Προτεινόμενα Θέματα Διπλωματικής Εργασίας

Από

[Incelligent Private Company, https://incelligent.net]

A. Application of Spatio-Temporal geometry forecasting models on telecommunication and/or transportation data

Introduction & Problem Statement:

Advanced geospatial analytics and data mediation platforms (such as Incelligen't product RAN.AI) can gather a vast amount of information for every telecommunication network event that occurs within an underlying telecommunication access network (2G, 3G, 4G, 5G ++). This information consists of network performance KPIs correlated in temporal labels (every a few seconds to every minute) as well as spatial labels that are either coarse (e.g. linked to a cell with a specific coverage area) or fine (localized via probe-assisted 3rd party localization enablers). Having this information processed through compression and aggregation pipelines, in conjunction with the application of spatial indexing techniques such as rasterization, generates ideal datasets for historical analysis of zones with detected performance characteristics / profiles. The study of how these problematic zones a) correlate with the current configuration of the cell network (antenna aspects, frequencies, bandwidth, interference) b) take form / shape based also in the underlying urban or rural topology as well as **c) evolve through time** via either i) no action, ii) random surrounding actions or iii) specific optimization actions can produce valuable components that can be a part of a predictive infrastructure deployment solution with great value on the CAPEX and OPEX phase of telecommunication infrastructures.

Required Skills:

- Strong programming skills, preferably in Python.
- Understanding of machine learning concepts, including supervised and unsupervised learning. Familiarity with Deep Learning frameworks, such as Tensorflow or PyTorch, for training and fine-tuning of Deep learning models.
- Knowledge of image processing techniques and algorithms to preprocess and analyze network KPI ground truth map data is considered a plus.
- Ability to conduct comprehensive literature reviews and stay updated with the latest advancements.
- Effective communication skills to present research findings and recommendations in a clear and concise manner.

References:

1. https://medium.com/stanford-cs224w/predicting-evolution-of-dynamic-graphs-7688eca1daf8 (general idea)
2. https://arxiv.org/abs/2003.00842 (gnn approach)
3. https://journals.aps.org/prmaterials/abstract/10.1103/PhysRevMaterials.6.103801 (stable diffusion approach)
4. https://github.com/gabrielspadon/ReGENN (example impl)

B. Quantization and Distillation of Large Models for Efficient Deployment in NLP or Computer Vision Tasks

Introduction:

As the size and complexity of deep neural network models continue to grow, deploying them on resource-constrained devices or in low-bandwidth environments becomes a significant challenge. This thesis proposal focuses on the quantization and distillation of large models, specifically in the domains of Natural Language Processing (NLP) or Computer Vision, with the aim of downsizing popular open-source models such as ResNet and BERT. The proposed research seeks to explore techniques that enable efficient deployment of these models by reducing their size while maintaining performance. The evaluation will be based on factors such as inference time, computational requirements (FLOPs), and relevant evaluation metrics.

Problem Statement:

Large-scale models, such as ResNet in Computer Vision and BERT in NLP, have achieved state-of-the-art performance but often come with substantial computational and memory requirements. These limitations hinder their practical deployment on devices with limited resources or in scenarios with strict latency constraints. Hence, there is a need for effective techniques to reduce the size and computational complexity of these models while preserving their performance.

Objective:

The primary objective of this thesis project is to investigate and develop techniques for the quantization and distillation of large models in NLP or Computer Vision. The specific goals of this project include:

1. Research and analyze existing methods for model quantization, which reduce the number of bits required to represent model weights and activations, and distillation, which transfers knowledge from large models to smaller models.
2. Select representative open-source models, such as ResNet for Computer Vision or BERT for NLP, and establish baselines for their performance in terms of accuracy, inference time, and computational requirements.
3. Develop and implement quantization methods to reduce the size of the selected models while maintaining a desirable level of performance.
4. Investigate distillation techniques to transfer knowledge from the large models to smaller, quantized models, ensuring minimal loss in performance.
5. Benchmark the quantized and distilled models against the baselines using evaluation metrics appropriate to the specific tasks, such as image classification accuracy or language understanding performance.
6. Assess the impact of the quantization and distillation on computational requirements, including FLOPs (floating-point operations) and memory footprint, to analyze the efficiency gains achieved.
7. Provide insights and recommendations on the trade-offs between model size, computational requirements, and performance for efficient deployment of large models in NLP or Computer Vision tasks.

