TAILOR

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

Automated AI v.1

Document type (nature)	Report
Deliverable No	7.1
Work package number(s)	7
Due Date	Due 31 August 2022
Responsible Beneficiary	ULEI, ID #7
Author(s)	Koen van der Blom
Publicity level	Public
Short description	Version 1 of report on novel insights, techniques, algorithms and tools developed within tasks T7.1, T7.2, T7.3, T7.4 and T7.5, and how they contribute towards trustworthy AI.

History			
Revision	**Date**	**Modification**	**Author**
1.0	2022-12-08	first version	Koen van der Blom

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Michela Milano	UNIBO	18/11/2022
Barry O'Sullivan	UCC	2/12/2022

Table of Contents

Summary of the report

This report gives an overview of novel insights, techniques, algorithms and tools in the area of automated AI (AutoAI) and how they contribute towards trustworthy AI. More specifically, an overview is given for each of the five areas of AutoAI detailed in T7.1-T7.5 of the TAILOR project.

Organisation

The following experts from the TAILOR consortium have been involved in the writing of this report, based on materials collected across a broad range of project partners:

Partner ID / Acronym	Name	Role
#7, ULEI	Koen van der Blom Jan van Rijn Mitra Baratchi Holger Hoos	WP7, T7.2, T7.3 Leader
#17, ALU-FR	Eddie Bergman	T7.1 Leader
#12, TUE	Pieter Gijsbers Joaquin Vanschoren	T7.5 Leader
#10, UNIBO	Andrea Borghesi	Participant
#34, NEO-UMA	Francisco Chicano Manuel López-Ibáñez Zakaria Abdelmoiz DAHI	Participant
#3, INRIA	Marc Schoenauer	T7.4 Leader
#14, JSI	Sašo Džeroski	Participant
#33, NKUA	Vana Kalogeraki Dimitrios Gunopulos	Participant
#35, PUT	Krzysztof Krawiec	Participant
#43, UPV	Alfons Juan	Participant
#23, CRIL	Daniel Le Berre	Participant

1. Introduction

The increasingly widespread adoption of AI in recent years necessitates more efficient use of AI tools. AutoAI aims to automate (parts of) the construction and setup of AI systems. As a result, the adoption of AI should require less expertise and thus allow more people to benefit from it. In this report we discuss recent developments in AutoAI, focused on five main areas. Following that, advances in a few other areas are also highlighted and some recently developed AutoAI benchmarks are discussed.

Firstly, the most developed area of AutoAI is automated machine learning (AutoML). Evenso, adoption for specific applications can still be challenging. As such, Section 2 (*AutoML in the wild*) focuses on how to facilitate its use in practice, e.g., by providing tools that visualise parts of the procedure and results.

Secondly, while AutoML is the area within AutoAI that has received most attention, it has been primarily focused on supervised machine learning, and within that area, primarily on tabular data. In Section 3 (*Beyond standard supervised learning*) developments are covered that aim to advance AutoML to new contexts such as data science and reinforcement learning.

Thirdly, with the goal of AutoAI requiring less expertise also comes the responsibility to include warning systems that automatically detect (or avoid) issues with a (deployed) AI system. To this end Section 4 (*Self-monitoring AI systems*) covers advances in areas such automated neural network verification.

Fourthly, the development of AutoAI has been almost exclusively focused on single-objective applications (e.g., minimising a loss function). In practice, however, many applications benefit from taking into account multiple objectives, which is the focus of Section 5 (*Multi-objective AutoAI*). The inclusion of fairness objectives is particularly important.

Fifthly, AI systems are often trained from scratch, requiring significant amounts of data and compute power. However, these resources are not always available, making more efficient approaches not only nice, but also a necessity. To this end, Section 6 (*Ever-learning AutoAI*) goes into meta-learning and related techniques to automatically construct AI systems from far fewer samples than standard AI systems need to be trained.

All of these areas also contribute to the larger goal of trustworthy AI in the TAILOR project. Facilitating the use of AutoML in practice helps to further democratise the use of ML, as does the extension of AutoML to other areas beyond standard supervised learning. Self-monitoring AI systems help to improve the trust in AI by including checking and warning mechanisms in AI systems. The development of multi-objective AutoAI is vital to account for different aspects of fairness and trustworthiness. Finally, ever-learning AutoAI helps both to democratise AI by making deployment to new situations more efficient and to quickly adapt systems when trustworthiness issues are detected.

In the context of the TAILOR project, work resulting from this is highlighted by marking the references in **bold**. In addition to work by individual partners, a number of these works are the result of collaborations between multiple TAILOR partners. This report is, however, not limited to work developed within the TAILOR consortium and also highlights some important other recent developments.

2. AutoML in the wild [T7.1, ALU-FR]

Goal: Facilitate the usability of machine learning by non-machine-learning-experts who have data and a clear target to predict, but who are not familiar enough with machine learning to know which neural architecture or machine learning pipeline to use, and how to set its hyperparameters.

2.1. Automated algorithm configuration

To facilitate non-experts in core areas of machine learning, we have made great strides towards really creating "*Auto"* AI, tools that aim to flexibly replace the parts of AI that require human expertise, while leaving non-machine-learning-experts to focus on what matters, defining their own problem to which to apply AI. This is a core goal to enable the use of AutoML in the wild.

A notable tool that has succeeded in being introduced into the wild is [FeuEtAl22] AutoSklearn: *Hands-free AutoML around sklearn* (~40,000 downloads per month) which removes the need for a non-machine-learning-export to configure or optimise individual algorithms, finding the model pipeline that suits their problem at hand.
A counterpart to this for Deep Learning is [ZimEtAl21] Auto-Pytorch: *Hands-free AutoDL around pytorch* and also see the subsection for work towards automating reinforcement learning in a similar fashion.

There are also continuing advances in the state of the art in this domain by utilising previously gained information in a meta-learning fashion, significantly reducing required time to find suitable and performant models for each new problem.

- **Öztürk, Ekrem; Ferreira, Fabio; Jomaa, Hadi S.; Schmidt-Thieme, Lars; Grabocka, Josif; Hutter, Frank, "Zero-shot AutoML with Pretrained Models", International Conference on Machine Learning (ICML), 2022.**

The algorithms' automatic configuration analysis can allow users to compare the quality of the configurations and their influence on the training and test instances. Therefore, a tool named ACVIZ has been proposed to analyse the automatic configuration of algorithms with irace [LopEtAl16] providing a representation of the configuration process which allows extracting insightful information such as the evolution of the configuration over time. In addition, having the test data at hand allows also to display the performance of each configuration on the test instances.

- **[LopEtAl16]** M. López-Ibáñez, J. Dubois-Lacoste, L. Pérez Cáceres, T. Stützle, and M. Birattari, "The irace package: Iterated racing for automatic algorithm configuration," Operations Research Perspectives, vol. 3, pp. 43–58, 2016.

- **ACVIZ: A tool for the visual analysis of the configuration of algorithms with irace. de Souza , M., Ritt, M., López-Ibáñez, M., & Pérez Cáceres, L. Operations Research Perspectives, 8:100186, 2021.**

Automatic configuration is a widely spread and successful approach for algorithm tuning. However, it is computationally expensive, considering the need for multiple configuration evaluations over multiple instances. To reduce the running time when dealing with decision problems, capping techniques are used to discard low-quality configurations. Such an approach cannot be directly applied to optimisation problems where the objective is to optimise the cost of the best solution encountered. To cope with this, an automatic algorithm configuration approach has been proposed based on new capping methods that use the precedent executions to determine a performance envelope, which later will be used to evaluate new executions and cap those not fulfilling the envelope conditions.

