Propositions de Mémoires

Master en Sciences Informatiques
Année 2022-2023

ALAIN BUYS[1], ALEXANDRE DECAN[2], BRUNO QUOITIN[3],
HADRIEN MELOT[4], JEF WIJSEN[5], SOUHAIB BEN TAIEB[6],
STEPHANE DUPONT[7], TOM MENS[8], VERONIQUE BRUYERE[9],
NICOLAS GILLIS[10], MICKAEL RANDOUR[11], XAVIER SIEBERT[12],
et SEWERYN DYNEROWICZ[13]

20 septembre 2022

1. Alain.BUYS@umons.ac.be
2. Alexandre.DECAN@umons.ac.be
3. Bruno.QUOITIN@umons.ac.be
4. Hadrien.MELOT@umons.ac.be
5. Jef.WIJSEN@umons.ac.be
6. Souhaib.BENTAIEB@umons.ac.be
7. Stephane.DUPONT@umons.ac.be
8. Tom.MENS@umons.ac.be
9. Veronique.BRUYERE@umons.ac.be
10. Nicolas.GILLIS@umons.ac.be
11. mickael.randour@umons.ac.be
12. Xavier.SIEBERT@umons.ac.be
13. Seweryn.DYNEROWICZ@umons.ac.be

Table des matières

Chapitre 1

Mémoires (Mons)

1.1 Anonymisation des données personnelles.

Service	Réseaux et télécommunications – Systèmes d'Information
Directeur	ALAIN BUYS [1] et JEF WIJSEN [2]

Description

Le Règlement Général sur la Protection des Données (RGPD) garantit le droit des personnes à contrôler la façon dont les entreprises gèrent leurs données personnelles. Normalement, les entreprises concernées ne sont pas autorisées à diffuser ces informations à l'extérieur. D'autre part, il existe des incitents (notamment économiques) à l'exploitation statistique de ces données par des tiers. Par exemple, l'industrie pharmaceutique est fort logiquement intéressée par les résultats de l'administration de ses produits aux patients d'hôpitaux. Le RGPD suggère que l'anonymisation (si elle est irréversible) des données permet de sortir du cadre du règlement mais dit peu de choses sur les techniques à utiliser. Différentes techniques d'anonymisation ont été proposées et différents concepts sont apparus, notamment *k-anonymity*, *l-diversity* et plus récemment *differential privacy*. Le but de ce mémoire est d'une part de faire le point sur les méthodes existantes et d'autre part, d'essayer de déterminer dans quelle mesure elles répondent aux besoins légaux actuels (irréversibilité, ...).

Exigences ou prérequis

Un intérêt pour la cryptographie et les bases de données, ainsi que les mathématiques sous-jacentes.

1. Alain.BUYS@umons.ac.be
2. Jef.WIJSEN@umons.ac.be

1.2 Sujet dans le domaine « Courbes elliptiques en cryptographie »

Service	Réseaux et télécommunications
Directeur	ALAIN BUYS [3]

Description

Sujet à définir en concertation avec l'étudiant.

Exigences ou prérequis

Un intérêt pour la cryptographie et les mathématiques discrètes.

3. Alain.BUYS@umons.ac.be

1.3 Sujet dans le domaine « Cryptographie quantique »

Service	Réseaux et télécommunications
Directeur	ALAIN BUYS [4]

Description

Sujet à définir en concertation avec l'étudiant.

Exigences ou prérequis

Un intérêt pour la cryptographie et la physique mise en œuvre.

4. Alain.BUYS@umons.ac.be

1.4 Ordonnancement de communications UWB-TSCH à l'aide d'un solveur SMT

Service	Réseaux et Télécommunications
Directeur	BRUNO QUOITIN [5]

Description

L'objectif de ce mémoire est d'expérimenter l'ordonnancement de communications à l'aide d'un solveur SMT (*Satisfiability Modulo Theories*). Un solveur SMT présente l'avantage de permettre de vérifier des contraintes exprimées par exemple en logique du premier ordre, au contraire d'un solveur SAT qui sera limité à la logique booléenne. Un solveur SMT permet également de manipuler des données plus complexes que des booléens, comme p.ex. des entiers, à l'aide de « théories ». Dans ce mémoire, il est conseillé d'utiliser le solveur open-source Z3 [1].

Les communications à ordonnancer, sont celles d'un réseau UWB-TSCH [2]. Il s'agit d'un réseau utilisant des communications sans fil à très large bande (*Ultra WideBand*) combinant multiplexage temporel (*Time Slotted*) et saut de fréquences (*Channel Hopping*). Dans un tel réseau, les communications sont organisées à l'intérieur d'une matrice à deux dimensions appelée *slotframe* dont chaque cellule est indexée par le temps (numéro de timeslot) et par le canal fréquentiel. Etablir le contenu de cette *slotframe* peut être vu comme un problème d'optimisation où il s'agit d'établir toutes les communications, sans que celles-ci n'entrent en conflit. Il faudra par exemple empêcher que deux communications différentes soient prévues dans une même cellule ou qu'un noeud doive être actif dans deux canaux simultanément. L'objectif d'optimisation sera de limiter la taille de la *slotframe*. Dans le contexte UWB-TSCH, deux types de communications doivent être ordonnancées : des échanges de message permettant l'estimation de la distance entre les protagonistes et le rapatriement de ces estimations vers un point de collecte, le long d'un arbre recouvrant.

Exigences ou prérequis

Intérêt pour les systèmes embarqués et les réseaux informatiques.

Références

[1] Leonardo de MOURA et Nikolaj BJØRNER. « Z3 : An Efficient SMT Solver ». In : *Tools and Algorithms for the Construction and Analysis of Systems*. Sous la dir. de C. R. RAMAKRISHNAN et Jakob REHOF. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 337-340. ISBN : 978-3-540-78800-3.

[2] Maximilien CHARLIER, Bruno QUOITIN et David HAUWEELE. « Challenges in Using Time Slotted Channel Hopping with Ultra Wideband Communications ». In : *Proceedings of the International Conference on Internet of Things Design and Implementation*. IoTDI '19. Montreal, Quebec, Canada : ACM, 2019, p. 82-93. ISBN : 978-1-4503-6283-2. DOI : 10.1145/3302505.3310071. URL : http://doi.acm.org/10.1145/3302505.3310071.

5. Bruno.QUOITIN@umons.ac.be

1.5 Identifying IoT nodes by spying on their radio signal

Service	Réseaux et Télécommunications
Directeur	BRUNO QUOITIN [6]

Description

The goal of this project is to investigate how listening to the radio signal transmitted by IoT nodes could be used to help in their identification. The hypothesis behind this idea is that two different nodes, although manufactured with the same microprocessor and radio transceiver models, will behave slightly differently with regards to their transmitted signals, hence allowing identification. The question is are those differences significantly different to allow reliable identification ? Moreover, what are the differences/features such an identification system should look for in the radio signal ?

This project will focus on radio communications using the LoRa modulation. It is based on *Chirp Spread Spectrum* (CSS), a technique that sends radio signals whose frequency varies linearly with time. Information is encoded in the signal starting frequency. Although the LoRa modulation is a proprietary technique and its details are not publicly disclosed, it has been fully reverse-engineered and radio signal processing blocks exist in tools such as GNU radio [1, 2].

There is already a bit of scientific litterature that exists on the topic of LoRa *radiofrequency fingerprinting*, so the project should first aim at discovering and understanding the corresponding body of knowledge [3, 4]. Then, from a practical point of view, it is envisioned that the LoRa radio signal will be captured with a *Software Defined Radio* (SDR), a dedicated equipement able to sample the radio signal within a given frequency band. The project should aim at familiarization with such equipment and the obtained timeseries of I/Q samples. In the end, a system able to identify devices based on the captured radio signal should be designed, based on a technique documented in the scientific litterature, for example using differential constellation diagrams [4].

Exigences ou prérequis

Intérêt pour les réseaux informatiques, les télécommunications, le traitement du signal et la cybersécurité.

Références

[1] Alexandre MARQUET, Nicolas MONTAVONT et Georgios Z. PAPADOPOULOS. « Towards an SDR implementation of LoRa : Reverse-engineering, demodulation strategies and assessment over Rayleigh channel ». In : *Computer Communications* 153 (2020), p. 595-605. ISSN : 0140-3664. DOI : https://doi.org/10.1016/j.comcom.2020.02.034. URL : https://www.sciencedirect.com/science/article/pii/S0140366419314665.

