Relazione attività di ricerca

La prima parte del mio dottorato di ricerca si è svolto all'interno del gruppo di sistemi complessi (PhySyCom) dell'Università di Bologna, guidato dal prof. Bazzani. Con il gruppo PhySyCom ho potuto approfondire le mie competenze algoritmiche e ho studiato approfonditamente le potenzialità dei linguaggi di programmazione a basso livello come il C e il C++. Le mie competenze di programmazione sono state formate anche grazie alla partecipazione a *Eighth I.N.F.N. International School on architectures, tools and methodologies for developing efficient large scale scientific computing applications* ed al workshop *Intel-Code Modernization Workshop Rome*, durante i quali ho potuto studiare le tecniche di calcolo parallelo e distribuito.

Inizialmente mi sono occupato principalmente di analisi su dati di traffico automobilistico, forniti dalla regione Emilia Romagna, e su modelli a network, per l'ottimizzazione della rete peritale italiana di Unipol Assicurazioni, la quale ha finanziato il mio primo anno di dottorato. Per quanto riguarda i dati di traffico, l'obiettivo del lavoro era quello di riuscire a predire, con sufficiente anticipo, l'insorgenza di ingorghi stradali e quindi procedere a dare "l'allarme" agli automobilisti. Con questo progetto ho iniziato ad utilizzare e sviluppare i primi algoritmi di deep learning per l'analisi dati, i quali sono stati applicati fornendo buoni risultati, sui quali stiamo scrivendo un articolo. Questi risultati sono stati presentati alla conferenza *Problems in discrete dynamics: from biochemical systems to rare events, networks, clustering and related topics - II Edition* del 2017. Il lavoro svolto con Unipol Assicurazioni, invece, pur avendo portato a buoni risultati non si è potuto concretizzare in una pubblicazione, vista la sensibilità dei dati coinvolti. Nel mentre, in collaborazione con il centro INFN-CNAF, ho lavorato sull'ottimizzazione di una pipeline di bioinformatica, studiandone le performance di calcolo in termini di occupazione di memoria, tempistiche e consumi energetici. Questo lavoro mi ha permesso di partecipare alla conferenza *EuroPar2018* in cui ho presentato il mio lavoro e relativo articolo (*Cross-Environment comparison of a bioinformatics pipeline: perspectives for hybrid computations* [2].

A cavallo tra il primo e il secondo anno di dottorato ho iniziato a spostare la mia attività di ricerca su argomenti più attinenti al campo della biofisica e teoria dei network, lavorando con il gruppo Biophys dell'Università di Bologna, guidato dal prof. Castellani e dal prof. Remondini. In particolare, ho iniziato a lavorare sulla ricostruzione della struttura 3D delle proteine a partire dalla relativa mappa di contatto. Questo problema è stato da me affrontato mediante lo studio della matrice laplaciana associata al network e la relativa analisi spettrale. Utilizzando gli autovettori della matrice laplaciana come coordinate degli amminoacidi, ho potuto ottenere buoni risultati, i quali però sono ancora in fase di revisione e non si sono ancora concretizzati in una pubblicazione.

Durante il secondo anno la mia attività di ricerca è stata finanziata dal progetto *Venice*, svolto in collaborazione con il comune di Venezia, Canon Inc., Telecom Italia e Fabbrica Digitale. Il progetto si proponeva di studiare il flusso pedonale all'interno della città di Venezia. A questo scopo ho potuto ulteriormente approfondire le mie conoscenze sui modelli a rete neurale per l'object detection e studiare ulteriori tecniche di programmazione parallela. In particolare, ho studiato e sviluppato reti feed forward a grande dimensionalità

(es. Yolo, ImageNet, ResNet152, ...) e modelli a spin glass applicati a reti neurali (rFBP e rSGD). Oltre alle basi teoriche di questi algoritmi, buona parte del mio lavoro è stato speso nell'implementare ed ottimizzare questi metodi mediante tecniche di calcolo parallelo multi-threading (OMP) e distribuito (MPI). Per alcune applicazioni si è resa necessaria l'implementazione di kernel in CUDA e OpenCL per sfruttare acceleratori GPU.

Utilizzando una di queste reti a deep learning, implementata su schede Jetson TX2 con annesse telecamere, siamo riusciti a monitorare in tempo reale il flusso pedonale della città di Venezia, andando a contare (con anche informazioni direzionali sul moto) le persone all'interno di sequenze video. In questo progetto mi sono occupato anche della gestione delle telecamere e della manutenzione del software da remoto, implementando pipeline che fossero in grado di eseguire questi compiti autonomamente. Questo progetto è tuttora in corso, ma il mio contributo si è fermato all'analisi dei flussi pedonali ottenuti attraverso i dati di telefonia mobile. In questo lavoro ho applicato una versione modificata dell'algoritmo per l'analisi di livelli di espressione genica (DNetPRO), sviluppato durante il mio lavoro di tesi magistrale, per la ricostruzione dei percorsi pedonali all'interno della città, ottenendo ottimi risultati. Le analisi svolte si sono concretizzate nell'articolo *Unraveling pedestrian mobility on a road network using ICTs data during great tourist events* [6].