- **Marcelo De Souza, Marcus Ritt, and Manuel López-Ibáñez. Capping Methods for the Automatic Configuration of Optimization Algorithms. Computers & Operations Research, 139:105615, 2022.**

Automatic configuration of sampling algorithms based on Markov Chain Monte Carlo (MCMC) methods can allow users to answer inferential queries (arising in several domains including statistical physics, computational biology and machine learning) in a more efficient way. A new algorithm called LSB has been introduced to automate the configuration of MCMC when sampling from discrete distributions, as well as to reduce the number of query evaluations required to draw samples from such distributions. The key contribution is based on introducing mutual information as an objective criterion to learn the proposal distribution.

- **Emanuele Sansone. LSB: Local Self-Balancing MCMC in Discrete Spaces. International Conference on Machine Learning (ICML), 2022.**

2.2. Hyperparameter optimization

One of the greatest difficulties with AutoML is Hyperparameter Optimization (HPO). This is the result of the vast number of choices a non-machine-learning-expert is confronted with to apply ML to their problem. HPO addresses this and is the backbone of any modern AutoML system. The goal is to sensibly explore the parameter space and to find performant parameter settings. Before the introduction of HPO (and automated HPO in particular), this was only available to those with tremendous compute power, or the scarcely available expert knowledge.

One popular tool which is the culmination of many years of research is SMAC3: *A Versatile Bayesian Optimization Package for HPO* [LinEtAl22] (~50,000 downloads per month) which makes many different HPO methods available that are accessible to both researchers and users in the wild alike. We continue to extend these ideas regularly with recent advances being made in allowing humans with expert ML knowledge to use these automated tools efficiently:

- **Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, Luigi Nardi, “PiBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization”, ICLR 2022**

However, although these HPO tools improve efficiency, they still require significant computing power, limiting the use of AutoML to only those with sufficient resources. A common approach to reduce the required computational resources is to reuse information gained from previous data, using *ever-learning AutoAI* to make these approaches practical for everyday users:

- Feurer, Matthias; Letham, Benjamin; Hutter, Frank; Bakshy, Eytan, “Practical Transfer Learning for Bayesian Optimization”, arXiv:1802:02219v3 [stat.ML], 2022.

There is also ongoing work on building new tools in the regards that advance state of the art:

- **[NooEtAl21]** Awad, Noor; Mallik, Neeratyoy; Hutter, Frank, “DEHB: Evolutionary Hyberband for Scalable, Robust and Efficient Hyperparameter Optimization”, IJCAI 2021.
    - https://github.com/automl/DEHB

One essential part of successful AutoML in the wild is to allow non-machine-learning-experts to inquire into the details of the automation process. To this end tools are developed to peer into the black-box optimization process and monitor the optimization process:

- **[RenEtAl22] DeepCAVE: *An Interactive Analysis Tool for HPO in AutoML***
    - https://github.com/automl/DeepCAVE
    - https://arxiv.org/abs/2206.03493

2.3. Neural architecture search

Identifying an adequate configuration of Generative Artificial Neural Networks (GANs) is tricky and challenging considering that GANs train simultaneously two deep neural networks. So, unlike other neural networks, GANs have a larger number of parameters to be configured which increases the tuning complexity. Considering these facts, a new parallel/distributed methodology has been devised for GANs hyperparameters’ configurations. The proposal applies the iterated racing approach taking advantage of parallel/distributed computing for the efficient use of resources during the configuration.

- Esteban, M., Toutouh, J., Nesmachnow, S. (2021). Parallel/Distributed Intelligent Hyperparameters Search for Generative Artificial Neural Networks. In: Jagode, H., Anzt, H., Ltaief, H., Luszczek, P. (eds) High Performance Computing. ISC High Performance 2021. Lecture Notes in Computer Science, vol 12761. Springer, Cham. https://doi.org/10.1007/978-3-030-90539-2_20

2.4. Real-world applications

Decision-making in economics is a complex task that goes beyond the capacities of manual and empirical methods. Artificial intelligence is intended to provide a more efficient alternative to such primitive techniques and contribute to the decision-making process. Considering these facts, an AI-based approach has been devised for performing both automatic-nation-wise profiling and prediction based on economic features. The profiling is achieved using unsupervised learning through K-means where hyperparameters such as the initial centroids are automatically optimised/configured through evolutionary computation. On the other hand, the prediction is done via supervised learning built-up on long-short term memory neural networks.

- **Dahi, Z.A., Luque, G., Alba, E. (2022). A Machine Learning-Based Approach for Economics-Tailored Applications: The Spanish Case Study. In: Jiménez Laredo, J.L., Hidalgo, J.I., Babaagba, K.O. (eds) Applications of Evolutionary Computation. EvoApplications 2022. Lecture Notes in Computer Science, vol 13224. Springer, Cham. https://doi.org/10.1007/978-3-031-02462-7_36**

In real-world applications such as traffic light scheduling problems, solutions require the simulation under various scenarios in such a way that good solutions should be robust against several scenarios as well as achieving good performance on the fitness function. The irace technique [LopEtAl16] has already proven its efficiency for such tasks. However, the irace operators for solution generation were designed for configuring algorithm parameters with different data types, while evolutionary algorithms possess promising operators for numerical optimisation that could help sample new solutions from the best found in the search. Therefore, a hybridisation has been carried out of irace's elitist iterated racing mechanism with evolutionary operators from differential evolution and the genetic algorithm.

- **Cintrano, C., Ferrer, J., López-Ibáñez, M., Alba, E. (2021). Hybridization of Racing Methods with Evolutionary Operators for Simulation Optimization of Traffic Lights Programs. In: Zarges, C., Verel, S. (eds) Evolutionary Computation in Combinatorial Optimization. EvoCOP 2021. Lecture Notes in Computer Science(), vol 12692. Springer, Cham. https://doi.org/10.1007/978-3-030-72904-2_2**

When responding to a pandemic situation, policy makers rely on forecasts of the spread. In the context of the current COVID-19 pandemic, various sophisticated epidemic and machine learning models have been used for forecasting. In a recently accepted paper [TetEtAl22] we study the role of open data along with AutoML systems in acquiring high-performance forecasting models for COVID-19. We adapted the AutoML framework auto-sklearn to the time series forecasting task and introduced two variants for multi-step ahead COVID-19 forecasting which we refer to as (a) multi-output and (b) repeated single output forecasting. We studied the usefulness of anonymized open mobility data sets (place visits, and the use of different transportation modes) in addition to open mortality data.

- **[TetEtAl22] J. Tetteroo, M. Baratchi and H. H. Hoos, "Automated Machine Learning for COVID-19 Forecasting," in IEEE Access, vol. 10, pp. 94718-94737, 2022, doi: 10.1109/ACCESS.2022.3202220.**

Furthermore, a systems framework for incorporating ML techniques for privacy preservation in a mobile cloud environment has been developed. This is a particularly important application due to the prevalence of cloud-based solutions in several applications.

- **[TomEtAl22a]** Dimitrios Tomaras, Michail Tsenos, Vana Kalogeraki, Practical Privacy Preservation in a Mobile Cloud Environment. MDM 2022, June 2022
- **[TomEtAl22b]** A framework for supporting privacy preservation functions in a Mobile Cloud Environment. Dimitrios Tomaras, Michail Tsenos, Vana Kalogeraki, MDM 2022, June 2022.

Finally, a flexible streaming framework that enables the application of learning techniques for predicting demand and dynamically incorporating such predictions to optimise resource management in real-time has been recently completed.

- **[TseEtAl22]** Michail Tsenos, Aristotelis Peri, Vana Kalogeraki,AMESoS: A Scalable and Elastic Framework for Latency Sensitive Streaming Pipelines, DEBS 2022, June 2022.
- **[GiaKal22]** Thanos Giannakopoulos and Vana Kalogeraki, An Elastic and Scalable Topic-based Pub/Sub System using Deep Reinforcement Learning. DAIS 2022, June 2022.

3. Beyond standard supervised learning [T7.2, ULEI]

Goal: Expand the scope of AutoML, an important special case of AutoAI, to diverse and rich learning settings.