[2] Joachim TAPPAREL, Orion AFISIADIS, Paul MAYORAZ, Alexios BALATSOUKAS-STIMMING et Andreas BURG. « An Open-Source LoRa Physical Layer Prototype on GNU Radio ». In : *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. 2020, p. 1-5. DOI : 10.1109/SPAWC48557.2020.9154273.

[3] Guanxiong SHEN, Junqing ZHANG, Alan MARSHALL, Linning PENG et Xianbin WANG. *Radio Frequency Fingerprint Identification for LoRa Using Spectrogram and CNN*. 2021. DOI : 10.48550/ARXIV.2101.01668. URL : https://arxiv.org/abs/2101.01668.

[4] Y. JIANG, L. PENG, A. HU et al. « Physical layer identification of LoRa devices using constellation trace figure ». In : *EURASIP Journal on Wireless Communications and Networking* (223 2019). DOI : 10.1186/s13638-019-1542-x.

6. Bruno.QUOITIN@umons.ac.be

1.6 Support de multiples racines dans un réseau 6TiSCH

Service	Réseaux et Télécommunications
Directeur	BRUNO QUOITIN [7] et JEREMY DUBRULLE
Attribué à	Lionel GOFFAUX

Description

6TiSCH [1] est une architecture de communication pour les réseaux d'objets industriels. Elle est standardisée par l'IETF et combine un protocole de routage (RPL) à la technologie radio IEEE 802.15.4 avec le mode d'accès TSCH. Dans cette architecture, la racine du réseau joue des rôles multiples. Elle agit comme racine de l'arbre établit par RPL, organisant les communications multi-sauts, mais également comme source d'horloge pour le multiplexage temporel mis en oeuvre par TSCH.

Ce projet s'intéresse aux mécanismes permettant de déployer des réseaux 6TiSCH utilisant de multiples racines. Celles-ci permettent de rendre le réseau plus robuste en cas de panne de l'une d'entre elles voire d'équilibrer la charge de trafic ou encore d'étendre la couverture géographique du réseau. Une des difficultés de la mise en oeuvre d'un déploiement multi-racines est de permettre la synchronisation temporelle entre celles-ci. Plusieurs approches pourront être considérées telles que l'usage d'une source d'horloge commune externe (serveur NTP ou signal GNSS [2, 3]) ou un fonctionnement maître/esclave, une racine agissant comme source d'horloge (en utilisant un protocole tel que PTP [4]). Dans ces derniers cas, des communications entre les racines doivent être établies au travers d'un réseau *backbone* reposant par exemple sur Ethernet ou du Wi-Fi.

Un prototype du projet devra être implémenté dans le RTOS Contiki-OS et testé sur des systèmes embarqués de type Zolertia Re-Mote qui seront déployés dans le laboratoire Réseaux. Une validation du système développé et une étude de performance devront être réalisées. Cette dernière devrait permettre notamment de mesurer l'erreur sur le temps maintenu par les différentes racines, d'étudier les facteurs influençant cette erreur et de déterminer si elle peut être rendue suffisamment faible que pour permettre un fonctionnement sans faille d'un réseau à racines multiples.

Exigences ou prérequis

Intérêt pour l'architecture des ordinateurs.

Références

[1] Pascal THUBERT. *An Architecture for IPv6 over the TSCH mode of IEEE 802.15.4*. Internet-Draft draft-ietf-6tisch-architecture-29. Work in Progress. Internet Engineering Task Force, août 2020. 71 p. URL : https://datatracker.ietf.org/doc/html/draft-ietf-6tisch-architecture-29.

[2] So-Young HWANG, Dong-Hui YU et Ki-Joune LI. « Embedded System Design for Network Time Synchronization ». In : *Embedded and Ubiquitous Computing*. Sous la dir. de Laurence T. YANG, Minyi GUO, Guang R. GAO et Niraj K. JHA. Berlin, Heidelberg : Springer Berlin Heidelberg, 2004, p. 96-106. ISBN : 978-3-540-30121-9.

[3] Roman LIM, Balz MAAG et Lothar THIELE. « Time-of-Flight Aware Time Synchronization for Wireless Embedded Systems ». In : *Proceedings of the 2016 International Conference on Embedded Wireless Systems and Networks*. EWSN '16. Graz, Austria : Junction Publishing, 2016, p. 149-158. ISBN : 9780994988607.

[4] « IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems ». In : *IEEE Std 1588-2008 (Revision of IEEE Std 1588-2002)* (2008), p. 1-269. DOI : 10.1109/IEEESTD.2008.4579760.

7. Bruno.QUOITIN@umons.ac.be

1.7 Requêtes spécifiques à la théorie extrémale des graphes

Service	Algorithmique
Directeur	HADRIEN MELOT [8]
Attribué à	Thomas Lavend'Homme

Description

PHOEG est un système d'aide à la découverte en théorie extrémale des graphes, developpé au sein du Service d'Algorithmique.

Ce mémoire contient deux parties principales. Une première partie consiste à développer une API permettant l'interface entre le portail web de PHOEG et sa base de données (PostgreSQL), de telle sorte à permettre à des théoriciens des graphes de réaliser des requêtes sans avoir à programmer.

La deuxième partie consiste à étudier les propriétés extrémales d'un ou plusieurs invariants de graphes (le choix des invariants se fera en accord avec le directeur), en partant d'articles récents.

8. Hadrien.MELOT@umons.ac.be

1.8 Détermination et visualisation automatique d'équations liées à des graphes

Service	Algorithmique
Directeur	HADRIEN MELOT [9]

Description

La théorie extrémale des graphes s'intéresse à déterminer des bornes sur des invariants de graphes. Il existe quantité d'invariants qui permettent de décrire les propriétés des graphes, comme le nombre chromatique, le nombre de stabilité, le diamètre, etc.

Le Service d'Algorithmique développe des outils d'aide à la découverte dans ce domaine et une des tâches est de déterminer, à l'aide d'une vision géométrique de problèmes liés aux graphes, des généralisations pour tout n (où n est le nombre de sommets des graphes considérés) de valeurs d'invariants sous la forme d'équations, à partir de valeurs connues pour des petites valeurs de n.

Après une familiarisation avec les outils, et en concertation avec l'étudiant, des objectifs secondaires peuvent également être déterminés.

Exigences ou prérequis

Intérêt marqué pour l'algorithmique et la théorie des graphes.

9. Hadrien.MELOT@umons.ac.be

1.9 *Consistent Query Answering* avec des clés étrangères et primaires

Service	Systèmes d'Information
Directeur	JEF WIJSEN [10]

Description

Étudier et implémenter le problème étudié et résolu dans la publication suivante :

> Miika Hannula et Jef Wijsen. "A Dichotomy in Consistent Query Answering for Primary Keys and Unary Foreign Keys". In : *PODS'22 : Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Philadelphia, PA, USA, June 12-17, 2022*. ACM, 2022.

Une version étendue se trouve en « open access » sur arXiv [1].

Abstract. Since 2005, significant progress has been made in the problem of Consistent Query Answering (CQA) with respect to primary keys. In this problem, the input is a database instance that may violate one or more primary key constraints. A repair is defined as a maximal subinstance that satisfies all primary keys. Given a Boolean query q the question then is whether q holds true in every repair. So far, theoretical research in this field has not addressed the combination of primary key and foreign key constraints, despite the importance of referential integrity in database systems. This paper addresses the problem of CQA with respect to both primary keys and foreign keys. In this setting, it is natural to adopt the notion of symmetric-difference repairs, because foreign keys can be repaired by inserting new tuples. We consider the case where foreign keys are unary, and queries are conjunctive queries without self-joins. In this setting, we characterize the boundary between those CQA problems that admit a consistent first-order rewriting, and those that do not.

Exigences ou prérequis

Bien comprendre les notions suivantes vues en *Bases de données I* et *Bases de données II* : clé primaire, clé étrangère, dépendance fonctionnelle, requête conjonctive, logique des prédicats, Datalog avec négation stratifiée.

Références

[1] Miika HANNULA et Jef WIJSEN. « A Dichotomy in Consistent Query Answering for Primary Keys and Unary Foreign Keys ». In : *CoRR* abs/2203.13475 (2022).

10. Jef.WIJSEN@umons.ac.be

1.10 Le clustering hiérarchique

Service	Systèmes d'Information
Directeur	JEF WIJSEN [11]

Description

Dasgupta et Long [1] on prouvé le résultat suivant en fournissant un algorithme.

