

Thèse CIFRE Modélisation data driven pour la conception de systèmes complexes

INTRODUCTION

Cette thèse CIFRE sera réalisée à Toulouse de la cadre de la participation de Liebherr à l'institut d'intelligence artificielle ANITI. L'étudiant(e) aura ainsi accès à l'environnement industriel et ses enjeux ainsi qu'à l'excellence académique en matière d'IA.

Présentation ANITI

ANITI, Artificial and Natural Intelligence Toulouse Institute, est l'institut interdisciplinaire d'intelligence artificielle de Toulouse. L'activité d'ANITI repose sur 3 grands piliers : recherche scientifique, formation et contribution au développement économique.

Sa spécificité est de développer une nouvelle génération d'intelligence artificielle dite hybride, associant de façon intégrée des techniques d'apprentissage automatique à partir de données et des modèles permettant d'exprimer des contraintes et d'effectuer des raisonnements logiques.

Rassemblant environ 200 chercheur.e.s issu.e.s des universités, écoles d'ingénieurs et organismes de recherche scientifique et technologique de Toulouse et sa région, et d'une cinquantaine de partenaires. Les secteurs d'application stratégiques ciblés sont la mobilité et les transports et la robotique/cobotique pour l'industrie du futur.

Présentation LIEBHERR

Cette thèse sera réalisée en collaboration avec les équipes Datalab et Contrôle et Modélisation Dynamique.

Liebherr Aerospace Toulouse produit des systèmes de gestion et de conditionnement d'air pour l'aéronautique. Dans ce contexte, le Data Lab a pour mission de développer des méthodes et outils dans le domaine data analytics pour supporter la démarche entreprise de réduction des coûts de développement, d'optimisation des procédés industriels et développement de nouveaux services digitaux.

L'équipe Contrôle Dynamique est en charge de la conception et la validation des lois de contrôles des systèmes dynamiques.

RESUME DU PROJET DE RECHERCHE

Pour le développement des systèmes d'air Liebherr développe des modèles numériques sur la base de résolution d'équations. Ces modèles ont différentes échelles de représentativité : composant ou système.

Il s'agit de modèles dynamiques qui ont deux principaux objectifs :

- Définir et valider les algorithmes de contrôle des systèmes embarqués
- Améliorer le choix des paramètres de design des composants après intégration dans le système

Toutefois, ces méthodes de modélisation nécessitent la mise en équation des phénomènes. Cela peut s'avérer difficile pour certains problèmes physiques très complexes. C'est ainsi le cas pour les systèmes pile à combustible ou cycle vapeur décrits ci-après.

Liebherr dispose d'installation qui permettent de réaliser des essais des composants et des systèmes dans des environnements représentatifs. Des données d'essai contenant de nombreuses informations sur le comportement des produits sont donc disponibles. Naturellement, ces données sont coûteuses à acquérir et sont donc généralement disponibles en petite quantité.

Ce projet vise ainsi à développer un ensemble de méthodologies IA qui permettront de développer des modèles dynamiques basés sur des données dans ce contexte données rares et ainsi se doter des capacités d'optimisation du contrôle et du design même dans des cas de physiques complexes.

Le projet comportera 3 étapes

- Création de modèles données : apprentissage automatique sur les cas d'application proposés par Liebherr
- Intégration de connaissances physiques : amélioration des modèles données par prise en compte de connaissances physiques partielles
- Plan d'expérience séquentiel : définition de nouvelles données à acquérir pour optimiser la représentativité du modèle données

DESCRIPTION DETAILLEE

Le plan de travail de cette thèse s'articule en 3 étapes

Etape 1 : création de modèles données

L'objet de cette première partie est d'optimiser l'usage ou d'améliorer les techniques d'apprentissage automatique adaptées aux séries temporelles afin de créer les meilleurs modèles à partir des données qui seront mises à disposition par Liebherr. Il s'agit donc d'une part de disposer des fondations pour se doter de la capacité à créer des modèles à partir de données uniquement et d'autre part disposer de modèles de référence qu'il s'agira d'améliorer lors des étapes suivantes de projet. L'évaluation de la qualité sera définie *a priori* en fonction de l'usage final que l'on souhaite en faire. La définition de ces critères d'évaluation est une activité préalable qui permettra de mesurer la qualité des modèles en fonction de leur objectif d'utilisation.

Les données seront de différentes natures : essais sur banc, essais en vol et éventuellement données d'opérations en service. S'agissant des essais nous disposerons de quelques dizaines d'essais pouvant durer de quelques dizaines de minutes à quelques heures. Les fréquences d'échantillonnage varient entre 1 Hertz et 10 Hertz. À chaque instant t , les capteurs mesurent les variables d'intérêt. Il s'agit donc de séries temporelles. Le modèle à apprendre va avoir en entrée une ou plusieurs séries temporelles et va devoir fournir en sortie une ou des séries temporelles. L'objectif est donc d'apprendre la relation entre ces séries en entrée et en sortie, relation qui correspondra au comportement d'un composant ou système.