3.1. Spatio-temporal data for earth observation

Recently, there have been advancements in the development of AutoML techniques that help easy adoption of AI techniques for spatio-temporal data. One focus area has been on special pre-processing tasks in the pipeline of machine learning tasks from such data. In **[ArpEtAl21],** a technique is presented powered by AutoML for spatial interpolation (i.e., filling the missing values). Currently, the work is being extended in collaboration with researchers from the European Space Agency for recovering the data of the cloud-obscured images.

Another important preprocessing task in the machine learning pipeline of spatio-temporal data (e.g., satellite images) is super-resolution (SR). This is the image processing approach to increase the resolution of images. In [WasEtAll22] AutoSR4EO is proposed, an AutoML method for automatically constructing a neural network for SR. This is achieved by designing a search space based on state-of-the-art methods in SR and incorporating transfer learning from available EO datasets. This research is performed in collaboration with researchers of the European Space Agency.

- **[ArpEtAl21] Arp L, Baratchi M, Hoos H, VPint: value propagation-based spatial interpolation, Data Mining and Knowledge Discovery (2022).**
- [WasEtAll22] Wasala J, Baratchi M, Marselis S. Arp L, Longepe N, Hoos H, AutoSR4EO: An AutoML Approach to Super-Resolution for Earth Observation. Submitted to WACF 2022.

3.2. Automated reinforcement learning

Reinforcement Learning (RL) has been trending towards the adoption of AutoAI approaches, which gave rise to the area of AutoRL. This extends the domain of AutoAI to the non-supervised learning setting but there is still much work to make this practical. There is a recent comprehensive survey on the topic as well as open challenges to guide future work in AutoRL:

- Parker-Holder, Jack; Rajan, Raghu; Song, Xingyou; Biedenkapp, André; Miao, Yingjie; Eimer, Theresa; Zhang, Baohe; Nguyen, Vu; Calandra, Roberto; Faust, Aleksandra; Hutter, Frank; Lindauer, Marius, "Automated Reinforcement Learning (AutoRL): A Survey and Open Problems", Journal of Artificial Intelligence Research (JAIR), vol. 74, pp. 517-568, 2022.
- Zhang, Baohe; Rajan, Raghu; Pineda, Luis; Lambert, Nathan; Biedenkapp, André; Chua, Kurtland; Hutter, Frank; Calandra, Roberto, "On the Importance of Hyperparameter Optimization for Model-based Reinforcement Learning", Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)'21, 2021.

There has been steady progress in this domain, namely to reduce reinforcement learning tremendous compute usage but also to facilitate useful, cross-domain reinforcement learning agents.

- Franke, Jörg K H; Köhler, Gregor; Biedenkapp, André; Hutter, Frank, "Sample-Efficient Automated Deep Reinforcement Learning", International Conference on Learning Representations (ICLR) 2021, 2021.
- Adriaensen, Steven; Biedenkapp, André; Shala, Gresa; Awad, Noor; Eimer, Theresa; Lindauer, Marius; Hutter, Frank, "Automated Dynamic Algorithm Configuration", arXiv:2205.13881, 2022.

3.3. AutoML for text-classification

An AutoML approach for text classification has been developed which jointly optimises sparse representations and document embeddings, to achieve high accuracy and low resource usage.

- **Blaž Škrlj, Matej Martinc, Nada Lavrač & Senja Pollak (2021) autoBOT: evolving neuro-symbolic representations for explainable low resource text classification. Machine Learning, 110: 989–1028. https://doi.org/10.1007/s10994-021-05968-x**

This AutoML approach considers the use of evolutionary algorithms to jointly optimise various sparse representations of a given text and two types of document embeddings.

3.4. AutoAI/ML in benchmarking optimization algorithms

In the context of developing AutoAI approaches for a broader range of AI tasks, an ontology has been developed for describing experiments in benchmarking optimization algorithms, and integrated with benchmarking tools widely used in the optimization community.

- **Kostovska, Ana, Vermetten, Diederick, Doerr, Carola, Džeroski, Sašo, Panov, Panče, Eftimov, Tome (2021). OPTION : optimization algorithm benchmarking ontology. GECCO '21, Proceedings of the Genetic and Evolutionary Computation Conference, pp. 239-240. ACM, https://doi.org/10.1145/3449726.3459579**

Furthermore, an ontology for describing experiments in benchmarking optimization algorithms is being developed, i.e., applying optimization approaches to different optimization problems, with the aim of facilitating the storing, sharing, and reuse of the results of such experiments. Significant progress has been made on the development of the ontology and integrating it with benchmarking tools, such as IOHprofiler, described in the following journal submission.

- Kostovska, Ana, Vermetten, Diederick, Doerr, Carola, Džeroski, Sašo, Panov, Panče, Eftimov, Tome (2021). OPTION : OPTImization algorithm benchmarking ONtology. IEEE Transactions on Evolutionary Computation. Under review.

In a broader context, the use of ontologies for developing automated approaches has been considered in different areas of AI and, more broadly, computer science: S. Džeroski gave an invited talk at the CEC'21 Workshop on Good Benchmarking Practices for Evolutionary Computation.

- Džeroski, Sašo. Ontologies for Open Computer Science. Invited talk at the CEC'21 Workshop on Good Benchmarking Practices for Evolutionary Computation. https://sites.google.com/view/benchmarking-network/home/activities/cec-2021-workshop

Finally, richer representations of optimization algorithms (e.g., modules in CMA-ES variants) and optimization problem instances (ELA features) have been developed to predict and explain the corresponding performance.

- Kostovska, Ana, Vermetten, Diederick, Džeroski, Sašo, Doerr, Carola, Korošec, Peter, Eftimov, Tome (2022). The importance of landscape features for performance prediction of modular CMA-ES variants. Proceedings of the Genetic and Evolutionary Computation Conference, pp. 648-656. ACM, https://doi.org/10.1145/3512290.3528832

3.5. Symbolic regression

In the task of symbolic regression (SR), the goal is to synthesise, given a training set of examples, a model represented as an evaluable expression (usually a mathematical expression comprising algebraic operators and transcendental functions) that maps the values of the input variables to the output variable(s). By constraining the syntax of SR models, one can make sure that they conform with the underlying domain knowledge (e.g., concerning the physical interpretation of variables, admissible dependencies between them, etc.). Crucially, SR models are transparent by being usually compact and expressed in terms of symbols with known semantics. Apart from solving typical regression tasks, SR can be used to learn models for AutoML and AutoAI (e.g., SR models that predict a system's performance based on its hyperparameters). The leading methodology for solving SR problems is genetic programming, a variant of evolutionary algorithm in which a working population of candidate SR models is gradually improved using search operators.

The methods commonly used to solve SR problems assume that the input and output variables are unconstrained. This is however rare in the real world; for instance the output variable may be required to be bounded to avoid damage to a controlled hardware component. Also, ignoring the existing constraints devoids the SR algorithm of potentially valuable domain knowledge. To address this issue, [BlaKra] introduced the class of constrained SR problems and proposed a genetic programming algorithm that solves them using a Satisfiability Modulo Theories (SMT) solver, which is used to formally verify the candidate solutions. The essence of the method consists in collecting the counterexamples resulting from model verification and using them to improve search guidance. The method is exact: upon successful termination, the produced model is guaranteed to meet the specified constraints. The effectiveness of the proposed method has been compared against standard constraint-agnostic machine learning regression algorithms on a range of benchmarks, and demonstrated that it outperforms them on several performance indicators.

In a related study on SR **[KraKos22]**, an approach to SR has been proposed that maintains, apart from the population of candidate solutions, a library of small partial models (fragments of expressions). This makes it possible to identify the promising components and guide search using two mechanisms in parallel: the conventional quality of fit to the training set, and matching contexts with subprograms using a gradient-based mechanism. In experimental assessment, the approach significantly outperformed the control setups. Maintaining partial models in efficient data structures prevents redundancy and lessens the demand for computational resources, in particular memory.

