Prolégomènes aux réseaux de neurones multirésolution

Vincent Lostanlen, université de New York

1er janvier 2020

Résumé

Les réseaux convolutifs profonds (convnets) occupent une place prépondérante en reconnaissance des formes, et notamment en écoute artificielle (*machine listening*). Si l'utilisation de convolutions bidimensionnelles dans le plan temps–fréquence a fourni des résultant probants en classification de sons, la perspective de les employer directement sur le domaine de la forme d'onde (*raw waveform*), sans a priori de représentation temps–fréquence, permettrait à l'avenir de tirer pleinement parti du paradigme ń de bout en bout ż (*end to end*). Toutefois, l'apprentissage statistique d'une représentation de type spectrogramme soulève des problèmes de repliement de spectre (*aliasing*), de scalabilité, et de reconstruction de phase. Dans ce document, je propose d'y subvenir en initiant une démarche de recherche mêlant étroitement approximations multirésolution (MRA) et apprentissage profond. J'esquisse les principaux défis théoriques et pratiques restant à relever pour entrainer des réseaux de neurones multirésolution (MuReNN) sur des tâches d'écoute artificielle, supervisées ou non.

1 Rappels sur les réseaux de neurones

1.1 Réseaux de neurones en boucle ouverte

D'AlphaGo à Google Hello, le réseau convolutif (*convnet*) est en pointe des innovations récentes en reconnaissance des formes dans les signaux et images, connues du grand public sous le nom d'ń intelligence artificielle ż. L'idée maitresse du convnet, par rapport à un réseau de neurones densément connecté, réside dans le partage de poids synaptique (*weight sharing*). Étant donné un vecteur $\boldsymbol{x}$ de dimension $N > 1$, l'activation $\boldsymbol{y}$ des neurones du premier niveau résulte d'une application linéaire (ou affine) $\mathbf{W}$ suivie d'une non-linéarité ponctuelle ρ. En omettant par simplicité le terme constant et en choisissant N pour dimension de $\boldsymbol{y}$, on obtient

$$\forall t \in \mathbb{Z}_N,\ \boldsymbol{y}[t] = (\rho \circ \mathbf{W})(\boldsymbol{x})[t] = \rho\left(\sum_{t'=0}^{N-1} \mathbf{W}[t,t']\boldsymbol{x}[t']\right).$$

1.2 Réseaux de neurones convolutifs

Avec des conditions périodiques aux limites (toujours par simplicité), le partage de poids synaptique revient à faire de la matrice carrée $\mathbf{W}$ une matrice circulante, entièrement déterminée par sa première colonne $\boldsymbol{\psi}$; d'où

$$\mathbf{W}[t, t'] = \boldsymbol{\psi}\left[(t - t') \bmod N\right].$$

On modélise ainsi la propagation du signal $\boldsymbol{x}$ dans la couche neuronale (*layer*) comme une convolution rectifiée dont le noyau est $\boldsymbol{\psi}$, notée $\boldsymbol{y} = \rho(\boldsymbol{x} * \boldsymbol{\psi})$. Tandis que l'apprentissage de la matrice $\mathbf{W}$ requiert N^2 paramètres dans le cadre densément connecté $\boldsymbol{y} = \rho(\mathbf{W}\boldsymbol{x})$, le noyau $\boldsymbol{\psi}$ ne comporte que N paramètres indépendants au plus. On restreint habituellement le support de $\boldsymbol{\psi}$ à $T \ll N$ échantillons, ce qui accélère les calculs et réduit encore le nombre de degrés de liberté pendant l'apprentissage. En contrepartie de cette hypothèse de support compact, on juxtapose dans le même niveau du réseau des noyaux de convolution multiples $\boldsymbol{\psi}_\gamma$, indexés par une variable catégorielle discrète $\gamma \in \Gamma$. La réponse de la première couche se structure alors comme une matrice $\mathbf{Y}$ dont les entrées sont

$$\forall t \in \mathbb{Z}_N,\ \forall \gamma \in \Gamma,\ \mathbf{Y}[t, \gamma] = \rho\left(\boldsymbol{x} * \boldsymbol{\psi}_\gamma\right)[t] = \sum_{t'=0}^{N-1} \rho(\boldsymbol{x}[t']\boldsymbol{\psi}[(t - t') \bmod N]).$$

