PhD Course in Computer Science
XXXVII cycle, a.y. 2021/2022

n.1 scholarship funded by Istituto Consiglio Nazionale delle Ricerche (CNR-ISTI) linked to the project:*"Deep learning techniques for next generation context-aware recommendation and search systems "*.

Abstract

The PhD position will focus on novel machine/deep learning techniques for Information Retrieval (IR) and Recommender Systems (RS). As a PhD candidate you will perform cutting edge research in the field of deep learning techniques for next generation context-aware search and recommendation. The research will primarily focus on models for representing user preferences and contextual information and on practical neural IR/RS solutions exploiting these models effectively.

n.1 scholarship funded by ROCHE S.p.A. linked to research project:*"Structural causal models, static and dynamic causal networks for health and medicine".*

Abstract

An increasing availability of data in medicine and healthcare is demanding for effective and efficient methods to extract knowledge which can then be used for decision making under uncertainty. Artificial intelligence, machine learning and deep learning are developing different models to this extent. However, it is currently understood that decision making is not the same as using prediction to make decisions. Indeed, causal networks and structural causal models, together with the potential outcome framework are receiving increasing attention to make causal statements combining domain knowledge, whenever available, and data. Electronic health records, administrative data, geographic information, together with environmental and lifestyle data are collected every day, for hundred of thousands of patients. In particular, every data type originates from a specific source and to extract knowledge controlling for confounding effects is an extremely challenging task. The way in which data originated from different sources are linked and joined together strongly affect the quality of the knowledge which can be extracted by machine learning algorithms.
This project aims to formulate and solve the following problems:
1) Causal discovery; discovery of the causal structure of the data generating process from data originated from different data sources, static and/or temporal,
2) Decision making under uncertainty; develop mathematical models for decision making under uncertainty when the causal structure of the data generating process is learnt from data.
In particular, the research project will study, analyze and develop algorithms for structural learning of causal networks and structural causal models. Different setting will be taken into consideration; static, i.e., when the data generating process can be studied without taking into account time or sequence, temporal, i.e., the case where patients data span a given period of time where multiple observation are available. Furthermore, causal discovery will also concentrate the attention to setting condition where the causal sufficiency or no unmeasured confounders hypothesis applies, as well as when such a too strong assumption is relaxed and it is assumed that some unmeasured quantity affect the data generating process to be studied.

<u>High level training apprenticeship contracts</u>

n. 1 contract with Spazio Dati s.r.l.

Salary for apprendice:
25.183,20 gross per month, level 3 CCNL Commerce

Profile
Computer Science, Artificial Intelligence, Knowledge Representation

Title
Cracking the Long-tail Entity Barrier in the Development of Domain-specific Knowledge Graphs

Project
Natural Language Processing for Knowledge Base Construction

Building knowledge graphs from structured and unstructured sources is crucial to feed valuable data services and improve several downstream applications. This is the case at SpazioDati where a large amount of data about companies and related entities (persons, locations, property, etc.) are integrated into a knowledge graph that supports several applications (eg. identification of entity mentions in texts such as news, web pages, documents etc.., network analysis, entity similarity and more), including also those techniques applied to continuously update the graph and make it grow. Critical components of such services are entity linking services to disambiguate and annotate entity mentions in news and other textual sources, as well as representation learning models developed on top of the entities and textual documentation. While state ofthe-art techniques to build knowledge graphs deal well with popular entities, long-tail entities, e.g., littleknown companies and persons, which have a crucial business value, have highly ambiguous references and are associated with limited information, which makes their processing challenging. This project aims at advancing techniques to extract, link, integrate, and learn representations of long-tail entities to build domain-specific knowledge graphs effectively and incrementally, with vertical application on companyrelated information. The techniques defined by the candidate must consider constraints coming from the industry context and at the same time be developed and evaluated according to shared scientific methodologies so as to combine business impact and scientific relevance. Specific topics addressed in the thesis include:
• Methodologies to evaluate the performance of information extraction algorithms (entity extraction, typing and linking, focused relation extraction) for long-tail entities
• Methodologies to improve the performance of information extraction algorithms (entity extraction, typing and linking, focused relation extraction) for long-tail entities, possibly with humans-in-theloop
• Neural representation learning models to learn high-quality distributed representations of long-tail entities in knowledge graphs built from structured and unstructured data

n. 1 contract with Datasinc s.r.l.

Salary for apprendice:
40 weekly hours, 20.000€ gross

Title
The project aims at researching Natural Language Processing (NLP) and Computer Vision (CV) applied to the fields of Bank credit, legal and Real Estate (namely Fintech, Legaltech, Proptech and Insurtech).

Abstract

The challenges to overcome are many and as an example:

- Merge CV and NLP to better classify documents and upgrade the current OCR techniques / technologies;

- Explore new techniques of Data Linking and Information Extraction to create new SOTA models leveraging on the specific content applied to the domain;

- Create new semantic search engines able to accelerate the "digital transformation" of players within the specific fields of business both in terms of coherence of the documentation base and the retrieval of correlated available information

Posizioni con Percorso Executive

n.1 posizione riservata ai dipendenti di Noovle S.r.l. a Socio Unico Noovle S.p.A.

