Sujet de recherche :

Quantification d'incertitude dans les réseaux de neurones pour des applications critiques

24 octobre 2023

Porteurs : Mathieu Segond & Laurent Lefebvre (Framatome)
Nicolas Bousquet & Bertrand Iooss ou Merlin Keller (EDF)
mathieu.segond@framatome.com
laurent.lefebvre@framatome.com
merlin.keller@edf.fr

Stade de Maturité : formalisation mathématique mature, analyse bibliographique déjà effectuée, approche agnostique au formalisme mathématique, disponibilité de jeux de données d'applications critiques pour le secteur industriel

---

1 Contexte

Les réseaux de neurones sont des modèles d'apprentissage statistique capables de traiter des données complexes et d'apprendre des relations non linéaires entre des entrées et des sorties d'intérêt. Ils sont souvent utilisés dans des applications de classification ou de régression, de reconnaissance d'objets et de traitement du langage naturel.

Cependant, les réseaux de neurones, comme tous les algorithmes d'apprentissage peuvent être sujets aux incertitudes, en particulier en présence de données bruitées, de données incomplètes ou de données différentes de celles utilisées pour entraîner l'algorithme. L'incertitude peut se manifester de différentes manières, notamment par des erreurs de classification, des prédictions erronées ou des scores de confiance faibles.

La quantification d'incertitude est le processus d'estimation de l'incertitude associée à une mesure ou à une estimation. Elle est souvent utilisée dans les domaines de l'apprentissage statistique et de la science des données, où il est important de pouvoir prendre en compte l'incertitude des données et des modèles. La quantification d'incertitude peut donc être appliquée pour estimer l'incertitude d'un réseau de neurones. Cette incertitude peut ensuite être utilisée pour améliorer la robustesse du réseau aux données bruitées et incomplètes ou afin de prendre des décisions plus éclairées.

La littérature [10] regroupe généralement les méthodes d'estimation de l'incertitude en différents types, en fonction de la nature (déterministe ou stochastique) du modèle ou du nombre de propagation avant (ou inférence) :

Méthodes déterministes simples - Single deterministic methods (par exemple [16]) : Ces méthodes donnent une prédiction sur la base d'une seule inférence dans un réseau déterministe, et la quantification de l'incertitude est soit dérivée à l'aide de méthodes externes (ou extraites du modèle) soit directement prédite par le réseau.

Méthodes Bayésiennes - Bayesian methods (par exemple [3]) : Ces méthodes traitent les poids des réseaux de neurones comme des distributions de probabilité au lieu de valeurs fixes. Des techniques comme l'échantillonnage Monte-Carlo (Monte-Carlo par Chaîne de Markov ou Monte-Carlo Hamiltonien, etc.), l'approximation de Laplace ou l'inférence variationnelle sont utilisées pour approcher la distribution a posteriori (non connue) des poids. Cette dernière permet ensuite de capturer l'incertitude des poids du réseau, et donc l'incertitude de prédictions.

Dropout Monte-Carlo - Monte-Carlo Dropout [9] : Le dropout est une technique de régularisation couramment utilisée dans les réseaux de neurones. Pendant l'entraînement, le dropout met aléatoirement à zéro une fraction des neurones. En effectuant plusieurs inférences avec le dropout activé, la variabilité des prédictions peut être utilisée pour estimer l'incertitude.

Méthodes d'ensembles profonds - Deep Ensembles [12] : Ces méthodes consistent à entraîner plusieurs réseaux de neurones avec des initialisations ou des architectures différentes et à moyenner leurs prédictions. La variabilité entre les prédictions des différents réseaux peut être utilisée pour estimer l'incertitude. On peut noter que les ensembles profonds peuvent être considérés comme Bayésiens, car ils forment une somme de distributions *delta* qui approchent la loi a posteriori [18].

Lissage Gaussien des poids - Stochastic Weight Averaging Gaussian (SWAG) [15] : Cette méthode consiste à moyenner les poids du modèle obtenus au cours des dernières étapes de l'entraînement, lorsque les poids ont tendance à converger vers un plateau. Ce lissage permet de réduire le sur-apprentissage et d'améliorer la capacité du modèle à faire des prédictions plus confiantes. Pour faire des prédictions sur de nouvelles données, plusieurs inférences sont effectuées en utilisant chaque échantillon de poids. Les prédictions de ces échantillons sont ensuite moyennées pour obtenir la prédiction finale, la variance donne l'incertitude.

Ré-échantillonnage - Bootstrapping (par exemple [13]) : Le Bootstrapping est une technique de ré-échantillonnage où plusieurs ensembles de données sont créés en échantillonnant avec remise à partir de l'ensemble de données original. Chaque ensemble de données est ensuite utilisé pour entraîner un réseau de neurones distinct. La variabilité entre les prédictions de ces modèles peut être utilisée pour estimer l'incertitude.

