 000 images distribuées dans 21 et 30 classes de différentes scènes aériennes telles que l'aéroport, le terrain de baseball, les résidences denses, etc. Le troisième ensemble de données est la base de données de texture DTD qui est constituée d'une collection de 47 classes de texture avec 120 images par catégorie. Enfin, la dernière application concerne la reconnaissance de scènes d'intérieur avec le jeu de données Indoor. Ce dernier compte 67 catégories et un total de 15 620 images.

Pour l'évaluation des performances du modèle proposé, 80% des images sont utilisées pour l'apprentissage concernant les bases UC Merced, DTD et Indoor et les 20% restantes sont utilisées pour les tests, tandis que pour l'ensemble de données AID, seulement 20% des images sont utilisées pour l'apprentissage (optimiseur SGD, learning rate de 10^{-3}). Pour l'ensemble des expérimentations suivantes, le réseau de neurones convolutif considéré est le modèle VGG-16 pré-entrainé sur l'ensemble de données ImageNet. Par conséquent, les poids de toutes ses couches convolutives sont gelés. Seuls les paramètres des couches suivantes (1×1 convolution, MPN-COV, FC, etc.) sont appris. Pour l'approche d'ensemble, $N = 10$ sous-ensembles sont considérés et la convolution 1×1 est d'une profondeur $d = 256$, *i.e.* les matrices de covariance sont de dimension 256×256. Ces deux derniers paramètres ont été ajustés par des expériences préliminaires et restent fixes pour tous les jeux de données.

3.1 Étude d'ablation

Dans cette sous-partie, une étude d'ablation est effectuée afin d'évaluer la valeur ajoutée de chaque élément dans le modèle proposé basé sur la représentation au second-ordre des carte de caractéristiques d'un CNN. Le tableau 1 montre les performances de classification en termes de précision globale. Sur la première ligne, le modèle est un simple *fine tuning* d'un réseau VGG-16. La deuxième ligne consiste en une version multi-couches de ce réseau VGG-16. Dans les trois lignes suivantes, on utilise une représentation au second ordre des cartes de caractéristiques du CNN. Le MPN-COV original publié dans [8] correspond à la troisième ligne, tandis que notre proposition figure sur la dernière ligne où la stratégie d'apprentissage d'ensemble comprend à la fois une approche multi-couches et une

étape de représentation au second ordre à travers les matrices de covariance. Comme observé dans le tableau 1, l'approche multi-couches permet d'améliorer de façon constante la performance de classification. Un gain important de 1% à 8% est observé sur les quatre jeux de données. La représentation au second ordre des cartes de caractéristiques d'un CNN permet également d'améliorer la précision globale. Enfin, les meilleurs résultats sont obtenus pour l'architecture proposée lorsque ces deux derniers éléments sont utilisés dans une approche d'apprentissage d'ensemble.

3.2 Interprétation du modèle

Afin de mieux comprendre le potentiel de l'approche proposée, nous exploitons une méthode de cartographie de l'activation de classe pondérée par gradient (Grad-CAM) [11]. Grad-CAM est une technique de visualisation bien connue qui est utile pour comprendre comment un modèle a été amené à prendre une décision. Il consiste à produire des cartes thermiques représentant les zones de l'image qui ont été prises en compte dans les images d'entrée pour fournir la décision finale. Ainsi, les cartes thermiques indiquent l'importance de chaque pixel par rapport à la classe d'intérêt en augmentant ou en diminuant l'intensité de la valeur du niveau de gris. Ici, trois images de trois classes différentes de la base de données UC Merced sont utilisées, à savoir, les classes avion, résidentiel et forêt. Leurs cartes correspondantes sont indiquées dans la Figure 2. La probabilité de classification obtenue en sortie du Softmax est affichée en haut à droite de chaque image. Comme observé, l'architecture proposée (dernière ligne) permet de mieux se concentrer sur l'objet d'intérêt. Par exemple, avec le modèle proposé, le contour de l'avion est mieux délimité dans l'image de l'avion qu'avec un simple réseau VGG-16. En outre, pour les images appartenant aux classes résidentiel et forêt, l'attention est mieux distribuée sur les éléments caractéristiques de la scène tels que les maisons et les arbres. De plus, les probabilités de classification sont constamment améliorées lorsque chaque élément (approche multi-couches, représentation au second ordre, stratégie d'ensemble) de l'architecture proposée est utilisé, ce qui illustre l'intérêt de leur utilisation conjointe.