Required Skills:

- Strong programming skills, preferably in Python
- Familiarity with Deep Learning frameworks, such as Tensorflow or PyTorch and Huggin Face, for training and fine-tuning Language Model and other Deep learning models.

- Familiarity with techniques for model quantization, including weight quantization, activation quantization, and knowledge distillation.
- Ability to analyze and interpret evaluation metrics relevant to NLP or Computer Vision tasks, such as accuracy, precision, recall, or F1 score.
- Understanding of machine learning concepts, including model architecture and training processes for large models like ResNet and BERT.
- Ability to conduct comprehensive literature reviews and stay updated with the latest advancements in model quantization.
- Effective communication skills to present research findings and recommendations in a clear and concise manner.

C. Image-To-Image techniques to predict 5G Network Performance using satellite imaging

Introduction:

The ever-growing cellular technologies that allow for the world-wide personal telecommunication systems (namely 3G, 4G, 5th Generation of the 3GPP standards) become increasingly complex to design, plan and predict, making traditional deterministic engineering approaches that utilize domain expert knowledge partially accurate or obsolete. With the rise and the continuous validation of the C-NN family of deep learning models in various domains, opportunities for DL-assisted radio network rollout schemes are seen as candidates for the future industry standard and it may play a key role for the technology's next updates (6g and ongoing).

Problem Statement:

Cellular networks are positioned in key locations of an underlying geography aiming at providing a specific QoS standard (measured in a set of radio network KPIs) for an area of interest. This area usually contains a number of mobile terminals that use a selection of interchangeable technology versions (2G, 3G, 4G and onwards) according to a) the quality of each technology, b) the policy that is being applied from the mobile network operator and c) the capabilities of the terminal device to serve their demands for internet access and voice call / SMS communication. Network operators continue to improve their networks with a set of reconfiguration actions that affect the network experience of these mobile terminals. Examples of these actions are i) changing the configuration of a specific radio antenna element to "aim" at a different geographical location, ii) activation / de-activation of elements, iii) placement of a new technology layer – i.e a new set of cells that use a completely new technology "on top" of the existing previous. However, as the complexity of the multi-technology network layers rise, there is an increasing difficulty in predicting the KPI changes for any reconfiguration actions. In addition, the rollout of a new technology altogether is of key essence for the mobile operators because there is a very high-risk element for their technological investment (i.e. high ROI is expected within small periods of time). An engine that is able to use prior information (historical network KPI of previous technologies, network element configuration, previous re-configuration actions as well as imaging or rasterized geographical information) and can give accurate estimations of these KPIs for any given hypothetical action (with greater focus on new technology rollout) will prove of key importance and very high financial profit for telecommunication operators aiming at embracing such infrastructure updates.

Objective:

The primary objective of this thesis project is to test the applicability of c-NN based deep learning models for the execution of the new (or adjusted network's) performance KPIs. This system will be able to be integrated with the planning tools of an example telecommunication network operator and give him important insights for the outcome of future investments based on static configuration and/or image map data. Existing literature is pointing towards the direction of models, such as image-to-image, stable diffusion or other GAN-based families of models that can be "tweaked" in order to predict the network performance in the form of 2D, multi-KPI representation of the hypothetical new network's quality (e.g. the introduction of 6G on top of an existing 2-4-5G network) using such input. Key elements of this approach can be found in literature [1][2][3][4][5] that can transform an input image (or raw metadata / text) into other 2d representations based on their training set. Another model approach, [6] Is a GAN-based approach that can be used for DL-based enhancement of image data (i.e. increasing their resolution). For this thesis it can inspire the same approach that can allow for a more detailed 'drill-down' operation of the network KPIs within an area where we have very sparse, aggregate measurements.