1.3 Équivariance à la translation

Nonobstant les questions d'optimisation et de supervision procurant la famille de noyaux $(\boldsymbol{\psi}_\gamma)_\gamma$, l'équation ci-dessus rappelle verbatim le codage parcimonieux convolutif, dont la convergence est mieux comprise théoriquement que l'apprentissage profond en toute généralité. D'autre part, le partage des poids fait apparaitre une propriété d'équivariance à la translation : décaler le signal temporel $\boldsymbol{x}$ d'un échantillon occasionne un décalage d'une ligne dans $\mathbf{Y}$. Formellement, le générateur $\mathcal{T}_N : \boldsymbol{x} \mapsto (\boldsymbol{t} \mapsto \boldsymbol{x}[t + 1] \bmod N)$ du groupe cyclique C_N commute avec $(\rho \circ \mathbf{W})$. Cette propriété algébrique d'équivariance est très générale et s'étend aisément aux convolutions multidimensionnelles, voire à certains groupes non abéliens.

L'équivariance permet en outre d'itérer l'application d'opérateurs convolutifs et de non-linéarités ponctuelles, constituant ainsi un circuit en boucle ouverte (*feedforward*). Par exemple, avec trois niveaux de convolutions et en définissant ρ comme la fonction de valeur absolue, on écrit

$$\forall t \in \mathbb{Z}_N,\ \forall \gamma_3 \in \Gamma_3,\ \mathbf{Y}[t, \gamma_3] = \left| \sum_{\gamma_2} \left| \sum_{\gamma_1} |\boldsymbol{x} * \boldsymbol{\psi}^{m=1}_{\gamma_1}| * \boldsymbol{\psi}^{m=2}_{\gamma_1, \gamma_2} \right| * \boldsymbol{\psi}^{m=3}_{\gamma_2, \gamma_3} \right| [t].$$

2 Position du problème

2.1 Convolutions temporelles ou temps–fréquence ?

Jusqu'à 2015, tous les systèmes d'apprentissage profond pour l'écoute artificielle optimisaient les convolutions pour $m \geq 2$, via telle ou telle variante de la descente de gradient stochastique[1]. La première couche, en revanche, est définie a priori comme un banc de filtres sur une échelle logarithmique (ou mel, soit quasi-logarithmique), et demeure inchangée pendant l'apprentissage. Les filtres passe-bande $\boldsymbol{\psi}_{\gamma_1}^{m=1}$ s'apparentent ainsi à une famille d'ondelettes, et l'ensemble fini Γ se munit d'une loi de composition interne vérifiant $\boldsymbol{\psi}_{\gamma_1+j}^{m=1}(t) = \alpha^j \boldsymbol{\psi}_{\gamma_1}^{m=1}(\alpha^j t)$ avec $1 < \alpha \leq 2$ un facteur constant d'échelle.

2.2 Le convnet bien tempéré

Récemment, la distribution de routines accélérées pour la convolution à grand support T par transformée de Fourier rapide (cuFFT) a stimulé l'intérêt de la communauté pour la question de la reconnaissance dans la forme d'onde $\boldsymbol{x}$ elle-même (*raw audio*) [2] que dans le scalogramme en ondelettes $|\boldsymbol{x} * \boldsymbol{\psi}_{\gamma_1}^{m=1}|$. En effet, il est à parier, contrairement à l'a priori des ondelettes, que la largeur de bande critique idéale pour l'écoute automatique ne soit pas une fonction affine de la fréquence ; et, qui plus est, qu'elle dépende du domaine sonore considéré : musique de telle ou telle culture, bioacoustique aérienne ou sous-marine, etc. Il est donc urgent de concevoir ce que j'appellerai un ń convnet bien tempéré ż, c'est-à-dire capable d'épouser la discrétisation qui exploitera optimalement l'équivariance à la transposition fréquentielle comme une translation le long de l'axe γ_1.

