Invariance et stabilité par déformations des réseaux convolutionnels profonds

Alberto BIETTI[1], Julien MAIRAL[1]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK
38000 Grenoble, France
alberto.bietti@inria.fr, julien.mairal@inria.fr

Résumé – Le succès des réseaux convolutionnels profonds a souvent été attribué à leur capacité d'apprendre des représentations invariantes et multi-échelle de signaux naturels. Or, il manque encore une étude précise de ces propriétés, ainsi que leur lien avec les propriétés d'apprentissage. Ce travail étudie les représentations convolutionnelles profondes de signaux, leur invariance à des groupes de transformations, leur stabilité à l'action de difféomorphismes, et leur capacité à préserver le signal. L'étude est basée sur les méthodes à noyaux, en particulier un noyau convolutionnel hierarchique, ce qui nous permet de séparer la représentation des données de l'apprentissage, et de définir une notion de complexité canonique, la norme RKHS, qui contrôle à la fois la stabilité et la généralisation d'un modèle donné.

Abstract – The success of deep convolutional architectures is often attributed in part to their ability to learn multiscale and invariant representations of natural signals. However, a precise study of these properties and how they affect learning guarantees is still missing. In this paper, we consider deep convolutional representations of signals; we study their invariance to translations and to more general groups of transformations, their stability to the action of diffeomorphisms, and their ability to preserve signal information. This analysis is carried by introducing a multilayer kernel based on convolutional kernel networks and by studying the geometry induced by the kernel mapping. This allows us to separate data representation from learning, and to provide a canonical measure of model complexity, the RKHS norm, which controls both stability and generalization of any learned model.

1 Introduction

Les réseaux de neurones profonds ont obtenu des résultats impressionnants dans des tâches prédictives avec des données structurées et disponibles en grande quantité. En particulier, les réseaux convolutionnels profonds (appelés CNN, [5]) peuvent exploiter la structure localement stationnaire des images naturelles à plusieurs échelles grâce à leurs opérateurs de convolution, tout en obtenant de l'invariance par translations à l'aide d'opérations de moyennage, ou *pooling*. Toutefois, la nature précise de ces invariance et les espaces fonctionnels caractérisant ces réseaux convolutionnels sont mal compris, et ces modèles sont parfois simplement utilisés comme des boîtes noires qui fonctionnent grâce à des années d'améliorations empiriques.

Comprendre les propriétés intrinsèques à cette classe de fonctions prédictives reste une question fondamentale. Par exemple, une étude de la géométrie des représentations convolutionnelles peut mener à de nouvelles intuitions sur leur succès, ainsi qu'à des meilleures mesures de complexité, qui a leur tour peuvent produire des nouvelles méthodes de régularisation grâce à un contrôle rigoureux des variations des fonctions de prédiction. Lorsque les données sont des signaux naturels, une bonne façon d'étudier cela est d'étudier la stabilité des modèles à des translations ou des déformations.

Notre approche se base sur les méthodes à noyaux [8], qui permettent d'étudier séparément la représentation des données et les modèles prédictifs appris sur un jeu de données. En particulier, nous définissons un noyau convolutionnel hiérarchique similaire à [6]. D'une part, nous étudions les propriétés de la représentation correspondante, telles que la préservation du signal, l'invariance et la stabilité à l'action de difféomorphismes, dans un contexte similaire à celui du scattering de S. Mallat [7, 3]. D'autre part, nous analysons certaines fonctions de l'espace fonctionnel obtenu (espace de Hilbert à noyau reproduisant, ou RKHS), telles que des réseaux convolutionnels génériques avec des activations lisses et homogènes, et caractérisons leur complexité grâce à leur norme RKHS, qui peut être bornée en fonctions des paramètres. Cette norme peut ensuite garantir la stabilité de ces modèles, ainsi que leur capacité de généralisation.

2 Noyau et représentation

Les méthodes à noyaux cherchent à apprendre des fonctions non-linéaires en construisant un noyau défini positif $K(\cdot, \cdot)$, et en apprenant des fonctions de la forme $f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$, où $\mathcal{H}$ est l'espace de Hilbert à noyau reproduisant (RKHS) associé à K, et $\Phi(x) = K(x, \cdot) \in \mathcal{H}$ défini la représentation du noyau. La norme de Hilbert obtenue permet ainsi de contrôler la complexité d'un modèle $f \in \mathcal{H}$ et la stabilité de ses prédic-