Ho applicato l'algoritmo DNetPRO anche ad altre serie di dati biologici forniti da collaborazioni esterne (prof.ssa Minozzi con dati di RNASeq e prof.ssa Boccardi con profili di citochine di pazienti affetti da Alzheimer). Entrambi i lavori hanno dimostrato l'efficienza dell'algoritmo per l'estrazione di signature biologiche efficienti sia da un punto di vista statistico che biologico e si sono concretizzati in due pubblicazioni (*Combinatorial Discriminant Analysis applied to RNAseq data reveals a set of 10 transcripts as signatures of infection of cattle with Mycobacterium avium subsp. paratuberculosis* [5] e *Cognitive decline and Alzheimer's disease in the old age: sex influence on a "cytokinome signature"* [1]).

Ho applicato anche i modelli a spin glass utilizzati per le analisi sui dati di traffico (Replicated Focusing Belief Propagation) a dati di mutazioni del genoma (SNPs) forniti dal progetto COMPARE. Il lavoro ha fornito validi risultati che sono stati presentati alla conferenza CCS 2019 nel lavoro *Classification of Genome Wide Association data by Belief Propagation Neural network* [4]. Stiamo ancora lavorando alla loro pubblicazione e alla stesura del relativo articolo.

Alla stessa conferenza è stato presentato anche un mio secondo lavoro (CHIMeRA) (*Introducing the Complex Human Interactions in MEdical Records and Atlases Network - CHIMERA*) che ha coinvolto parte del mio terzo anno di dottorato. Una versione aggiornata del medesimo lavoro è stata presentata alla conferenza BioPhys & PlexNet 2019.

La mia attività di ricerca di questo ultimo anno è stata finanziata dal progetto INFN FiloBlu, svolto in collaborazione con l'Università Sapienza di Roma, l'ospedale Sant'Andrea di Roma e altri partner esterni. Il progetto si proponeva l'analisi di dialoghi medico-paziente ottenuti da chat mediche. Le conversazioni vengono raccolte attraverso un'APP e i messaggi testuali sono quindi raccolti in un database centrale. Mediante tecniche di natural language processing ed algoritmi di Machine Learning è possibile attribuire uno score a ciascun messaggio (mediante tecniche di "Sentiment Analysis"). Lo scopo è fornire un ranking ai messaggi dei pazienti, in modo da garantire al medico un'immediata visualizzazione di quelli più urgenti. Il mio contributo al progetto ha riguardato sia la realizzazione di un'ontologia dei termini medici relativi alle malattie in lingua italiana (che ha portato alla realizzazione di SymptomsNet, discusso nella tesi), che alla creazione di una pipeline che gestisse automaticamente la comunicazione tra le APP e il database centrale in cui vengono salvati e processati i messaggi. Nel dettaglio, ho realizzato un servizio che permettesse la lettura dei dati da un database centrale e fosse in grado di processarli in real-time al fine di fornire un responso immediato all'APP che regola la chat. Le tecniche di natural language processing studiate e sviluppate nel progetto FiloBlu sono state poi

sfruttate per la creazione del (sopra citato) multi-network CHIMeRA (Complex Human Interaction of Medical Records and Atlases).

Il progetto CHIMeRA si propone di unire in un'unica struttura a network diverse sorgenti di dati bio-medici (es. DisGeNet, i.e interazione tra geni e malattie, DrugBank, i.e interazione tra malattie e farmaci, HMDB, i.e interazione tra malattie e metaboliti, RxList, i.e interazione tra sintomi e malattie, etc.). I dati sono stati reperiti da sorgenti open-source online ed estratti mediante tecniche di web-scraping da siti pubblici. Gran parte del lavoro ha coinvolto l'organizzazione e la gestione di questi dati, uniformando le varie sorgenti di dati testuali per massimizzarne l'overlap. Il grafo ottenuto coinvolge più di 3.6×10^5 nodi (composti da 7 tipologie di nodo) e 3.8×10^7 links ed è quindi indispensabile utilizzare metodi ottimizzati per la sua elaborazione ed analisi. A tal fine l'intera struttura è stata convertita in un graph database utilizzando ArangoDB ed è stata predisposta una web-page per la realizzazione di query che permettano l'estrazione di porzioni di interesse. Le prime analisi svolte sul database CHIMeRA hanno riguardato la Acute Myeloid Leukemia in collaborazione con il progetto europeo HARMONY. Le analisi sono ancora in fase di sviluppo e questo progetto verrà portato avanti da me anche al seguito del dottorato.

