PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Marco Castelluccio

XXXI Cycle

Training and Research Activities Report – First Year

Tutor: Carlo Sansone – co-Tutor: Annalisa Verdoliva

1. Information

PhD Student: Marco Castelluccio
MS title: Computer Engineering – University of Naples Federico II
PhD cycle: XXXI – ITEE University of Naples Federico II
Fellowship type: PhD student without grant
Tutor: Carlo Sansone – **co-Tutor:** Annalisa Verdoliva

I received my MS degree (cum laude) in Computer Engineering from the Università degli Studi di Napoli Federico II in September 2015.

I've been a Software Engineer at Mozilla since November 2015, after two internships and a period of contracting with the same company during my graduate studies.

2. Study and Training activities

Courses

1. Course "Data Mining" – Carlo Sansone – September 2016
2. Course "Elaborazione di Segnali Multimediali" – Annalisa Verdoliva – September 2016

Seminars

1. The Five Tribes of Machine Learning (And What You Can Learn from Each) – Pedro Domingos (University of Washington) – November 2015 – Association for Computing Machinery (ACM)
2. Ur/Web: A Simple Model for Programming the Web – Adam Chlipala (MIT) – January 2016 – Mozilla

3. Best Practices in Software Benchmarking – Jan Vitek (Northeastern University), Joe Parker (Kew Royal Botanical Gardens), Simon Taylor (Lancaster University), Tomáš Kalibera (University of Kent), James Davenport (University of Bath), Edd Barrett (King's College London), Jeremy Bennett (Embecosm) – April 2016 – King's College London
4. The Deep Learning Phenomenon – Alison Lowndes (NVIDIA Corporation) – April 2016 – EPSRC Centre for Doctoral Training in Medical Imaging in collaboration with NVIDIA Corporation
5. Lies, Damned Lies and Software Analytics: Why Big Data Needs Thick Data – Margaret-Anne Storey (University of Victoria) – May 2016 – Association for Computing Machinery (ACM)
6. Catch Up on CUDA – Chris Mason (Acceleware) – June 2016 – NVIDIA Corporation
7. Large-Scale Deep Learning with TensorFlow for Building Intelligent Systems – Jeff Dean (Google Research) – July 2016 – Association for Computing Machinery (ACM)
8. Embracing Data Science in Your Organization – Christine Doig (Continuum Analytics) – September 2016 – Association for Computing Machinery (ACM)
9. Analysing a 2x2 Table – Angie Wade, Eirini Koutoumanou and Vicki Aldridge (University College London) – October 2016 – University College London
10. COW 48 – Tutorial on Causal Impact – Kay Brodersen (ETH Zurich and Google) and William Martin (University College London) – October 2016 – University College London
11. TensorFlow: A Framework for Scalable Machine Learning – Martin Wicke (Google) – October 2016 – Association for Computing Machinery (ACM)

External courses

1. Course "Data Mining – Classification" – June 2016 – Università degli Studi di Milano Bicocca

Year	Modules	Seminars	Research	Tot.
1	21 (20)	7 (5)	35 (35)	63 (60)
2	(9)	(6)	(42)	57 (60)
3	(0)	(5)	(55)	60 (60)
Tot.	**30 (30-70)**	**18 (10-30)**	**132 (80-140)**	**180 (180)**

3. Research activity

- **Convolutional Neural Networks for Classification of Remote Sensing Images** – Continued work on using convolutional neural networks for classification of remote sensing images started during the MSc thesis [1], analyzing results on additional datasets [2].

  Preparation of a paper for the JURSE2017 [3] conference, titled '*Training Convolutional Neural Networks for Semantic Classification of Remote Sensing Imagery*'.

- **Automating the Understanding of Groups of Crash Reports** – Many Software Engineering studies have focused on improving the bucketing of crash reports with different techniques (e.g. by using different distance metrics, like Levenshtein [4], custom stack-based metrics [5], by using information retrieval techniques [6,7], etc.).

  The focus of my work is instead on how to automatically describe the buckets' properties in the most useful way for developers. Understanding what makes a crash group meaningfully different than other groups is indeed very often useful for

debugging (and in some cases even enough for fixing the crash, e.g. by blocklisting a certain graphics card), but it involves a tedious and error-prone manual exploration of the database of crashes.