Theorem 1. *Take the cost of a clustering to be the largest radius of its clusters. Then, any data set in any metric space has a hierarchical clustering in which, for each k, the induced k-clustering has cost at most eight times that of the optimal k-clustering.*

Récemment, Sakib A. Mondal [2] a démontré que ce théorème reste vrai si l'on remplace *"at most eight times"* par *"at most six times"*. L'objectif de ce mémoire est d'étudier l'algorithme original et son amélioration.

Références

[1] Sanjoy DASGUPTA et Philip M. LONG. « Performance guarantees for hierarchical clustering ». In : *J. Comput. Syst. Sci.* 70.4 (2005), p. 555-569.

[2] Sakib A. MONDAL. « An improved approximation algorithm for hierarchical clustering ». In : *Pattern Recognit. Lett.* 104 (2018), p. 23-28.

11. Jef.WIJSEN@umons.ac.be

1.11 XPath Query Containment

Service	Systèmes d'Information
Directeur	JEF WIJSEN [12]

Description

Dans le cours de *Bases de données II*, les étudiants apprennent le langage XPath ainsi qu'un algorithme pour décider si deux requêtes conjonctives sont équivalentes. Une question naturelle est alors : existe-t-il un algorithme pour décider si deux requêtes XPath sont équivalentes ? Cette question fait l'objet de ce mémoire, en partant d'un article à ce sujet [1].

Exigences ou prérequis

Avoir bien réussi (et aimé) le cours de *Bases de données II*.

Références

[1] Thomas SCHWENTICK. « XPath query containment ». In : *SIGMOD Rec.* 33.1 (2004), p. 101-109.

12. Jef.WIJSEN@umons.ac.be

1.12 The double descent phenomenon in machine learning

Service	Big Data and Machine Learning
Directeur	SOUHAIB BEN TAIEB [13]

Description

The bias-variance trade-off is a fundamental concept in machine learning and statistics. The idea is that models of higher complexity have lower bias but higher variance. According to this theory, once model complexity passes a certain threshold, models "overfit" with the variance term dominating the test error, and hence from this point onward, increasing model complexity will only decrease performance (i.e., increase test error). Hence conventional wisdom in machine learning is that, once we pass a certain threshold, "larger models are worse". However, in the modern practice, very rich models such as neural networks are trained to exactly fit (i.e., interpolate) the data. Classically, such models would be considered over-fit, and yet they often obtain high accuracy on test data. The goal of this project is to study the *double descent* phenomenon, which describes generalization error that decreases, increases, and then again decreases, with increases in model flexibility. See the following relevant references : [1, 2, 3, 4, 5].

Exigences ou prérequis

Basic notions of probability, statistics, and machine learning.

Références

[1] Brady NEAL. « On the Bias-Variance Tradeoff : Textbooks Need an Update ». In : (déc. 2019). arXiv : 1912.08286 [cs.LG].

[2] Mikhail BELKIN, Daniel HSU, Siyuan MA et Soumik MANDAL. « Reconciling modern machine-learning practice and the classical bias-variance trade-off ». en. In : *Proceedings of the National Academy of Sciences of the United States of America* 116.32 (août 2019), p. 15849-15854.

[3] Preetum NAKKIRAN, Gal KAPLUN, Yamini BANSAL, Tristan YANG, Boaz BARAK et Ilya SUTSKEVER. « Deep Double Descent : Where Bigger Models and More Data Hurt ». Sept. 2019.

[4] Chiyuan ZHANG, Samy BENGIO, Moritz HARDT, Benjamin RECHT et Oriol VINYALS. « Understanding deep learning requires rethinking generalization ». In : *arXiv preprint arXiv :1611.03530* (2016).

[5] Zitong YANG, Yaodong YU, Chong YOU, Jacob STEINHARDT et Yi MA. « Rethinking Bias-Variance Trade-off for Generalization of Neural Networks ». In : *ICML*. 2020, p. 10767-10777. URL : http://proceedings.mlr.press/v119/yang20j.html.

13. Souhaib.BENTAIEB@umons.ac.be

1.13 Neural temporal point processes

Service	Big Data and Machine Learning
Directeur	SOUHAIB BEN TAIEB [14]

Description

Sequences of discrete events taking place at irregular time intervals are often generated from a plethora of human and natural phenomena. For instance, the activity of an user on a social media platform, the occurences of earthquakes at a specific location, or the recording of a patient's electronic health indicators in time can be recorded as streams of such asynchronous events. A common approach to describe these sequences of events is by considering them as realizations of a Temporal Point Process (TPP), a probabilistic framework allowing to evaluate the occurence probability of the next event given the history of the process (i.e. events that occurred in the past). Facing increasing data complexity and volume, simple parametric forms of TPPs, such as the Poisson or the Hawkes process, might lack sufficient flexibility to model complex temporal dynamics and potential interactions amongst events, leading to overall poor model performance. To paliate this issue, advances in the field of Neural Networks, and especially regarding Recurrent Neural Network (RNN) and Transformer architectures, have been recently incorporated to the framework of TPPs, in the hope of improving the model's expressiveness in generating relevant representations. However, such approaches, although promising, lack a common methodology regarding data processing, predictions, evaluation, and benchmarks, rendering comparison across models difficult. In the light of this issue, the goal of the project is to study recent Neural Network advances in the field of Temporal Point Processes, and to provide rigorous evaluation of the models' performance, while including fairness in the comparisons. Most of the research will be carried out using the Pytorch library in Python. See the following relevant references for more information : [1, 2, 3, 4]

Exigences ou prérequis

Basic notions of probability, statistics, and machine learning.

Références

[1] Jakob Gulddahl RASMUSSEN. *Lecture Notes : Temporal Point Processes and the Conditional Intensity Function*. Rapp. tech. Aalborg University, juin 2018.

[2] Oleksandr SHCHUR, Ali Caner TÜRKMEN, Tim JANUSCHOWSKI et Stephan GÜNNEMANN. « Neural Temporal Point Processes : A Review ». In : *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Avr. 2021, p. 4585-4593.

[3] Oleksandr SHCHUR, Marin BILOŠ et Stephan GÜNNEMANN. « Intensity-Free Learning of Temporal Point Processes ». In : *International Conference on Learning Representations (ICLR)*. 2020.

[4] Qiang ZHANG, Aldo LIPANI, Omer KIRNAP et Emine YILMAZ. « Self-Attentive Hawkes Processes ». In : *Proceedings of Machine Learning Research*. Juill. 2019, p. 11183-11119.

14. Souhaib.BENTAIEB@umons.ac.be

1.14 Uncertainty quantification with deep neural networks

Service	Big Data and Machine Learning
Directeur	SOUHAIB BEN TAIEB [15]

Description

Deep neural networks (DNNs) achieve state-of-the-art results in a large variety of applications. However, a major barrier to their deployment in critical applications is the need for a reliable uncertainty quantification. In the context of regression (i.e., when the target variable is continuous), strong assumptions are typically made on the distribution of the target variable. For example, using a normal distribution while the distribution is highly multimodal. Recent advances in generative modelling have led to effective and flexible probabilistic models parameterized by DNNs such as normalizing flows or energy-based models. The goal of this project is to study different ways to parameterize predictive distributions with DNN-based methods and to evaluate them empirically in terms of statistical accuracy and computational complexity. The research will preferably be carried out using PyTorch. See the following relevant references : [1, 2, 3, 4, 5, 6, 7, 8].

Exigences ou prérequis

Basic notions of probability, statistics, and machine learning.

Références

[1] 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 et Xiao Xiang ZHU. « A Survey of Uncertainty in Deep Neural Networks ». In : (juill. 2021). arXiv : 2107.03342 [cs.LG].

[2] Simon BACHSTEIN. « Uncertainty Quantification in Deep Learning ». Mém. de mast. Universitat Ulm, jan. 2019.

[3] Maria R CERVERA, Rafael DÄTWYLER, Francesco D'ANGELO, Hamza KEURTI, Benjamin F GREWE et Christian HENNING. « Uncertainty estimation under model misspecification in neural network regression ». In : (nov. 2021). arXiv : 2111.11763 [cs.LG].

[4] Ivan KOBYZEV, Simon J D PRINCE et Marcus A BRUBAKER. « Normalizing Flows : An Introduction and Review of Current Methods ». en. In : *IEEE Trans. Pattern Anal. Mach. Intell.* 43.11 (nov. 2021), p. 3964-3979.

[5] Conor DURKAN, Artur BEKASOV, Iain MURRAY et George PAPAMAKARIOS. « Neural Spline Flows ». In : (juin 2019). arXiv : 1906.04032 [stat.ML].

