Course Assignment

Course on Fairness, Accountability, Confidentiality and Transparency in AI
January 2021
University of Amsterdam

Maarten de Rijke
derijke@uva.nl

Maurits Bleeker
m.j.r.bleeker@uva.nl

Sami Jullien
s.jullien@uva.nl

Ana Lucic
a.lucic@uva.nl

1 INTRODUCTION

The objective of this course is understanding the *technical* aspects of each of the four topics (Fairness, Accountability, Confidentiality, Transparency), specifically existing algorithms, while also making a contribution to the community by releasing a Python package.

1.1 Fairness

Research on fairness primarily involves mitigating the algorithmic discrimination of individuals based on protected attributes such as gender or race. There are many different (oftentimes competing) fairness definitions resulting in a wide range of ways to frame the problem.

1.2 Accountability

Research on accountability is usually centred around identifying who is responsible for the (potentially incorrect or unjust) decision that is a result of an algorithmic prediction. As a result, this research is typically less focused on the algorithms themselves and instead places the focus on the *impact* of algorithms.

1.3 Confidentiality

Research on confidentiality examines how the privacy of individuals whose data is being used to develop ML models can be preserved. If the data is high-dimensional and there is a wide range of possible values per feature, the information is essentially identifiable, and therefore simple anonymization of personal identifiers such as name or email is not enough.

1.4 Transparency

Research on transparency involves interpreting the behaviour of complex models. This is typically done in either a global (interpreting the whole model) or local (interpreting individual predictions) manner. In this course we will primarily focus on the latter, which involves methods such as identifying important features, generating counterfactual examples, or finding prototypical examples of a particular class.

2 PROJECT DESCRIPTION

The lack of reproducibility has been an ongoing issue in academic research. The goal of this project is to assess the reproducibility of existing work by reimplementing an algorithm, replicating and/or extending the experiments from the corresponding paper, and detailing your findings in a report. In this assignment you will implement an existing FACT algorithm in groups of 4. We will follow the setup from the Machine Learning Reproducibility Challenge (MLRC) (`https://paperswithcode.com/rc2020`), and encourage you to participate in the challenge by submitting your work. The task description specifies: *"Essentially, think of your role as an inspector verifying the validity of the experimental results and conclusions of the paper. In some instances, your role will also extend to helping the authors improve the quality of their work and paper."* The full task description is available here: `https://paperswithcode.com/rc2020/task`.

There are two scenarios possible for this project:

(1) There already exists an open-source implementation of your selected paper. You are allowed to use this, but we will be aware of the fact that this implementation is available. Given the implementation:

? (a) The results you obtain are different as described in the paper (i.e. the paper is not reproducible). Your report should explain what these differences are and why they occur. You should also try to resolve the problem(s) and explain your rationale behind the choices you made, as well as describing your implementation process and the results you obtained.

? (b) The results are reproducible, meaning this method can now be used for further research. The experimental results are less robust when they do not scale beyond the original model, data(s) and domain(s) used in the paper. Are these results also reproducible for other domains, datasets, model (configurations), etc?

~~(2) There is no open-source implementation available, meaning your group needs to reimplement everything yourselves. What are the difficulties while reproducing this work and how have you solved them? Are the results similar as described in the paper? If not, why? If yes, is this work is reproducible for other domains, datasets, model (configurations).~~

If an open-source implementation exists, the result 'the paper is reproducible' is not enough for a good grade. Either you need to go beyond the original results by questioning the results on other domains, data, and/or model configurations, or you need to show that the results are not as in the paper and propose an alternative solution. This might be challenging for a four week project, so please keep this in mind when choosing a paper.

FACT-AI Course, January 2021, University of Amsterdam

If there is no open-source implementation, the report should explain in detail how and if the work is reproducible. The deadline for handing in the project on Canvas is **23:59 on 29 January 2021**, which is also the same as the deadline for submitting to the MLRC. This is intentional – formally participating in the challenge is a great opportunity to understand how ML research is done by interacting with reviewers and getting feedback on your work.

2.1 Report

To participate in the challenge, you need to claim a paper (from Section 4) and write a short proposal of your plan to reproduce the paper (see `https://paperswithcode.com/rc2020/registration`). Note that although MLRC includes all papers from top conferences, we are only focusing on papers about FACT topics.

To write the report, you will use the MLRC template: `https://www.overleaf.com/project/5f4e72de7681920001b208f9`. The objective of the report is to explain the results you obtained as well as the process behind the implementation. Your report should be **no more than 8 pages long (excluding references)**.

If you would like to receive feedback on an early draft of your report, you can email it to your TA by **23:59 on 20 January**. You will need to submit the final report via Canvas by **23:59 on 29 January 2021**.

