MODÈLES NEURONAUX DE QUANTIFICATION VECTORIELLE, APPRENTISSAGE CONTINU ET DISTRIBUTIONS NON-STATIONNAIRES

NEURAL MODELS FOR VECTOR QUANTIZATION, CONTINUAL LEARNING AND NON-STATIONARY DISTRIBUTIONS

Etablissement **Université de Lorraine**

École doctorale **IAEM - INFORMATIQUE - AUTOMATIQUE - ELECTRONIQUE - ELECTROTECHNIQUE - MATHEMATIQUES**

Spécialité **Informatique**

Unité de recherche **LORIA - Laboratoire Lorrain de Recherche en Informatique et ses Applications**

Encadrement de la thèse **Bernard GIRAU**

Co-Encadrant **Hervé FREZZA-BUET**

Financement du 01-10-2024 au 30-09-2027 *origine* **projet ANR SORLAHNA** *Employeur* **UNIVERSITE DE LORRAINE**

Début de la thèse le **1 octobre 2024**

Date limite de candidature (à 23h59) **16 mai 2024**

Mots clés - Keywords

réseaux de neurones, quantification vectorielle, cartes auto-organisées, apprentissage continu

neural networks, vector quantization, self-organizing maps, continual learning

Description de la problématique de recherche - Project description

Ce sujet fait partie du projet ANR SORLAHNA, dont les deux objectifs sont : 1) concevoir une méthodologie d'implantation matérielle suffisamment flexible pour des architectures neuronales versatiles de quantification vectorielle (VQ) topologique, sur la base d'un NoC (network on chip) capable d'instancier une topologie virtuelle dynamique sur un circuit programmable, et 2) définir des modèles de VQ topologique pour le recueil dynamique des données issues d'un réseau distribué de capteurs et destinées à un apprentissage machine. C'est sur ce second aspect que porte cette thèse.

Dans une première phase, l'étudiant analysera les différents algorithmes neuronaux de VQ topologique existants (SOM, NG), ainsi que leurs différentes variantes proposées en lien avec des données non stationnaires ou d'éventuelles implantations matérielles (DSOM, CSOM [4] , NP-SOM [5], PSOM [6], GNG, GWR, etc.). En interaction avec différents chercheurs et étudiants impliqués dans le projet Sorlahna, l'étudiant analysera le potentiel des méthodes de quantification vectorielle à traiter des problèmes de CL, en définissant notamment une première série de critères pour qualifier et quantifier les performances des algorithmes de VQ confrontés à un problème de CL (sur la base d'un ensemble prédéfini de situations artificielles contrôlées d'apprentissage continu). La plupart des travaux actuels sur le CL considèrent un problème dont la statistique est stationnaire, dont on pourrait extraire des échantillons de façon i.i.d, mais pour lesquels ces échantillons sont fournis, en pratique et pour des raisons techniques, de façon non i.i.d. L'exemple typique est celui d'une classification où l'on fournit au préalable les échantillons de certaines classes uniquement, puis les échantillons d'autres classes dans un second temps, alors que le but est bien de traiter un problème de classification sur toutes ces classes. Nos travaux visent à étendre le problème du CL à des cas où la statistique est intrinsèquement non-stationnaire, ce qui amène à considérer des problèmes dont la formalisation elle-même inclut une dimension temporelle. Les situations et critères issus de cette première phase devront donc être représentatifs de problèmes où le caractère non i.i.d. des données d'apprentissage est aussi bien dû au processus d'échantillonnage des données qu'à la distribution sous-jacente elle-même.

Dans une seconde phase, l'étudiant étudiera les méthodes et propriétés permettant de (ré)générer des données représentatives du problème à partir des seules informations issues de la VQ topologique. Cette question dépend directement de la capacité des modèles de VQ à 'bien' modéliser la densité de probabilité de l'espace d'entrée, mais elle devient plus complexe dans un contexte de CL et d'évolution dynamique de la distribution des données. La validation de la capacité (ré)générative de nos algorithmes de VQ topologique ouvrira la voie à la troisième phase des travaux, dans laquelle des protocoles de transmission dynamique des informations de VQ seront mis au point afin d'optimiser la récolte de données issues de capteurs physiquement distribués, lorsque l'information recherchée n'est pas la connaissance des valeurs exactes perçues, mais la connaissance de leur distribution, comme cela peut être le cas lors d'un apprentissage machine. Par ailleurs, la nature topologique de la VQ réalisée par nos modèles pourra être exploitée pour optimiser la transmission des informations décrivant les données perçues.
En résumé, cette thèse vise à généraliser le concept du Continual Learning au cas des distributions non-stationnaires pour les

algorithmes de quantification vectorielle avec préservation de topologie, dans un contexte où cette généralisation doit être compatible avec une implémentation matérielle de ces algorithmes.