Méthodes d'augmentation des données - Test-time augmentation methods [14] : Ces méthodes donnent la prédiction sur la base d'un seul réseau déterministe, mais elles utilisent les techniques d'augmentation de données afin de générer plusieurs prédictions qui sont utilisées pour évaluer l'incertitude.

Théorie des prédictions conformes - Conformal Prediction Theory (par exemple [1]) : Cette théorie propose un cadre pour la construction d'intervalles de prédiction ou de régions avec une validité garantie. Elle repose sur l'idée d'utiliser un ensemble d'apprentissage pour générer un ensemble de "scores de conformité" qui capturent la similarité entre un exemple de test et les exemples d'apprentissage. Ces scores de conformité sont utilisés pour quantifier l'incertitude associée à chaque prédiction. Dans le contexte des réseaux de neurones, l'approche conforme consiste à entraîner le réseau sur un ensemble d'apprentissage et à générer des scores de conformité pour chaque exemple de test. Ces scores de conformité peuvent être basés sur diverses mesures, telles que l'erreur de prédiction du réseau, la distance aux voisins les plus proches ou des distances statistiques.

On peut également citer d'autres techniques de quantification d'incertitude : l'approche masque-ensemble - Masksembles [6], les réseaux Bayésiens de neurones latent-posterior - Latent-Posterior Bayesian Neural Network [8], les agrégats d'ensemble - Batch Ensemble [17], l'approche hypermodèle - Hypermodel approach [7] et l'approche par couche - Depth uncertainty [2]. Des travaux de [4, 11] s'intéressent également aux garanties théoriques du comportement des prédictions d'un réseau de neurones basées sur les propriétés de Lipchitz sous-jacentes au réseau.

Toutes ces techniques proposent différentes manières d'estimer l'incertitude dans les réseaux de neurones, et le choix de la méthode dépend de l'application spécifique et des ressources disponibles. Il est important de noter qu'il existe des efforts de recherche en cours dans ce domaine, et de nouvelles techniques pourraient émerger dans le futur.

Pour une revue détaillée, nous conseillons les travaux de Gawlikowski et al. [10].

2 Problème scientifique

Il existe de multiples approches pour estimer l'incertitude de prédictions d'un réseau de neurones. Cependant, il reste de nombreuses questions en suspens si ces méthodes doivent être utilisées pour des applications critiques comme l'aéronautique, le nucléaire, le médical, la conduite autonome, etc., i.e. toutes applications qui peuvent avoir un im-

pact conséquent sur la vie humaine (à noter que l'Europe publiera un texte réglementaire fin 2023 nommé "AI Act" qui définira un cadre quant à l'utilisation de l'IA en particulier pour des applications critiques, cf. [5]).

En effet, il est important d'évaluer à la fois la qualité de prédictions d'un modèle (via des métriques de performance comme l'erreur moyenne, le taux de variance expliquée, la précision, etc.) ainsi que la qualité de l'incertitude estimée (via des métriques comme la log-vraisemblance, la probabilité de couverture de l'intervalle de prédiction, la largeur moyenne de l'intervalle de prédiction, etc.). Autrement dit, il est indispensable de s'assurer que la prédiction et l'estimation de l'incertitude du modèle sont de qualité i.e. il est nécessaire de savoir si on peut avoir confiance dans la prédiction et dans l'incertitude. Cette question est d'autant plus compliquée lorsque l'algorithme est utilisé avec des données éloignées de la base d'apprentissage. On parle de décalage de distribution - Distribution shift quand la distribution des nouvelles données est (partiellement) différente de celle des données d'apprentissage, ou hors distribution - Out Of Distribution quand les nouvelles données n'ont rien à voir avec les données d'apprentissage. Ces cas de figure apparaissent quand l'algorithme est déployé pour des scénarios "du monde réel" qui peuvent différer du cadre construit lors de la phase d'apprentissage.

Ainsi, dans le cadre de scénarios réels impliquant des données hors distribution, la question de la qualité de l'estimation de l'incertitude est d'autant plus cruciale.

La littérature mentionnée précédemment évoque quelques pistes de recherche concernant l'estimation de l'incertitude dans un contexte de données hors distribution (l'approche conforme par exemple). Mais un travail de recherche reste nécessaire dans un contexte de données hors distribution non connues à l'avance et pour lesquelles les méthodes actuelles de quantification d'incertitude sont peu robustes.