The system will aim to achieve the following specific goals:

1. Train and fine-tune a state-of-the-art c-NN based deep learning using anonymized real network data and generate 2-dimensional KPI maps
2. Expand the model using configuration-based parametrization (exogenic variables in the form of text) to change the generated maps (e.g. for the performance of a hypothetical new radio layer)
3. Expand on different approaches on the same learning task using other techniques found in literature, highlighting their key pros and cons
4. Tackle with deployment aspects of the solution, including packaging, containerization, utilization of hardware resources (e.g. CUDA passthrough) and production-ready optimizations
5. The system can be tested under confidential scope with real telecommunication operator hypothetical scenarios with partially disclosable results.

Required Skills:

- Strong programming skills, preferably in Python
- Familiarity with Deep Learning frameworks, such as Tensorflow or PyTorch and Huggin Face, for training and fine-tuning of convolutional neural networks and other Deep learning models.
- Knowledge of Image processing techniques and algorithms to preprocess and analyze network KPI ground truth map data.
- Understanding of machine learning concepts, including supervised and unsupervised learning, to build and evaluate the generative model.
- Ability to conduct comprehensive literature reviews and stay updated with the latest advancements in image generative model technologies.
- Effective communication skills to present research findings and recommendations in a clear and concise manner.

References:

[1] Generative Adversarial Networks for Image-to-Image Translation. Arun Solanki, Anand Nayyar, and Mohd Naved. Elsevier (2021)

[2] Image to Image Translation with Conditional Adversarial Networks. Phillip Isola Jun-Yan Zhu Tinghui Zhou Alexei A. Efros. 2018

[3] Unsupervised Image-to-Image Translation Networks. Ming-Yu Liu, Thomas Breuel, Jan Kautz. 2018

[4] Palette: Image-to-Image Diffusion Models. Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans David J Fleet, Mohammad Norouzi. 2022

[5] An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. 2021

[6] TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. Yifan Jiang, Shiyu Chang, Zhangyang Wang1. 2021

D. Responsible AI: Ensuring Ethical AI Development and GDPR Compliance

Introduction:

The rapid advancement of artificial intelligence (AI) technologies has brought about numerous ethical challenges and legal concerns. One of the key areas of concern is ensuring responsible AI development, particularly in terms of compliance with the General Data Protection Regulation (GDPR). This thesis proposal aims to investigate the ethical implications and legal frameworks surrounding responsible AI development, with a specific focus on GDPR compliance. By addressing these issues, the research aims to contribute to the development of ethical and legally compliant AI systems.

Problem Statement:

The proliferation of AI technologies has raised concerns about the potential misuse of personal data and the erosion of privacy rights. As AI systems increasingly rely on vast amounts of personal data, it becomes crucial to ensure that the development of AI is conducted responsibly and in compliance with relevant legal frameworks, such as the GDPR. However, the ethical implications and legal requirements for responsible AI development and GDPR compliance are still relatively unexplored and require thorough investigation.

Objective:

The primary objective of this research is to evaluate the ethical implications and legal frameworks surrounding responsible AI development, specifically focusing on GDPR compliance of a social network that is currently being developed. By examining the current state of AI development practices and legal obligations, this study aims to identify potential gaps and challenges in ensuring ethical AI development and compliance with the GDPR. Furthermore, the research seeks to propose recommendations and strategies to bridge these gaps and promote responsible AI development that aligns with legal and ethical standards.

Required Skills:

- A solid understanding of AI technologies, algorithms, and their applications.
- Familiarity with Data Privacy and GDPR.
- Strong research skills in legal domains are a plus, in order to analyze relevant legislation, regulations, and case law concerning AI and data privacy.
- Proficiency in English.