[6] Yang SONG et Diederik P KINGMA. « How to Train Your Energy-Based Models ». In : (jan. 2021). arXiv : 2101.03288 [cs.LG].

[7] Fredrik K GUSTAFSSON, Martin DANELLJAN, Radu TIMOFTE et Thomas B SCHÖN. « How to Train Your Energy-Based Model for Regression ». In : (mai 2020). arXiv : 2005.01698 [cs.CV].

[8] Sam BOND-TAYLOR, Adam LEACH, Yang LONG et Chris G WILLCOCKS. « Deep Generative Modelling : A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models ». en. In : *IEEE Trans. Pattern Anal. Mach. Intell.* PP (sept. 2021).

15. Souhaib.BENTAIEB@umons.ac.be

1.15 Extraction d'information du langage naturel

Service	Intelligence Artificielle
Directeur	STEPHANE DUPONT [16]

Description

Ce mémoire portera sur le développement et la comparaison de méthodes de traitement du langage naturel pour l'extraction d'information à partir de texte. Les méthodes recommandées comporteront les architectures de réseaux de neurones artificiels de type récurrent ainsi que ceux avec mécanismes d'attention.

La catégorie de techniques d'analyse des sentiments basée sur l'aspect (ABSA - Aspect-based sentiment analysis) sera exploitée. L'ABSA peut être utilisée par exemple pour analyser les commentaires des clients en associant des sentiments spécifiques à différents aspects d'un produit ou d'un service. Il s'agira ici de réaliser une étude comparative de méthodes de transfert d'apprentissage afin d'exploiter l'ABSA dans un autre cadre, en ciblant l'extraction d'information sur les opinions exprimées en lien avec les discussions sur des domaines d'actualité. Possibilité d'orienter le projet soit vers l'analyse des opinions, soit l'analyse de tentatives de harcèlement et d'intimidation ("bullying, trolling, etc.") verbales.

Exigences ou prérequis

Intérêt pour l'intelligence artificielle, l'analyse de texte et la compréhension du langage naturel.

16. Stephane.DUPONT@umons.ac.be

1.16 Vidéo 4D

Service	Intelligence Artificielle
Directeur	STEPHANE DUPONT [17]

Description

Les technologies de capture vidéo ne cessent de progresser, en résolution spatiale et temporelle (8K, 120fps, etc.) et en plage dynamique (HDR, etc.). Certaines technologies de capteurs permettent aussi d'obtenir une information de profondeur (3D). Nous sommes également dans une ère de développement de la photo computationnelle. Le logiciel et l'intelligence artificielle viennent en aval du capteur vidéo pour en augmenter les possibilités, accroitre sa résolution, inférer la 3D, etc.

Dans ce contexte, ce projet a pour objectif de réaliser un prototype de système de capture vidéo 4D. Plusieurs caméras seront utilisées pour suivre une action sous différents angles de vue et reconstruire une réalité volumétrique en 3D. Celle-ci offrira ensuite la possibilité de recréer des vidéos dans lesquelles les angles de prise de vue pourront ainsi être choisis en post-production. Il sera recommandé d'exploiter des approches par réseaux de neurones convolutifs spécialement dédiés et entrainés à estimer les correspondances entre les pixels provenant des différentes caméras, et à estimer les surfaces des sujets filmés.

Exigences ou prérequis

Intérêt pour l'intelligence artificielle, le traitement d'image et vidéo, la 3D.

17. Stephane.DUPONT@umons.ac.be

1.17 IA pour l'interprétation de l'environnement sonore

Service	Intelligence Artificielle
Directeur	STEPHANE DUPONT [18]

Description

Ces dernières années, les technologies de reconnaissance vocale ont considérablement progressé, notamment grâce aux avancées en IA et en apprentissage automatique end-to-end. La reconnaissance de sons environnementaux est également possible (sonnerie de téléphone, bruit de pas, de voiture, etc...).

Ce projet a pour objectif la réalisation d'un logiciel exploitant ces technologies et permettant la reconnaissance et la localisation de sons dans l'espace. Premièrement, le logiciel effectuera une reconnaissance en continu et en temps-réel des sons. Ceci reposera sur des algorithmes et modèles d'IA choisis après une étude des projets open source ainsi que des offres IA sur le cloud.

Diverses applications peuvent ensuite en découler, comme l'assistance aux personnes sourdes, ou l'adaptation de films, séries ou dessins animés pour ce public.

Exigences ou prérequis

Intérêt pour l'intelligence artificielle, l'audition par ordinateur.

18. Stephane.DUPONT@umons.ac.be

1.18 Comparaison entre modèle 3D et objet réel

Service	Intelligence Artificielle
Directeur	STEPHANE DUPONT [19]

Description

La vision par ordinateur permet la reconnaissance et la segmentation d'objets en 3D (segmentation sémantique). Sur base d'algorithmes d'IA de segmentation sémantique en 3D, ce TFE développera le cas particulier de la segmentation d'éléments structurels et décoratifs d'intérieurs de bâtiments, ainsi qu'une méthode de comparaison en temps-réel entre des modèles 3D et la réalité de terrain ainsi capturée. Ceci devrait pouvoir fonctionner sur tablette et à terme permettre des applications d'état des lieux et de suivi de chantier, assistées par l'IA.

Exigences ou prérequis

Intérêt pour l'intelligence artificielle, la modélisation 3D, les applications sur tablettes.

19. Stephane.DUPONT@umons.ac.be

1.19 Etude comparatives d'approches de traduction automatique du français vers la langue des signes

Service	Intelligence Artificielle
Directeur	STEPHANE DUPONT [20]

Description

Ce projet a pour objectif la réalisation d'un logiciel permettant la traduction du français vers le langage des signes. Une étude comparative des méthodes de traduction utilisant la modélisation de séries par réseaux de neurones (principalement RNN - réseaux de neurones récurrents - et Transformers - réseaux de neurones avec attention -) sera effectuée, et le travail impliquera également la conception d'une base de donnée d'apprentissage pour ces systèmes de traduction automatique, ainsi que le choix de normes de représentation du langage des signes.

Le résultat pourrait éventuellement (travail supplémentaire si le temps le permet) ensuite être converti en informations permettant de piloter les mouvements d'un personnage virtuel en 3D pour le rendu du discours en langage des signes.

Exigences ou prérequis

Intérêt pour l'intelligence artificielle, les technologies de traitement du langage naturel, et l'intégration logicielle.

20. Stephane.DUPONT@umons.ac.be

1.20 Etude de technique permettant l'apprentissage en continu sans oubli catastrophique de modèles de traitement du langage naturel

Service	Intelligence Artificielle
Directeur	STEPHANE DUPONT [21]

Description

Les modèles d'IA appelés "modèles de langage" sont à la base de nombreuses opportunités d'exploitation de l'IA dans des domaines très variés. Ces modèles, notamment les réseaux de neurones avec attention, peuvent être entrainés sur des corpus de textes colossaux et convergent vers une modélisation du langage mais qui embarque également une représentation codée (au sein des poids du réseau de neurone) de nombreux faits et relations.

Ces réseaux sont ensuite exploités en les spécialisant sur une domaine ou bien sur une problématique de compréhension du langage spécifique. Cette approche pose de nombreuses questions fondamentales. L'une d'entre elle est la tendance du modèle résultant à ne plus fonctionner correctement sur le domaine source, un phénomène connu sous le nom d'oubli catastrophique, et que l'on cherche souvent à éviter. Secundo, les faits mémorisés sous la forme d'une réseau de neurones sont souvent déconnectés de tout contexte, notamment temporel ou spatial, ou bien ce contexte peut aussi être tributaire de l'oubli catastrophique.

Ce travail visera a étudier les approches proposées dans la littérature afin de régulariser les processus d'apprentissage automatique afin de permettre un apprentissage en continu et réduire ces phénomènes défavorables à l'exploitation concrète de ces approches par des secteurs qui ne disposent pas de volumes de données quasi infinis.

Autres mots-clés pertinents : few-shot learning, curriculum learning, continual learning.

Exigences ou prérequis

Intérêt pour l'intelligence artificielle et les technologies de traitement du langage naturel.

21. Stephane.DUPONT@umons.ac.be

1.21 A comparative study of BERT Transformers for identifying bot accounts

Service	Service de Génie Logiciel
Directeur	TOM MENS [22] (doctorant : Pooya Rostami Mazrae)

Description

Transformers are deep learning models that adopt the mechanism of self-attention, differentially weighting the significance of each part of the input data. They are used primarily in the fields of natural language processing (NLP) and computer vision (CV). BERT, which stends for Bidirectional Encoder Representations from Transformers was introduced by Google in 2018 as a very powerful and novel approach for NLP. BERT is often achieving a very high performance for many NLP tasks. Since the original article by Devlin et al., many variants of the BERT algorithm have been proposed in the research litterature.

