DAIS ITA – BPP20

Distributed Analytics and Information Sciences

International Technology Alliance

Biennial Program Plan 2020

15th Jan 2021 – 20th Sept 2022 (Year 2)

Summary of Accomplishments

Written on 15st September 2021
Agreement W911NF-16-3-0001

Table of Contents

Summary of Accomplishments

Executive Summary

In this report covering the BPP20 year two period we have chosen to highlight certain key publications for each of the four projects. Please refer to the Science Library https://dais-legacy.org/science-library/) for all publications during this period and the whole of the DAIS ITA program so far. At the time of writing there are a number of "in flight" publications that have not yet been published to science library since they are under review, or as yet unpublished in their external conference or journal format, however for the BPP20 year two period we have confirmed 11 journal papers and 28 external conference papers, all of which have been published to the science library. Our estimate is that up to an additional 44 external conference papers will eventually be published for BPP20 that were created and submitted in the year two period, but will be published following the conclusion of the program. In addition, open source components continue to be maintained and will be added to, on github at https://github.com/dais-ita/. Additionally, the DAIS ITA Legacy Site (https://dais-legacy.org/) contains a selection of key technical achievements as well as historical program documents including the Biennial Program Plan (BPP) documents, and annual reports.

The contributions and highlighted results of each project are summarised below.

Project 7: Policy-enabled Dynamic Infrastructure

In year two researchers from Imperial College, Purdue University, Yale University, Southampton University and University College London have made significant contributions on different aspects of policy-enabled dynamic infrastructure for efficient management of the networked computing system across coalition members, (federated) policy learning methods that support transfer and domain adaptation between source and target domains of different coalition parties with limited access to data, and software architectures and solutions for management of coalition network infrastructures. Details are given below per task.

Task 7.1: Infrastructure Design & Distributed Control for Dynamic SDC

A team of researchers from Imperial College, Purdue University, and Yale University have developed a novel technique to decompose the state space of reinforcement learning (RL) as a way to overcome the state-explosion problem and improve the learning speed for certain problem settings. Specifically, the new technique decomposes the state space of such systems into subspaces, and a neural network can be implemented for each of the decomposed subspaces, before a combining neural network captures the interactions between state subspaces for the RL problems. The new technique is shown to learn faster than traditional RL by using an Alibaba dataset. The work will be presented at the ITA Closeout Event and submitted to the IEEE ICC 2022 (https://dais-ita.org/node/6220). Another team of researchers from Imperial College and IBM U.S. has investigated theoretical modelling and quantification of performance enhancement of distributed SDN architectures, exploring in particular how this is influenced by inter-domain synchronization levels and network structural properties (https://dais-ita.org/node/4663).

Enhanced results have recently been published in ACM/IEEE Trans. on Networking in 2021. In another effort, researchers from Imperial College have also developed, in collaboration with IBM U.S., a novel embedding technique for efficient reinforcement learning, which jointly embeds states and actions combining aspects of model-free and model-based reinforcement learning. The work has been accepted for publication at the CIKM 2021 in November 2021 (https://dais-ita.org/node/5815).

Efficient management is also enabled through robust learning of SDC control policies. The robustness of the SDC architecture is challenged by network fragmentation events where mobile nodes lose connectivity from the controllers and therefore they cannot receive new flow rules and update their policies. The situation can be improved if fragmentation events are predicted in time and mobile nodes self-organize and run a distributed mobile ad hoc network (MANET) protocol in order to preserve traffic routing even after fragmentation occurs. Researchers from Yale, in collaboration with IBM U.K. and ARL, have proposed a deep neural network approach for predicting network fragmentations. The network fragmentation prediction is based on prediction of path loss value at a next time instance from a given (time window) past sequence of path loss values. Results have been published in Q. Qin et al. "Learning-aided SDC Control in Mobile Ad Hoc Networks", SPIE 2021 (https://dais-ita.org/node/6230). To improve even further the use of Deep Neural Networks in tactical ad hoc networks where dynamic changes in topologies are the norm rather than the exceptions, researchers from Yale University, in collaboration with IBM U.S., have explored the use of graph attention networks (GATs) to provide generalisable mechanisms for performing inference over different network topologies, including topologies not seen during the training of these networks. This has been applied in particular on the task of network congestion prediction and results are given in K. Poularakis et al. "Generalizable and Interpretable Deep Learning in Networking for Network Congestion Prediction", to appear at ICNP 2021 (https://dais-ita.org/node/5890). The results of the above two publications were also used to prepare a demonstration for the forthcoming final Showcase event. At the core of the use Machine Learning for efficient resource utilisation is data collection. Coalition operations often involve systems with multiple self-interested stakeholders (e.g. services or users belonging to different domains). Resource allocation can be optimised also with respect to these self-interests. Researchers from Southampton have developed novel auction-based algorithms that incentivise truthful reporting in edge cloud resource allocation settings with multiple self-interested service consumers or providers. This novel approach combines reinforcement learning with incentive compatible mechanism design, ensuring that an allocation mechanism remains truthful despite adapting its behaviour based on historical data. Results have been published in S. Stein et al. "Strategy proof Reinforcement Learning for Online Resource Allocation", AAMAS 2020 (https://dais-ita.org/node/4969). Finally, RL has been proven to be versatile for system control, including resource and network management functions of SDC, but SDC may become fragmented due to natural component failures and man-made tactical reasons. In a separate effort, researchers from Imperial College, Purdue University and IBM UK have continued their investigation to enhance RL performance with help from transfer learning (TL) in the event of domain fragmentation and reconnection in SDC. Specifically, they showed that the joint RL-TL technique, with generative adversary network (GAN) to generate augmented data for RL, can significantly improve the learning process by almost two orders of magnitude while achieving about half of the maximum performance gain after the domain re-connection following fragmentation. Results of this work were presented at the SPIE 2021 Conference (https://dais-ita.org/node/5818). By extending the work for fragmentation of multiple SDC domains, the team also prepared a demonstration of this work for the forthcoming final Showcase event.