Une première technique envisagée pour un tel apprentissage est une forme de réseaux de neurones récurrents : les réseaux LSTM. Le principe de récurrence permet au réseau de calculer une sortie en fonction de l'information actuelle mais aussi en fonction de la précédente. Ce principe est particulièrement adapté sur des séries temporelles où il y a des dépendances avec les données précédentes. Cependant on peut avoir des dépendances proches (quelques points précédemment) ou beaucoup plus lointaines (quelques dizaines/centaines de pas de temps précédents).

Un défi majeur sera la quantité de données mise à disposition pour cet apprentissage. Il s'agira donc d'optimiser l'utilisation des techniques : architecture, hyper-paramètres, choix de la fonction coût ou encore type/génération de sous-séquence dans le cas des LSTM. Parmi les architectures à étudier, nous nous focaliserons notamment sur des architectures récurrentes qui, comme le filtre de Kalman, sont capables d'alterner deux opérations : la propagation temporelle d'une part, et l'analyse à partir d'observations. Ces étapes, qui peuvent être vues dans le cas du filtre de Kalman comme des transformations de densités de probabilités paramétriques, seront estimées en utilisant des données de fonctionnement du dispositif expérimental. Cette approche a été formalisée récemment dans le cadre ANITI sous la forme d'une approche de Bayes variationnelle qui rivalise avec les algorithmes d'assimilation de données de l'état de l'art [REF au travail de AF] sur des systèmes académiques.

Ces résultats prometteurs obtenus, nous désirons à présent confronter notre approche à des situations industrielles réelles. De nombreuses questions se posent alors, pour que la technique puisse répondre à l'objectif de prédiction :

1) Quelles architectures neuronales sont les plus adéquates pour capturer la physique sous-jacente ? La physique sous-jacente prenant souvent la forme d'équations aux dérivées partielles les réseaux convolutifs semblent des candidats à considérer lorsque des couplages locaux sont à représenter dans les propagations temporelles. Les réseaux complètement connectés seront bien sûr à privilégier pour les transferts globaux d'information, où des champs réceptifs larges sont désirés. Les réseaux profonds seront probablement incontournables pour modéliser les fortes non-linéarités.

2) Quels algorithmes d'apprentissage peuvent être utilisés pour entraîner ces réseaux « reproduisant » la physique ? Par analogie avec la résolution classique de problèmes physiques, des méthodes efficaces doivent être introduites pour les physiques complexes. En effet, les exigences en précision pour les problèmes physiques sont plus exigeants en précision (typiquement, des précisions relatives fiable sont demandée) que les problèmes de classification d'image qui sont de nature discrète. Il sera important de trouver les bonnes techniques d'optimisation pour ces problèmes, parfois très non-convexes. On pourra s'appuyer sur l'expérience de l'équipe APO sur la méthode des régions de confiance [REF] et sur les méthodes de régularisation.

3) Une question centrale concerne la quantification l'erreur de généralisation des réseaux récurrent produits et la génération de données d'apprentissage pour les réseaux. Quels outils mathématiques permettent de garantir le contrôle de l'erreur de généralisation. Il s'agira de combiner l'universalité des réseaux de neurones avec des résultats de convergence statistiques. Le projet Graine permettra d'aller au-delà et d'affiner ces résultats de convergence en les ajustant sur la problématique spécifique de la pile à hydrogène. Il sera alors possible de s'appuyer sur des considérations physiques pour guider le choix de la géométrie des données la plus pertinente pour l'apprentissage.

4) Il est fort probable que les techniques d'optimisation utilisées pour les problèmes usuels de classification (les méthodes d'optimisation du premier ordre) soient mises en défaut pour ces problèmes d'apprentissages où en plus de la non-linéarité, il faudra faire face à des besoins de précision élevée. Il est fort probable qu'une réflexion doive être menée pour trouver les meilleurs optimiseurs, ce qui pourra être fait en collaboration avec les chercheurs ANITI impliqués en optimisation de grande taille.

La validation sera faite classiquement sur un jeu de test dédié avec des critères définis préalablement mais aussi après intégration dans la chaîne de traitement qui permettra d'utiliser le modèle pour son but applicatif.

Etape 2 : intégration de connaissances physiques

Le deuxième volet de ce travail consiste à augmenter la qualité du modèle données en prenant en compte des connaissances physiques. Plusieurs approches sont envisageables et seront naturellement très liées à la nature du problème à résoudre.