E. Benchmarking and Evaluation of Prompt Engineering Methods with LLMs

Introduction:

Prompt engineering has emerged as a crucial aspect of Language Learning Models (LLMs). It involves the development of effective prompts that can influence the performance of a model. However, a significant gap exists in the evaluation and benchmarking of various prompt engineering methods. This proposal aims to develop benchmark infrastructures and evaluation metrics for assessing the effectiveness of different prompt engineering methods, but, most importantly, to contribute to the field of LLMs by providing a much-needed evaluation framework for prompt engineering methods.

Problem Statement:

The field of prompt engineering for LLMs lacks a standardized approach to evaluate and compare different methods. While several prompt engineering techniques have been proposed and implemented, their comparative effectiveness remains unclear. This lack of clarity can hinder the optimization of LLMs and limit their performance. There is a pressing need to understand the impact of these techniques on LLM performance, and a robust benchmarking and evaluation infrastructure would greatly facilitate this understanding.

Objective:

The study aims to provide a comprehensive evaluation and benchmarking framework for prompt engineering methods in LLMs. This research will address the following objectives:

- Develop a benchmark infrastructure that can accommodate different prompt engineering methods.
- Formulate evaluation metrics that accurately measure the effectiveness of these methods.
- Compare and analyze the impact of various prompt engineering techniques on LLM performance.
- Provide insights and recommendations for optimal prompt engineering practices in LLMs.

The research will involve a thorough review of the literature to identify existing prompt engineering methods and their evaluation techniques. This will be followed by the development of a benchmark infrastructure that can host these methods. The evaluation metrics will be formulated based on factors influencing LLM performance. Comparative analysis will be conducted using statistical methods to understand the impact of different techniques.

Required Skills:

- Willingness to understand and study Language Learning Models and prompt engineering techniques.
- Decent experience with programming languages, preferably Python, for developing the benchmark infrastructure.
- Ability to design and implement effective evaluation metrics.
- Statistical analysis skills for comparing and interpreting the results.
- Proficiency in English for constructing well defined prompts.

F. Mobility/localization-based clustering for mobile network applications

Problem statement:

The ever-growing demand for location-based services has increased the need for advanced analytics that exploit positioning and mobility/flow monitoring information for various use cases. In the case of mobile networks, this information can be exploited to inspect, visualise the actual network KPIs creating key insights, as well as for network planning and optimization. In this context, by understanding crowd movement, people flows, identifying trajectories and traffic profiles (i.e. pedestrian vs. car, high/low velocity etc.), a network operator can detect areas/points of interest, problematic areas or key areas that display some uniques and/or common KPI profiles.
Representative such work can be found in [1].

Proposed Solution:

The proposed thesis targets the improvement of clustering areas using such information combined also with network KPI data (related to coverage and quality) and improve upon existing work by the company, as indicatively presented in [2], [3].
The dataset will include real-world network operator data of a large scale area coupled with geolocation information (latitude, longitude, user device identifier, timestamp) of several user equipment devices for the selected area of analysis. Access to inhouse simulator can also be considered for result comparison/validation purposes.
It is expected that the candidate will apply selected deep learning techniques that will fit the use case. Indicatively -and not limited to- trajectory modelling and clustering using RNNs with LSTMs/GRUs, sequence to sequence (Seq2Seq) auto-encoders and variants, coupled with standard clustering techniques.

Tech stack/ requirements: Python programming skills, Strong SQL knowledge, knowledge/ experience in ML libraries (e.g. scikit) and at least one python deep learning framework (TensorFlow, Keras, PyTorch)
Knowledge of code version control via Git is considered a plus
Knowledge of mobile network KPIs not mandatory, but considered a plus

Deliverables:

- Study of algorithmic approach, research of previous relevant work
- Algorithm implementation/ code (prototyping level)
- the result should be demonstrated with large number of real-world and/or simulated trajectories in the context of the telecom sector.