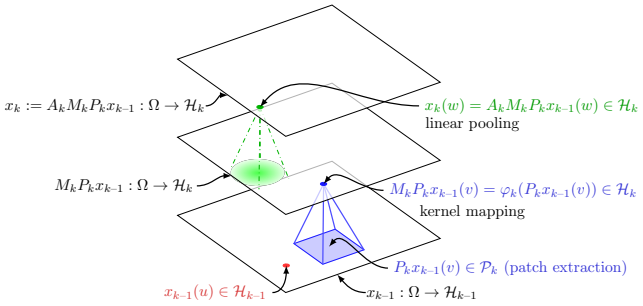

FIGURE 1 – Construction de x_k à partir de x_{k-1} à l'aide des opérateurs P_k, M_k et A_k.

tions :

$$|f(x) - f(x')| \leq \|f\|_{\mathcal{H}} \cdot \|\Phi(x) - \Phi(x')\|_{\mathcal{H}}, \qquad (1)$$

pour deux signaux x et x' donnés. Cette partie introduit la représentation convolutionnelle étudiée, et donc le noyau correspondant, qui généralise le noyau hiérarchique de [6].

On considère un signal continu x_0 dans $L^2(\mathbb{R}^d, \mathcal{H}_0)$, avec par exemple $d = 2$ et $\mathcal{H}_0 = \mathbb{R}^3$ pour une image RGB. La représentation est ensuite construite en appliquant une séquence d'opérateurs décrits ci-dessous, produisant une séquence de "feature maps" $x_k \in L^2(\mathbb{R}^d, \mathcal{H}_k)$, voir Figure 1.

Extraction de patch. Cet opérateur extrait des patch de x_{k-1}, avec une forme S_k compacte (par exemple une boîte centrée, comme dans la Figure 1). On définit alors $P_k : L^2(\Omega, \mathcal{H}_{k-1}) \to L^2(\Omega, \mathcal{P}_k)$, avec $\mathcal{P}_k = \mathcal{H}_{k-1}^{S_k}$, par

$$P_k x_{k-1}(u) = (v \mapsto x_{k-1}(u+v))_{v \in S_k} \in \mathcal{P}_k.$$

Cet opérateur est linéaire et préserve la norme L^2.

Mapping du noyau. Ensuite, chaque patch de x_{k-1} est envoyé dans un RKHS $\mathcal{H}_k$ à l'aide d'une fonction non-linéaire (ou "mapping") $\varphi_k : \mathcal{P}_k \to \mathcal{H}_k$ associée à un noyau défini positif K_k défini sur les patch. On peut alors définir l'opérateur M_k par

$$M_k P_k x_{k-1}(u) := \varphi_k(P_k x_{k-1}(u)) \in \mathcal{H}_k.$$

Nous utilisons des noyaux de la forme

$$K_k(z, z') = \|z\| \|z'\| \kappa_k\left(\frac{\langle z, z' \rangle}{\|z\| \|z'\|}\right), \qquad (2)$$

avec $\kappa_k : [-1, 1] \to \mathbb{R}$ de la forme

$$\kappa_k(u) = \sum_{j=0}^{+\infty} b_j u^j \quad \text{avec} \forall j, b_j \geq 0, \quad \kappa_k(1) = 1, \quad \kappa_k'(1) = 1.$$

Des exemples de choix possibles sont le noyau exponentiel (Gaussien sur la sphère) ou le noyau polynomial inverse (voir [6, 10, 11]). Sous ces hypothèses, φ_k est non-expansive et préserve la norme, et de même pour M_k.

Pooling. Enfin, la couche x_k est obtenue à l'aide d'un moyennage local, ou pooling, pour atténuer les hautes fréquences et produire de l'invariance locale par translations. On applique un opérateur linéaire de convolution A_k avec un filtre Gaussien à l'échelle σ_k, $h_{\sigma_k}(u) := \sigma_k^{-d} h(u/\sigma_k)$, avec $h(u) = (2\pi)^{-d/2} \exp(-|u|^2/2)$. Ensuite, on a

$$\begin{aligned} x_k(u) &= A_k M_k P_k x_{k-1}(u) \\ &= \int_{\mathbb{R}^d} h_{\sigma_k}(u - v) M_k P_k x_{k-1}(v) dv \in \mathcal{H}_k. \end{aligned}$$

On peut alors montrer que $\|A_k\| \leq 1$.