Task 7.2: Federated Policy Learning and Management

The second main objective of Project 7 is the development of approaches for federated learning of both local and global policies, and federated policy management that enable the composition of policies learned at local parties. At the core of policy transfer is the fundamental problem of transfer learning: how models trained to learn policies in one source domain can be transferred to a related by different target domain, in particular when there is limited access to data in the target domain. Addressing this problem would allow a local coalition party to leverage models learned by other source parties on source datasets (i.e., source models) when learning their own model in the presence of limited data. During the second year, the team of 7.2 has progressed with work on generative adversarial networks (GANs) approach for creating a domain-invariant mapping of the source and target datasets. Researchers from IBM U.S., in collaboration with Purdue, ARL, and Imperial College, have applied the GAN-based methodology to the problem of adapting a classifier, trained to recognize heavy-weight and light-weight vehicles in an urban environment based their sounds, for the same classification task for vehicles in an artic environment and in a desert environment. Experiments have been carried out using the Acoustic-seismic Classification Identification DataSet which has a lot of acoustic data for vehicles in a normal environment but very few for the desert and artic environments. Results from these experiments show that the GAN-based approach is more accurate compared to baseline approaches. Results are reported in a demo/presentation in the September 2021 final DAIS meeting. Researchers from Purdue have extended the GAN-based approach to support the privacy of the source dataset. The new approach is based on a differentially private (DP) adversarial domain-adaptation (DP-ADA) workflow, which also supports the case in which the source and target datasets have different feature dimensions. The workflow is organized according to the following two steps: (i) Organization A has already created a dataset of labeled data. This dataset is referred to as the source dataset. Organization A then trains a generative adversarial network (GAN) on the source dataset with DP guarantees. In the workflow, this GAN is trained using the DP-CGAN approach, which is able of generating a synthetic dataset mimicking the probability distribution of the source dataset with DP guarantees. Organization A then provides the trained GAN to Organization B. (ii) Organization B uses the GAN provided by Organization A to generate a synthetic source dataset to mimic the source dataset. It then performs adversarial DA leveraging its own small, labeled target dataset. The approach has been evaluated on intrusion detection datasets; the results show that it achieves high accuracy compared with two baseline approaches even for a small value of the privacy budget (e.g., epsilon = 0.18) and a small labeled target dataset A paper reporting these results is currently under submission.

As part of Task 7.2, the team has carried out research on applying security policies to constrain explorations made by reinforcement learning (RL) agents managing software defined coalitions (SDCs). Researchers from Purdue, in collaboration with IBM U.S. and Imperial College, have developed a framework by which a RL agent is able to speed up/slow down network flows based on a dynamic assessment of potential risks of each flow. The approach has been instantiated to software defined networks, a special case of SDCs. The evaluation results show that the security constrained RL technique is able to optimize the transmission rate while at the same time achieving good results in identifying malicious flows. The results are reported in the paper titled “A Security-

Constrained Reinforcement Learning for Software Defined Networks", IEEE ICC 2021 (https://dais-ita.org/node/6222). Researchers from Purdue have extended such a framework to support 'intelligent policies' for traffic engineering (routing) in order to both maximize functionality gain and minimize security risk. The extended framework, referred to as STE-SDN, has been instantiated in a simulated SDN environment of 5000 nodes with different security services and three attack classes: DDoS, Web-based and Brute-Force attacks. The framework has been analysed using the CICIDS-17 dataset in terms of performance and effectiveness in mitigating security risks. The results show that STE-SDN reduces the security risk on average in all the scenarios considered. With respect to performance, the experiments show that when using STE-SDN the optimal latency loss achieved is higher than when STE-SDN is not used. This is the cost that one has to pay for better network security. However, from the results one can see that for many of the episodes the cost is not that high as alternate secure candidate paths by the RL agent have similar latency to the optimal insecure candidate paths. Thus, as the functionally reward is also part of the reward function, STE-SDN is able to "intelligently" optimize the latency or functionality for these episodes but not at the cost of security. A paper reporting these results is currently under submission.

Project 8: Federated Learning for Coalition Analytics

A summary of all three project 8 tasks is included below.

Task 8.1: Distributed Online Learning with Multiple Learners

P8.1 has been working on the following areas in Year 2.

1. A Gang of Adversarial Bandits. In this work we considered the problem of recommendation to users which are connected in an (e.g. social) network. Specifically we have a set of items in which to recommend to users. On each trial we must recommend one item to a specific given user and afterwards receive the user's rating on the item. The goal is to get a high cumulative rating for the selected items. Our assumption is that users that are (strongly) linked in the given network are likely to enjoy the same items and hence give similar ratings. We suppose the non-stochastic setting. We give two learning algorithms, GABA-I and GABA-II, for this problem: both extremely fast and with good regret bounds. GABA-II is faster than GABA-I but has a slightly worse regret bound. This work is under review for NeurIPS 2021.

2. Cooperative Bandits with Constrained Feedback. In this work, we study a scenario where M agents cooperate together to solve the same instance of a K-armed stochastic bandit problem. Agents have limited access to a local subset of arms and are asynchronous with different gaps between decision-making rounds. The goal is to find the global optimal arm and agents are able to pull any arm, however, agents can only observe the reward when the selected arm is local. The challenge is a tradeoff for agents between pulling a local arm with observable feedback, or pulling external arms without feedback and relying on others' observations that occur at different rates. We develop AAE-LCB, a two-stage algorithm that prioritizes pulling local arms following an active arm elimination policy, and switches to other arms only if all local arms are dominated by some external arms. We analyze the regret of AAE-LCB and show it matches the regret lower

bound up to a K factor. This work has been submitted to NeurIPS 2021 (https://dais-ita.org/node/6167).