This subject is part of the ANR SORLAHNA project, whose two objectives are: 1) to design a sufficiently flexible hardware implementation methodology for versatile neural architectures of topological vector quantization (VQ), based on a NoC (network on chip) able to instantiate a dynamic virtual topology on a programmable circuit, and 2) to define topological VQ models for the dynamic acquisition of data recorded from a distributed network of sensors and intended for machine learning. This thesis focuses on this second aspect.

In a first phase, the student will analyze the different existing topological VQ neural algorithms (SOM, NG), as well as their different variants proposed for non-stationary data or possible hardware implementations (DSOM, CSOM [4], NP-SOM [5], PSOM [6], GNG, GWR, etc.). In interaction with different researchers and students involved in the Sorlahna project, the student will analyze how vector quantification methods can deal with CL problems, in particular by defining a first series of criteria to qualify and quantify the performances of the VQ algorithms confronted to a CL problem (based on a predefined set of controlled artificial situations for continuous learning). Most current work on CL considers a problem whose statistics are stationary, from which samples could be extracted in an i.i.d manner, but for which these samples are provided in a non-i.i.d manner, for practical or technical reasons,. The typical example is that of a classification where we first provide samples from certain classes only, then samples from other classes in a second step, whereas the goal is to deal with a classification problem on all these classes. Our work aims to extend the CL problem to cases where statistics are intrinsically non-stationary, which leads to consider problems whose formalization itself includes a temporal dimension. The situations and criteria resulting from this first phase must therefore be representative of problems where the non-i.i.d. character of training data is due to both the data sampling process and the underlying distribution itself.

In a second phase, the student will study the methods and properties that make it possible to (re)generate data only from the information encoded in the topological VQ, such data having to be statistically representative of the problem. This question directly depends on the ability of VQ models to “well” model the probability density of the input space, but it becomes more complex in a context of CL and dynamic evolution of the data distribution. Validation of the (re)generative capacity of our topological VQ algorithms will pave the way for the third phase of the work. In this phase, protocols for the dynamic transmission of VQ information will be defined in order to optimize data acquisition from physically distributed sensors, when the information that is sought is not the knowledge of the exact values perceived, but the knowledge of their distribution, as it can be the case during machine learning tasks. Furthermore, the topological nature of the VQ performed by our models can be exploited to optimize the transmission of information describing the perceived data.
In summary, this thesis aims to generalize the concept of Continual Learning to the case of non-stationary distributions for vector quantization algorithms with topology preservation, in a context where this generalization must be compatible with a hardware implementation of these algorithms.

Thématique / Contexte

La quantification vectorielle (VQ) consiste à modéliser la densité de probabilité d'un espace d'entrée (souvent connue grâce à un large ensemble d'échantillons) avec un ensemble fini de vecteurs prototypes, de telle sorte que n'importe quel point de l'espace d'entrée puisse être associé de manière satisfaisante à un prototype. Un modèle de VQ est dit topologique lorsqu'il projette simultanément une structure de voisinage sur les prototypes.

Cette structure apprise est liée à la topologie, inconnue, de la variété d'où sont tirés les échantillons. Cet apprentissage étend les algorithmes de quantification vectorielle en leur conférant des propriétés dites 'd'auto-organisation'.
L'apprentissage continu (continual learning, CL) se différencie de l’apprentissage automatique plus classique dans la mesure où les données sont fournies au fil de l’eau, sans la classique hypothèse i.i.d. ([variables aléatoires] indépendantes et identiquement distribuées). En effet, dans un cadre CL, la statistique des données présentées au fil de l'eau varie avec le temps (on parle du passage d'une tâche à une autre, par exemple).