References:

[1] LOCUS Project, https://www.locus-project.eu/

[2] Y. Filippas, A. Margaris and K. Tsagkaris, "Deep Learning Approaches for Mobile Trajectory Prediction,"*2021 IEEE Globecom Workshops (GC Wkshps)*, Madrid, Spain, 2021, pp. 1-6, doi: 10.1109/GCWkshps52748.2021.9682164.

[3] A. Margaris, I. Filippas, K. Tsagkaris, Hybrid Network–Spatial Clustering for Optimizing 5G Mobile Networks. Appl. Sci. 2022, 12, 1203. https://doi.org/10.3390/app12031203

G. Improving localization accuracy using mobility information for mobile network applications

Problem statement:

The ever-growing demand for location-based services has increased the need for advanced analytics that exploit positioning and mobility/flow monitoring information for various use cases. In the case of mobile networks, network planning and optimization can be based on such analytics.

However, low localization accuracy whether global navigation satellite system (GNSS)-based or ML/AI-based can lead to misleading insights and subsequent bad decisions. Related work on positioning and relevant analytics can be found in [1].

Proposed Solution:

The proposed thesis targets the improvement of localization accuracy by understanding crowd movement/ people flows, identifying trajectories and traffic profiles (i.e. pedestrian vs. car, high/low velocity etc.).

In this context, a localization result can be improved based on the knowledge extracted and this network optimization decisions to be more reliable. The solution will build upon in-house knowledge of such cases and existing published and unpublished work, as indicatively presented in [2], [3].

The dataset will include real-world network operator data of a large scale area coupled with geolocation information (latitude, longitude, user device identifier, timestamp) of several user equipment devices for the selected area of analysis, but also access to an in-house simulator can be given for results comparison and validation purposes.

It is expected that the candidate will apply selected deep learning techniques that will fit the use case. Indicatively -and not limited to- trajectory modelling using RNNs with LSTMs/GRUs, etc.

Tech stack/ requirements: Python programming skills, Strong SQL knowledge, knowledge/ experience in ML libraries (e.g. scikit) and at least one python deep learning framework (TensorFlow, Keras, PyTorch)

Knowledge of code version control via Git is considered a plus

All documents should be written in English

Deliverables:

- Study of algorithmic approach, research of previous relevant work
- Algorithm implementation/ code (prototyping level)
- the result should be demonstrated with large number of real-world and/or simulated trajectories in the context of the telecom sector.

References:

[1] LOCUS Project, https://www.locus-project.eu/

[2] Y. Filippas, A. Margaris and K. Tsagkaris, "Deep Learning Approaches for Mobile Trajectory Prediction,"*2021 IEEE Globecom Workshops (GC Wkshps)*, Madrid, Spain, 2021, pp. 1-6, doi: 10.1109/GCWkshps52748.2021.9682164.

[3] A. Margaris, I. Filippas, K. Tsagkaris, Hybrid Network–Spatial Clustering for Optimizing 5G Mobile Networks. Appl. Sci. 2022, 12, 1203. https://doi.org/10.3390/app12031203

H. Agentic Chunking: Empowering LLMs for text segmentation

Introduction & Problem Statement:

Large Language Models (LLMs) currently lead the charge in technological innovation. Their remarkable proficiency in understanding and generating language has facilitated the realization of numerous groundbreaking applications in computer science. From text-to-code conversion to document summarization, and from virtual assistants to language translation, LLMs have paved the way for unprecedented advancements. However, despite their remarkable abilities, these powerful tools face certain limitations, such as the substantial demand for computational resources and constraints on query length. These challenges are particularly evident in Question-Answering systems, where accessing extensive external information sources can prove daunting. In response, the architecture of Retrieval-Augmented Generation (RAG) has been proposed, aiming to bridge the gap between LLMs and external data sources. Nevertheless, despite significant progress, critical questions and areas for improvement persist.