Progetto

The objective of the research is to analyze and deepen conversational AI systems by extending the different techniques proposed in recent years in the state of the art. In particular, we will study how these AI systems automate and scale consumer interactions not only on messaging channels but also on the different platforms that provide certain interaction flows between customer and service.
The main interest in this research area comes from the strong need that companies exhibit in recent years in the contact center context, as they are facing increasing pressure to achieve complex management goals with few resources.
In fact, Conversational AI systems are designed to provide a solution to this issue. It is expected that 20% of customer interactions will be handled by conversational AI agents by 2022.
The main issues that will be addressed are assisting customers, assisting agents and unlocking knowledge

Titolo/ Argomento della Tesi
Conversational AI far customer services

n. 3 posizioni riservate ai dipendenti presso Social Things srl

Posizione 1:

Progetto

The purpose of this research is to analyze and apply nlp techniques for automated creation of chatbots used in enterprise customer services. A fundamental part of the process of creating a chatbot is to define what the customer is looking for when interrogating a company's chatbot. The main idea is to introduce in this phase an automatic tool that allows to extract from any dataset of requests to the cs the possible intents (topic) that the customer can ask. The problem is then identified as a clustering one. Thus, in NLP, this is usually divided into three steps: Embedding, Dimensionality Reduction and Clustering. For each of these steps, in literature there are different methods of implementation, which are more or less appropriate depending on the input dataset. The objective is therefore to define a methodology that allows to automate the selection of methods used, in order to obtain the best possible result of text division. Thus, our intermediate goals are:

Year 1:
1 month: study of the state of the art for embedding techniques, dimensionality reduction, clustering, and goodness-of-fit evaluation of unsupervised clustering.
2 months: definition of a clustering pipeline that fits the characteristics of the input dataset.

4 months: development of a prototype.
1 month: optimization of the prototype.
2 months: study of the efficiency of the prototype according to the variation of the input dataset

Year 2:
4 months: optimization of internal algorithms through parameter tuning methodologies.
4 months: application of innovative methods and/or development of new methodologies for the three steps of the process.
2 months: study of automation of pipeline method selection as the input dataset changes.

Year 3:
4 months: system experimentation
4 months: final algorithm development
2 months: final algorithm evaluation
1 months: final algorithm evaluation

Titolo/Argomento della Tesi
Studies and Application of NLP techniques for text classification

Posizione 2:

Progetto
The aim of the project is to analyse and develop new techniques for the evaluation of the explanability of intelligent models, starting from the currently existing methodologies for the evaluation of User Experience. This research is necessary to improve user awareness and efficiency in the use of these systems, making them more usable to users who use them in an unconscious way. It happens frequently that non-expert users see these systems as a sort of black box that provides and recommends to them what they actually need but without fully understanding the mechanisms. This conception of the model leads to the inefficiency of the model itself. Therefore, by studying the state of the art of XAI models and evaluation methodologies we want to analyse the approaches used till now and improve them through the creation of new evaluation metrics. In particular, the project aim is to evaluate the usability of the system for different types of users who use XAI systems: AI Experts, Data Experts and AI Novices. The evaluation methods considered to evaluate the interpretability and debugging of the model concern the evaluation of: Model Performance, Explainer Fidelity, and Model Trustworthiness. The evaluation methods considered to evaluate algorithm transparency, user trust towards the system, BIAS mitigation and privacy awareness concern the evaluation of: user mental model, usefulness and satisfaction, user trust and reliance towards the system and Task Performance. The case study analysed concerns the Explainability of recommendation systems, with application to the recommendation system used by the e-learning platform WhoTeach for the creation of online courses.
Our intermediate goals are therefore:

Year 1:
2 months: study of the state of the art of XAI systems and evaluation methods for these systems
1 month: Identification of evaluation methods for User Experience of explainability of recommendation systems.
3 months: Creation of methods for evaluating the User Experience of the explainability of recommendation systems.
3 months: study and application of the methods of evaluation on a real system through simulated experimentation and on real users

Year 2:

1 month: Analysis of the obtained results
2 months: Optimization of the proposed methodology
3 months: Study and application of the evaluation methods on a real system through simulated experimentation and on real users.
1 month: Analysis of the results obtained and optimization of the methodology.
2 months: Optimization of the proposed methodology

Year 3:
3 months: validation of the proposed methodology.
3 months: Verification of the goodness of the proposed methodology
3 months comparison with other evaluation methodologies currently in use.

Titolo/Argomento della Tesi
Analysis and development of methodologies for the evaluation of the explanability of intelligent models

Posizione 3:

Progetto

The objective of this research is to investigate, and study rule-based recommendation systems with the aim of developing a natively explainable system. In recent years, users have been faced with complex decision problems in choosing a particular item that represents their interest. The complexity of the problem relates to the enormous amount of information available whose interpretation is not always immediate for a user. Therefore, the latest research in this domain tries to optimize the decision phase that each user must face. To obtain the best possible results, attempts are being made to go beyond evaluation metrics alone, but increasingly to use the concept of explainability of a model. This concept is always present in structures based on decision trees at the basis of the system to be implemented. It makes it possible to increase the user's knowledge of the motivation that led to the choices and decisions made by the recommendation system and thus makes the user aware of the choices made. A user faced with a recommendation system that provides this motivation is more likely to reuse it. In the research path we will try to study systems based on decision trees to learn the best possible structures to implement the desired software. Current recommendation systems will be analyzed to check for innovative and efficient techniques and the optimization that can be applied will be extrapolated from them. In detail, the analysis will focus on models with a primary structure of the decision tree.