4 Conclusion

Cet article a introduit un modèle d'ensemble d'apprentissage profond basé sur la représentation au second ordre de cartes de caractéristiques d'un CNN. En considérant une approche multi-couches, ce modèle a permis de représenter des images à différents niveaux d'abstraction. Ensuite, pour modéliser les dépendances entre ces cartes d'activation, une étape de *covariance pooling* a été utilisée. En outre, une architecture d'ensemble a été proposée pour obtenir de meilleures performances prédictives. Les résultats expérimentaux sur quatre jeux de données ont confirmé le potentiel du modèle proposé pour diverses applications de traitement d'images telles que la classification de scènes de télédétection, la classification de textures et la reconnaissance de scènes intérieures.

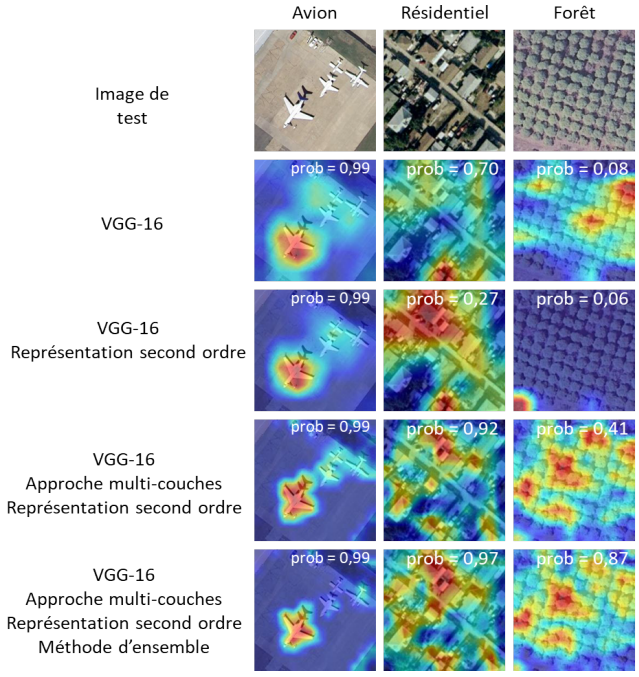

FIGURE 2 – Cartes thermiques obtenues par Grad-CAM représentant les classes d'activation sur 3 images de la base de données UC Merced.

Références

[1] Y. Le Cun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems 2*, D. S. Touretzky, Ed., 1990, pp. 396–404.

[2] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition ?" 2018.

[3] C. Ionescu, O. Vantzos, and C. Sminchisescu, "Matrix backpropagation for deep networks with structured layers," in *IEEE ICCV*, 2015, pp. 2965–2973.

[4] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE TGRS*, vol. 56, no. 12, pp. 6899–6910, Dec 2018.

[5] D. Acharya, Z. Huang, D. P. Paudel, and L. V. Gool, "Covariance pooling for facial expression recognition," *CoRR*, vol. abs/1805.04855, 2018.

[6] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *2019 IEEE CVPR*, 2019, pp. 3019–3028.

[7] S. Akodad, L. Bombrun, J. Xia, Y. Berthoumieu, and C. Germain, "Ensemble learning approaches based on covariance pooling of cnn features for high resolution remote sensing scene classification," *Remote Sensing*, vol. 12, no. 20, 2020.

[8] Q. Wang, J. Xie, W. Zuo, L. Zhang, and P. Li, "Deep CNNs meet global covariance pooling : Better representation and generalization," *CoRR*, vol. abs/1904.06836, 2019.

[9] I. L. Dryden, A. Koloydenko, and D. Zhou, "Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 1102 – 1123, 2009.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] 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 *2017 IEEE ICCV*, 2017, pp. 618–626.