The goal of this masters thesis is threefold :

1. To understand and explain in a comprehensive manner the inner workings of BERT and its many derivatives like AlBERT, RoBERTa, and DistilBERT.
2. To empirically validate BERT on the NLP-based problem of "bot" detection. "Bots" are automated agents that are used by software developers to carry out repetitive activities during collaborative development on social coding platforms such as GitHub. Examples of bot usage can be very diverse, including continuous integration services, automated code quality analysis, verifying licences, welcoming newcomers, monitoring dependencies and vulnerabilities, helping in code reviewing, checking for long periods of inactivity, and so on. Bots manifest themselves as machine accounts, but are difficult to distinguish from human accounts. Golzadeh et al. have proposed a classification model that analyses the comments associated to issues and pull requests in GitHub repositories to distinguish bot accounts from human accounts. The goal would be to apply BERT on this specific problem, and to verify to which extent BERT can outperform the approach proposed by Golzadeh et al.
3. To evaluate and compare different variants of BERT on the aforementioned bot detection problem, in order to select the most appropriate variant for this task.

Related works

— Devlin, Jacob, et al. "BERT : Pre-training of deep bidirectional transformers for language understanding." 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Association for Computational Linguistics, 2019, pp. 4171-4186
https ://doi.org/10.18653/v1/N19-1423
— Lan, Zhenzhong, et al. "AlBERT : A lite BERT for self-supervised learning of language representations." arXiv preprint arXiv :1909.11942 (2019).
— Liu, Yinhan, et al. "RoBERTa : A robustly optimized BERT pretraining approach." arXiv preprint arXiv :1907.11692 (2019).
— Sanh, Victor, et al. "DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter." arXiv preprint arXiv :1910.01108 (2019).
— Golzadeh, Mehdi et al. "A ground-truth dataset and classification model for detecting bots in GitHub issue and PR comments." Journal of Systems and Software, Volume 175, 2021
https ://doi.org/10.1016/j.jss.2021.110911

Exigences ou prérequis

Intérêt pour l'intelligence artificielle, le génie logiciel, le développement logiciel, l'apprentissage par machine, les réseaux de neurones.

22. Tom.MENS@umons.ac.be

1.22 Une analyse empirique de la duplication de code dans les CI/CD workflows sur GitHub

Service	Service de Génie Logiciel
Directeur	TOM MENS [23]

Description

Les pratiques d'intégration et de déploiement continus (CI/CD) [1] s'ancrent de plus en plus dans les procédés de développement logiciel collaboratif. Ces deux pratiques visent à accélérer le rythme de production et d'évolution du logiciel, tout en assurant une qualité accrue. Elles intègrent typiquement l'automatisation de la construction du logiciel, l'application de tests logiciels, l'analyse de la qualité logicielle, la détection de défauts et la production de releases, le tout au fur et à mesure de l'évolution du logiciel.

GitHub est la plate-forme principale pour le développement collaboratif des logiciels open source. GitHub Actions est le mécanisme de CI/CD qui est intégré dans GitHub (`https://docs.github.com/en/actions`). Afin d'automatiser le "workflow" d'un processus de développement dans un dépôt GitHub, il convient à créer une ou plusieurs fichiers YAML au sein du répertoire .github/workflows du dépôt. Chaque workflow est typiquement décomposé en jobs et steps.

Notre hypothèse de travail est qu'on peut trouver pas mal de "duplication de code" entre les fichiers workflow de différents dépots GitHub, en supposant qu'un façon habituelle de créer un nouveau workflow consiste à copier-coller un workflow existant d'ailleurs.

Ce mémoire consistera donc à :

— Étudier la littérature (état de l'art) sur les techniques et outils existants permettant de détecter du code dupliqué et du code cloné (e.g., [2, 3, 4]).

— Développer en Python (de préférence) un outil en ligne de commande Linux permettant de détecter de manière automatique du code dupliqué entre des workflows hébergés dans une ou plusieurs dépots GitHub.

— Effectuer une analyse empirique et quantitative de la présence du "code dupliqué" dans les workflow. Est-il commun de trouver du code dupliqué au sein d'un meme dépot, entre différents dépots du meme propriétaire, entre dépots de différents propriétaire ? A quel niveau se situe le code dupliqué ? Y a-t-il moyen d'éviter du code dupliqué en réutilisant des "actions" ? Pour effectuer cette analyse empirique, plusieurs milliers de dépôts GitHub et leurs workflows correspondantes doivent être analysées et comparées.

Exigences ou prérequis

Une excellente connaissance en programmation et développement logiciel.

Références

[1] Mojtaba SHAHIN, Muhammad ALI BABAR et Liming ZHU. « Continuous Integration, Delivery and Deployment : A Systematic Review on Approaches, Tools, Challenges and Practices ». In : *IEEE Access* 5 (2017), p. 3909-3943. DOI : 10.1109/ACCESS.2017.2685629.

[2] Chanchal K. ROY, James R. CORDY et Rainer KOSCHKE. « Comparison and evaluation of code clone detection techniques and tools : A qualitative approach ». In : *Science of Computer Programming* 74.7 (2009), p. 470-495. ISSN : 0167-6423. DOI : https://doi.org/10.1016/j.scico.2009.02.007. URL : https://www.sciencedirect.com/science/article/pii/S0167642309000367.

23. Tom.MENS@umons.ac.be

[3] Qurat Ul AIN, Wasi Haider BUTT, Muhammad Waseem ANWAR, Farooque AZAM et Bilal MAQBOOL. « A Systematic Review on Code Clone Detection ». In : *IEEE Access* 7 (2019), p. 86121-86144. DOI : 10.1109/ACCESS.2019.2918202.

[4] G. SHOBHA, Ajay RANA, Vineet KANSAL et Sarvesh TANWAR. « Code Clone Detection—A Systematic Review ». In : *Emerging Technologies in Data Mining and Information Security*. Sous la dir. d'Aboul Ella HASSANIEN, Siddhartha BHATTACHARYYA, Satyajit CHAKRABATI, Abhishek BHATTACHARYA et Soumi DUTTA. Singapore : Springer Singapore, 2021, p. 645-655. ISBN : 978-981-33-4367-2.

1.23 Analysing the technical lag and versioning strategies in the GitHub Actions ecosystem of software development workflows

Service	Service de Génie Logiciel
Directeur	TOM MENS [24]

Description

Ce sujet de mémoire s'inscrit dans le cadre d'un projet de recherche FNRS intitulé "Analyse empirique, recommandations et améliorations pour les écosystèmes d'automatisation des flux de développement logiciel"

Les pratiques d'intégration et de déploiement continus (CI/CD) [1] s'ancrent de plus en plus dans les procédés de développement logiciel collaboratif. Ces deux pratiques visent à accélérer le rythme de production et d'évolution du logiciel, tout en assurant une qualité accrue. Elles intègrent typiquement l'automatisation de la construction du logiciel, l'application de tests logiciels, l'analyse de la qualité logicielle, la détection de défauts et la production de releases, le tout au fur et à mesure de l'évolution du logiciel.

GitHub est la plateforme de développement collaboratif la plus répandue dans la communauté open source. GitHub Actions est l'outil de CI/CD intégré dans GitHub (`https://docs.github.com/en/actions`). Afin d'automatiser un processus de développement dans un dépôt GitHub, il convient à créer un "workflow", sous la forme d'un fichier YAML au sein du répertoire .github/workflows du dépôt. Un des avantages principaux de GitHub est qu'il y a un marché de "actions" réutilisables qui facilitent la création de workflows complexes avec très peu d'effort (`https://github.com/marketplace`).

Les actions sont versionnées, et peuvent être utilisées dans un workflow à l'aide du mot clé uses :. Par exemple, uses : actions/checkout@v3. L'annotation @v3 précise quelle version de l'action doit être utilisée.

Les Actions ont tendance à utiliser un numéro de version en trois composants Major.Minor.Patch (par exemple version 1.3.0), à l'instar du principe de "semantic versioning" (`https://semver.org`) [2]. Un incrément du composant majeur ne garantie pas de "backward compatibility". Un incrément du composant mineur est sensé de rester compatible même si cette nouvelle version peut incorporer des nouvelles fonctionnalités. Un incrément du composant Patch correspond à des simple corrections des bogues ou failles de sécurité. Ainsi, un utilisateur d'une Action peut rester à jour sans casser la compatibilité s'il utilise toujours la dernière version mineure et patch.