2.2 Final Code Submission

The final submission of your implementation should be in a private GitHub repository with all the information, code and data needed to test your implementation. Any commits you make to your repository after the deadline will be ignored. All implementations requiring a deep learning framework **must be done in PyTorch**. Please set your repository up in a clean and reasonable way with the following components:

- Environment configuration.
- IPython notebook detailing all results in the report. Please ensure that it is possible to simply run all cells and obtain the results without any issues. Make sure that only the code for generating the results is present in the notebook. The model(s) and all the other files needs to be generate the results should be in separated files. It should function as some kind of API.
- Instructions for how to run your implementation.
- Dataset(s) used in the experiments.
- All required scripts for testing the implementation.

look at
reports
published
by
MLRC
instead

Examples of some of last year's projects can be found here: `https://github.com/uva-fact-ai-course/uva-fact-ai-course`.

We also want your code to be reproducible. Please take a look at the following resources for suggestions and best practices on producing reproducible code: (1) `https://github.com/paperswithcode/releasing-research-code`, and (2) `https://www.cs.mcgill.ca/~ksinha4/practices_for_reproducibility/`.

2.3 Presentation

The final part of the project is a 10 minute presentation on your findings. This should essentially be a summary of your written report and will take place during the last week of the course, on 29 January 2020 (exact times to be scheduled).

2.4 Grading

You will be graded according to the Grading Matrix provided in this document. If you submit to the MLRC, you will get an extra 0.5 point. Please keep in mind that if you submit, the paper will be publicly available on OpenReview and therefore anyone can see it. We will also award another 0.5 point the 10 groups with the best code implementations.

3 LOGISTICS

Please complete the following steps **by 18:00 on 4 January**:

(1) Choose your group for the project. There should be a maximum of 4 students per group. All communication about the project should take place with the entire group.
(2) Discuss with your group which papers you would like to implement from Section 4.
(3) Create a private GitHub repository for your project. All communication will be handled (and logged) via issues in this repository.
(4) **One person per group** needs to fill out the following Google Form: `https://forms.gle/sppyYvcf9FCx2LXv8`. We will do our best to take everyone's paper preferences into account but given the number of students taking this course, we simply cannot guarantee that you will be assigned one of your top papers.

Each group will get 2x 20 minute online Practicums with their TA each week. The Practicum times for each paper are listed

4 PAPERS TO BE REPRODUCED

You will implement **one** of the following papers with your group. There can be a **maximum of 4 groups** working on the same paper.

- M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni. Fairness by learning orthogonal disentangled representations. In *ECCV*, 2020
- D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. Parameterized explainer for graph neural network. In *NeurIPS*, 2020
- W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps. Towards visually explaining variational autoencoders. In *CVPR*, 2020
- P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. H. Chi. Fairness without demographics through adversarially reweighted learning. In *NeurIPS*, 2020
- P. Li, H. Zhao, and H. Liu. Deep fair clustering for visual learning. In *CVPR*, 2020
- A. K. Mohankumar, P. Nema, S. Narasimhan, M. M. Khapra, B. V. Srinivasan, and B. Ravindran. Towards transparent and explainable attention models. In *ACL*, 2020
- Y. Zhu, X. Yu, M. Chandraker, and Y.-X. Wang. Private-knn: Practical differential privacy for computer vision. In *CVPR*, 2020
- L. Xiang, H. Ma, H. Zhang, Y. Zhang, J. Ren, and Q. Zhang. Interpretable complex-valued neural networks for privacy protection. In *ICLR*, 2020
- S. Li, B. Hooi, and G. H. Lee. Identifying through flows for recovering latent representations. In *ICLR*, 2020

Table 1: Lecture schedule for the course.

	Mon 4 Jan	**Wed 6 Jan**	**Fri 8 Jan**	**Mon 25 Jan**
9:00 - 10:00	Fairness lecture	Transparency lecture	Confidentiality lecture	Accountability lecture
10:00 - 11:00	Reproducibility lecture	Guest lecture	Discussion group #3	Guest lecture
11:00 - 12:00		Discussion group #2	Paper dissection (Sami) #3	
12:00 - 13:00		Paper dissection (Maartje) #2		
13:00 - 14:00				
14:00 - 15:00	Discussion group#1			
15:00 - 16:00	Paper dissection (Maurits) #1			

- G. Plumb, J. Terhorst, S. Sankararaman, and A. Talwalkar. Explaining groups of points in low-dimensional representations. In *ICML*, 2020
- M. O'Shaughnessy, G. Canal, M. Connor, M. Davenport, and C. Rozell. Generative causal explanations of black-box classifiers. In *NeurIPS*, 2020
- D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton. Learning to deceive with attention-based explanations. In *ACL*, 2020
- J. Fisher, A. Mittal, D. Palfrey, and C. Christodoulopoulos. Debiasing knowledge graph embeddings. In *EMNLP*, 2020

5 LECTURES

There are four lecture blocks for this course, one for each topic. The schedule can be found in Table 1. Each lecture block will consist of (i) a general lecture on the topic, (ii) a student discussion group on a prominent paper, and (iii) a "paper dissection" session, where a TA will go over the same prominent paper. The purpose of the discussion group is to learn how to pick apart research papers, since this is an important part of reimplementing existing work (and an important part of research in general). To prepare for the discussion group, you are expected to read the paper in advance and contribute to the discussion. In the corresponding paper dissection session, a TA will go over the same paper you discussed in the discussion group, to give an overview of the papers' strengths and weaknesses.