2.3 Instabilités numériques des convnets monorésolution

Pour autant, la construction d'une représentation parcimonieuse (*sparse*) pour les signaux naturels requiert des propriétés mathématiques précises quant au dictionnaire temps–fréquence : principe d'incertitude de Heisenberg, quadrature de phase, noyau reproduisant, moments dissipants, bornes de Riesz, incohérence mutuelle. Malheureusement, à cause d'un certain manque de dialogue avec l'analyse harmonique computationnelle, les recherches actuelles en apprentissage profond *stricto sensu*ignorent, pour la grande partie, de telles considérations. Par conséquent, les réseaux convolutifs sont exposés, a fortiori pour $T \sim 10^3$, à des instabilités numériques, par exemple le repliement de spectre (*aliasing*), les caractéristiques doublons (*duplicate features*), et la dissipation du gradient (*vanishing gradient*). Ces instabilités sont sources de surapprentissage, de gaspillage de ressources, et d'artefacts imprévisibles, à plus forte raison dans les modèles génératifs profonds.

3 Projet d'orientation de recherche

3.1 Synergie entre approximations multirésolution et réseaux convolutifs

Je propose d'améliorer le schéma numérique de discrétisation des réseaux de neurones convolutifs à l'aune de la théorie des approximations multirésolution (MRA) [3]. Ici, l'objectif de la MRA ne sera pas d'assurer une stricte orthonormalité de la famille de filtres $(\boldsymbol{\psi}_\gamma)_\gamma$ apprise par le convnet, mais plutôt d'exploiter le fait que ces filtres ont tendance à converger vers des passe-bande de facteur de qualité $Q \gg 1$ dépendant continûment de leur fréquence centrale. Ainsi, il est possible de projeter le signal d'entrée $\boldsymbol{x}$ sur des espaces d'approximation successifs $\mathbf{V}_j$ dont la résolution 2^{-j} décroit exponentiellement.

3.2 Allègement paramétrique

L'idée-clé est donc d'allouer un nombre $M_j \sim Q$ de filtres à l'échelle $j \in \mathbb{N}$, puis de les faire opérer sur les coefficients de détail de l'échelle j échantillonnés à la fréquence de Nyquist $2^{-j}\pi$, et ce récursivement pour tout $j \leq J = \Theta(\log T)$. On obtiendra ainsi un réseau de neurones multirésolution (*MUltiREsolution Neural Network* ou MuReNN) équivalent à un convnet traditionnel dont les filtres ont un support T en résolution fine ($j = 0$) et s'échelonnent à raison de M_j filtres par octave à l'échelle j. Pourtant, en vertu des sous-échantillonnages dyadiques successifs intervenant dans la pyramide multirésolution ($\mathbf{V}_j \subset \mathbf{V}_{j+1}$ pour tout j), le support discret de $\boldsymbol{\psi}_\gamma$ dans un MuReNN est de l'ordre de $Q \ll T$ et est indépendant de j. En prenant $M_j = Q = -\log \alpha$ constant par simplicité, on trouve $\Theta(Q^2 \log T)$ paramètres à optimiser dans un MuReNN contre $\Theta(QT \log T)$ dans un convnet à une dimension, d'où un allègement de $\Theta(T/Q)$, typiquement de l'ordre de 10^2. Cet allègement permettra, je l'espère, d'entrainer des MuReNN à vaste champ réceptif T et facteur de qualité Q élevé, quand bien même le nombre d'exemples annotés demeurerait très inférieur à $QT \log T$.

3.3 Accélération algorithmique

Avec des convolutions de Winograd et une méthode d'*overlap-add*, la complexité théorique du MuReNN est de $\Theta(QT \log Q)$ par couche. En comparaison, la complexité théorique du convnet est de $\Theta(QT \log^2 T)$, soit un facteur d'accélération de $\Theta(\log^2 T / \log Q)$ typiquement de l'ordre de 10^2. Ceci permettra d'envisager le portage de MuReNN sur des systèmes embarqués à ressources énergétiques limitées, tels que des téléphones portables alimentés par batterie [4].

