Ministero dell'Università e della Ricerca

Segretariato Generale
Direzione generale per il coordinamento e la valorizzazione della ricerca e dei suoi risultati

Ufficio III

Relazione Scientifica Intermedia - Seconda Annualità
PRIN 2017 - protocollo: 2017WR7SHH

Principal Investigator

FERRAGINA Paolo
(cognome) **(nome)**

Università di PISA
(Università/Ente)

Risultati conseguiti

AMBITO DI VALUTAZIONE	RISPOSTA (spuntare)	DESCRIZIONE (max 3.000 caratteri spazi inclusi)
1) Personale appositamente da reclutare (titolare di contratti a tempo determinato, assegni di ricerca, borse di dottorato). **Sono stati stipulati contratti dal Gruppo di Ricerca?** ***Specificare, per ogni contratto, la data di attivazione, la tipologia di contratto e la durata. Segnalare altresì eventuali rescissioni o interruzioni, evidenziando le relative motivazioni.***	SI	Si è acquisita, a partire da 1 Luglio 2021, la Dottoranda Mariella Bonomo (Dottorato in ICT-UniPA) sui fondi del progetto per i rimanenti 18 mesi di borsa. L'assegno di ricerca del Dott. Manuel Striani presso UniUPO finanziato dal progetto è terminato il 31 agosto 2021. Il 10 Settembre 2021 sarà bandito un nuovo assegno di ricerca annuale presso UniUPO totalmente finanziato dal progetto. E' stato bandito il 30 agosto 2021 un assegno di

		ricerca annuale, a valere interamente sui fondi del progetto, per attività di ricerca da svolgersi presso l'Unità UniPI sul tema delle strutture dati learned e compresse. Il Dott. Alessandro Petrini ha preso servizio presso UniMI il 1/12/2020 con un assegno di ricerca finanziato dal progetto, durata un anno estendibile a due. La Dott.ssa Jessica Gliozzo sta usufruendo del primo anno di borsa (scadenza 1 novembre 2021) dei tre previsti dal dottorato europeo congiunto fra il Dipartimento di Informatica di UniMI e il Joint Research Center Ispra, di cui il progetto finanzia il primo anno.
2) Attrezzature, strumentazioni e software di nuovo acquisto. **Sono stati effettuati acquisti in tale ambito da parte del Gruppo di Ricerca?** ***Specificare la tipologia di bene acquistato e il relativo uso nell'ambito del progetto.***	SI	UniMI ha acquistato un Lenovo TC M80t – Tower a ulteriore supporto degli esperimenti condotti per l'implementazione ed esecuzione di modelli basati su reti neurali profonde, di strutture dati learned e di classificatori succinti.
3.1) Attività di divulgazione dei risultati. **E' stata sviluppata tale attività in ambito di convegni, seminari, ecc.?** ***Specificare per ogni partecipazione a convegno: il titolo del convegno, data, luogo, il titolo della ricerca presentata e il nome dello speaker. Se tale attività non è stata svolta, illustrarne la motivazione.***	SI	CURATELE, CONTRIBUTI A CONFERENZA, E SEMINARI SU INVITO Special Issue di Briefing in Bioinformatics, "Integrative Bioinformatics and Omics Data Sources Interoperability in the Next-Generation Sequencing Era, Briefings in Bioinformatics, 22, (1), 2021. Co-Guest Editor: S.E. Rombo VLDB Workshops (DMAH 2020), 31/8-4/9 2020, online. "Prediction of lncRNA-Disease Associations from Tripartite Graphs". Speaker: M. Bonomo 12th Intl Joint Conf. on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2020), 2-4 Nov. 2020, Budapest, Ungheria. "Identifying the k Best Targets for an Advertisement Campaign via Online Social Networks". Speaker: M. Bonomo 8-hours seminar on "Compact and learned data structures" given to the students of the course on Algorithm Engineering at UniPI, A.Y. 2020/21. After the seminar, we organized a coding challenge on compressed and learned data structures. Organizers: G. Vinciguerra and P. Ferragina LADSIOS workshop co-located with VLDB '21. Online: 16/8/21. "The design of learning-based compressed data structures". Speaker: G. Vinciguerra

Seminar at the University of Melbourne. Online: 12/4/21. “A tutorial on learning-based compressed data structures”. Speaker: G. Vinciguerra

Seminar at Univ. de Lille and Inria. Online: 19/3/21. “Theory and practice of learning-based compressed data structures”. Speaker: G. Vinciguerra