Inspired by the STUCCO [8,9] and CIGAR [10] data mining algorithms, I've devised an algorithm to automate the understanding of groups of crash reports.

- **Emprical Study of Uplifts in Mozilla Firefox** – Collaboration with the École Polytechnique de Montréal on studying uplifts (backports) in the Mozilla Firefox software.

  Mozilla Firefox has adopted a rapid-release model [11,12], called the "train model". Uplifts are high-priority changes (e.g. changes to fix top crashes, serious regressions, compatibility problems with widely used web sites or addons, etc.) that "jump" the channels (e.g. from the "nightly" channel to the "beta" channel, instead of "nightly" -> "aurora" -> "beta"), instead of stabilizing for 18 weeks before being released to the public.

  The aim is to understand the properties of these critical changes vs normal changes; understand which uplifts introduced bugs (by using the SZZ algorithm proposed in [13]) and why; with the ultimate goal of building a model to predict the riskiness of an uplift.
  Currently targeting The 14th International Conference on Mining Software Repositories (MSR '17) [14].

[1] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land Use Classification in Remote Sensing Images by Convolutional Neural Networks", arXiv preprint.

[2] G.S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maitre, "Structrual High-Resolution Satellite Image Indexing", in Processings of the ISPRS, TC VII Symposium Part A: 100 Years ISPRS—Advancing Remote Sensing Science, Vienna, Austria, 5–7 July 2010.

[3] http://jurse2017.com/.

[4] Tejinder Dhaliwal, Foutse Khomh, Ying Zou, "Classifying Field Crash Reports for Fixing Bugs: A Case Study of Mozilla Firefox", in Proceedings of the 2011 27th IEEE International Conference on Software Maintenance (ICSM '11), pp. 333-342.

[5] Y. Dang, R. Wu, H. Zhang, D. Zhang, and P. Nobel, "ReBucket: a method for clustering duplicate crash reports based on call stack similarity", in Proceedings of the 34th International Conference on Software Engineering (ICSE '12), pp. 1084-1093.

[6] J. Lerch, and M, Mezini, "Finding Duplicates of Your Yet Unwritten Bug Report", in Proceedings of the 2013 17th European Conference on Software Maintenance and Reengineering (CSMR '13). IEEE Computer Society, Washington, DC, USA, 69-78.

[7] J. C. Campbell, E. A. Santos, and A. Hindle, "The unreasonable effectiveness of traditional information retrieval in crash report deduplication", in Proceedings of the 13th International Conference on Mining Software Repositories (MSR '16). ACM, New York, NY, USA, 269-280.

[8] S.D. Bay, and M.J. Pazzani, "Detecting change in categorical data: Mining contrast sets", in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99), pp. 302–306, San Diego, U.S.A., August 1999.

[9] S.D. Bay, and M.J. Pazzani, "Detecting group differences: Mining contrast sets", Data Mining and Knowledge Discovery, 5(3):213–246, 2001.

[10] R. J. Hilderman, and T. Peckham, "A Statistically Sound Alternative Approach to Mining Contrast Sets", in Proceedings of the 4th Australasian Data Mining Conference (AusDM), 2005.

[11] http://mozilla.github.io/process-releases/draft/development_overview/.

[12] M. V. Mäntylä, B. Adams, F. Khomh, E. Engström, and K. Petersen, "On rapid releases and software testing: a case study and a semi-systematic literature review", Empirical Softw. Engg. 20, 5 (October 2015), 1384-1425.

[13] J. Śliwerski, T. Zimmermann, and A. Zeller, "When do changes induce fixes?", in Proceedings of the 2005 international workshop on Mining software repositories (MSR '05). ACM, New York, NY, USA, 1-5.

[14] http://2017.msrconf.org/.

4. Products

M. Castelluccio, G. Poggi, C. Sansone, L. Verdoliva – Land Use Classification in Remote Sensing Images by Convolutional Neural Networks – https://arxiv.org/abs/1508.00092

M. Castelluccio, G. Poggi, C. Sansone, L. Verdoliva – Training Convolutional Neural Networks for Semantic Classification of Remote Sensing Imagery – JURSE2017 (submitted)