Un apprentissage machine classique, qui repose sur l’hypothèse d’échantillons i.i.d., va dans ce cas adapter son apprentissage à la nouvelle distribution des données, considérant qu'il faut adapter le modèle à une statistique des échantillons qui a changé sans plus tenir compte de la distribution précédemment apprise. Cette adaptation est ce qui est qualifié comme un « oubli catastrophique ». Le CL est un domaine de l’apprentissage machine qui vise à produire des algorithmes qui peuvent apprendre de plus en plus de choses en ligne sans souffrir du caractère non i.i.d des données fournies. Ce point soulève des difficultés algorithmiques particulières. Par exemple, alors que les approches neuronales ont souvent des taux d'apprentissage qui décroissent, pour explorer d'abord la statistique, et consolider ensuite la représentation de cette statistique qui n'a pas changé, les algorithmes capables de s'adapter en permanence aux changements ont à garder une flexibilité constante, au détriment de leur stabilité. Les difficultés peuvent être aussi architecturales, avec la nécessité d'adapter la structure même du réseau de neurones.

L’équipe BISCUIT du LORIA s’intéresse aux paradigmes de calcul distribués à grain fin (modèles d’inspiration neuronale par exemple), considérant notamment leur capacité à s’instancier sur support matériel numérique lorsque des gains en vitesse de calcul et en consommation d'énergie sont visés. Le lien avec l’apprentissage automatique se fait au niveau des approches neuronales de quantification vectorielle (VQ) topologique car ces approches sont des exemples de systèmes à grain fin dont le calcul distribué réalise une forme d’auto-organisation.

L’équipe BISCUIT du LORIA et l’équipe MEA de l’IJL ont entamé une collaboration (projet ANR SORLAHNA) autour de la question de l’implémentation électronique de méthodes de VQ topologique (cartes auto-organisatrices SOM [1], gaz neuronaux NG [2], ...) pour

l'apprentissage continu (Continual Learning, CL [3]). Dans ce projet, l'accent est mis sur deux axes de recherche : 1) l'implémentation matérielle efficace de ces algorithmes même en cas de changements dynamiques de leur topologie sous-jacente, et 2) la question encore trop peu abordée du rôle que peuvent jouer les algorithmes de VQ topologique suffisamment adaptatifs dans le cadre de l'apprentissage continu.

Références bibliographiques

[1] T. Kohonen. Self-organized formation of topologically correct feature maps in Biological Cybernetics, vol. 43(1), pp. 59–69, 1982.
[2] B. Fritzke. A growing neural gas network learns topologies In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, Advances in Neural Information Processing Systems 7, pages 625 632. MIT Press, Cambridge MA, 1995.
[3] L. Wang, X. Zhang, H. Su, and J. Zhu. A comprehensive survey of continual learning : Theory, method and application. arXiv, 2023.
[4] B. Girau and A. Upegui. Cellular Self-Organising Maps - CSOM. 13th Int. Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM), 2019.
[5] Y. Bernard et al. NP-SOM: network programmable self-organizing maps. IEEE 30th Int. Conf. on Tools with Artificial Intelligence (ICTAI), 2018.
[6] A. Upegui et al. Pruning Self-Organizing Maps for Cellular Hardware Architectures. 12th NASA/ESA Conference on Adaptive Hardware and Systems (AHS), 2018.

Précisions sur l'encadrement - Details on the thesis supervision

Comité de suivi individuel de thèse tel que défini par l'ED IAEM

Conditions scientifiques matérielles et financières du projet de recherche

Contrat doctoral type, Université de Lorraine. Voir arrêté https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000046820745.

Objectifs de valorisation des travaux de recherche du doctorant : diffusion, publication et confidentialité, droit à la propriété intellectuelle,...

Publications dans des conférences et revues internationales. Réalisation de démonstrateurs.

Profil et compétences recherchées - Profile and skills required

Le candidat doit avoir l'équivalent d'un Master en informatique, de préférence dans une spécialité liée à l'intelligence artificielle, aux réseaux de neurones et/ou au calcul numérique distribué. Une expérience de la conception logicielle est requise, et des bases solides en probabilités et statistiques seront appréciées. Le candidat doit parler couramment l'anglais et/ou le français.
The candidate should have the equivalent of a Master's degree in Computer Science, preferably in a specialty related to artificial intelligence, neural networks and/or distributed numerical computation. Experience in software design is required, and a solid knowledge of probability and statistics will be appreciated. The candidate must be fluent in English and/or French.

Dernière mise à jour le 9 avril 2024