- [BlaKra] Błądek, I., Krawiec, K., Counterexample-Driven Genetic Programming for Symbolic Regression with Formal Constraints, (under review in IEEE Transactions on Evolutionary Computation).
- **[KraKos22] Krawiec, K., Kossiński, D., "Compositional Genetic Programming for Symbolic Regression". In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. GECCO '22. Boston, Massachusetts: Association for Computing Machinery, 2022, pp. 570–573. isbn: 9781450392686. doi: 10.1145/3520304.3529077 url: https://doi.org/10.1145/3520304.3529077**

3.6. Automated Data Science

This WP has also connected AutoML to the more ambitious and broader goal of working towards automated data science. There is joint work involving three partners that appeared as:

- **Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, and Christopher K.I. Williams Automating Data Science, CACM, March 2022, Vol 65 No 3**

Data science covers the full spectrum of deriving insight from data, from initial data gathering and interpretation, via processing and engineering of data, and exploration and modelling, to eventually producing novel insights and decision support systems. Given the complexity of data science projects and related demand for human expertise, automation has the potential to transform the data science process. This CACM article (which was prominently featured on the cover of the journal) provides a review of this newly emerging field, and characterises the challenges in data exploration, data engineering, model building and model exploitation.

On February 2 to 4, KU Leuven organised a research camp on automated data science as part of the ERC AdG Synth project of Luc De Raedt, and it also featured talks by Tijl De Bie, Holger Hoos and Sumit Gulwani. This event was also connecting to the TAILOR project, especially to WP7 on AutoAI. For more details, see: https://synth.cs.kuleuven.be/sites/synth.cs.kuleuven.be/files/

4. Self-monitoring AI systems [T7.3, ULEI]

Goal: Automatically detect when an AI system (such as a classifier, predictor or reasoning engine obtained from an AutoAI system) gets 'off-track' and can no longer be used safely and reliably.

4.1. Neural network verification

When included in an automated system, robustness verification can be utilised in a self-monitoring setting, but for this to work the verification process has to be sufficiently fast. In new work on robustness verification **[KonEtAl22]**, automated algorithm configuration is utilised to speed up the verification of neural networks. The authors found that due to the heterogeneity of the instances, there is no single configuration that convincingly outperforms other configurations. As a solution, a portfolio of complementary configurations is employed. The algorithm configurator Hydra [LinEtAl10] is extended to build a portfolio of complementary configurations that can be run in parallel. Once one of these configurations solves an instance, the whole portfolio can be terminated, which significantly speeds up the runtime of the packages MIPVerify and Venus (even when considering the additional cost for running multiple configurations).

- **[KonEtAl22] Matthias König, Holger Hoos and Jan N. van Rijn, Speeding Up Neural Network Robustness Verification via Algorithm Configuration and an Optimised Mixed Integer Linear Programming Solver Portfolio, Machine Learning Journal, 2022.**
- **[LinEtAl10]** Xu, Lin, Holger Hoos, and Kevin Leyton-Brown. "Hydra: Automatically configuring algorithms for portfolio-based selection." Twenty-Fourth AAAI Conference on Artificial Intelligence. 2010.

4.2. Safe learning and optimisation

Safe learning attempts to avoid the evaluation of strategies, solutions or policies that can induce an irreversible loss (e.g. machine malfunctioning, etc.). A systematic literature review of safe reinforcement learning was already published in 2015 [GarFer15]. Although, it did not consider new techniques that have been proposed in active learning/optimisation. To cope with this, a systematic review has been done of missing algorithms in several domains such as reinforcement learning, Gaussian process regression and classification, evolutionary computing and active learning, including the functioning mechanism of each class of techniques and how they are ultimately connected.

- **[GarFer15]** García, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. J. Mach. Learn. Res. 16(1), 1437–1480 (2015).
- **Kim, Y., Allmendinger, R., López-Ibáñez, M. (2021). Safe Learning and Optimization Techniques: Towards a Survey of the State of the Art. In: Heintz, F., Milano, M., O'Sullivan, B. (eds) Trustworthy AI - Integrating Learning, Optimization and Reasoning. TAILOR 2020. Lecture Notes in Computer Science, vol 12641. Springer, Cham. https://doi.org/10.1007/978-3-030-73959-1_12**

4.3. Dynamic Algorithm Configuration for Pseudo-Boolean Solving

Dynamic Algorithm Configuration (DAC) allows one to change the current configuration of a system at runtime, and to learn when to do so. In this work, we applied DAC to the specific case of Pseudo-Boolean solving with the solver Sat4j. Recent work shows that various strategies can be proposed to get the best of the particular case of Pseudo-Boolean constraints. Those strategies are complementary. This work is focussed on a particular strategy: the bumping strategy during conflict analysis. Initial results on benchmarks from the same family were promising: the proposed dynamic strategy performed better than the best static strategy on each benchmark. They were presented at the first TAILOR WP7 workshop. So far, results on more diverse benchmarks are not conclusive. The data exchanged between the solver and DAC needs to be further explored. All code related to this work is available in a dedicated branch on Sat4j git repository: https://gitlab.ow2.org/sat4j/sat4j/-/tree/DAC

5. Multi-objective AutoAI [T7.4, INRIA]

Goal: Develop multi-objective AutoAI methods to automatically determine the best tradeoffs between performance and other objectives, e.g., derived from the six dimensions of trustworthiness.

Most real-world problems are multi-objective (e.g., optimise quality and minimise cost). This goes for AutoML too, especially in the realm of TAILOR: beside pure performance (e.g., classification accuracy in supervised learning, optimal objective in optimization, …), objectives for Trustworthiness should also simultaneously be tackled: Computational cost (i.e., sustainability), robustness, explainability, privacy preservation, …). The goal of Multi-Objective AutoAI is to propose a set of Pareto-optimal settings (choice of algorithmic

building blocks and/or hyperparameters) that allow the user to decide at runtime (and at no cost) on which objective they will put emphasis.
Several works have been carried on by TAILOR partners, and several papers have also been published outside TAILOR, but worth considering for anyone working in the area, and that we will hence briefly describe in the last section of this report.

5.1. MO-CASH

Early time series classification (EarlyTSC) is a special multi-objective classification task involving the prediction of a class label based on partial observation of a given time series. Most EarlyTSC algorithms consider the trade-off between accuracy and earliness as two competing objectives, using a single dedicated hyperparameter. To deal with the challenging task of optimising two conflicting objectives in early time series classification, we propose MultiETSC, a system for multi-objective algorithm selection and hyperparameter optimisation (MO-CASH) for EarlyTSC. MultiETSC can potentially leverage any existing or future EarlyTSC algorithm and produces a set of Pareto optimal algorithm configurations from which a user can choose a posteriori.

- **[OttEtAl21] Ottervanger G, Baratcni M., Hoos H., MultiETSC: automated machine learning for early time series classification, Data Mining and Knowledge Discovery, 2021.**

5.2. Federated learning

Federated learning cooperatively trains a server-based model using local models running on edge devices, while preserving data privacy. Such a training paradigm implies substantial communication between the server and workers, which can hinder the efficiency. Reducing the communication without affecting the training accuracy is conflictual. Moreover, the literature investigates the model's compression and communication rounds separately, although they jointly contribute to the communication overhead. Therefore, an approach has been proposed to solve the communication overload in federated learning as a multiobjective problem where the solution consists of the hyperparameter optimisation of both the model's and paradigm's training settings.