La gran parte del terzo anno l'ho spesa sull'implementazione di diverse librerie di calcolo parallelo per lo sviluppo e la prototipizzazione di reti neurali. In particolare ho sviluppato la libreria Byron, interamente scritta in linguaggio C++, e NumPyNet, in Python. Queste librerie sono state utilizzate in diversi progetti di tesi magistrale da studenti che ho seguito e sono state da me applicate per l'analisi di immagini mediche. In particolare mi sono concentrato sulla realizzazione di reti per Super Resolution, Object Detection (seguendo quanto fatto sul progetto *Venice*) e Image Segmentation.

Nel dettaglio, ho ottimizzato due reti per la super risoluzione e le ho applicate ad immagini di risonanza magnetica (MRI) cerebrale. Non è stato possibile effettuare un addestramento dedicato per questa tipologia di immagini per motivi di tempo e reperibilità dei dati e per questo è stata utilizzato un addestramento su immagini non di tipo biomedico: ciò nonostante la qualità delle immagini ottenute (quantificata dai parametri standard della super risoluzione come PSNR e SSIM) è risultata più che soddisfacente ed è tuttora oggetto di studio. Inoltre, è stato possibile testare l'efficienza della rete YOLO per l'object detection (ottimizzata all'interno della libreria Byron): è stato dimostrato un aumento di oltre il 70% nell'individuazione di piccoli oggetti e soprattutto di persone all'interno di folle. Un primo prototipo di rete basato sull'architettura U-Net è stato sviluppato per la segmentazione di immagini TAC del femore. Le immagini sono state pre-processate ed annotate mediante una pipeline semi-automatica di image processing ed i risultati preliminari hanno dimostrato l'efficienza di questo modello per la segmentazione di immagini mediche. I risultati ottenuti da queste applicazioni sono ampiamente descritte all'interno del mio progetto di tesi e saranno argomento delle mie prossime pubblicazioni.

Durante tutti e tre gli anni ho portato avanti il mio progetto di tesi magistrale basato sull'algoritmo DNetPRO (il quale mi ha anche permesso di vincere il premio INFN Giulia Vita Finzi). Tale algoritmo è stato argomento anche di un terzo articolo (oltre a quelli sopra citati) che al momento è stato caricato su BioRxiv [3] e costituisce la base della prima parte del mio progetto di tesi.

Oltre a questi progetti, durante questi anni, ho collaborato a diversi altri lavori e soprattutto ho sviluppato ed ottimizzato diversi algoritmi. Non tutti hanno già trovato un'applicazione nei progetti in cui il gruppo è stato coinvolto, ma sono tutti pubblicati sulla mia pagina Github. Ogni codice è provvisto di una documentazione sull'installazione e di una pipeline di continuous integration che ne garantisca l'utilizzo su diversi sistemi operativi e versioni del software (https://github.com/Nico-Curti/).

Bibliography

[1] V. Boccardi, L. Paolacci, D. Remondini, E. Giampieri, G. Poli, N. Curti, R. Cecchetti, A. Villa, C. Ruggiero, S. Brancorsini, and P. Mecocci. Cognitive decline and alzheimer's disease in the old age: identified a sex specific "cytokinome signature". Oct 2019.

[2] N. Curti, E. Giampieri, A. Ferraro, C. Vistoli, E. Ronchieri, D. Cesini, B. Martelli, C. Duma Doina, and G. Castellani. Cross-environment comparison of a bioinformatics pipeline: Perspectives for hybrid computations. *Springer, Cham, Euro-Par 2018: Parallel Processing Workshops*, 11339, 2019.

[3] N. Curti, E. Giampieri, G. Levi, G. Castellani, and D. Remondini. Dnetpro: A network approach for low-dimensional signatures from high-throughput data. *bioRxiv*, 2019.

[4] D. Dall'Olio, N. Curti, G. Castellani, A. Bazzani, and D. Remondini. C++ implementation, optimization and application of the focusing belief propagation algorithm, 2019.

[5] M. Malvisi, N. Curti, D. Remondini, G. Gandini, F. Palazzo, G. Pagnacco, J. L. Williams, and G. Minozzi. Combinatorial discriminant analysis applied to rnaseq data reveals a set of 10 transcripts as signatures of infection of cattle with mycobacterium avium subsp. paratuberculosis. *Animals (MDPI)*, submitted, 2019.

[6] C. Mizzi, A. Fabbri, S. Rambaldi, F. Bertini, N. Curti, S. Sinigardi, R. Luzi, G. Venturi, M. Davide, G. Muratore, A. Vannelli, and A. Bazzani. Unraveling pedestrian mobility on a road network using icts data during great tourist events. *EPJ Data Science*, 7(1):44, Oct 2018.