Stanford Compression Workshop. Online: 25/2/21. “Learning-based compressed data structures (poster)”. Speaker: G. Vinciguerra

23rd SIAM Symp. on Algorithm Engineering and Experiments (ALENEX). Online: 5/1/21. “A ‘learned’ approach to quicken and compress rank/select dictionaries”. Speaker: A. Boffa

Dalhousie Univ. Online: 26/3/21. “Wheeler Graphs: A framework for BWT-based data Structures”. Speaker: G. Manzini

Huawei Workshop "Compute and Storage Technology 2020". Online: 3/12/20. “Learned data structures, challenges for storage systems''. Speaker: P. Ferragina

25th Intl Conf. on Pattern Recognition (ICPR). 10-15/1/21, Milano. “Compression strategies and space-conscious representations for deep neural networks”. Speaker: Marco Frasca

3rd Workshop on Reproducible Research in Pattern Recognition (RRPR). 11/1/21, Milano. “Reproducing the Sparse Huffman Address Map Compression for Deep Neural Networks”. Speaker: Dario Malchiodi

NOTA: Il Progetto segue una politica di promozione dei giovani, p.e. privilegiando la loro partecipazione come speaker a congressi internazionali. Secondo regolamento MUR, alcuni di essi non si possono rendicontare, ma i seguenti loro contributi a congresso sono menzionati in quanto le ricerche presentate sono scaturite da ricercatori del presente progetto.

AAI4H 2020. Speaker: M. Randazzo (contributor S. Rombo).

ALENEX 2021. Speaker: Massimiliano Rossi (contributor G. Manzini).

DCC 2021. Speaker: Marco Oliva (contributor G. Manzini).

		WABI 2021. Speaker: Garance Gourdel (contributor G. Manzini).
3.2) Attività di divulgazione dei risultati. E' stata sviluppata tale attività nell'ambito della pubblicazione di lavori su riviste? *Specificare per ogni pubblicazione peer-reviewed su rivista: gli autori, il titolo del lavoro, il nome della rivista, il volume, l'anno della pubblicazione, il codice DOI e il tipo di open-access. Se tale attività non è stata svolta, illustrarne la motivazione.*	SI	Umberto Ferraro Petrillo, Francesco Palini, Giuseppe Cattaneo, Raffaele Giancarlo. FASTA/Q data compressors for MapReduce-Hadoop genomics: space and time savings made easy, BMC Bioinformatics, Vol 22, 2021. Gold Open Access. DOI: https://doi.org/10.1186/s12859-021-04063-1 Umberto Ferraro Petrillo, Francesco Palini, Giuseppe Cattaneo, Raffaele Giancarlo. Alignment-free Genomic Analysis via a Big Data Spark Platform, Bioinformatics, Vol 37, pp. 1658-16765, 2021 Open Access al sito del Publisher in 6 mesi. DOI: https://doi.org/10.1093/bioinformatics/btab014 Green Open Access Preprint: https://arxiv.org/abs/2005.00942 Domenico Amato, Giosuè Lo Bosco, Riccardo Rizzo. CORENup: a combination of convolutional and recurrent deep neural networks for nucleosome positioning identification, BMC Bioinformatics, Vol. 21, 2020. Gold Open Access. DOI: https://doi.org/10.1186/s12859-020-03627-x Paolo Ferragina, Fabrizio Lillo, and Giorgio Vinciguerra. On the performance of learned data structures. Theoretical Computer Science, vol 871, pp. 107-120, 2021. Gold Open Access. DOI: https://doi.org/10.1016/j.tcs.2021.04.015 Filippo Geraci, Giovanni Manzini: EZcount: An all-in-one software for microRNA expression quantification from NGS sequencing data. Comput. Biol. Medicine 133: (2021) DOI: https://doi.org/10.1016/j.compbiomed.2021.104352 Green Open Access, Postprint su pagina del progetto. Lavinia Egidi, Felipe A. Louza, Giovanni Manzini: Space Efficient Merging of de Bruijn Graphs and Wheeler Graphs. Algorithmica, (2021) DOI: https://doi.org/10.1007/s00453-021-00855-2 Green Open Access, Postprint su pagina del progetto.

Relazione tecnica

Breve descrizione delle attività svolte da ciascuna unità di ricerca, nel periodo di riferimento.
Evidenziare, inoltre, con riferimento all'intero Gruppo di Ricerca:
a) se ci sono stati cambiamenti (aggiunte/eliminazioni o spostamenti temporali) rispetto al previsto, illustrando le principali motivazioni;
b) quale sia il reale progresso verso gli obiettivi previsti, indicando, altresì, gli eventuali risultati ottenuti;
c) come i risultati già ottenuti verranno sfruttati nell'ambito delle attività in corso di svolgimento, o se sia possibile prevederne uno sfruttamento diretto (brevetti, immissione di prodotti sul mercato, ecc);
d) se sono sopraggiunte particolari difficoltà che mettano a rischio il conseguimento degli obiettivi minimi previsti.