3. Distributed bandit learning. We focus on large-scale learning in distributed systems, and consider a scenario where multiple agents cooperate together to solve the same instance of a multi-armed stochastic bandit problem. The agents have limited access to a local subset of arms and are asynchronous with different gaps between decision-making rounds. The goal is to find the local optimal arm for each agent since agents are able to pull local arms. The challenge is a tradeoff between identifying optimal local arms and reducing the communication complexity of algorithms. For this heterogeneous multiagent setting, we propose two respective algorithms, CO-UCB and CO-AAE. Both algorithms are proven to attain the order-optimal regret. In addition, a careful selection of the valuable information for cooperation, CO-AAE achieves a low communication complexity. Last, numerical experiments verify the efficiency of both algorithms. We analyze the regret of the proposed algorithm and show it matches the regret lower bound up to a constant factor related to the number of arms. In addition to the above problem, we also investigated a learning problem in distributed systems with constrained information. Specifically, we consider a model where agents have limited access to a local subset of arms, and are asynchronous with different action rates. The goal is to find the global optimal arm and agents are able to pull any arm, however, they can only observe the reward when the selected arm is local. The challenge is a tradeoff for agents between pulling a local arm with observable feedback, or pulling external arms without feedback and relying on others' observations that occur at different rates. We propose AAE-LCB, a two-stage algorithm that prioritizes pulling local arms following an active arm elimination policy, and switches to other arms only if all local arms are dominated by some external arms. We analyze the regret of AAE-LCB and show it matches the regret lower bound up to a K factor.

Task 8.2: Agile Analytics Enabled by Decentralized Continuous Learning in Coalitions

P8.2 has been working on technologies supporting agile analytics, including machine learning techniques for rapid training and adaptation of analytics models and the use of learning-based approaches for resource-efficient execution of analytics applications.

The work in Year 2 has been focused in the following areas:

1. Efficient Federated Learning. We have extended our model pruning work to include a complete convergence analysis and a complete set of experiments. The paper has been submitted to the IEEE Transactions on Neural Networks and Learning Systems (https://dais-ita.org/node/6162). We have also extended our work on accelerated gradient methods for smooth convex optimization when gradients are inexact. Inexact gradients appear, for example, in large-scale machine learning problems where high-precision gradients are expensive to compute due to the limited system resources. A draft of the paper is available at https://dais-ita.org/node/6163. In addition, an earlier work on adaptive federated learning led by IBM US, Imperial, PSU, and ARL received the IEEE Communications Society Leonard G. Abraham Prize

2. Robust Solutions to Constrained Optimization Problems by LSTM Networks. Many technical issues for communications and computer infrastructures can be formulated as optimization problems. Gradient-based iterative algorithms have been widely utilized to solve these problems.

Much research focuses on improving the iteration convergence. However, when system parameters change, it requires a new solution from the iterative methods. Therefore, it is helpful to develop machine-learning solution frameworks that can quickly produce solutions over a range of system parameters. We have proposed a learning approach to solve non-convex, constrained optimization problems. Two coupled Long Short Term Memory (LSTM) networks are used to find the optimal solution. Numerical experiments using a dataset from Alibaba reveal that the relative discrepancy between the generated solution and the optimum is less than 1% and 0.1% after 2 and 12 iterations in the trained LSTM networks, respectively. A paper has been submitted to IEEE MILCOM 2021 (https://dais-ita.org/node/6219).

3. Joint Cardinality Reduction (CR), Dimensionality Reduction (DR), and Quantization (QT). We furthered our study of enhancing the proposed approximate k-means algorithms by incorporating QT into a pipeline of CR and DR methods. Previously, by analyzing the impact of QT on the overall approximation error, we formulated an optimization problem to jointly configure the CR, DR, and QT methods, so as to minimize the communication cost under a given bound on the approximation error. In this quarter, we finished evaluations on two real datasets, which demonstrated that incorporating suitably configured QT into the data reduction pipeline can further reduce the communication cost by up to 60% in the centralized setting and 10% in the distributed setting compared to using CR and DR alone, without adversely affecting the computation complexity at data sources or the quality of the solution. The work is under submission to TPDS (https://dais-ita.org/node/5872).

4. Online Resource Allocation Using Distributed Bidding Approaches. We have improved the scalability and flexibility of our resource allocation algorithm. Since a realistic system may consist of thousands of small jobs, we implement server-side clustering to improve the speed of job allocation in cases where using the knapsack algorithm alone would be a hindrance. In addition, we are exploring a pre-emption case, in which jobs may be evicted from servers sometime after allocation to allow more profitable jobs to be accepted. A paper has been accepted to IEEE MASS 2021.

Task 8.3: Cognitive Workflows: Goal Directed Distributed Analytics Using Semantic Vector Spaces

P8.3 has been working toward the vision of cognitive workflows in which the properties of Semantic Vector Spaces (SVS) can be combined with the Vector Symbolic Architectures (VSA) to semantically represent services and service workflows in the coalition setting. The Military relevance of this task is in a future coalition context where analytics applications can be automatically composed from assets and services that may be distributed across the coalition network and owned by different coalition partners. The question we are addressing is the possibility to discover the required component services and compose the necessary workflows to perform distributed analytics tasks.