Nous considérons qu'un workflow possède un "technical lag" si les versions des Actions utilisés dans ce workflow ne sont pas à jour. Ce mémoire vise à érudier et analyser empiriquement le lag technique dans l'écosystème de GitHub Actions, en se basant sur le framework conceptuel du technical lag qui a été introduit et analysé empiriquement dans [3, 4].

Exigences ou prérequis

Avoir un fort intérêt en analyse de données, programmation Python, open source, et développement logiciel.

Références

[1] Mojtaba SHAHIN, Muhammad ALI BABAR et Liming ZHU. « Continuous Integration, Delivery and Deployment : A Systematic Review on Approaches, Tools, Challenges and Practices ». In : *IEEE Access* 5 (2017), p. 3909-3943. DOI : 10.1109/ACCESS.2017.2685629.

[2] Alexandre DECAN et Tom MENS. « What do package dependencies tell us about semantic versioning ? » In : *IEEE Transactions on Software Engineering* 47.6 (2019), p. 1226-1240.

24. Tom.MENS@umons.ac.be

[3] Alexandre DECAN, Tom MENS et Eleni CONSTANTINOU. « On the evolution of technical lag in the npm package dependency network ». In : *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE. 2018, p. 404-414.

[4] Ahmed ZEROUALI, Tom MENS, Jesus GONZALEZ-BARAHONA, Alexandre DECAN, Eleni CONSTANTINOU et Gregorio ROBLES. « A formal framework for measuring technical lag in component repositories—and its application to npm ». In : *Journal of Software : Evolution and Process* 31.8 (2019), e2157.

1.24 Identifying bot accounts in GitHub repositories based on their activity feed

Service	Service de Génie Logiciel
Directeur	TOM MENS [25]

Description

The exact contents of this thesis proposal still needs to be written. The current text is just an unstructured draft of ideas.

GitHub accounts can be used to analyse the activities of the contributors to git repositories hosted on GitHub. These activities can be quite diverse, including issues, pull requests, commits, code reviews, discussions, commenting. The activity feed of the last 30 days of each GitHub account can easily be retrieved using the GitHub API (`https://docs.github.com/en/graphql`).

Accounts in GitHub can represent either human contributors or bots (machine accounts). Bots [1, 2] are automated agents that support repetitive activities on behalf of some human contributor.

The goal of this masters thesis is to develop a classification model to distinguish bot accounts from human accounts based on information that can be extracted from their recent activity feed on GitHub. (This feed does not only contain the kinds and numbers of activities that have been carried out by the account, but also their order, their timestamp, their contents, and lots of other potentially relevant information).

Our hypothesis is that such a classification model is likely to provide either more accurate or faster results then existing bot detection models such as BoDeGHa that was recently developed in our lab [3].

The classification model should be created using some ML or AI approach that is the most appropriate for this goal. The performance and accuracy of the model should be evaluated on the basis of a ground-truth dataset of bots, and compared against state-of-the-art bot detection models such as BoDeGHa.

Exigences ou prérequis

Références

[1] Carlene LEBEUF, Margaret Anne STOREY et Alexey ZAGALSKY. « Software Bots ». In : *IEEE Software* 35.1 (2017), p. 18-23. ISSN : 07407459. DOI : 10.1109/MS.2017.4541027.

[2] Mairieli WESSEL, Bruno Mendes DE SOUZA, Igor STEINMACHER, Igor S. WIESE, Ivanilton POLATO, Ana Paula CHAVES et Marco A. GEROSA. « The power of bots : Understanding bots in OSS projects ». In : *ACM Conference on Human-Computer Interaction*. 2018. DOI : 10.1145/3274451.

[3] Mehdi GOLZADEH, Alexandre DECAN, Damien LEGAY et Tom MENS. « A ground-truth dataset and classification model for detecting bots in GitHub issue and PR comments ». In : *Journal of Systems and Software* 175 (mai 2021). DOI : https://doi.org/10.1016/j.jss.2021.110911.

25. Tom.MENS@umons.ac.be

1.25 Apprentissage de machines de Moore

Service	Informatique Théorique
Directeur	VERONIQUE BRUYERE [26]

Description

L'*apprentissage d'automates* trouve de nombreuses applications en traitement de la parole, en traduction automatique, en vérification et synthèse de systèmes informatiques, en biologie computationelle, en data mining, et même en musique (voir le survey sur le sujet écrit par de la Higuera [1])

Dans le cadre de ce mémoire, on s'intéresse à l'*apprentissage passif*. Dans le cas des automates, un automate déterministe de taille minimale est appris automatiquement à partir d'un ensemble donné de mots classés comme appartenant ou n'appartement pas au langage accepté par l'automate (voir par exemple la thèse de Daniel Neider [2]). Ce mémoire se concentre sur l'apprentissage passif de *machines de Moore* qui sont des automates étendus avec des entrées et des sorties. Dans un premier temps, l'étudiant étudiera les algorithmes d'apprentissage proposés dans l'article récent [3] et proposera une implémentation afin de reproduire les résultats expérimentaux qui y sont présentés. Dans un second temps, l'étudiant s'intéressera à l'apprentissage d'un modèle d'automate proche des machines de Moore : les machines de Mealy (voir par exemple [4]) qui semble avoir été plus étudié et pour lequel on dispose de librairies. Le but sera de comparer les techniques d'apprentissage pour ces deux familles de machines.

Exigences ou prérequis

Algorithmique et Structures de données. Calculabilité et Complexité.

Références

[1] Colin de la HIGUERA. « A bibliographical study of grammatical inference ». In : *Pattern Recognit.* 38.9 (2005), p. 1332-1348. DOI : 10.1016/j.patcog.2005.01.003. URL : https://doi.org/10.1016/j.patcog.2005.01.003.

[2] Daniel NEIDER. « Applications of automata learning in verification and synthesis ». Thèse de doct. RWTH Aachen University, 2014. URL : http://darwin.bth.rwth-aachen.de/opus3/volltexte/2014/5169.

[3] Georgios GIANTAMIDIS, Stavros TRIPAKIS et Stylianos BASAGIANNIS. « Learning Moore machines from input-output traces ». In : *Int. J. Softw. Tools Technol. Transf.* 23.1 (2021), p. 1-29. DOI : 10.1007/s10009-019-00544-0. URL : https://doi.org/10.1007/s10009-019-00544-0.

[4] Muzammil SHAHBAZ et Roland GROZ. « Inferring Mealy Machines ». In : *FM 2009 : Formal Methods, Second World Congress, Eindhoven, The Netherlands, November 2-6, 2009. Proceedings*. Sous la dir. d'Ana CAVALCANTI et Dennis DAMS. T. 5850. Lecture Notes in Computer Science. Springer, 2009, p. 207-222. DOI : 10.1007/978-3-642-05089-3_14. URL : https://doi.org/10.1007/978-3-642-05089-3%5C_14.

26. Veronique.BRUYERE@umons.ac.be

1.26 Apprentissage d'automates en présence d'un professeur inexpérimenté

Service	Informatique Théorique
Directeur	VERONIQUE BRUYERE [27]

Description

Le thème de ce mémoire concerne l'*apprentissage d'automates*, domaine qui trouve de nombreuses applications en traitement de la parole, en traduction automatique, en vérification et synthèse de systèmes informatiques, etc (voir par exemple le survey écrit par de la Higuera [1]). Dans le cadre de l'apprentissage *actif*, un professeur connaît un langage régulier L et un élève tente de construire un automate acceptant ce langage L en posant des questions au professeur du type : "est-ce que ce mot appartient à L ?", "est-ce que cet automate accepte L ?".

Dans le cadre de ce mémoire, on se focalise sur le cas d'un professeur qui n'a pas toujours la réponse à la question posée. Par exemple, il ne peut pas répondre si oui ou non un mot donné appartient au langage L. Le point de départ du mémoire sera l'article [2] qui présente différentes techniques d'apprentissage lorsqu'on a affaire à un professeur inexpérimenté.

Dans un premier temps, on propose de comprendre certains de ces algorithmes ainsi que leur complexité. On demande ensuite d'implémenter ces algorithmes. Dans certains cas, l'implémentation pourra s'appuyer sur des librairies existentes sur les automates, dans d'autres cas, on utilisera des SAT solveurs.