Una descrizione delle attività e dei software sviluppati è presente nella pagina del progetto learned.di.unipi.it. Il 12.3.2021 si è svolto online il terzo meeting.

UniMI ha continuato lo studio di tecniche di compressione di reti neurali convoluzionali, identificando un nuovo metodo di quantizzazione probabilistico dei pesi della rete, e nuove strutture dati per la compressione di matrici (derivate da reti neurali) che hanno ottenuto miglioramenti nel caso di matrici sparse e quantizzate. Questa linea di ricerca ha prodotto un lavoro sottoposto a una rivista di riferimento, è disponibile su arXiv e sul sito del progetto, e due pubblicazioni a conferenza internazionale. L'unità ha inoltre sviluppato strutture dati learned per l'indicizzazione di stringhe (in collaborazione con UniPI) e filtri di Bloom learned.

UniPA ha continuato il progetto e la realizzazione di piattaforme software per Learned Binary Search, la classificazione, compressione (multi-choice o basata su BWT) e analisi di BigData genomici via rappresentazioni succinte su Hadoop e Spark. Il software è disponibile sul sito del progetto e lavori preliminari su tali attività sono su arXiv o sul sito del progetto. I risultati indicati l'anno scorso come reports, sono stati pubblicati. Si sono avute collaborazioni con le altre unità.

UniUPO ha continuato lo studio di strutture dati efficienti mediante la combinazione di metodi di compressione che integrano grammatiche con altre tecniche di compressione. Queste strutture dati ibride sono state applicate all'indicizzazione di sequenze di DNA, e presentate a conferenze internazionali. In collaborazione con UniPI, queste tecniche ibride hanno consentito di migliorare la compressione di matrici di grandi dimensioni usate in Machine Learning. I risultati ottenuti sono stati pubblicati su due riviste e presentati a tre conferenze; altri sono in corso di sottomissione.

UniPI ha continuato lo studio di strutture dati compresse e learned, progettando tecniche per la compressione di vettori binari con supporto di operazioni rank/select (il codice è disponibile su GitHub). I risultati sono stati presentati alla conferenza SIAM ALENEX '21, e una loro versione estesa è stata sottomessa ad ACM Trans. on Algorithms. L'unità ha anche esteso la libreria PGM-index per indicizzare input n-dim e per interrogazioni quali orthogonal range e (approximate) k-nearest neighbour. L'interesse industriale del PGM è testimoniato dalle oltre 530 "stelle" su GitHub, dalla prima pagina su HackerNews del 25/1/21 (con 580 punti e 120 commenti), e decine di condivisioni su Twitter. L'unità continua la collaborazione con gli altri partner sulla compressione di matrici (UniUPO), sulle strutture dati learned per stringhe (UniMI), e sui modelli per la predizione di serie temporali (UniUPO e UniPA).

Domanda A) Nessun cambiamento rilevante rispetto al previsto, nonostante la situazione pandemica continui a limitare la collaborazione tra sedi e i viaggi per la disseminazione in presenza dei risultati.

Domanda B) Le attività del progetto hanno avuto un concreto sviluppo dimostrato dai risultati ottenuti. Si segnalano collaborazioni internazionali prestigiose sui temi del progetto: con ricercatori di Harvard University, Karlsruhe Institute of Technology, Johns Hopkins University, Centro di Genomica Computazionale del TJ Watson IBM, Dalhousie, Cile, Florida.

Domanda C) Alcuni software e metodi possono diventare "prodotti" se saranno disponibili ulteriori finanziamenti e

collaborazioni con aziende. I più maturi per industrial transfer sono: FADE (Bioinformatics '21), FASTDOOPPC (BMC Bioinformatics '21), PFP (J. Comp. Biol. '20) e PGM-index (VLDB '20, ICML '20, TCS '21). Si segnala anche una domanda di brevetto italiana sui temi del progetto, co-inventata da P. Ferragina, G. Manzini, e G. Vinciguerra, depositata il 28/5/21 con numero 102021000014069.

Domanda D) Malgrado emergenza Covid, i proponenti affermano che dovrebbero riuscire a realizzare gli obiettivi del progetto.

17/09/2021 10:55