Titolo/Argomento della Tesi
Reimplementation and optimization of a rule-based recommendation system

n. 1 posizione riservata ai dipendenti di X- Next

Profilo
Computer Science, Machine Learning

Progetto

High Flow Anomaly Detection

Progettazione, sviluppo e validazione di algoritmi di machine learning per l'Identificazione di anomalie a partire da detector di raggi X in condizioni di flusso elevato. Le tecniche di imaging mediante raggi X trovano

numerose applicazioni in ambito industriale, medicale e nel campo della sicurezza. Tradizionalmente queste tecniche si basano su sistemi di rivelazione a energia integrata (energy integration detector, EID), che misurano solamente l'energia totale del fascio policromatico di fotoni che incide sul detector. Negli ultimi anni si sta sviluppando una nuova tipologia di rivelatori a semiconduttore (photo-counting detector, PCD) in grado di discriminare i singoli fotoni del fascio. Questi ultimi, interagendo con il materiale semiconduttore, generano al suo interno una carica elettrica proporzionale alla loro energia. Dalla lettura dell'impulso elettrico prodotto da ciascuno di questi eventi di interazione risulta perciò possibile risolvere energeticamente ogni fotone. L'azienda Xnext sfrutta questa tecnologia e l'informazione aggiuntiva che restituisce per effettuare analisi chimico-fisiche dei materiali scansionati mediante raggi x.
Tuttavia, le prestazioni di PCD sono condizionate da fattori strutturali e da fenomeni fisici che possono degradare il segnale acquisito. Tra questi fenomeni, il pulse pileup risulta essere uno dei più critici. Esso consiste nella sovrapposizione di impulsi elettrici dovuta all'interazione quasi simultanea di due o più fotoni con il detector. Il rivelatore, per limiti fisici intrinseci, genera degli impulsi di dimensione temporale finita, nell'ordine delle centinaia di nanosecondi. Nel caso in cui due o più fotoni interagiscano con il semiconduttore entro questo intervallo di tempo, non verranno adeguatamente risolti in numero ed energia dal sistema senza un modello in grado di descrivere il fenomeno. Considerato che il pulse pileup risulta tanto più rilevante quanto maggiore è il flusso di fotoni incidenti sul detector e che un alto flusso è una condizione necessaria per il raggiungimento di un numero di conteggi adeguato in sistemi che lavorano in tempo reale su linee di produzione industriali, questo fenomeno risulta limitante. Questo progetto di tesi si propone di recuperare l'informazione riguardante il numero e l'energia dei fotoni incidenti sul detector a partire dal segnale distorto. Attualmente esistono algoritmi in grado di scartare eventi sovrapposti, in modo tale da salvaguardare l'informazione energetica dei fotoni incidenti, a forte scapito tuttavia del numero di conteggi misurati. Altri algoritmi si basano su modelli analitici della distorsione da pulse pileup, che sfruttano le informazioni riguardanti la modalità di risposta del detector, ma risultano poco applicabili real-time a causa della loro complessità temporale. La forte diffusione dei metodi di Machine Learnir:ig e i notevoli risultati che la loro applicazione ha portato in numerosi campi, tra i quali il processamento di segnali, stanno facendo guadagnare interesse a queste tecniche che iniziano a essere utilizzate anche per la risoluzione della problematica del pulse pileup. Nel caso specifico di Xnext, approcci data-driven risultano essere particolarmente vantaggiosi grazie alla grande disponibilità di dati non solo reali ma anche sintetici, realizzati attraverso un simulatore che è stato sviluppato internamente. In questo lavoro di tesi si vogliono investigare diverse tecniche di deep learning (tra le quali fully connected networks, convolutional neural networks e recurrent neural networks), tecniche di ensemble e altri metodi supervisionati (come gli alberi decisionali o le sum-product networks) applicati ai segnali prodotti dalla tecnologia di Xnext. Il progetto di ricerca si pone l'obiettivo di individuare una o più soluzioni che siano in grado di mitigare l'effetto pileup precedentemente descritto, dando riscontro di un effettivo miglioramento del segnale attraverso una validazione su dati simulati. Oltre a ciò, i modelli risultati efficaci verranno valutati anche sulla base della loro efficienza in termini di tempi di inferenza, tenendo in considerazione una loro possibile implementazione nel sistema di acquisizione reale di Xnext, con le restrizioni in termini di hardware e operatività real-time che esso implica.

Titolo/ Argomento della Tesi
Algoritmi di machine learning per la ricostruzione di segnali da detector di raggi x in condizioni di alto flusso