- **Morell, J.Á., Dahi, Z.A., Chicano, F., Luque, G., Alba, E. (2022). Optimising Communication Overhead in Federated Learning Using NSGA-II. In: Jiménez Laredo, J.L., Hidalgo, J.I., Babaagba, K.O. (eds) Applications of Evolutionary Computation. EvoApplications 2022. Lecture Notes in Computer Science, vol 13224. Springer, Cham. https://doi.org/10.1007/978-3-031-02462-7_21**

5.3. Meta-learning for multi-label classification

In multi-target prediction (MTP), which includes tasks such as multi-label classification (MLC), hierarchical multi-label classification (HMLC) and multi-target regression. Many measures are used to assess the performance of predictive models: A case in point is MLC, where at least 20 performance measures are in use. This makes MTP tasks excellent use cases for multi-objective AI. In this context, we are conducting empirical comparative studies of MTP methods on MTP tasks to generate performance data, for which we then perform meta-learning. The meta-models predict multiple aspects of performance and could be used, for example, for multi-objective algorithm selection (according to multiple performance measures).

A comprehensive empirical investigation of a wide range of MLC methods has been carried out on a wealth of datasets from different domains: 26 methods on 42 benchmark datasets using 20 evaluation measures. This study is the most comprehensive experimental work for the task of MLC performed thus far. In a nutshell, it identifies a subset of 5 methods that should be used in baseline comparisons: RFPCT (Random Forest of Predictive Clustering Trees), RFDTBR (Binary Relevance with Random Forest of Decision Trees), ECCJ48 (Ensemble of Classifier Chains built with J48), EBRJ48 (Ensemble of Binary Relevance built with J48) and AdaBoost.

- **Jasmin Bogatinovski, Ljupco Todorovski, Saso Dzeroski and Dragi Kocev. (2022) Comprehensive Comparative Study of Multi-Label Classification Methods. Expert Systems with Applications, 203: 117215. https://doi.org/10.1016/j.eswa.2022.117215**

Furthermore, a comprehensive meta-learning study of data sets and methods for multi-label classification (MLC) has been recently conducted, analysing 40 MLC data sets by using 50 meta-features describing different properties of the data. The main findings of this study are: (1) The meta-models show that the most important meta-features describe the label space. (2) The meta-features describing the relationships among the labels tend to occur a bit more often than the meta-features describing the distributions between and within the individual labels. (3) The optimization of the hyper-parameters can improve the predictive performance, but the extent of the improvements does not always justify the resources used in this context.

- **Jasmin Bogatinovski, Ljupco Todorovski, Saso Dzeroski and Dragi Kocev. (2022) Explaining the Performance of Multi-label Classification Methods with Data Set Properties. International Journal of Intelligent Systems. https://doi.org/10.1002/int.22835**

Finally an ontology-based online catalogue of MLC datasets originating from various application domains following the FAIR principles has been compiled. Based on the extensive experimental study mentioned above, the catalogue extensively describes many MLC datasets with comprehensible meta-features, MLC-specific semantic descriptions, and different data provenance information. The MLC data catalogue is available at: http://semantichub.ijs.si/MLCdatasets

- **Ana Kostovska, Jasmin Bogatinovski, Saso Dzeroski, Dragi Kocev, and Pance Panov. (2022) A catalogue with semantic annotations makes multilabel datasets FAIR. Scientific Reports 12: 7267. https://doi.org/10.1038/s41598-022-11316-3**

5.4. MO hyperparameter optimization

The following recently published papers have contributed to the advancement of the field, and it seems important to take them into account in any further work on Multi-Objective

AutoAI. However, as they have no connection with TAILOR partners, we will only briefly cite them here, for the sake of completeness.

The most important paper in the area certainly is the recent survey of techniques for Multi-Objective Hyper-Parameter Optimization (MOHPO):

- Multi-Objective Hyperparameter Optimization -- An Overview. Karl, Florian ; Pielok, Tobias ; Moosbauer, Julia ; Pfisterer, Florian ; Coors, Stefan ; Binder, Martin ; Schneider, Lennart ; Thomas, Janek ; Richter, Jakob ; Lang, Michel ; Garrido-Merchán, Eduardo C. ; Branke, Juergen ; Bischl, Bernd. arXiv:2206.07438

This paper focuses on MOHPO (i.e., not dealing with Multi-Objective pipeline design), and surveys both Evolutionary and Bayesian Optimization methods. The primary objective is always the performance, but several secondary objectives are proposed, like operating conditions, prediction time, sparseness, fairness, interpretability and robustness. This paper has a comprehensive list of citations (with 272 distinct references!) and we will not discuss here papers that are cited in this survey. However, several papers that are of interest to the Multi-Objective AutoAI community are not cited there, either because they were published later, or because they do not fall into the focus of this survey.
The three following papers deal with Neural Architecture Search using evolutionary approaches, and indeed do not pertain to HPO. They all use variation operators (crossover and mutations) to directly modify the architecture of the Deep Network of interest, and this allows the exploration of a larger search space than parameterized search.

- Tackling Neural Architecture Search With Quality Diversity Optimization. Lennart Schneider, Florian Pfisterer, Paul Kent, Juergen Branke, Bernd Bischl, Janek Thomas OpenReview AutoML 2022
- GenExp: Multi-Objective Pruning for Deep Neural Network Based on Genetic Algorithm Xu, K., Zhang, D., An, J., Liu, L., Liu, L., & Wang, D. Neurocomputing 451 (2021): 81-94.
- Evolutionary neural cascade search across supernetworks Alexander Chebykin, Tanja Alderliesten, Peter A. N. Bosman. GECCO 2022

The following paper does not perform AutoAI per se, but very originally performs landscape analysis from the point of view of multi-objective exploration, looking for trade-offs between cost and accuracy.

- Cost-vs-accuracy of sampling in multi-objective combinatorial exploratory landscape analysis Raphaël Cosson, Bilel Derbel, Arnaud Liefooghe, Sébastien Verel, Hernan Aguirre, Qingfu Zhang, Kiyoshi Tanaka. GECCO 2022

6. Ever-learning AutoAI [T7.5, TU/e]

Goal: Ensure that AutoAI gets better over time, producing better models with less data, and avoids the computational overhead of starting from scratch for any new use case, or change in scenario.

6.1. Meta-learning

Meta-learning is the field of research that deals with learning across datasets. Among several approaches, two very popular methods are MAML (Finn et al., 2017), which aims to learn an initialization for Neural Networks that works well across datasets and can be easily and rapidly fine-tuned on new datasets, and Prototypical Networks (Snell et al., 2017), which builds both a metric space in which prototypical examples of new classes can be built and classification done by computing distances to these prototypes. A more expressive approach than MAML is the LSTM-metalearner (Ravi & Larochelle 2017), which does not only learn the initialization, but also the optimization procedure (learning an optimizer is clearly an AutoAI task). Finn et al. have shown that MAML performs better than the LSTM-metalearner. This is interesting, as the LSTM meta-learner is more expressive and should therefore intuitively perform better. Huisman et al. (2022) proposed various hypotheses why this could be the case, and developed TURTLE, a novel meta-learning approach that outperforms both methods.

- Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International conference on machine learning, ICML'17, pp. 1126-1135. PMLR.
- Jake Snell, Kevin Swersky, Richard Zemel. Prototypical Networks for Few-shot Learning. In NIPS 2017.
- Ravi, S., & Larochelle, H. (2017). Optimization as a model for few-shot learning. In Proc. International Conference on Learning Representations, ICLR'17.
- Huisman, M., Plaat, A. & van Rijn, J.N. Stateless neural meta-learning using second-order gradients. Mach Learn (2022). https://doi.org/10.1007/s10994-022-06210-y

6.2. Meta-Deep Learning Challenges

The successful application of deep neural networks often requires large amounts of data and computing resources, restricting its success to domains where this is available. Meta-learning methods can help tackle this problem by transferring knowledge from related tasks, thus reducing the amount of data and computing resources needed to learn new tasks. The first MetaDL challenge, a NeurIPS 2021 challenge, was a competition on few-shot learning, which attracted over 15 teams that made over 100 code submissions. The lessons learned include that learning good representations is essential for effective transfer learning, and are described in length in the following paper.