Construction multi-couche. Enfin, nous construisons la représentation multi-couche en enchaînant les opérateurs ci-dessus plusieurs fois. En pratique, la taille des patch et du pooling augmentent exponentiellement avec k, en maintenant σ_k du même ordre que $\sup_{c \in S_k} |c|$, pour augmenter le niveau d'invariance et la taille des champs réceptifs. La représentation à n couches est alors donnée par

$$x_n := A_n M_n P_n A_{n-1} M_{n-1} P_{n-1} \cdots A_1 M_1 P_1 x_0. \qquad (3)$$

On peut ensuite définir un noyau de prédiction à partir de deux signaux x_0 et x_0' :

$$\mathcal{K}_n(x_0, x_0') = \langle x_n, x_n' \rangle = \int_{u \in \mathbb{R}^d} \langle x_n(u), x_n'(u) \rangle du. \qquad (4)$$

Le RKHS de $\mathcal{K}_n$ contient alors toutes les fonctions de la forme $f(x_0) = \langle w, x_n \rangle$ avec w dans $L^2(\mathbb{R}^d, \mathcal{H}_n)$.

Discrétisation et préservation du signal. En pratique, le signal d'entrée est discret, et on peut donc considérer qu'il s'agit d'un sous-échantillonnage d'un signal continu $x_0 = A_0 x$, où x est un signal de départ et A_0 un opérateur d'anti-aliasing, qui correspond par exemple au filtrage local d'un capteur de caméra numérique. On peut ensuite construire des feature map discrètes $\bar{x}_k$ de façon analogue à la construction continue, et il est usuel d'effectuer un sous-échantillonnage des feature map après l'opération de pooling, avec un facteur de l'ordre de l'échelle du filtre Gaussien du pooling.

On peut alors montrer qu'il est possible de reconstruire $\bar{x}_{k-1}$ à partir de $\bar{x}_k$ dès que le facteur de sous-échantillonnage est plus petit que la taille du patch [1]. Cette reconstruction est possible car le RKHS $\mathcal{H}_k$ contient des fonctions linéaires qui permettent d'effectuer des "mesures" linéaires arbitraires des patch de $\bar{x}_{k-1}$. Il faut noter que cette opération de reconstruction n'est pas pratique, et peut être instable car elle se base sur des déconvolutions. Cela montre néanmoins que la représentation préserve l'information du signal d'origine, ce qui permet autorise éventuellement d'apprendre un modèle qui discrimine à partir de détails précis du signal.

3 Invariance et stabilité

Au delà de l'invariance par translation, il est souhaitable d'avoir une représentation qui est stable à des petites déformations locales. Par exemple, pour certaines tâches de classification, la

FIGURE 2 – (Haut) petites déformations obtenues à partir d'une même image du chiffre 5. (Bas) différentes images du chiffre 5 peuvent être vues comme des déformations l'une de l'autre.

sortie est préservée lorsqu'on effectue une petite déformation du signal d'entrée, tel qu'une image d'un chiffre (voir Figure 2). Dans ce cas, il est utile d'apprendre à l'aide de représentations stables. Ces déformations peuvent être décrites par un difféomorphisme C^1, $\tau : \Omega \to \Omega$, et on définit l'opérateur linéaire L_τ par $L_\tau x(u) = x(u - \tau(u))$. Notre caractérisation de stabilité est similaire à celle introduite par Mallat [7] : la représentation $\Phi(\cdot)$ est *stable* à l'action des difféomorphismes s'il existe des réels positifs C_1 et C_2 tels que

$$\|\Phi(L_\tau x) - \Phi(x)\| \leq (C_1 \|\nabla\tau\|_\infty + C_2 \|\tau\|_\infty)\|x\|, \qquad (5)$$

où $\nabla\tau$ est le Jacobien de τ, $\|\nabla\tau\|_\infty := \sup_{u\in\Omega} \|\nabla\tau(u)\|$, et $\|\tau\|_\infty := \sup_{u\in\Omega} |\tau(u)|$. La quantité $\|\nabla\tau(u)\|$ mesure la taille des déformations en u, et on suppose $\|\nabla\tau\|_\infty \leq 1/2$ comme dans [7]. Pour que Φ soit quasi-invariante par translation, on aimerait avoir C_2 petit (car on a $\nabla\tau = 0$ lorsque τ est une translation). Quand $\nabla\tau$ est non-nul, le difféomorphisme dévie d'une translation, et donne alors des déformations locales contrôlées par $\|\nabla\tau\|_\infty$.