Dans le cas de modèle système, il peut être envisagé d'utiliser des modèles dits physique connus et disponibles de sous-ensembles du système. Cette première voie est donc une approche hybride intelligence artificielle et modélisation physique. Par exemple, il est possible de construire des réseaux récurrents dont sous-systèmes implantent des opérations « imitant » une physique pré-supposée agissant dans un espace latent. Ces opérations pourraient modéliser des phénomènes physiques répertoriés (une diffusion, une advection, une réaction) en implantant des schémas classiques et connus issu du domaine de la résolution numérique d'équations aux dérivées partielles. Cette approche permettrait d'introduire une forme d'explicabilité (de l'architecture) du réseau de Neurones, puisque son principe de construction ne seraient plus uniquement algébriques et intégrerait les types de physique désiré sous forme de sous-systèmes couplé. Si cette voie paraît naturelle, elle introduit un certain nombre de questions :

- 1) Comment réaliser l'apprentissage statistique des sous-systèmes couplés ? Peut-on le réaliser par sous-système, ou une approche globale, plus optimale, devra être menée ? Une difficulté de l'approche globale est bien sûr liée au grand nombre de paramètres sur lequel porte l'échantillonnage et optimisation.
- 2) Si une approximation précise des sous-systèmes représente un intérêt applicatif en soi, et est à considérer, comment prendre ces spécifications de performance dans l'approche globale ?
- 3) Comment orchestrer la convergence de l'apprentissage sur les sous-systèmes couplés ? Est-il possible ou nécessaire de traiter les sous-systèmes d'une manière asymétrique (certain devant converger en priorité) ?
- 4) Si l'approche globale est favorisée en raison de son optimalité, peut-on utiliser des approches locales simplement pour accélérer le processus d'apprentissage (comme technique de pré-conditionnement) ?
- 5) Si les physiques de certains sous-systèmes sont connues, mais coûteuses à simulées, est-il intéressant de les remplacer par des modèles de substitution pour réduire les coûts de calcul.

Une deuxième voie sera d'adapter les architectures des réseaux afin de permettre de préserver des propriétés physiques des phénomènes à modéliser. Il s'agira ici de généraliser à des contraintes générales des techniques souvent utilisées au niveau des fonctions d'activation notamment pour imposer le signe ou plus généralement la plage de variation de certaines quantités. On pourrait forcer par exemple la conservation de certaines grandeurs au niveau de l'architecture même du réseau.

Enfin on cherchera à guider l'apprentissage. A titre d'exemple, il pourrait être possible pour le système VCS de reconstruire une variable additionnelle modélisant l'état de la phase du fréon au cours du temps. La prise en compte de cette variable comme variable explicative permettrait d'améliorer la qualité de l'inférence. Cette approche s'apparente au « feature engineering » qui permet plus généralement d'introduire des observables dont les quantités à prédire sont des fonctions faciles à découvrir pour le réseau.

Etape 3 : plan d'expérience séquentiel

Les modèles créés à l'étape 1 et améliorés à l'étape 2 seront limités par la quantité d'information contenues par les données utilisées pour l'apprentissage. Ces données sont obtenues en réalisant une première campagne d'essais ou à partir de données d'exploitations. L'objectif premier n'est pas nécessairement de fournir la matière pour créer un modèle. Ces essais ont une représentativité limitée liée à la taille ou la complexité des systèmes, la variété des conditions d'opérations à reproduire et le nombre d'essais réalisés qui doit être minimisés (accès aux ressources d'essais et coût des essais).

L'amélioration de la précision et de la représentativité des modèles passera par la génération de données additionnelles. Compte tenu des coûts d'acquisition de ces données, il est fondamental d'identifier les zones où le modèle doit être améliorés, identifier la valeur *a priori* qu'apporterait un nouvel essai et enfin définir la stratégie optimale d'essais additionnels à réaliser (i.e. scénario temporel, conditions). Des méthodes du type apprentissage actif seront étudiées.

La maîtrise d'une telle méthodologie apparait majeure dans la capacité à créer des modèles data-driven de qualité. Elle apporterait également une grande valeur dans la mise au point de tout modèle qui nécessite également des données d'essais pour ajuster des degrés de liberté et optimiser la fidélité du modèle.

PROFIL et DETAILS ADMINISTRATIFS

Domaine de formation et niveau : Data Science, Master ou d'un diplôme d'ingénieur

Rémunération : 32 k€ brut

Durée et période souhaitées 3 ans à partir d'Octobre 2021

Service / Département : Liebherr Data Lab IT – Site de Toulouse

LIEBHERR-AEROSPACE TOULOUSE S.A.

RÉF. : ... / ... / ... / ... DATE : 19/07/2021 PAGE : 4/4

Contacts

Nicolas Canouet / nicolas.canouet@liebherr.com / 05 67 73 24 74

Serge Gratton / serge.gratton@toulouse-inp.fr

Fabrice Gamboa / fabrice.gamboa@math.univ-toulouse.fr

Site internet : www.liebherr.com