The work in Year 2 has been focused in the following areas:

1. Mathematical Properties of Semantic Vector Space for Cognitive Workflow (Cardiff University, IBM Research Europe, IBM Research US, ARL, Dstl). This subtask has continued the investigation of hierarchical semantic service composition using VSA and service representation

as VSA vectors using Open Web standards. Building on previous work using dense hyper-vector representations (i.e. 10Kbit binary vectors). The work has shown how to describe verifiably trustable services as VSA vectors in support of multi-party operations A paper entitled "Enabling Discoverable Trusted Services for Highly Dynamic Decentralized Workflows", was presented at the: 15th Workshop on Workflows in Support of Large-Scale Science conference. Held on 11th November 2020. https://dais-ita.org/node/5474. A video presentation that describes our work on VSA work has been produced. See https://dais-legacy.org/1a11/ We also have a video presentation which explains how VSA can be used for unambiguous communication exchange in coalition operations, by using semantic symbolic vectors where coalition partners may use different terminology/language to describe the same assets. See https://dais-legacy.org/1a04/. A major achievement during the year has been the development of a complete analytical formulation of a sparse vector representation of symbolic vectors and an approach for representing real valued vectors in the sparse vector representation. The work has shown that the sparse vector representation is better suited to low energy neuromorphic processing (see below). A paper describing these results is in preparation and is planned to be submitted to a special issue of Frontiers in Neuroscience in September.

2. Distributed Cognitive workflow (Cardiff University, IBM Research Europe). This subtask has focused on the constructing of Compositional Plan Vectors (CPV) using reinforcement learning. The idea is that since most environments have some level of inherent structure to them, where subtasks can be completed in a certain order to complete a larger goal. The work has shown that if we harness the underlying structure of an environment by optimising for compositionality, then we can improve the training time of multi-task reinforcement learning. We further demonstrated that if we store prior experiences in a manner that allows for reuse as self-produced expert demonstrations, then we can optimise for semantic similarity of tasks, further improving efficiency of training. Results from our experiments strongly support this hypothesis and we have shown that this works up to the equivalent of MT-10 benchmark standard. A paper entitled "Reinforcement Learning with Compositional Plan Vectors and Trajectory Experience Replay" has been produced and submitted for publication. See: https://dais-ita.org/node/5904.

3. Edge Efficient Cognitive Workflows (Cardiff University, IBM Research Europe, Purdue University, ARL). This subtask has been investigating how VSA operations can be performed in energy constrained environments using new emerging devices that can perform 'In Memory' and/or neuromorphic (i.e. spiking neural network) processing. Energy savings of the order of x100 are achievable in comparison to equivalent standard hardware implementations. A paper describing the use of a PCM memory device for edge of network operations was presented at the SPIE 21 conference: Energy efficient in-memory computing to enable decentralised service workflow composition in support of multi-domain operations. See: https://dais-ita.org/node/6152. A video presentation that describes the work on energy efficient VSA using in memory processing has been produced. See: https://dais-legacy.org/1f01/ Using a sparse vector representation we have shown that Our work on SNN's has also continued to investigate the possibility of using deep SNN's to perform the spike based back propagation and learning in deep architectures as reported in https://dais-ita.org/node/3321 This type of SNN learning could be exploited in the Plan Arithmetic subtask in 8.3.2 where SNN's of this type could be used to perform the CNN functions required to learn the policy for the reinforcement learning. This would potentially provide a future energy efficient approach for online reinforcement learning at the network edge. The work has

demonstrated an elegant neuromorphic simulation of a VSA clean-up memory and shown that the energy requirements of this model are like that for dense vectors using in memory processing on a phase change memory device (PCM). A video describing this work has been produced for the DAIS legacy website. See: https://dais-legacy.org/1f02/

4. UK Transition Activities (IBM UK, Cardiff,Dstl). The following transition activities are/have been undertaken in collaboration with Dstl:

- Dynamic reconfiguration of battlespace communications plans. Whilst this task was completed in Year 1, the results were incorporated into the SPIE 21 paper as a representative case study of using VSA in a representative TacCIS environment. See: https://dais-ita.org/node/6152 . A video presentation of this work has been produced. See https://dais-legacy.org/1a02/.
- NATO Federated Mission Networks (FMN) Service Discovery. In this transition task we have demonstrated how the VSA approach for defining and orchestrating distributed service workflows can be used to perform agile command-and-control tasks in a TacCIS type environment. Using an example scenario, we have shown how a commander can issue a command encoded as a VSA vector that can perform a sequence of complex discovery and orchestration tasks. A demonstration of this capability has been developed using a scenario shows how a group of military assets with different required characteristics can be discovered and commanded to communicate via a common FMN chat service, using their geo-position to determine the most appropriate units to participate in the mission. A final report (including a link to a video of the demonstration) is available. See: https://dais-ita.org/node/6153
- Knowledge Graph Embedding. Generating semantic vector spaces has been a key element of the work and this transition task reviewed a wide range of techniques that can be used to learn vector embeddings of knowledge graphs. Details of the military application cannot be made public but a paper summarising the unclassified work in this transition task was submitted to SPIE 21. See: https://dais-ita.org/node/5751

Project 9: Agile Composition for Coalition Environments

The overall research context for Project 9 during Year 2 was on maintaining a reliable working relationship among partners in a coalition despite differences and attacks on cohesion and analytics by external and internal adversaries. We highlight below select activities and findings from each of the two tasks under this project.

Task 9.1: Interpretability of Neural Networks in Distributed & Contested Environments under Incomplete Trust

In BPP20 year 2, researchers in the task had several key accomplishments.