Hypothèses pour la stabilité. Pour étudier la stabilité de la représentation (3), on suppose que le signal d'entrée x_0 s'écrit $x_0 = A_0 x$, où A_0 est un opérateur de pooling initial à l'échelle σ_0, ce qui permet de contrôler les hautes fréquences du signal à la première couche. Comme dit dessus, cette hypothèse est raisonnable en pratique, vue que les signaux considérés sont typiquement discrets, et A_0 peut être vu comme une opération d'anti-aliasing intrinsèque à l'acquisition du signal. De plus, σ_0 peut être pris arbitrairement petit, et donc nos résultats ne perdent pas de généralité. On cherche donc à étudier la stabilité de la représentation

$$\Phi_n(x) := A_n M_n P_n A_{n-1} M_{n-1} P_{n-1} \cdots A_1 M_1 P_1 A_0 x.$$

Une quantité importante dans notre étude est la taille relative maximale d'un patch par rapport à la résolution de la couche précédente, qu'on suppose finie, donnée par

$$\kappa := \max_k \frac{\sup_{c\in S_k} |c|}{\sigma_{k-1}}. \qquad (6)$$

Borne de stabilité. On peut remarquer que les opérateurs P_K, M_k et A_k commutent avec l'opérateur de translation L_c (où $c \in \mathbb{R}^d$ est un vecteur de translation fixe). On obtient alors l'invariance par translation

$$\begin{aligned}\|\Phi_n(L_c x) - \Phi_n(x)\| &= \|L_c \Phi_n(x) - \Phi_n(x)\| \\ &\leq \|L_c A_n - A_n\| \|x\| \leq C_2 \frac{|c|}{\sigma_n} \cdot \|x\|,\end{aligned}$$

où la dernière inégalité est une conséquence de [7, Lemme 2.11], avec C une constante positive. Lorsque τ est un difféomorphisme avec $\nabla\tau \neq 0$, L_τ ne commute plus avec P_k et A_k, mais on peut montrer la borne suivante sur le commutateur $[L_\tau, P_k A_{k-1}]$ (avec $[A, B] = AB - BA$), grâce au fait que la taille des patch est adaptée à la résolution σ_{k-1} à travers la quantité κ ci-dessus :

$$\|[P_k A_{k-1}, L_\tau]\| \leq C_{1,\kappa} \|\nabla\tau\|_\infty,$$

pour une constante positive $C_{1,\kappa}$ qui croît avec κ^{d+1} [1]. Nous obtenons alors le résultat de stabilité suivant.

Theorem 1 (Stabilité [1]) *Si $\|\nabla\tau\|_\infty \leq 1/2$, on a*

$$\|\Phi_n(L_\tau x) - \Phi_n(x)\| \leq \left(C_{1,\kappa}\,(1+n)\,\|\nabla\tau\|_\infty + \frac{C_2}{\sigma_n}\|\tau\|_\infty \right) \|x\|.$$

Si σ_n sont fixés, par exemple par un objectif voulu d'invariance par translation, et que les tailles de patch et échelles de pooling augmentent exponentiellement avec un facteur de l'ordre de κ (ce qui permet de vérifier (6) ainsi que la préservation du signal dans le cas discret), alors le nombre de couches n est logarithmique en σ_n et κ, alors que $C_{1,\kappa}$ croît avec κ^{d+1}. Ainsi, notre résultat préconise un choix de κ aussi petit que possible pour une meilleure stabilité, et donc des patch de petite taille (par exemple 3x3, un choix courant dans les architectures récentes [9]). Pour préserver le signal, le sous-échantillonnage doit alors être plus petit que cette taille de patch, donnant lieu à des architectures profondes.

Invariance à d'autres groupes de transformations. Il est possible d'étendre notre étude d'invariance à des groupes de transformations plus riches que les translations, par exemple les roto-translations, qui comportent à la fois une translation et une rotation. Ceci nécessite une nouvelle construction adaptée pour que les opérateur P_k, M_k et A_k commutent avec ces nouvelles transformations (voir [1, 4]). Pour le cas du groupe de roto-translation, on peut alors obtenir une représentation invariante aux roto-translations, tout en étant stable au sous-groupe de translations [1].

4 Fonctions du RKHS et leur complexité

Le noyau $\mathcal{K}_n$ défini à l'Équation (4) donne lieu à un espace fonctionnel $\mathcal{H}$ contenant des fonctions de la forme $f(x) = \langle f, \Phi(x)\rangle$, qui préservent donc les propriétés d'invariance et stabilité de la représentation $\Phi(\cdot)$ étudiées dans la section 3, à un un facteur près, la norme RKHS $\|f\|_\mathcal{H}$, selon (1). Cette norme contrôle alors à la fois la stabilité et la généralisation d'un modèle prédictif f donné. Dans cette section, nous montrons que $\mathcal{H}$ contient des réseaux convolutionnels génériques à activations lisses et étudions leur norme.