- **El Baz et al, Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification, Proceedings of the NeurIPS 2021 Competition and Demonstration Track, PMLR 176, 2022. (Leiden University, INRIA and TUE). https://proceedings.mlr.press/v176/el-baz22a/el-baz22a.pdf**

These results were the basis for a new challenge for meta-learning, called Cross-Domain Meta-DL, that is co-organized by the same TAILOR partners, and run as a TAILOR challenge, within TAILOR WP2 (and the following text is redundant with the complete description of this challenge in Deliverable 2.3, summarised here for completeness). The

Cross-Domain Meta-DL has been accepted as a NeurIPS 2022 challenge, and is described in detail in a comprehensive white paper (reference below), and detailed instructions to participate are available as a tutorial on the challenge web site. While the previous challenge focused on within-domain few-shot learning problems, with the aim of learning efficiently N-way k-shot tasks (i.e., N class classification problems with k training examples each), this competition challenges the participants to solve “any-way” and “any-shot” problems drawn from various domains chosen for their humanitarian and societal impact (healthcare, ecology, biology, manufacturing, …). The goal is to meta-learn a good model that can quickly learn tasks from a variety of domains, with any number of ways (within the range 2-20) and any number of shots (within the range 1-20). After a public phase using 10 public datasets, the feedback phase started on 1 July, on 10 hidden other datasets. The final phase will take place from 1 Sep. on, during which the last submission of each participant from the feedback phase will be blind-tested on 10 new hidden datasets to rank the participants. The winners will be announced on 1 Oct.

- **NeurIPS’22 Cross-Domain MetaDL competition: Design and baseline results. Dustin Carrión-Ojeda, Hong Chen, Adrian El Baz, Sergio Escalera, Chaoyu Guan, Isabelle Guyon, Ihsan Ullah, Xin Wang, and Wenwu Zhu. https://drive.google.com/file/d/145t-KVmHNIFCweiljbPwimmAXMvHHf7e/view**

6.3. Meta-learning datasets

Despite the popularity of the meta-learning field, progress is held back by a lack of good, challenging, and computationally feasible meta-datasets that enable us to accurately assess the generalisation abilities of meta-learning algorithms. To remedy this, Meta-Album has been introduced, an extensible multidomain meta-dataset, including (so far) 40 image classification datasets from 10 different domains. This is part of a long-term effort to create a publicly available and growing meta-dataset. Meta-Album was specifically designed to facilitate meta-learning research in the cross-domain few-shot setting, which is more realistic than commonly used evaluation protocols. In practical real-world applications, tasks may come from various domains, include classes not drawn at random, but stemming from a class hierarchy, and include any number of classes and/or examples per class. We hope that this dataset will allow us to create better and more practically useful meta-learning techniques. Moreover, we are extending it to also allow continual learning settings, especially class-incremental and domain-incremental learning tasks.

All datasets and Open Source code is available at https://meta-album.github.io/. A paper giving all details about how this dataset of datasets has been built is currently submitted to NeurIPS and as such available on OpenReview. Further details about Meta-Album are presented in Deliverable 2.3 (foundational benchmarks and challenges).

- **Meta-Album: Multi-domain Meta-Dataset for Few-Shot Image Classification. Ihsan Ullah, Dustin Carrion, Sergio Escalera, Isabelle M Guyon, Mike Huisman, Felix Mohr, Jan N. van Rijn, Haozhe Sun, Joaquin Vanschoren, Phan Anh Vu. *Under review at the NeurIPS 2022 Datasets and Benchmarks Track.***

6.4. Speeding up AutoML

When hyperparameter optimization of a machine learning algorithm is repeated for multiple datasets it is possible to transfer knowledge to an optimization run on a new dataset. We developed a new hyperparameter-free ensemble model for Bayesian optimization that is a generalisation of two existing transfer learning extensions to Bayesian optimization and establish a worst-case bound compared to vanilla Bayesian optimization.We demonstrate that our contributions substantially reduce optimization time compared to standard Gaussian process-based Bayesian optimization and improve over the current state-of-the-art for transfer hyperparameter optimization.

- Practical Transfer Learning for Bayesian Optimization. Feurer, Matthias; Letham, Benjamin; Hutter, Frank; Bakshy, Eytan. arXiv:1802:02219v3, 2022.

Neural architecture search (NAS) methods have successfully enabled the automated search of neural architectures in various domains. However, most techniques start from scratch with every new task. Recently, techniques have been proposed that generalise across tasks. For instance, MetaNAS warm-starts the well-known DARTS technique by learning over a group of tasks what would be a good initialization of the NAS search space (leveraging the MAML method), hence significantly speeding up the search for good neural architectures on similar tasks. However, this method tends to be unstable, leading to very different architectures even with very similar starting conditions. To resolve this, we introduced a combination of regularisation techniques (dropout on skip-connections and max-W regularisation with power cosine annealing), which yields better-performing architectures while also making them smaller (with fewer parameters).

- Meta-learning of neural architectures for few-shot learning. Elsken, T., Staffler, B., Metzen, J. H., and Hutter, F. CVPR 2020
- **Regularized Meta-Learning for Neural Architecture Search. Rob van Gastel, Joaquin Vanschoren. AutoML Conference 2022 Late-breaking papers track**

One potentially disruptive area of research we have looked into is using the popular Transformer based architectures for generating predictions on tabular data, an area where classical tree-based machine learning models have been the default choice. By leveraging nothing but simulated data for meta-learning, these new methods allow all training time to be offloaded as pre-training, giving us near instant predictions for new, unseen data that are competitive with the current state of the art from AutoML tools. We believe these pretrained transformers are very promising in allowing users to have strong predictions with no training compute cost, greatly democratising machine learning.

- **[HolEtAl22] Meta-Learning a Real-Time Tabular AutoML Method For Small Data. Noah Hollmann, Samuel Müller, Katharina Eggensperger, Frank Hutter. NeurIPS 2022.**
  - https://arxiv.org/abs/2207.01848

6.5. Continuous AutoML

AutoML has been used successfully in settings where the learning task is assumed to be static. In many real-world scenarios, however, the data distribution will evolve over time, also known as concept drift. Most AutoML methods will not be able to adapt to such changes. We designed an adaptive Online Automated Machine Learning (OAML) system that can effectively design online pipelines in dynamic environments depending on exactly what kind of drift occurs, thus automating pipeline design for online learning while continuously adapting to data drift. It searches the complete pipeline configuration space of online learners, including preprocessing algorithms and ensembling techniques. It outperforms existing AutoML techniques as well as existing online learning techniques on most of the available real-world and synthetic tasks. It is integrated in GAMA, a very extensible AutoML tool developed by TUE. This work has been released on arXiv and is currently under review in the machine learning journal (MLJ).

- **Online AutoML: An adaptive AutoML framework for online learning. Bilge Celik, Prabhant Singh, Joaquin Vanschoren. https://arxiv.org/abs/2201.09750**

7. Other work

Beyond work relating strongly to the tasks defined in TAILOR work package 7, several other interesting developments are discussed in this section.

7.1. Hardware Dimensioning of AI algorithms

In recent years, Artificial Intelligence (AI) has become widespread in many domains. While the adoption of AI techniques is on the rise, a big challenge still has no simple answer: determining the right hardware (HW) architecture and configuration (e.g., HW on premises or cloud resources) – also referred to as hardware dimensioning. This issue is significantly exacerbated by the difficulty of anticipating the behaviour of an AI algorithm on different HW architectures, and by potential constraints both on the available budget and on the quality of the solution. An automated way to match algorithms, user constraints and HW resources would be welcomed for AI practitioners. To address this issue HADA **[*DeFEtAl22*]** was proposed, an automated approach for HW dimensioning based on the Empirical Model Learning (EML) framework [*LomEtAl17*], where ML models are embedded within an optimization problem to enable decision making over complex real-world systems. The idea is to learn the relationships between the AI algorithm performances and HW resources via ML models. In this way, complex aspects of the problem can be approximated with surrogates (data-driven models) rather than being explicitly or analytically expressed.