1. Expanding explainability beyond images [Cardiff, IBM-UK, IBM-US]: We worked on expanding our work on explanations to input modalities beyond images. Specifically, we developed strategies for explaining video models. We observed that applying traditional mechanisms for explaining video models lead to a lot of visual clutter (due to spatial relevance across frames). In this work we developed a selective relevance technique to filter out the spatial relevance and only highlight the relevant temporal relevance used for classifying videos. The preprint version of the work is on arXiv and currently under submission to BMVC, 2021 (https://dais-ita.org/node/6217). We have extended selective relevance to work with models that work with both audio and video modalities as well (https://dais-ita.org/node/5136). More recently, we have started to explore the role of attention in explaining large-scale transformer and BERT models which have become very popular for NLP tasks. As part of a joint effort between 9.1 and 7.1, we have developed novel technique for explaining transformer models (https://dais-ita.org/node/5820) and we are currently exploring how effective our technique is for explaining classification tasks on the Reddit dataset.

2. Quantifying Uncertainty in Explanations and Exploiting Explanations as a Side Channel [ARL, IBM-UK, IBM-US, UCLA]: We investigated ways to quantify uncertainty in explanations. As models are being increasingly accompanied with explanations, there is a need to quantify how uncertain these explanations are, and this would be useful in calibrating trust on the model. Some of these ideas were previously presented as a Patterns journal paper (https://dais-ita.org/node/5158). An initial exploration of various to quantify uncertainty in explanation was also published in SPIE, 2021 (https://dais-ita.org/node/6149). More recently, we explored the use of explanations as a side-channel for efficient black-box attacks and investigated corresponding mitigation strategies. A variety of explanation methods have been proposed in recent years to help users gain an insight into the results returned by neural networks, which are otherwise complex and opaque black-boxes. However, the explanations give rise to a potential side-channel which an adversarial user can leverage for attacks. In particular, common explanation methods that highlight input dimensions according to their importance or relevance to the result also leak information about gradients. This safety concern of providing post-hoc explanation information has not been explored in the community. In this work, we first show that such information can be easily exploited to significantly improve the efficiency of an attacker seeking to mount an adversarial input attack, which is concerning as explanations play a particularly important role in safety-critical systems. We investigate in depth how one can leverage explanation information to attack neural models and show that the commonly used explanation techniques such as Gradient, Integrated Gradients, SmoothGrad, can give attackers an advantage in mounting successful black-box attacks by reducing the queries. Second, we propose a differential privacy (DP)-based strategy to defend against these attacks effectively. By evaluating the trade-off between utility and privacy of explanations, we show our DP-based defense can reduce the information leakage and maintain their ability to inform the human user. Finally, we empirically demonstrate the risk of bypassing the DP-based defense using an adaptive attack based on averaging multiple explanations. We show that the defense is less easily circumvented if the explanation is a non-linear transformation of the gradient. We recently submitted a paper based on this work (https://dais-ita.org/node/6214) to IEEE Oakland S&P.

3. Learning to detect malicious clients in federated learning [IBM-USA, UCLA]: Recent years have seen the increasing attention and popularity of federated learning (FL), a distributed

learning framework for privacy and data security. However, by its fundamental design, federated learning is inherently vulnerable to model poisoning attack: a malicious client may submit the local updates to influence the weights of the global model. Therefore, to detect malicious clients against model poisoning attacks in federated learning is useful in safety-critical tasks. Since the current proposed techniques are either less reliable or easily bypassed, we propose to learn an anomaly detector on the central server side and use it to detect unseen model poisoning attacks in federated learning. The detector can achieve a 99.8% detection AUC score while enjoying longevity as the model converges. In addition, our approach generalizes well and achieves high detection performance against unforeseen attacks. Our ongoing direction is to incorporate Secure Multi-party Computation (SMPC) to enable detection of malicious clients while not revealing individual model updates to protect user privacy.

4. Towards imperceptible query-limited adversarial attacks with Perceptual Feature Fidelity Loss [UCLA]: Recently, there has been a large amount of work towards fooling deep-learning-based classifiers, particularly for images, via adversarial inputs that are visually similar to benign examples. However, researchers usually use Lp-norm minimization as a proxy for imperceptibility, which oversimplifies the diversity and richness of real-world images and human visual perception. In this work, we propose a novel perceptual metric utilizing the well-established connection between the low-level image feature fidelity and human visual sensitivity, where we call it Perceptual Feature Fidelity Loss. We show that our metric can robustly reflect and describe the imperceptibility of the generated adversarial images validated in various conditions. Moreover, we demonstrate that this metric is highly flexible, which can be conveniently integrated into different existing optimization frameworks to guide the noise distribution for better imperceptibility. We show that our approach can achieve a 109% higher SSIM increment and similar PSNR value (+ 2dB) using the same query budget. The metric is particularly useful in the challenging black-box attack with limited queries, where the imperceptibility is hard to achieve due to the non-trivial perturbation power. A paper based on this work is under submission (https://dais-ita.org/node/5885).

5. Video Classification using Concept Bottlenecks via Automatic Concept Extraction [ARL, Imperial College, UCLA] (joint work with 10.2 and 10.3): Recent efforts in interpretable deep learning models have shown that concept-based explanation methods enable reasoning about extracted high-level visual concepts from images, e.g., identifying the wing color and beak length of a bird for species classification. However, concept-based explanation architectures for image classification rely on a pre-defined set of concepts and do not generalize to video classification. We present an automatic concept extraction method that identifies a rich set of concepts from natural language descriptions of videos that obviates the need for researchers to manually define concepts in domains in which they may not be experts. We also propose a generalized, end-to-end deep learning architecture that combines a video classification model with an attention-based concept layer for explainable video classification. To demonstrate our method's viability, we constructed a dataset combining an existing baseball video dataset with short, crowd-sourced natural language explanations and labels. The accuracy of our interpretable model is competitive with the popular methods for video classification -- an essential step in integrating human explanations and knowledge with the predictive capabilities of deep learning.