Exigences ou prérequis

Il faut aimer l'algorithmique et les structures de données. Il est conseillé de suivre le cours de *Calculabilité et complexité*.

Références

[1] Colin de la HIGUERA. « A bibliographical study of grammatical inference ». In : *Pattern Recognit.* 38.9 (2005), p. 1332-1348. DOI : 10.1016/j.patcog.2005.01.003. URL : https://doi.org/10.1016/j.patcog.2005.01.003.

[2] Martin LEUCKER et Daniel NEIDER. « Learning Minimal Deterministic Automata from Inexperienced Teachers ». In : *Leveraging Applications of Formal Methods, Verification and Validation. Technologies for Mastering Change - 5th International Symposium, ISoLA 2012, Heraklion, Crete, Greece, October 15-18, 2012, Proceedings, Part I*. Sous la dir. de Tiziana MARGARIA et Bernhard STEFFEN. T. 7609. Lecture Notes in Computer Science. Springer, 2012, p. 524-538. DOI : 10.1007/978-3-642-34026-0_39. URL : https://doi.org/10.1007/978-3-642-34026-0%5C_39.

27. Veronique.BRUYERE@umons.ac.be

1.27 Regroupement de données dans des sous-espaces linéaires

Service	Mathématique et Recherche opérationnelle (FPMs)
Directeur	NICOLAS GILLIS [28]

Description

Le problème de regroupement (clustering) a pour objectif de classifier des données dans différents ensembles (on supposera que chaque élément de ces données est représenté par un vecteur de $\mathbb{R}^m$). Par exemple, le clustering permet de classifier un ensemble de documents en fonction du sujet traité, ou encore de classifier des images en fonction de leur type (paysage, portrait, etc.) ou de ce qu'elles contiennent (p.ex., chien vs. chat, motos vs. voitures, etc.). Dans ce travail, on propose de se concentrer sur un problème de clustering particulier : on va supposer que les éléments d'un même sous-ensemble appartiennent à un sous-espace linéaire de faible dimension (ce problème est appelé subspace clustering en anglais). En termes mathématiques, si $x_1, x_2, \ldots, x_n \in \mathbb{R}^m$ sont des points appartenant au même sous-ensemble, alors on va supposer qu'il existe un petit ensemble de vecteurs $y_1, y_2, \ldots, y_r \in \mathbb{R}^m$ (où $r \ll n$) tels que pour tout $1 \leq i \leq n$:

$$x_i = \sum_{k=1}^{r} \alpha_i(k) y_k + n_i \text{ pour des poids appropriés } \alpha_i \in \mathbb{R}^r, \text{ et du bruit } n_i \in \mathbb{R}^m.$$

Dans ce TFE, on propose d'étudier différents algorithmes pour résoudre ce problème. En fonction des préférences de l'étudiant, ces algorithmes pourront être basés sur l'optimisation convexe (en particulier, l'optimisation linéaire) et/ou sur des métaheuristiques (optimisation combinatoire). On appliquera ces algorithmes à des problèmes de classifications de documents et d'images ; voir ci-dessous pour deux exemples d'applications.

Fig. 1. Motion segmentation: given feature points on multiple rigidly moving objects tracked in multiple frames of a video (top), the goal is to separate the feature trajectories according to the moving objects (bottom).

Fig. 2. Face clustering: given face images of multiple subjects (top), the goal is to find images that belong to the same subject (bottom).

FIGURE 1.1 – Image provenant de l'article : Elhamifar and Vidal, Sparse Subspace Clustering : Algorithm, Theory, and Applications, IEEE Trans. on Pattern Analysis and Machine Intelligence 35(11) : 2765-2781, 2013.

28. Nicolas.GILLIS@umons.ac.be

1.28 La factorisation positive de matrices et ses applications

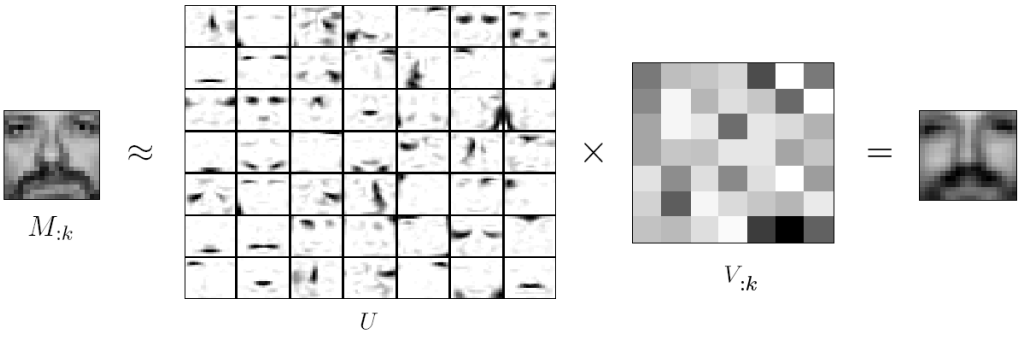
Service	Mathématique et Recherche opérationnelle (FPMs)
Directeur	NICOLAS GILLIS [29]

Description

Ce travail de fin d'étude a pour objet le problème de la factorisation positive de matrices (en anglais : nonnegative matrix factorization – NMF) qui peut être défini comme suit : étant donné une matrice M positive (c'est-à-dire une matrice dont toutes les entrées sont positives et on écrira $M \geq 0$) ayant p lignes et n colonnes et un rang de factorisation $r \ll \min(p, n)$, on désire trouver deux matrices U (p lignes, r colonnes) et V (r lignes, n colonnes) également positives et telles que $M \approx UV$. Une première application de la NMF est la compression de M : en effet, si r est suffisamment petit (précisément, si $r < \frac{np}{n+p}$), le nombre d'entrées dans U et V ($= pr + rn$) est plus petit que le nombre d'entrées dans M ($= np$).

La NMF comporte un grand nombre d'applications [30], notamment

— En analyse d'images, la NMF permet d'extraire automatiquement des caractéristiques communes et localisées de ces images ; voir ci-dessous où la NMF a été appliquée à un ensemble d'images de visages.



— En analyse de textes, la NMF permet d'identifier automatiquement différentes catégories/sujets (c'est-à-dire des ensembles de mots qui apparaissent simultanément dans un sous ensemble de textes) et de classifier les textes dans ces différentes catégories.

Dans ce TFE, il y aura la possibilité d'étudier différents aspects de ce problème :

1. (Algorithmes) Mise au point d'algorithmes efficaces pour calculer des factorisations positives $(U, V) \geq 0$ étant donné une matrice $M \geq 0$.
2. (Applications) Utiliser la NMF pour différentes applications, et la comparer à d'autres techniques existantes.
3. (Théorie) Malgré que la NMF soit un problème très difficile, il est possible de mettre au point des algorithmes efficaces dans certains cas particuliers, et de prouver leur efficacité.

29. Nicolas.GILLIS@umons.ac.be
30. Voir par exemple le livre récent bit.ly/bookNMF.

1.29 Logique et apprentissage

Service	Mathématiques Effectives
Directeur	MICKAEL RANDOUR [31]

Description

La logique mathématique constitue un puissant cadre de raisonnement pour analyser les systèmes complexes. C'est la pierre angulaire des méthodes formelles – un ensemble d'outils permettant d'étudier des garanties fortes sur des systèmes complexes. L'apprentissage automatique est un paradigme de plus en plus présent, dont les applications paraissent sans limite. Il est principalement basé sur des approches statistiques à l'efficacité sans pareille. Ce projet participe à un effort récent visant à combiner le raisonnement formel issu de la logique et la puissance de l'apprentissage automatique. D'un côté, introduire de l'apprentissage dans les méthodes formelles permet de résoudre de nombreux problèmes de performance, tout en conservant des garanties fortes. De l'autre, la logique offre des fondations mathématiques pour évaluer la qualité de processus d'apprentissage et leurs résultats, et d'ainsi garantir leur fiabilité et leur robustesse, à une époque où l'apprentissage et l'IA sont au coeur d'applications critiques.

Exigences ou prérequis

Intérêt pour les méthodes formelles, la théorie des jeux et/ou l'IA.