Toutefois, la difficulté majeure de l'implantation efficace du MuReNN résidera dans la récursion sur j, là où un convnet peut tirer parti d'un parallélisme SIMD sur les transformées de Fourier rapides inverse des produits

$\widehat{\mathbf{W}\boldsymbol{x}}(\omega,\gamma) = \widehat{\boldsymbol{x}}(\omega)\widehat{\boldsymbol{\psi}}_\gamma(\omega)$. Il faudra donc que nous veillions à formaliser l'adéquation architecturealgorithme du MuReNN en amont des expériences d'apprentissage profond auquel il sera destiné.

3.4 Arbre dual et transformée en ondelettes discrète

La transformée en ondelettes complexe par arbre dual (DT-CWT) apparait comme l'outil d'approximation multirésolution idoine pour les MuReNN [5]. En effet, avec la méthode de l'arbre dual, les deux paires de filtres en quadrature conjugués produisent conjointement une ondelette discrète à support compact et à valeurs complexes dont la partie imaginaire approxime la transformée de Hilbert de sa partie réelle, soit un délai d'un quart d'échantillon (*q-shift*). Autrement dit, la projection de $\boldsymbol{x}$ sur les coefficients de détails de chaque échelle j est quasiment un signal analytique, dont la transformée de Fourier est supportée sur $\mathbb{R}_+$. En partageant les poids synaptiques de chaque filtre réel $\boldsymbol{\psi}_\gamma$ appris par le MuReNN à travers les deux parties de l'arbre dual, on garantit que chaque couche convolutive du MuReNN produit une sortie analytique, sans dégénérescence de phase ni repliement de spectre.

4 Perspectives futures

4.1 Contrôle du délai de groupe

Il est très important que l'opérateur convolutif $\mathbf{W}$ soit à phase linéaire, c'est-à-dire à délai de groupe constant. Dans le cas d'un appris dans le domaine temporel, il est difficile de contrôler les déphasages éventuels entre filtres $\boldsymbol{\psi}_\gamma$. En revanche, avec une discrétisation MuReNN, il suffit d'assurer que la réponse de chaque $\boldsymbol{\psi}_\gamma$ soit symétrique pour éviter ces déphasages. En effet, la transformée de Fourier de $\boldsymbol{\psi}_\gamma$ est alors à valeurs réelles. Puisque la transformée en ondelettes complexe par arbre dual est elle-même à phase approximativement linéaire, il en va de même après multiplication par les filtres $\widehat{\boldsymbol{\psi}_\gamma}$. Afin d'assurer cette symétrie, je propose simplement d'opérer un partage de poids synaptique entre moitié gauche et moitié droite du noyau de convolution $\widehat{\boldsymbol{\psi}_\gamma}$. Ceci réduit à $Q/2$ le nombre de degrés de liberté à l'apprentissage, et revient à définir le MuReNN comme une base de cosinus discrète. Je reconnais cependant que, dans le cas du traitement de la parole, l'asymétrie de l'enveloppe temporelle est importante, comme en témoigne les expériences classique de masquage psychoacoustique. La question d'autoriser les MuReNN à apprendre des filtres types Gammatone sans toutefois causer de délai de groupe est assez délicate.

4.2 Auto-supervision MuReNN : *Flip it and reverse it !*

Une seconde difficulté de l'entrainement des convnets temporels, et que l'on retrouve dans les MuReNN, réside dans la question de construire une tonotopie, c'est-à-dire un homéomorphisme entre fréquence physique ω et hauteur perçue

γ. Il serait intéressant de voir si on pourrait produire une brisure spontanée de symétrie (comme en physique des particules) à partir d'une initialisation gaussienne indépendante et identiquement distribuée. Pour inciter le MuReNN à respecter cette tonotopie, je propose d'envisager des tâches prétextes d'apprentissage auto-supervisé. L'une des hypothèses que j'ai consisterait à opérer un renversement des M_j bandes de fréquences dans chaque octave j : les gammes montantes deviennent descendantes et vice-versa (même si la notion de grave et d'aigu à travers les octaves reste conservée).