6. Model Selection for Deployment from a Rashomon set of Models [IBM-USA, UCLA]: Deploying a machine learning model for any particular task faces the following key challenges: 1) The test data in the deployment scenario will not closely resemble the distribution of the training data. 2) Identifying the best or a subset of models to deploy given a Rashomon Set of Models. Multiple machine learning models can be trained each with different hyperparameters and there exists a set of these models that have the training performance close to the best performing model for that given task. This set of models is called a 'Rashomon set'. Previous works have suggested that an ensemble of models can be used to overcome these challenges but in a deployment setting, where the test data's distribution is different from the training data, a few bad performing models in the ensemble will bring down the overall performance of the ensemble. Our hypothesis is that an ideal model for deployment would have learned the distinguishing feature of the test dataset in the training phase. To identify the best model we make use of post-hoc explanation methods propose two different approaches: (i) Identify the model that has the most similar rank of important features between the test data and the training dataset; and, (ii) Instance-based model selection, where we select the best model from a set of models for each test sample. The model with the high prediction confidence and similar rank of important features between the test sample and training dataset will be selected. To evaluate our proposed approaches, we train text classification models on 20 Newsgroup dataset and deploy them to classify Reddit posts to their correct subreddits. This research is still on-going.

Task 9.2: Network intelligence from negative ties

In this final period of activity, numerous accomplishments have been achieved, and our team made substantial progress on expanding our understanding of negative, conflictual network ties in our research. In collaboration with Penn State and Cardiff, we developed a new approach to identify network motifs in graphs with the use of exponential random graph models. We use this statistical method to detect network motifs in both positive and negative networks to identify "network signatures" that differentiate between cooperative and conflictual networks. One article, forthcoming in *Applied Network Science*, found unique combinations of dyad, triad, and tetrad trends that distinguish among 24 networks https://dais-ita.org/node/6201. An additional paper systematically compares positive and negative multiple online and offline networks, finding reciprocity, but not transitivity, common in both negative and positive networks. It was presented at the meetings of the American Sociological Association, August 2021; a revised version is submitted for journal review https://dais-ita.org/node/5419, .

In other work (IBM US, Penn State, Cardif, and ARL), we introduce a new modelling technique, based on Graph LSTMs (Long Short-term Memory), that improves predictability of who and when will utter curse words next in social media data. Evaluations on the Twitter dataset show that considering network effects improves prediction performance by over 30%, in comparison with traditional statistical models such as the Hawkes Point Process. This paper was presented mid-January at the international network conference, NASN, https://dais-ita.org/node/5877.

Another paper, "From Social Networks to Negative Ties – Refining Analysis for Conflict and Adversarial Interaction" (Braines, Whitaker, Felmlee, McMillan, Julien, Giammanco) was presented at the SPIE 2021; this work applied the Cogni-Sketch environment to expose machine learning and analytic techniques for the earlier work on Indian terrorism data -> https://dais-ita.org/node/5813.

New techniques to characterize the difference between edges in triads have been introduced, giving a new perspective through which the dissemination (or containment) potential of induced substructures can be assessed. This represents a collaboration between Cardiff and Penn State https://dais-ita.org/node/6207. We introduce a new local edge-centrality measure that signifies the importance an edge plays within induced triads for a directed network. We observe that an edge can play one of two roles in providing connectivity within any particular triad, based on whether the edge supports connectivity to the third node or not. We call these alternative states overt and covert. As an edge may play alternative roles in different induced triads, this allows us to assess the local importance of an edge across multiple induced substructures. We introduce theory to count the number of induced triads in which an edge is overt and covert. Using 34 data sets derived from public sources, we show how the presence of overt and covert edges can be used to profile diverse real-world networks. This work is submitted for consideration at ASONAM 2021, IEEE/ACM International Conference on Advances in Social Network Analysis and Mining.

Also based on collaboration between Cardiff and Penn State, a journal paper has been submitted concerning understanding the characteristics of COVID-19 misinformation communities (https://dais-ita.org/node/6208). This demonstrates the utility of substructure census and motifs in distinguishing behavioural patterns associated with misinformation. The technique avoids recourse to semantic analysis and they identify the underlying substructures that are instrumental in characterizing key behavioural interactions. This has been revised for consideration by the Elsevier Online Social Networks and Media Journal.

Finally, collaboration between Cardiff, Penn State and IBM US has taken place to consider multi-scale user migration on Reddit (https://dais-ita.org/node/6209). This also represents a collaboration with T10.2 (Preece, Cardiff) and the work addresses users moving across virtual spaces within or across platforms. This paper provides a set of methods to help study migration at two scales: micro-scale involving individual users moving between spaces over relatively short time periods (between posts), and macro-scale involving groups of users moving over relatively longer time periods (changing their posting habits). This was presented at the Workshop on Cyber Social Threats at the 15th International AAAI Conference on Web and Social Media (ICWSM 2021).

Project 10: Instinctive Analytics in a Coalition Environment

In this project we pursue the organization, integration and autonomy of both human and machine agents to fulfill coalition objectives pertaining to multi-domain scenarios with context and situational awareness.

Task 10.1: Coherence in Coalitions: understanding internal group behavior and dynamics in complex multi-domain environments

In task 1 we have focused on understanding the function and operation of coalition-based groups in terms of their coherence and ability to make effective decisions.

