Interprétabilité des Réseaux de neurones récurrents

Encadrant :Sylvain Lamprier (Sylvain.Lamprier@lip6.fr)

Compétence souhaitées : des compétences minimales en apprentissage statistique et optimisation sont requises.Un bon niveau de programmation est demandé, la maîtrise d'une librairie de *deep learning* (Torch. TensorFlow, etc.) est un plus.

Depuis quelques années, l'emploi de réseaux de neurones récurrents (RNNs) pour la modélisation de séquences d'observations s'est très largement développé. L'objet du stage est d'explorer la possibilité d'extraire des automates probabilistes des représentations apprises par des RNNs sur différentes séries temporelles observées, afin d'en améliorer l'interprétabilité, tout en limitant au maximum la perte en efficacité prédictive des modèles considérés.

Contexte

D'une manière générale, le problème de prédiction dans les séquences revient à un problème de modélisation de dépendances temporelles, dans lequel la distribution de probabilités de chaque observation dépend de l'ensemble des observations qui la précèdent. Malheureusement, le nombre de dépendances à considérer devient très rapidement prohibitif lorsque la taille des séquences augmente, ce qui induit des problèmes de complexité et de généralisabilité des modèles appris. Traditionnellement, les modèles à dépendances temporelles se restreignent alors à la considération d'un nombre limité d'observations passées, suivant une hypothèse de Markov stipulant que la distribution de chaque observation ne dépend que d'un passé de taille n (modèles n-grammes). Cette approximation permet l'établissement de modèles efficaces pour de nombreuses tâches impliquant des séries temporelles (modélisation de la langue, de données financières, de signaux issus de capteurs physiques, etc.), mais conduit à de faibles performances lorsque d'importantes dépendances temporelles de long terme existent dans les données (par exemple, lorsque les données sont très bruitées et que le passé immédiat ne comporte que peu d'information ou lorsque des événements rares impactent fortement la séquence d'observations). Les réseaux de neurones récurrents apportent une réponse à ce genre de limitation par leur capacité à considérer des historiques de plus grande taille, notamment depuis l'apparition de modèles tels que les modèles LSTM ou GRU [1] qui permettent de conserver une mémoire à long terme dans leur représentation de l'état courant du système.

L'idée générale des réseaux récurrents est de définir des cycles entre les connections de leurs différents neurones, de manière à être capable de traiter des séquences de tailles variables, avec des paramètres appris à la fois concernant 1) l'impact des nouvelles observations sur l'état courant, 2) la gestion de la mémoire des observations passées et 3) les prédictions sur les observations futures. Ces modèles maintiennent une représentation continue de l'état du système permettant l'intégration des nouveaux événements et la prise en compte de l'historique à plus ou moins long terme pour la détermination des distributions sur les observations à venir. Ce genre de modèles est à comparer avec les modèles type Markov caché (HMM) qui supposent également l'existence d'états latents du système permettant d'expliquer les événements observés, mais se limitent à la considération d'un ensemble d'états fini (représentation discrète des états possibles du système). Néanmoins, le temps requis pour l'apprentissage des HMMs par des algorithmes de programmation dynamique type Viterbi, ainsi que l'espace de stockage des probabilités de transition entre états, augmentent quadratiquement avec le nombre d'états définis pour représenter le système. En outre, le nombre d'états nécessaires augmente exponentiellement avec la taille de l'historique que l'on souhaite considérer [1].
Pour toutes ces raisons, les réseaux de neurones récurrents paraissent fournir un cadre bien plus puissant pour la modélisation de séries temporelles complexes que les HMMs. D'un autre côté,

l’emploi d’une représentation continue des états telles que celle qu’ils manipulent rend difficile l’interprétabilité des dépendances apprises, contrairement à des modèles plus explicatifs tels que les HMMs. Cette difficile interprétabilité peut être un frein à l’utilisation de ces modèles, les prédictions proposées pouvant paraître quelque peu obscures pour l’utilisateur peu averti. Une nouvelle directive européenne impose en outre que l’ensemble des décisions prises concernant un usager doivent pouvoir lui être expliquées de manière claire [2]. Bien que le cadre de cette directive semble encore assez flou, sujet à interprétations, elle souligne en tout cas le besoin de définir des modèles statistiques dont les distributions qu’ils encodent puissent être clairement présentées et analysées. Une application concrète qui requiert des modèles explicatifs tels que les automates probabilistes produits par les HMMs concerne la retro-ingénierie logicielle dynamique, pour laquelle le problème est de produire un diagramme comportemental à partir de traces issues d’exécutions multiples d’un système, afin d’en représenter son fonctionnement de manière concise.

Description du projet

L’objectif de ce stage est d’étudier la possibilité d’extraire des représentations explicites, type automate probabiliste, de modèles de réseaux de neurones récurrents appris a priori. L’idée est d’essayer de tirer parti de la puissance expressive des RNNs pour l’apprentissage de dépendances temporelles complexes, tout en se ramenant à la définition de modèles à états discrets, bien plus interprétables par l’usager. On espère conserver les bénéfices d’un apprentissage continu malgré une discrétisation du modèle a posteriori. Le projet s’articulera autour des grandes étapes suivantes :

- État de l’art autour de la modélisation de séries temporelles, notamment par HMMs et RNNs, ainsi que sur les approches de la littérature pour servir l’interprétabilité des modèles statistiques type réseaux de neurones ;
- Définition d’une première approche de discrétisation qui consiste en l’application d’une grille sur l’espace de représentation. L’hypothèse est que les distributions prédictives définies selon les zones de l’espace respectent une certaine régularité qu’il est possible d’exploiter. Selon les probabilités de prédiction des différentes cellules de la grille et des effets des observations sur la représentation du système, il est alors possible de déterminer des transitions probabilistes entre un ensemble d’états finis. On évaluera l’approche par comparaison avec des méthodes type HMMs pour des tâches de modélisation du langage et/ou de retro-ingénierie logicielle.
- L’hypothèse qui consiste à dire que l’espace de représentation peut être découpé efficacement de manière régulière selon une grille n’est certainement pas réaliste. Il existe très probablement des zones dans lesquelles les distributions évoluent plus fortement que dans d’autres. Il s’agit alors d’essayer de mieux répartir les états discrets sur l’espace de représentation, de façon à se maximiser l’homogénéité des distributions sur les cellules définies en fonction des positions des états. Différentes approches pourront être envisagées, notamment par des algorithmes itératifs de déplacement des états dans l’espace ou par divisions successives de l’espace de représentation selon une mesure type entropie exprimant la divergence des distributions encodées au sein des différentes cellules.

[1] Zachary Chase Lipton: A Critical Review of Recurrent Neural Networks for Sequence Learning. CoRR abs/1506.00019 (2015)
[2] Bryce Goodman, Seth Flaxman: EU regulations on algorithmic decision-making and a "right to explanation". CoRR abs/1606.08813 (2016)