There are two exceptions:

- The second lecture block (on Transparency) will instead include a lecture on Reproducibility, since this is a major component of the assignment.
- The Accountability lecture block will have a guest lecture instead of a paper dissection session.

5.1 Papers for Dissections

The paper dissection sessions will cover the following papers:

- **Paper dissection #1 - Fairness:** M. Hardt, E. Price, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *NeurIPS 2016*, 2016. `https://papers.nips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf`
- **Paper dissection #2 - Transparency:** M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *KDD 2016*, 2016. `https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf`
- **Paper dissection #3 - Confidentiality:** M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM SIGSAC 2016*, 2016. `https://arxiv.org/pdf/1607.00133.pdf`

5.2 Questions to answer in the discussion group

We suggest you look into the following questions for the discussion group with other students:

- Get the main contributions from the introduction. Does this seem like a good field to lead research on? Do the premises seem correct?
- Literature review: Do you feel like the authors have been thorough enough and are not ignoring a field that you are aware of? This is the trickiest part to read, as it often is beyond your own knowledge.
- Methodology: Understand the algorithm introduced by the authors. Are they overselling it in the intro? What does it bring, what is new there?
- Experiments: Are the metrics realistic? Are the improvements as good as claimed by the authors? Is it easily reproducible?
- After reading the paper, do you agree with the abstract? Which aspects of this work might be difficult to reproduce?

6 EXTRA READING MATERIAL

Below is a list of prominent papers in the FACT domain, if you're interested in further reading.

6.1 Fairness

- L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed Impact of Fair Machine Learning. In *ICML 2018*
- C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The Variational Fair Autoencoder. In *ICLR 2016*, 2016

6.2 Accountability

- I. D. Raji and J. Buolamwini. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *AIES 2019*, 2019

- J. Mena Roldan, O. Pujol Vila, and J. Vitria Marca. Dirichlet Uncertainty Wrappers for Actionable Algorithm Accuracy Accountability and Auditability. In *FAT* 2020*, 2018

6.3 Confidentiality

- M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman. Differentially Private Fair Learning. In *ICML 2019*, 2019
- H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning Differentially Private Recurrent Language Models. In *ICLR 2018*, 2018

6.4 Transparency

- S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *NeurIPS 2017*, 2017
- M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. In *ICML 2017*, 2017

REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM SIGSAC 2016*, 2016.

[2] J. Fisher, A. Mittal, D. Palfrey, and C. Christodoulopoulos. Debiasing knowledge graph embeddings. In *EMNLP*, 2020.

[3] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *NeurIPS 2016*, 2016.

[4] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman. Differentially Private Fair Learning. In *ICML 2019*, 2019.

[5] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. H. Chi. Fairness without demographics through adversarially reweighted learning. In *NeurIPS*, 2020.

[6] P. Li, H. Zhao, and H. Liu. Deep fair clustering for visual learning. In *CVPR*, 2020.

[7] S. Li, B. Hooi, and G. H. Lee. Identifying through flows for recovering latent representations. In *ICLR*, 2020.

[8] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed Impact of Fair Machine Learning. In *ICML 2018*.

[9] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps. Towards visually explaining variational autoencoders. In *CVPR*, 2020.

[10] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The Variational Fair Autoencoder. In *ICLR 2016*, 2016.

[11] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *NeurIPS 2017*, 2017.

[12] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. Parameterized explainer for graph neural network. In *NeurIPS*, 2020.

[13] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning Differentially Private Recurrent Language Models. In *ICLR 2018*, 2018.

[14] J. Mena Roldan, O. Pujol Vila, and J. Vitria Marca. Dirichlet Uncertainty Wrappers for Actionable Algorithm Accuracy Accountability and Auditability. In *FAT* 2020*, 2018.

[15] A. K. Mohankumar, P. Nema, S. Narasimhan, M. M. Khapra, B. V. Srinivasan, and B. Ravindran. Towards transparent and explainable attention models. In *ACL*, 2020.

[16] M. O'Shaughnessy, G. Canal, M. Connor, M. Davenport, and C. Rozell. Generative causal explanations of black-box classifiers. In *NeurIPS*, 2020.

[17] G. Plumb, J. Terhorst, S. Sankararaman, and A. Talwalkar. Explaining groups of points in low-dimensional representations. In *ICML*, 2020.

[18] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton. Learning to deceive with attention-based explanations. In *ACL*, 2020.

[19] I. D. Raji and J. Buolamwini. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *AIES 2019*, 2019.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *KDD 2016*, 2016.

[21] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni. Fairness by learning orthogonal disentangled representations. In *ECCV*, 2020.

[22] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. In *ICML 2017*, 2017.

[23] L. Xiang, H. Ma, H. Zhang, Y. Zhang, J. Ren, and Q. Zhang. Interpretable complex-valued neural networks for privacy protection. In *ICLR*, 2020.

[24] Y. Zhu, X. Yu, M. Chandraker, and Y.-X. Wang. Private-knn: Practical differential privacy for computer vision. In *CVPR*, 2020.