- Work on the interplay between cognitive dissonance and social networks has been developed through a computational model that assesses effects through measures of tolerance and conviction for computational agents (https://dais-ita.org/node/6205). This paper considers networks with up to 10,000 nodes and is featured in IEEE Transactions on Computational Social Systems. As far as we can establish, this is one of the first papers to address this topic through computational modelling of large groups. This work has involved Cardiff, Yale, Dstl and ARL.
- Cardiff in collaboration with Yale has developed computational insights into the evolution of identity fusion, which represents a personal alignment with a group. This concept is important because it offers the current best explanation as to why individuals can become empowered to act selflessly for a group (e.g., the devoted actor concept). It is one of the first computational models to study the evolution of this concept and provides evidence that sensitivity to hypocrisy may play an important role in a group context. This has been published in Scientific Reports (Nature). See https://dais-ita.org/node/6204.
- Southampton, in collaboration with IBM UK and ARL, have further built upon the BPP20 work concerning competitive influence maximisation (https://dais-ita.org/node/5355 and https://dais-ita.org/node/5356). This has involved developing the problem in the presence of negative ties and under nonlinear budget constraints. The work draws collaborations across Southampton and ARL. This effort has recently resulted in the following publication (https://dais-ita.org/node/6228) titled "Shadowing and shielding: effective heuristics for continuous influence maximisation in the voting dynamics", published in PLoS ONE.
- Progress on the issue of problem solving through coalitions has also taken place. A paper "Optimizing the efficiency of collective decision making in groups" (https://dais-ita.org/node/6229) has been published in proceedings of SPIE2021, with authors - Malgorzata Turalska, Rosie Lickorish, Geeth De Mel, Liam Turner, Roger M. Whitaker, representing a collaboration between ARL, IBM UK and Cardiff. Led by ARL (Turalska), a further manuscript is being prepared summarizing work on collective problem solving adopting the NK-landscape framework, underlying the necessity of a memory mechanism for efficient search of solutions in complex spaces. The targeted publication venue is PNAS.

Task 10.2: Learning and Inferencing in Neuro-Symbolic Hybrids for Uncertainty Aware Human-Machine Situational Understanding

In task 2 we address the need to rapidly integrate machine analytic components in a way which (1) is aware of uncertainties; (2) exploits synergies; and (3) supports human decision makers. Our objective is to achieve a step change in free-flowing composition of uncertainty-aware human-agent and agent-agent information analytics. Building on the first year of BPP 2020, the second year has seen further significant success in our neuro-symbolic artificial intelligence (AI) research.

Previously, Cardiff, UCLA and ARL developed AI architectures that can learn and understand complex events, enhancing the trust and coordination between human and machine needed to successfully complete battlefield missions. This work addresses the challenge of sharing relevant knowledge between coalition partners about complex events, i.e., compositions of primitive activities connected by known spatial and temporal relationships. For such events, the training data available for machine learning is typically sparse. Two different approaches were developed by the team to enable learning at the neural layers by propagating gradients through the logic layer. The first, Neuroplex – published at SynSys 2020 (https://dais-ita.org/node/5382) – uses a neural surrogate for the symbolic layer. The second, DeepProbCEP – published at ICLP 2020 – uses DeepProbLog to propagate the gradients (https://dais-ita.org/node/5340). In year two, following feedback from the Peer Reviewers. we performed benchmarking experiments for the two approaches. We have shown that both Neuroplex and DeepProbCEP significantly outperform neural-only approaches, especially when training with small amounts of training data. DeepProbCEP achieves slight improvements on performance over Neuroplex while, using a logic layer makes DeepProbCEP slower when training compared to Neuroplex (manuscript under preparation for submission to *Expert Systems with Applications*).

Finally, in year 2 we created an integrated demonstration of multiple task 10.2 scientific outcomes together with elements from 9.1. This demonstration, Adapting AI systems to Recognise New Patterns of Distributed Activity, backed-up with open source code for its science and technology components, shows how our human-agent architecture, realised through the Cogni-Sketch environment (https://dais-ita.org/node/6145), enables rapid integration and adaptation of neuro-symbolic and deep neural network-based AI services processing a diverse range of multi-modal data. A paper describing the demonstration, Coalition Situational Understanding Via Adaptive, Trusted and Resilient Distributed Artificial Intelligence Analytics, was accepted for publication at the NATO IST-190 Symposium on *Artificial Intelligence, Machine Learning and Big Data for Hybrid Military Operations (AI4HMO).*

Task 10.3: NSPL – A Neural-Symbolic Learning of Generative Policies in Coalition Environments