31. mickael.randour@umons.ac.be

1.30 Synthèse Multicritère de Systèmes Réactifs : Fondations, Algorithmes et Outils

Service	Mathématiques Effectives
Directeur	MICKAEL RANDOUR [32]

Description

Nous vivons l'ère de l'intelligence ambiante : nous sommes entourés de systèmes (informatiques) réactifs (SIRs) qui interagissent continuellement avec leur environnement via des instructions utilisateur, des senseurs, etc. Leur exactitude est souvent cruciale, soit pour des raisons de sécurité (p.ex. ABS), soit pour des contraintes de production de masse (p.ex. smartphones). Néanmoins, leur conception est complexe et sujette aux erreurs. La vérification formelle et la synthèse sont deux réussites de l'informatique, qui ont pour but la construction automatique de contrôleurs fiables pour les SIRs. Beaucoup de techniques se fondent sur la théorie des jeux, modélisant les interactions entre le SIR et son environnement sous forme de jeu compétitif. Au cours de la dernière décennie, le domaine a évolué des spécifications Booléennes vers celles dites quantitatives, donnant naissance à des modèles capables de décrire la performance de SIRs. Cependant, les modèles actuels ne permettent de représenter qu'un seul aspect quantitatif (ou qualitatif) à la fois : ils ne tiennent pas compte de leurs interactions et des compromis en résultant. De tels compromis peuvent apparaître entre différentes ressources (p.ex. diminuer le temps de réponse requiert plus de puissance de calcul et de consommation énergétique) mais aussi entre différents modèles comportementaux (p.ex. performance dans le pire des cas vs. en moyenne). Ces interactions sont au coeur des scénarios réels et demandent que les développeurs décident de l'équilibre entre différents aspects. Il est donc nécessaire de fournir des modèles et des outils capables de représenter ces interactions pour que la synthèse soit efficace en pratique. L'objectif de ce projet est de mettre en place la prochaine génération de techniques de synthèse en établissant les bases formelles, les algorithmes et les outils nécessaires pour changer de paradigme, depuis les modèles qualitatifs et quantitatifs monocritères vers les multicritères.

Exigences ou prérequis

Intérêt pour les méthodes formelles, la théorie des jeux et/ou l'IA.

32. mickael.randour@umons.ac.be

1.31 Contrôleurs pour la Synthèse de Systèmes Réactifs : une Perspective Stratégique

Service	Mathématiques Effectives
Directeur	MICKAEL RANDOUR [33]

Description

Nous vivons l'ère de l'intelligence ambiante : nous sommes entourés de systèmes (informatiques) réactifs (SIRs) qui interagissent continuellement avec leur environnement via des instructions utilisateur, des senseurs, etc. Leur exactitude est souvent cruciale, soit pour des raisons de sécurité (p.ex. ABS), soit pour des contraintes de production de masse (p.ex. smartphones). Néanmoins, leur conception est complexe et sujette aux erreurs. La vérification formelle et la synthèse sont deux réussites de l'informatique, qui ont pour but la construction automatique de contrôleurs fiables pour les SIRs. Beaucoup de techniques se fondent sur la théorie des jeux, modélisant les interactions entre le SIR et son environnement sous forme de jeu compétitif. Le domaine a évolué des spécifications booléennes vers les quantitatives, amenant des modèles capables de décrire la performance de SIRs. Les recherches récentes se concentrent sur les interactions entre différents aspects quantitatifs (ou qualitatifs) et les compromis en résultant. De tels compromis peuvent apparaître entre différentes ressources (p.ex. diminuer le temps de réponse requiert plus de puissance de calcul et de consommation énergétique) mais aussi entre différents modèles comportementaux (p.ex. performance dans le pire des cas vs en moyenne). Ces interactions sont au coeur des scénarios réels et demandent que les développeurs décident de l'équilibre entre différents aspects. Mon groupe de recherche est l'un des précurseurs sur la synthèse multicritère. L'objectif de ce projet est de questionner le concept central de stratégie, actuellement basé sur des modèles inspirés des automates finis, et servant de canevas formel pour des contrôleurs implémentables. Mon but est d'augmenter la compréhension théorique et l'utilité pratique de ce concept abstrait via l'étude systématique de modèles de stratégie alternatifs. Ce projet présage des avancées fondamentales vers la synthèse multicritère réellement applicable.

Exigences ou prérequis

Intérêt pour les méthodes formelles, la théorie des jeux et/ou l'IA.

33. mickael.randour@umons.ac.be

1.32 Équité en apprentissage actif

Service	Mathématique et Recherche opérationnelle (FPMs)
Directeur	XAVIER SIEBERT [34]

Description

Le paradigme de l'apprentissage statistique traditionnel (« passif ») consiste à apprendre un concept, par exemple un classifieur, obtenu à partir de données étiquetées. Dans plusieurs applications, telle que la classification d'images, le processus d'étiquetage des données constitue en soi une tâche difficile, ce qui nous contraint à utiliser d'autres techniques d'apprentissage. L'une des plus prometteuses est l'apprentissage actif, dont le but principal est de réduire l'effort d'étiquetage [1], en choisissant séquentiellement les données à étiqueter, de manière à ce que les plus informatives (selon un critère bien défini) soient sélectionnées. Dans ce cas, le nombre de données étiquetées requis pour apprendre un concept est souvent nettement plus petit que celui obtenu dans le cadre de l'apprentissage passif, où les données sont étiquetées de manière aléatoire.
D'autre part, plusieurs travaux récents en apprentissage statistique ont mis en évidence les problèmes d'éthique et d'équité (*fairness*) auxquels sont confrontés la plupart des algorithmes de classification [2],[3]. Bien qu'atteignant de bonnes performances de classification, ils ne se soucient pas de l'aspect éthique et il apparaît clairement que plusieurs groupes sont parfois discriminés (e.g., race, genre, etc...).Il serait donc intéressant d'avoir des algorithmes performants qui soient le plus justes possible, afin de traiter les individus (i.e., les données) de la même manière, indépendamment de leur appartenance à un groupe spécifique.

Il existe plusieurs modélisations théoriques de l'équité en apprentissage statistique. L'objectif de ce projet sera de les comprendre afin d'identifier celle qui pourrait au mieux s'appliquer au contexte de l'apprentissage actif. Il serait ensuite question, sous certaines hypothèses, de construire un algorithme d'apprentissage actif qui prenne en considération l'équité [4].

Exigences ou prérequis

Intérêt pour la théorie de l'apprentissage statistique.

Références

[1] Sanjoy DASGUPTA. « Two faces of active learning ». In : *Theoretical computer science* 412.19 (2011), p. 1767-1781.

[2] Ninareh MEHRABI, Fred MORSTATTER, Nripsuta SAXENA, Kristina LERMAN et Aram GALSTYAN. « A survey on bias and fairness in machine learning ». In : *ACM Computing Surveys (CSUR)* 54.6 (2021), p. 1-35.

[3] Solon BAROCAS, Moritz HARDT et Arvind NARAYANAN. « Fairness in machine learning ». In : *Nips tutorial* 1 (2017), p. 2.

[4] Yiting CAO et Chao LAN. « Active Approximately Metric-Fair Learning ». In : *The 38th Conference on Uncertainty in Artificial Intelligence*. 2022.

34. Xavier.SIEBERT@umons.ac.be

1.33 Analyse de la robustesse du classifieur KNN aux attaques antagonistes.

Service	Mathématique et Recherche opérationnelle (FPMs)
Directeur	XAVIER SIEBERT [35]

Description

Une attaque antagoniste consiste à fournir à un modèle de classification un exemple intentionnellement modifié, avec la possibilité d'induire une classification erronée. Ce phénomène a des applications dans plusieurs domaines de l'intelligence artificielle, notamment dans le domaine de la santé, la robotique, la finance, la conduite autonome (Fig. 1.2).

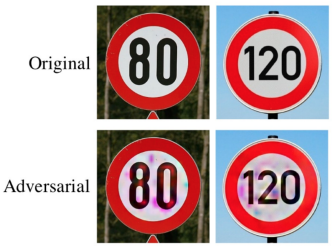


FIGURE 1.2 – Exemples de panneaux routiers originaux (haut) et après une attaque antagoniste (bas).

Dans ce projet, le modèle de classification des K-nn (ou K-plus proches voisins) fera l'objet d'une étude de la robustesse aux attaques antagonistes, afin d'identifier dans quelles situations il est possible d'obtenir simultanément un taux optimal de prédiction et de robustesse [1].

Exigences ou prérequis

Intérêt pour la théorie de l'apprentissage statistique et/ou l'IA.

Références

[1] Yizhen WANG, Somesh JHA et Kamalika CHAUDHURI. « Analyzing the robustness of nearest neighbors to adversarial examples ». In : *International Conference on Machine Learning*. PMLR. 2018, p. 5133-5142.

35. Xavier.SIEBERT@umons.ac.be