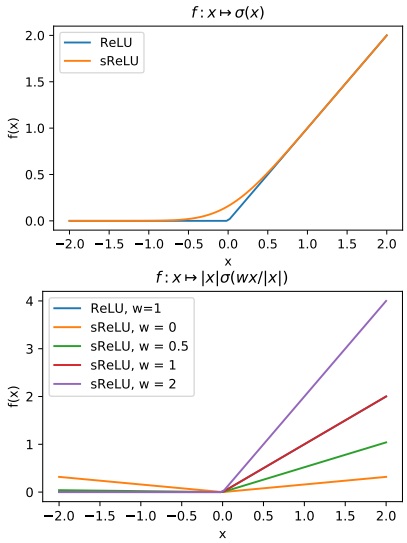

FIGURE 3 – Comparaison des fonctions obtenues avec une ReLU et sa version lissée sReLU. (Haut) cas non-homogène de [10, 11]. (Bas) notre cas homogène, pour différents paramètres w. Pour $w \geq 0.5$, on voit que sReLU et ReLU sont indistinguables.

Fonctions des RKHS $\mathcal{H}_k$ sur les patch. Lorsque le noyau K_k est de la forme (2), on peut montrer que le RKHS $\mathcal{H}_k$ associé contient des fonctions de la forme suivante [1, 10, 11] :

$$f_g(z) = \|z\|\sigma(\langle g, z\rangle/\|z\|), \tag{7}$$

où σ est une fonction d'activation lisse de la forme $\sigma(u) = \sum_{j=0}^{\infty} a_j u^j$. À une homogénéisation près, ces fonctions ressemblent à des fonctions usuellement utilisées pour définir un neurone dans un réseau de neurones. À noter que l'activation la plus répandue, la ReLU, est homogène mais non lisse, et donc ne peut pas être utilisée ici, mais que des approximations de celle-ci peuvent être utilisées [10, 11], et mènent alors à des fonctions très proches de la ReLU après homogénéisation (voir Figure 3). La norme de f_g peut être bornée comme suit :

$$\|f_g\| \leq C_\sigma(\|g\|^2),$$

où C_σ est croissante et dépend des a_j et b_j qui définissent l'activation σ et le noyau K_k, respectivement.

Réseaux convolutionnels dans le RKHS $\mathcal{H}$. On considère un CNN usuel avec des filtres $W_k(u) \in \mathbb{R}^{p_k \times p_{k-1}}$ pour $u \in S_k$ (p_k est le nombre de feature maps à la couche k), construit de manière analogue à l'architecture définie dans la section 2, mais où le mapping du noyau est remplacé par un opérateur linéaire W_k, suivi d'une activation lisse comme dans (7). La dernière couche est suivie d'un produit scalaire avec un paramètre $w_{n+1} \in L^2(\mathbb{R}^d, \mathbb{R}^{p_k})$ pour obtenir une prédiction scalaire, utilisée par exemple dans une tâche de classification ou de régression. On peut alors montrer que ce CNN, noté f_W, est contenu dans le RKHS $\mathcal{H}$, et borner sa norme comme suit [1] :

$$\|f_W\|^2 \leq \|w_{n+1}\|^2 \, C_\sigma^2(\|W_n\|_2^2 \, C_\sigma^2(\ldots C_\sigma^2(\|W_1\|_F^2)\ldots),$$

où $\|\cdot\|_2$, $\|\cdot\|_F$ sont les normes spectrales et de Frobénius.

5 Conclusion

Nous avons présenté une étude de l'invariance et de la stabilité par déformations des réseaux convolutionnels profonds, ainsi que leur lien avec la complexité de ces modèles, à l'aide des méthodes à noyaux. Notre étude permet de mieux comprendre les propriétés associées à différents choix d'architectures de réseaux convolutionnels, et suggère de nouvelles façons de régulariser qui prennent mieux en compte la complexité du modèle. Cela est important par exemple dans un cadre où il y a peu de données disponibles, ou lorsqu'on souhaite obtenir un modèle robuste à des perturbations adversariales [2].

Références

[1] A. Bietti and J. Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research*, 20(25) :1–49, 2019.

[2] A. Bietti, G. Mialon, D. Chen, and J. Mairal. A kernel perspective for regularizing deep neural networks. *preprint arXiv :1810.00363*, 2019.

[3] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Transactions on pattern analysis and machine intelligence (PAMI)*, 35(8) :1872–1886, 2013.

[4] T. Cohen and M. Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, 2016.

[5] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4) :541–551, 1989.

[6] J. Mairal. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[7] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10) :1331–1398, 2012.

[8] B. Schölkopf and A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. 2001.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2014.

[10] Y. Zhang, J. D. Lee, and M. I. Jordan. ℓ_1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning (ICML)*, 2016.

[11] Y. Zhang, P. Liang, and M. J. Wainwright. Convexified convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2017.