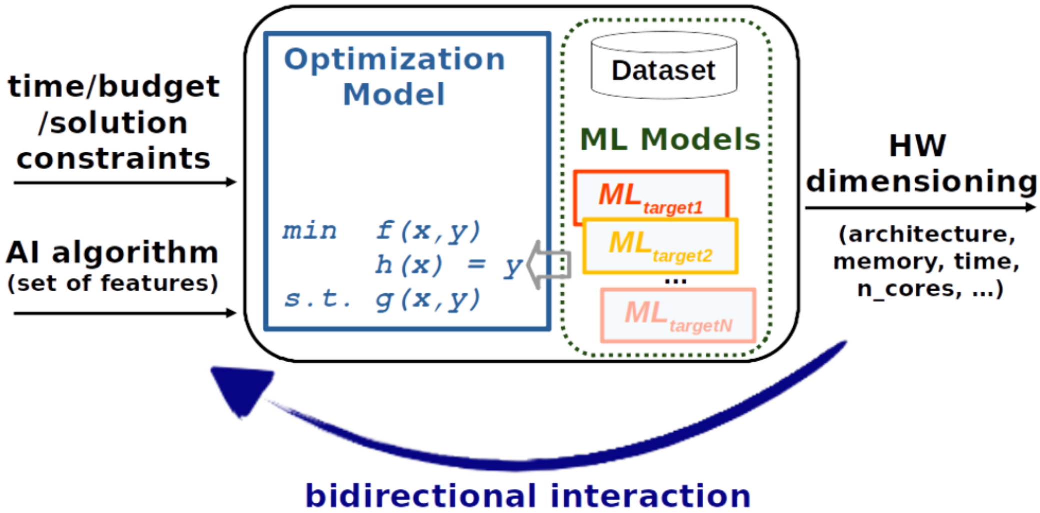

Fig. 7.1: Scheme of HADA. x are the decision variables; y are the observed variables; f(x,y) is the objective function; g(x,y) is the set of user-defined and domain knowledge constraints; h(x) is the approximation of the complex behaviour (the encoding of ML models)

As shown in Fig. 1, HADA can be represented as a black-box that receives as input a set of features describing an AI algorithm and some user-defined constraints (e.g. budget limits, time constraints and required solution quality), and it produces as output the optimal HW resource dimensioning needed to run the algorithm. HADA surpasses standard ML as it enables bidirectional interaction between an AI algorithm, its performance, and HW dimensioning: given a specific HW architecture, it can be used to automatically obtain the most suitable algorithm for a target task and its optimal parameters. Another key advantage of HADA is that, while an initial benchmarking phase is needed to build the data set, once the ML models are trained, they can be reused in the optimization phase on different data instances and different user-defined constraints. This represents a great advantage w.r.t. traditional black-box optimization methods such as surrogate-based methods. A byproduct of this approach is that, given a fixed HW architecture and configuration as input constraints, it is possible to obtain the most suitable algorithm together with its parameters.

The proposed approach was evaluated on two different online algorithms grounded on an energy management system, and a single HW resource (RAM memory). The experiments showed that the approach is flexible and robust, as demonstrated by the validation on unseen instances. In future works the authors also plan to extend the approach to deal with different target algorithms and domains, and to consider heterogeneous HW resources. This will require benchmarking multiple algorithms on different architectures in order to obtain accurate ML models, which could then be embedded in the optimization problem for HW dimensioning and algorithm fine-tuning. In this direction, active and transfer learning strategies will be explored to reduce the size of the required training set (e.g., [*BorEtAl20*]), or domain knowledge could be injected in the ML models, for instance to improve their explainability. In another direction there are plans to tackle more directly the unavoidable uncertainty implicit in the ML models, which are currently addressed through chance constraints.

7.2. Machine Learning and Language Processing

A range of interesting work has been conducted in this area and is briefly outlined in the following.

Streaming speech translation

The cascade approach to Speech Translation (ST) is based on a pipeline that concatenates an Automatic Speech Recognition (ASR) system followed by a Machine Translation (MT) system. Nowadays, state-of-the-art ST systems are populated with deep neural networks that are conceived to work in an offline setup in which the audio input to be translated is fully available in advance. However, a streaming setup defines a completely different picture, in which an unbounded audio input gradually becomes available and at the same time the translation needs to be generated under real-time constraints. In this work, we present a state-of-the-art streaming ST system in which neural-based models integrated in the ASR and MT components are carefully adapted in terms of their training and decoding procedures in order to run under a streaming setup. In addition, a direct segmentation model that adapts the continuous ASR output to the capacity of simultaneous MT systems trained at the sentence level is introduced to guarantee low latency while preserving the translation quality of the complete ST system. The resulting ST system is thoroughly evaluated on the real-life streaming Europarl-ST benchmark to gauge the trade-off between quality and latency for each component individually as well as for the complete ST system.

- **J. Iranzo, J. Jorge, P. Baquero, J. A. Silvestre, A. Giménez, J. Civera, A. Sanchis, A. Juan. Streaming cascade-based speech translation leveraged by a direct segmentation model Neural Networks, 142, pp. 303–315, 2021.** doi:10.1016/j.neunet.2021.05.013

Simultaneous translation systems

Simultaneous machine translation has recently gained traction thanks to significant quality improvements and the advent of streaming applications. Simultaneous translation systems need to find a trade-off between translation quality and response time, and with this purpose multiple latency measures have been proposed. However, latency evaluations for simultaneous translation are estimated at the sentence level, not taking into account the sequential nature of a streaming scenario. Indeed, these sentence-level latency measures are not well suited for continuous stream translation, resulting in figures that are not coherent with the simultaneous translation policy of the system being assessed. This work proposes a stream level adaptation of the current latency measures based on a re-segmentation approach applied to the output translation, that is successfully evaluated on streaming conditions for a reference IWSLT task.

- **J. Iranzo, J. Civera, A. Juan. Stream-level Latency Evaluation for Simultaneous Machine Translation. In Findings of the ACL: EMNLP 2021, pp. 664–670, Punta Cana (Dominican Republic), 2021.** doi:10.18653/v1/2021.findings-emnlp.58
- https://github.com/jairsan/Stream-level_Latency_Evaluation_for_Simultaneous_Mach [...]

Live streaming speech recognition

Although Long-Short Term Memory (LSTM) networks and deep Transformers are now extensively used in offline ASR, it is unclear how best offline systems can be adapted to work with them under the streaming setup. After gaining considerable experience in this regard in recent years, a recent study (see below) shows how an optimised, low-latency streaming decoder can be built in which bidirectional LSTM acoustic models, together with general interpolated language models, can be nicely integrated with minimal performance degradation. In brief, our streaming decoder consists of a one-pass, real-time search engine relying on a limited-duration window sliding over time and a number of ad hoc acoustic and language model pruning techniques. Extensive empirical assessment is provided on truly streaming tasks derived from the well-known LibriSpeech and TED talks datasets, as well as from TV shows on a main Spanish broadcasting station.