Notably, the effective segmentation of large documents into smaller, manageable chunks remains a crucial aspect for enhancing information retrieval capabilities. In particular, the absence of efficient chunking in information retrieval often results in the loss of valuable content relevant to the query at hand. To tackle this issue, various chunking strategies have been put forth, utilizing techniques to measure sentence similarity or identify text structure, resulting in chunks that frequently lack crucial information mainly due to their reliance on rule-based generation.

Objective:

The aim of this project is to address information loss during retrieval by introducing Agentic chunking strategies based on LLMs. Instead of strictly following rules or metrics for text chunking, agents utilizing LLMs will analyze the document and generate chunks that resemble human perception. To achieve this, we will utilize text summarization and recursive techniques to develop an efficient and cost-effective chunking strategy based on LLMs. The specific goals of the project are outlined as follows:

1) Benchmark existing text chunking strategies
2) Evaluate the text summarization and chunking capabilities of state-of-the-art LLM models
3) Develop and implement various LLM-based agentic chunking strategies
4) Conduct extensive evaluations of the implemented strategies using several LLM models and their quantized versions
5) Assess the impact of fine-tuning the LLM for the specific task
6) Offer insights and valuable feedback on the different agentic chunking strategies and compare their performance against conventional chunking methods

Required Skills:

- Understanding of Natural Language Processing (NLP), Generative Artificial Intelligence and Machine Learning (ML) principles
- Expert in Python

- Familiarity with Deep Learning frameworks such as TensorFlow or PyTorch, as well as HuggingFace
- Nice to have experience with model quantization, fine-tuning, and knowledge distillation
- Capability to analyze and interpret evaluation metrics relevant to NLP tasks
- Effective communication skills to present research findings and recommendations clearly and concisely

I. Code generation in repository-level though LLMs

Introduction & Problem Statement:
The rise of Large Language Models (LLMs) has marked a significant shift in the technological landscape, as their capacity to translate natural language into code facilitates direct communication between humans and computers. This specialized category of LLMs, known as Code Language Models (CLMs), is already integrated into numerous products, aiding programmers in their daily tasks, including text-to-code conversion, code cleaning and documentation, code translation, code explanation, and more. However, certain tasks still rely on manual intervention, primarily due to the immense size of the code that CLMs must process to fulfill user requests. Particularly, requests involving code parsing at the repository level demand CLMs with extensive context windows and substantial computational resources. To circumvent parsing the entire repository, leveraging additional information about the repository's structure can help pinpoint the location of interest, thereby guiding the CLM to where the new code should be placed. However, such structural information is not always readily available or sufficient, necessitating the generation of this information for each request.

Objective:
The objective of this project is to identify code files within a repository that are relevant to a specific coding task. Since parsing the entire repository beforehand to understand its structure is often insufficient or infeasible, the developed solution must identify the relevant code files based on factors such as documentation (if available), library imports or function calls, variable names, and more. The specific goals of the project are outlined as follows:

1) Conduct a comprehensive literature review and benchmark existing techniques for code generation at the repository level
2) Evaluate the performance of existing data structures for efficiently representing the structure of a code repository
3) Develop and implement an LLM-based solution for identifying the corresponding code files for each request
4) Integrate state-of-the-art CLM models into the existing system and benchmark their performance in coding tasks at the repository level
5) Fine-tune the LLM to enhance its performance for the specific task
6) Provide insights and valuable feedback on the usage of LLMs and CLMs in repository-level coding

Required Skills:

- Understanding of Natural Language Processing (NLP), Generative Artificial Intelligence, and Machine Learning (ML) principles
- Proficiency in Python programming
- Experience with Deep Learning frameworks like TensorFlow or PyTorch, as well as HuggingFace
- Understanding of fundamental software engineering principles
- Nice to have experience with LLM fine-tuning
- Capable of analyzing and interpreting evaluation metrics relevant to NLP tasks
- Excellent communication skills to present research findings and recommendations clearly and concisely