In Task 3 we explore how to enable coalition systems and devices to operate with minimal human intervention in highly heterogeneous, and dynamic contexts whilst maintaining a level of security and interpretability, to guarantee robust distributed analytics. Building on the achievements of the first year, in the second year we have consolidated, extended and integrated our scientific advancements neural-symbolic AI along four different directions: (1) neural-symbolic learning for learning interpretable models from unstructured data, (2) scaling up our symbolic machine learning system for learning interpretable models, (3) embedding logical inference into continuous space and (4) enabling explainability predictive models through automated concept extraction. Previously, Imperial, IBM UK, IBM US and ARL developed a neural-symbolic learning approach, called FF-NSL, for performing interpretable learning from noisy raw data (https://arxiv.org/abs/2012.05023). The underlying idea is to enable parties from a coalition operations to use their pre-trained deep learning models to extract features from the local surrounding environment and use the predictions generated by these models as contextual information for a symbolic machine learning system to learn interpretable decision making models (based on policies) that are robust to noise. The integration of deep learning and symbolic machine learning opened the question on how effective the interpretable learned models are when applied to contextual situations that are different from the one used to train the deep learning models, and for which the deep learning models would give wrong predictions. In year two, we have studied in-depth the robustness of our FF-NSL architecture in the presence of data distributional shift. We have developed a new data-set that is particularly challenging for pure deep learning architectures and been able to demonstrate (1) the advantage of using deep learning architectures that are sensitive to uncertainty when extracting features for the symbolic machine learning system and (2) the robustness of interpretable models learned by our FF-NSL system in maintaining their accuracy closer to 100% when applied to data with 70%-80% perturbation. We also extended the class of models that can be learned by our symbolic machine learning system FastLAS, in particular learning models about predictions that are not directly observed. Our approach FastLAS2 (https://dais-legacy.org/1c08/) for solving expressive interpretable machine learning task has been published at IJCAI 2021 is publicly available as open source. We have then developed in Year 2 a demo that integrates all these achievements. This demo, "Adapting AI with neural-symbolic learning and its application to generative policies in distributed coalition operations" shows how devices within a coalition operation can autonomously learn their behaviour, expressed as policies, through local communication to nearby devices, and use their learned generative policy models to rapidly come online into a mission despite contextual changes in which the operations are performed. The learning uses unstructured data which are processed using pre-trained deep learning models. The output of these models together with other relevant structured contextual information is used by our FastLAS system for intepretable machine learning to learn generative policy models expressed as Answer Set Grammars (developed in BBP18). We also extended the field of neural-symbolic AI by integrating in an end-to-end fashion the deep learning network training with logical constraints expressed as logic programs. The end-to-end training regime builds upon our previous year achievement on performing logical inference in continuous vector space (KR 2020), and has given rise to a novel Adaptive AI system for Human-Machine Federated Decision Making. Finally, joint collaboration between Tasks 10.2 and 10.3 has addressed the forth aspect of neural-symbolic AI, that of automatically extracting concepts from unstructured data to improve the explainability of predicted labels of complex video activities. The core idea is to automatically extract from natural language descriptions of videos, rich set of concepts that can be used in an end-to-end deep learning architecture. This architecture combines video classification model with an attention-based concept layer for explainable video classification. We have demonstrated that the accuracy of our learned interpretable model is competitive with SoTA methods for video classification and that the learned concepts provide a viable means for explaining the predicted labels.

Additional details

Additional details about the BPP20 period plans and goals can be found at https://dais-legacy.org/dais/historical_docs/files/BPP20-Technical-Volume.pdf. A summary of accomplishments and publications for the BPP20 and historical program documents including the Biennial Program Plan (BPP) documents, and annual reports can be found at https://dais-legacy.org/history/.

In the above summary of accomplishments, we have chosen to highlight certain key publications for each of the four projects. The next section of this report is a summary of publications occurring during the BPP20 year two period. Please refer to the Science Library (https://dais-legacy.org/science-library/) for all publications during this period and the whole of the DAIS ITA program so far.

Summary of Publications

Introduction

The data and statistics in this document are summarised from the publicly available DAIS ITA Science Library [https://dais-legacy.org/science-library/]. In some cases there are publications listed in this document that are not yet published to Science Library, e.g. because they have not yet been published in their external conference or journal venue. Eventually, all successfully accepted papers will be published to Science Library, even after the BPP20 period has concluded. The papers counted in this summary document will tally closely to those reported throughout the BPP20 period in the Quarterly Progress Reports (QPRs) but they are reported separately here as a single succinct summary as the repeating periodic nature of the QPR document means that papers are regularly reported in numerous quarters as they progress from submitted to accepted and published, plus each QPR only deals with the papers relevant to that quarter. The purpose of this stand-alone summary document is therefore to provide a short and simple overall summary for the period, backed up by detailed statistics on science library and more detailed contextual reporting in each of the QPRs.

Overall Summary

The BPP20 period (20 months) saw 28 journal papers published at external peer reviewed journals, and 67 conference papers published at peer reviewed external conferences or workshops. This combines the BPP20 year one period (12 months) which saw 17 journal papers published at external peer reviewed journals, and 39 conference papers published at peer reviewed external conferences or workshops and the BPP20 year two period which saw 11 journal papers published at external peer reviewed journals, and 28 conference papers published at peer reviewed external conferences or workshops. This compares to 6 journal papers and 90 conference/workshop papers in the IPP period (16 months), and 36 journal papers along with 182 conference/workshop papers in the BPP18 period (24 months). It should be noted that papers are "counted" in the period that they are published, and this may not be the same period as when the effort was spent in writing the paper, especially for journal papers which may require more than 6 months for publication. With this important caveat in mind, the publication data is broken down by project in the table below for each project and period. The IPP and BPP18 data is included purely for comparative purposes.

In addition to the confirmed publications we are aware of for 44 another possible external publications including papers not yet confirmed as published, and those found in Google Scholar but not reported in CENSE. This is likely a maximum value if all submitted papers are accepted and published as some publications will be rejected or be included in venues which do not publish proceedings.   We estimate the actual number of additional external publications to be around 30.

Period		Journal	External Conference
BPP20 (Year 2) Jan-21 to Sept21	BPP20 P7	1	2
	BPP20 P8	4	10
	BPP20 P9	2	11
	BPP20 P10	4	5
	Cross-project	-	-
	Total	**11**	**28**
BPP20 (Year 1) Jan-20 to Jan-21	BPP20 P7	8	8
	BPP20 P8	5	16
	BPP20 P9	3	3
	BPP20 P10	1	12
	Cross-project	-	-
	Total	**17**	**39**
BPP18 Jan-18 to Jan-20	BPP18 P1	13	28
	BPP18 P2	4	31
	BPP18 P3	4	25
	BPP18 P4	9	25
	BPP18 P5	3	47
	BPP18 P6	3	16
	Cross-project	0	10
	Total	**36**	**182**
IPP Sep 2016 – Jan-2018	IPP P1	1	10
	IPP P2	0	23
	IPP P3	3	8
	IPP P4	0	13
	IPP P5	0	7
	IPP P6	2	18
	Cross-project	0	11
	Total	**6**	**90**

Table 1: Overall confirmed external publication summary (by project)