4.3 MuReNN en spirale

Dans le cadre d'une application au traitement de la musique, je propose de construire un "MuReNN en spirale", c'est-à-dire partageant les filtres $\boldsymbol{\psi}_\gamma$ à travers les octaves, s'adaptant ainsi automatiquement aux inégalités de tempérament du corpus musical analysé [6]. On pourrait le faire tourner, en mode supervisé ou non, sur des corpus musicaux de différentes époques ou cultures, et regarder le tempérament appris comme les fréquences centrales empiriques des $\boldsymbol{\psi}_\gamma$, sous hypothèse d'équivalence d'octave. Mon espoir est de trouver tempérament égal sur de la musique de piano seul, une solmisation en sargam sur de la musique hindoustanie, en maqam sur de la musique arabo-andalouse. En somme, un "convnet bien tempéré" (*well-tempered convnet*) comme voulu.

4.4 Diffusion MuReNN

Quand on aura bien compris le premier niveau des MuReNN, on pourra passer au second, à l'instar du passage des MFSC aux spectre de diffusion profonde (*deep scattering spectrum*) par Andén et Mallat. Avec $\rho : z \mapsto |z|$ pour non-linéarité ponctuelle, chaque couche du MuReNN aura pour but de démoduler les oscillations de la partie analytique de $(\boldsymbol{x} * \boldsymbol{\psi}_\gamma)$ pour tout γ. J'espère que la composition profonde de couches MuReNN atteindra des échelles d'invariance grandissant exponentiellement avec le nombre de non-linéarités ponctuelles, à la manière d'un réseau de diffusion en ondelettes (*wavelet scattering network*). Il serait intéressant de comparer les propriétés théoriques des MuReNN avec celles, aujourd'hui maitrisées, de la diffusion en ondelettes.

4.5 Décodeur MuReNN

Enfin, on pourra envisager d'intégrer le MuReNN à des modèles génératifs profonds, comme des réseaux génératifs adversariaux, des autoencodeurs variationnels, ou des transformeurs. En l'absence de non-linéarité ρ, c'est très facile : il suffit, pour la partie réelle comme pour la partie imaginaire, d'appliquer les opérateurs adjoints $\boldsymbol{\psi}_\gamma^*$, puis enfin les transformées en ondelettes discrètes inverses correspondant aux deux branches de l'arbre dual. Mais quand on n'a qu'une version moyennée et sous-échantillonnée du module, il faut précéder cette opération par un sur-échantillonnage et surtout une reconstruction de phase. Il faudra réfléchir à une méthode de reconstruction de phase pertinente.

Le Griffin–Lim accéléré donne des résultats intéressants mais l'idéal, notamment pour des applications en musique par ordinateur et en synthèse de parole (*text to speech*), serait de pouvoir générer le signal sans artefacts en boucle ouverte, ceci sans besoin de rétroaction.

Références

[1] Monika Dörfler, Thomas Grill, Roswitha Bammer, and Arthur Flexer. Basic filters for convolutional neural networks applied to music : Training or design ? *Neural Computing and Applications*, pages 1–14.

[2] Neil Zeghidour. *Learning representations of speech from the raw waveform.* PhD thesis, Université PSL, 2019.

[3] Stéphane Mallat. A theory for multiresolution signal decomposition : The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (7) :674–693, 1989.

[4] Nicholas D Lane, Sourav Bhattacharya, Akhil Mathur, Petko Georgiev, Claudio Forlivesi, and Fahim Kawsar. Squeezing deep learning into mobile and embedded devices. *IEEE Pervasive Computing*, 16(3) :82–88, 2017.

[5] Ivan W. Selesnick, Richard G. Baraniuk, and Nicholas G Kingsbury. The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6) :123–151, 2005.

[6] Vincent Lostanlen and Carmine-Emanuele Cella. Deep convolutional networks on the pitch spiral for musical instrument recognition. In *Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference*, 2016.