3 Applications

On souhaite pouvoir estimer l'incertitude de prédictions de réseaux de neurones pour des applications critiques, comme en sûreté nucléaire. En effet, ce type d'algorithme est extrêmement performant et pourrait améliorer la sûreté de centrales nucléaires, notamment s'il était utilisé pour reconstruire des fonctions importantes pour la sûreté intégrées au système de protection ou de surveillance. En outre, les réseaux de neurones sont parfaitement adaptés aux systèmes embarqués utilisés dans les centrales.

Néanmoins, pour pouvoir prétendre un jour à utiliser des réseaux de neurones dans le secteur nucléaire, il conviendra de répondre à de nombreuses questions qui seront notamment soulevées par l'autorité de sûreté nucléaire.

L'une de ses questions concernera la qualité de l'incertitude estimée, en phase d'apprentissage, de validation et pour des scénarios impliquant des données réelles qui seront hors distribution.

Une autre question concernera l'utilisation de ces méthodes dans des applications embarquées pour lesquelles il existe des contraintes matérielles, de capacité de mémoire, de puissance de calculs, etc. qui limiteront peut-être les méthodes de quantification d'incertitude.

Le travail de recherche aura pour but de répondre à ces questions pour des applications critiques.

Il est à noter que les problématiques précédentes sont rencontrées par beaucoup de secteurs industriels. Le travail de recherche, ici appliqué au secteur nucléaire, pourra donc être utilisé par d'autres acteurs, comme l'aéronautique, le spatial, le médical, etc.

4 Intérêt des Parties

EDF et FRAMATOME souhaitent s'associer pour co-porter ce sujet de thèse (précédé d'un stage), avec des chercheurs académiques membres des Établissements Paris Saclay, car ils partagent des intérêts forts en termes de maîtrise technique et d'application à des cas d'étude. Plus précisément :

- **Pour FRAMATOME** : FRAMATOME souhaite, par le biais d'une application sur un démonstrateur industriel, améliorer le benchmark des méthodes de quantification d'incertitudes (et sélectionner des méthodes appropriées pour les cas industriels) afin d'accompagner des études de *bilan de marges de fonctionnement*. En effet, le licensing d'outils technologiques ne peut être réalisé sans une information forte apportée à la gestion des incertitudes.
- **Pour EDF** : la montée en compétence sur les nouvelles approches de la quantification d'incertitude pour l'usage des réseaux de neurones, renforcée par la manipulation de démonstrateur industriel, est importante pour préparer l'usage des outils de demain. Des applications fortes portent notamment sur l'accélération des études de sensibilité faites par méta-modèle de type réseau de neurones, avec des applications en thermohydraulique des composants et systèmes.

5 Type de collaboration souhaitée

1. Stage bibliographique en co-tutelle à partir de **mars 2024**, pouvant être mené dans un laboratoire académique, puis thèse
2. Recherche d'un partenaire académique appartement aux Etablissements de l'Université Paris Saclay
3. Demande de cofinancement COFUND DeMythif.ai

Références

[1] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022.

[2] Javier Antorán, James Urquhart Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks, 2020.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks, 2015.

[4] Patrick L. Combettes and Jean-Christophe Pesquet. Lipschitz certificates for layered network structures driven by averaged activation operators. *SIAM Journal on Mathematics of Data Science*, 2(2) :529–557, 2020.

[5] European Commision. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206`.

[6] Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation, 2021.

[7] Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Ian Osband, Zheng Wen, and Benjamin Van Roy. Hypermodels for exploration, 2020.

[8] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Severine Dubuisson, and Isabelle Bloch. Encoding the latent posterior of bayesian neural networks for uncertainty quantification, 2021.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation : Representing model uncertainty in deep learning, 2016.

[10] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks, 2022.

[11] Kavya Gupta, Fateh Kaakai, Beatrice Pesquet-Popescu, Jean-Christophe Pesquet, and Fragkiskos D. Malliaros. Multivariate lipschitz analysis of the stability of neural networks. *Frontiers in Signal Processing*, 2, 2022.

[12] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[13] Juho Lee, Yoonho Lee, Jungtaek Kim, Eunho Yang, Sung Ju Hwang, and Yee Whye Teh. Bootstrapping neural processes, 2020.

[14] Dmitry Molchanov, Alexander Lyzhov, Yuliya Molchanova, Arsenii Ashukha, and Dmitry Vetrov. Greedy policy search : A simple baseline for learnable test-time augmentation, 2020.

[15] Florian Seligmann, Philipp Becker, Michael Volpp, and Gerhard Neumann. Beyond deep ensembles : A large-scale evaluation of bayesian deep learning under distribution shift, 2023.

[16] Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning, 2019.

[17] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble : An alternative approach to efficient ensemble and lifelong learning, 2020.

[18] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization, 2022.