- **J. Jorge, A. Giménez, J. A. Silvestre, J. Civera, A. Sanchis, A. Juan. Live Streaming Speech Recognition Using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30 , pp. 148–161, 2021. doi:10.1109/TASLP.2021.3133216**

2020 Speech-to-text challenge
A recently published paper describes the automatic speech recognition (ASR) systems built by the MLLP-VRAIN research group of Universitat Politècnica de València for the Albayzín-RTVE 2020 Speech-to-Text Challenge, and includes an extension of the work consisting of building and evaluating equivalent systems under the closed data conditions from the 2018 challenge. The primary system (p-streaming_1500ms_nlt) was a hybrid ASR system using streaming one-pass decoding with a context window of 1.5 seconds. This system achieved 16.0% WER on the test-2020 set. We also submitted three contrastive systems. From these, the system c2-streaming_600ms_t is highlighted, which, following a similar configuration as the primary system with a smaller context window of 0.6 s, scored 16.9% WER points on the same test set, with a measured empirical latency of 0.81 ± 0.09 s (mean ± stdev). That is, state-of-the-art latencies were obtained for high-quality automatic live captioning with a small WER degradation of 6% relative. As an extension, the equivalent closed-condition systems obtained 23.3% WER and 23.5% WER, respectively. When evaluated with an unconstrained language model, 19.9% WER and 20.4% WER were obtained; i.e., not far behind the top-performing systems with only 5% of the full acoustic data and with the extra ability of being streaming-capable. Indeed, all of these streaming systems could be put into production environments for automatic captioning of live media streams.

- **P. Baquero, J. Jorge, A. Giménez, J. Iranzo, A. Pérez, G. Garcés, J. A. Silvestre, J. Civera, A. Sanchis, A. Juan. MLLP-VRAIN Spanish ASR Systems for the Albayzin-RTVE 2020 Speech-To-Text Challenge: Extension. Applied Sciences, 12 (2), pp. 804, 2022. doi:10.3390/app12020804**

Streaming machine translation
Simultaneous machine translation has recently gained traction thanks to significant quality improvements and the advent of streaming applications. Simultaneous translation systems need to find a trade-off between translation quality and response time, and with this purpose multiple latency measures have been proposed. However, latency evaluations for

simultaneous translation are estimated at the sentence level, not taking into account the sequential nature of a streaming scenario. Indeed, these sentence-level latency measures are not well suited for continuous stream translation, resulting in figures that are not coherent with the simultaneous translation policy of the system being assessed. This work proposes a stream-level adaptation of the current latency measures based on a re-segmentation approach applied to the output translation, that is successfully evaluated on streaming conditions for a reference IWSLT task.

- **J. Iranzo, J. Civera, A. Juan. From Simultaneous to Streaming Machine Translation by Leveraging Streaming History. In Proc. 60th Annual Meeting of the Association for Computational Linguistics Vol. 1: Long Papers (ACL 2022), pp. 6972–6985, Dublin (Ireland), 2022. doi:10.18653/v1/2022.acl-long.480**
- https://github.com/jairsan/Speech_Translation_Segmenter

8. Benchmarks

The AutoML benchmark
A benchmarking tool has been developed by TAILOR partner TUE to evaluate AutoML frameworks for tabular data and carried out large scale experiments. See Section 6 of the AutoAI benchmark report v2 (D7.6) for more details, including an outline for future work.

Automatic Speech Recognition benchmark
Europarl-ASR is a new large speech and text corpus of parliamentary debates including 1300 hours of transcribed speeches and 70 million tokens of text in English extracted from European Parliament sessions. Section 8 of the benchmark report v2 (D7.6) provides more details.

9. Conclusion

Based on the status of AutoAI discussed in this report there are a number of clear and specific steps that can be taken in the near future to extend this work, and the TAILOR consortium is well positioned to make these a reality. We briefly highlight a few examples. Related to AutoML in the wild, ongoing work on HPO, automated configuration and NAS tools will help to improve how efficiently and widely they can be used. In the area of self-monitoring AI systems the development of a neural network verification benchmark would benefit the rigour in the field and help to advance it. For every-learning AutoAI, further extending the Meta-Album benchmark library to also allow continual learning settings, especially class-incremental and domain-incremental learning tasks, will help to cover a larger spectrum of problems in meta-learning.

References

[ArpEtAl21] Arp L, Baratchi M, Hoos H, VPint: value propagation-based spatial interpolation, Data Mining and Knowledge Discovery (2022).

[BorEtAl20] A. Borghesi, G. Tagliavini, M. Lombardi, L. Benini, M. Milano, Combining learning and optimization for transprecision computing, in: Proceedings of the 17th ACM International Conference on Computing Frontiers, 2020, pp. 10–18.

[DeFEtAl22] De Filippo A, Borghesi A, Boscarino A, Milano M. HADA: An automated tool for hardware dimensioning of AI applications. Knowledge-Based Systems. 2022 Sep 5;251:109199.

[FeuEtAl22] Feurer, Matthias; Eggensperger, Katharina; Falkner, Stefan; Lindauer, Marius; Hutter, Frank, Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning, accepted at JMLR 2022.

[GiaKal22] Thanos Giannakopoulos and Vana Kalogeraki, An Elastic and Scalable Topic-based Pub/Sub System using Deep Reinforcement Learning. DAIS 2022, June 2022.

[HolEtAl22] Hollmann Noah, Muller Samual, Eggensperger Katharina, Hutter Frank, “Meta-Learning a Real-Time Tabular AutoML Method For Small Data”, arXiv:2207.01848

[LinEtAl22] Lindauer, Marius; Eggensperger, Katharina; Feurer, Matthias; Biedenkapp, André; Deng, Difan; Benjamins, Carolin; Ruhkopf, Tim; Sass, René; Hutter, Frank, “SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization”, Journal of Machine Learning Research (JMLR) -- MLOSS, vol. 23, no. 54, pp. 1-9, 2022.

[LomEtAl17] Lombardi M, Milano M, Bartolini A. Empirical decision model learning. Artificial Intelligence. 2017 Mar 1;244:343-67.

[LopEtAl16] López-Ibáñez, Manuel, et al. "The irace package: Iterated racing for automatic algorithm configuration." Operations Research Perspectives 3 (2016): 43-58.

[OttEtAl21] Ottervanger G, Baratcni M., Hoos H., MultiETSC: automated machine learning for early time series classification, Data Mining and Knowledge Discovery, 2021.

[RajEtAl21] Rajan, Raghu; Diaz, Jessica Lizeth Borja; Guttikonda, Suresh; Ferreira, Fabio; Biedenkapp, André; von Hartz, Jan Ole; Hutter, Frank, “MDP Playground: A Design and Debug Testbed for Reinforcement Learning” In: arXiv:1909.07750, 2021.

[RenEtAl22] Sass, René; Bergman, Eddie; Biedenkapp, André; Hutter, Frank; Lindauer, Marius, DeepCAVE: An Interactive Analysis Tool for Automated Machine Learning, Workshop on Adaptive Experimental Design and Active Learning in the Real World ICML 2022.

[SouEtAl21] de Souza, Marcelo; Ritt, Marcus; López-Ibáñez, Manuel; Pérez Cáceres, Leslie; “ACVIZ: A tool for the visual analysis of the configuration of algorithms with irace”, Operations Research Perspectives, Volume 8, 2021.

[TetEtAl22] Tetteroo J, Baratchi M, Hoos H, Automated Machine Learning for COVID-19 Forecasting. Accepted to IEEE Access.

[TomEtAl22a] Dimitrios Tomaras, Michail Tsenos, Vana Kalogeraki, Practical Privacy Preservation in a Mobile Cloud Environment. MDM 2022, June 2022

[TomEtAl22b] A framework for supporting privacy preservation functions in a Mobile Cloud Environment. Dimitrios Tomaras, Michail Tsenos, Vana Kalogeraki, MDM 2022, June 2022.

[TseEtAl22] Michail Tsenos, Aristotelis Peri, Vana Kalogeraki,AMESoS: A Scalable and Elastic Framework for Latency Sensitive Streaming Pipelines, DEBS 2022, June 2022.

[WasEtAll22] Wasala J, Baratchi M, Marselis S. Arp L, Longepe N, Hoos H, AutoSR4EO: An AutoML Approach to Super-Resolution for Earth Observation. Submitted to WACF 2022.

[ZimEtAl21] Zimmer, Lucas; Lindauer, Marius; Hutter, Frank, “Auto-Pytorch: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL”, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-1, 2021.