Course Assignment

Course on Fairness, Accountability, Confidentiality and Transparency in AI
January 2020
University of Amsterdam

Maarten de Rijke
derijke@uva.nl

Maurits Bleeker
m.j.r.bleeker@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. In this assignment, we will narrow the focus on Fairness and Transparency.

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 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 Fairness or Transparency algorithm in groups of 4. 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 requires creativity, which might be challenging for a four week project.

If there is no open-source implementation, the report should explain in detail how and if the work is reproducible. The deadline for the project is **23:59 on 31 January 2020**.

2.1 Report

This project also includes a written component based on the guidelines from the ACM Artifact Review and Badging process [15]. 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 5 pages** long and have the following sections:

(1) Abstract: Summary of your work and the main results.
(2) Introduction: Brief overview of the field (Fairness or Transparency) and the task your algorithm is solving.
(3) Method: Explain the algorithm in detail and how your group implemented it. If you chose a paper with an existing implementation, explain how you extended it.
(4) Experimental setup: Describe your experimental settings and the rationale behind the choices your group made.
(5) Results: Detail the experimental results.
(6) Discussion: Explain your findings. How do these results match up with what was reported in the paper? What are the advantages and disadvantages of this method? How can its shortcomings be overcome? What other experimental settings should be tested?
(7) Broader Implications: Connect to papers in Section 5.
(8) Conclusion: Summarize your findings and choose an ACM badge to award the paper.

FACT-AI Course, January 2020, University of Amsterdam

See [16] for an example.[1] You may put extra results, minor details, etc. in an appendix but all of the major findings should be in the main report.

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 22 January**. You will need to submit the final report via Canvas by **23:59 on 31 January 2020**.

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.

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 31 January 2020 (exact times to be scheduled, during your tutorial sessions).

2.4 Grading

You will be graded according to the Grading Matrix provided in this document. The best implementations per paper will be released as a Python package as a contribution to the field. The group with the best implementation per paper will also receive an extra 0.5 point on their final grade. Since there are 10 papers to choose from, there is a maximum of 10 groups that can obtain the extra 0.5 point.

3 LOGISTICS

Please complete the following steps **by 18:00 on 6 January**:

(1) Choose your group for the project. There should be a maximum of 4 students per 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/FtRD4tJx94VoAv1x6`. We will do our utmost 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.

You will have two Practicums per week where you can ask your TA questions about your paper. You need to go to the Practicums that correspond to your paper (see attached Practicum Schedule sheet).

At the start of the course, to help you with any questions you may have, we will have open office hours in Village.AI (the mint green container building in the Startup Village) from 12:00–14:00 on January 6.

4 PAPERS TO BE REPRODUCED

You will implement **one** of the following papers with your group. There can be a **maximum of 5 groups** working on the same paper.

4.1 Transparency

- O. Li, H. Liu, C. Chen, and C. Rudin. Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. In *AAAI 2018*, 2018
- A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *NeurIPS 2018*, 2018
- S. Srinivas and F. Fleuret. Full-Gradient Representation for Neural Network Visualization. In *NeurIPS 2019*, 2019
- D. Alvarez-Melis and T. S. Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In *NeurIPS 2018*, 2018
- S. Jain and B. C. Wallace. Attention is not Explanation. In *NAACL 2019*, 2019

4.2 Fairness

- A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of Attention: Amortizing Individual Fairness in Rankings. In *SIGIR 2018*, 2018. Some prior knowledge of Information Retrieval is required.
- B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.
- S. Bordia and S. R. Bowman. Identifying and reducing gender bias in word-level language models. In *ACL 2019*, 2019. Language modelling.
- A. Amini, A. Soleimany, W. Schwarting, S. Bhatia, and D. Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *AAAI 2019*, 2019. Understanding of Variational Auto Encoders is required.
- T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS 2016*, 2016.

[1] This paper won the Best Reproducibility Paper Award at the European Conference on Information Retrieval in 2019.

5 READING MATERIAL

Below is a list of prominent papers in the FACT domain. Your report should connect to some of these works in the Broader Implications section.

5.1 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

5.2 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

5.3 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

5.4 Transparency

- S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *NeurIPS 2017*, 2017
- M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *KDD 2016*, 2016
- M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. In *ICML 2017*, 2017

REFERENCES

[1] D. Alvarez-Melis and T. S. Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In *NeurIPS 2018*, 2018.

[2] A. Amini, A. Soleimany, W. Schwarting, S. Bhatia, and D. Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *AAAI 2019*, 2019.

[3] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of Attention: Amortizing Individual Fairness in Rankings. In *SIGIR 2018*, 2018.

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS 2016*, 2016.

[5] S. Bordia and S. R. Bowman. Identifying and reducing gender bias in word-level language models. In *ACL 2019*, 2019.

[6] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *NeurIPS 2018*, 2018.

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

[8] S. Jain and B. C. Wallace. Attention is not Explanation. In *NAACL 2019*, 2019.

[9] O. Li, H. Liu, C. Chen, and C. Rudin. Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. In *AAAI 2018*, 2018.

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

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

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

[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. of Computing Machinery. Artifact review and badging.

[16] H. Oosterhuis and M. de Rijke. Optimizing Ranking Models in an Online Setting. *ECIR 2019*, 2019.

[17] 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.

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

[19] S. Srinivas and F. Fleuret. Full-Gradient Representation for Neural Network Visualization. In *NeurIPS 2019*, 2019.

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

[21] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.

Practicum Schedule

	Paper		Practicums			TA
Topic	Title	Link	Day	Time	Location	
Transparency	DL for Case-based Reasoning through Prototypes	https://arxiv.org/abs/1710.04806	Tuesday	15 - 16	SP C4.203 GIS-Studio	Weeks 1&2: Phillip, Weeks 3&4: Andreas
			Friday	9 - 10	SP G3.13	
Transparency	Full-gradient Representations for NN Visualization	https://arxiv.org/abs/1905.00780	Tuesday	17 - 18	SP C4.203 GIS-Studio	Leon
			Wednesday	17 - 18	SP C4.203 GIS-Studio	
Transparency	Explanations Based on the Missing	https://arxiv.org/abs/1802.07623	Tuesday	14 - 15	SP C4.203 GIS-Studio	Morris
			Friday	11 - 12	SP G3.13	
Transparency	Towards Robust Interpretability with Self-explaining NNs	https://arxiv.org/abs/1806.07538	Thursday	11 - 12	SP C4.203 GIS-Studio	Simon
			Friday	13 - 14	SP G3.13	
Transparency	Attention is not Explanation	https://arxiv.org/abs/1902.10186	Wednesday	15 - 16	SP C4.203 GIS-Studio	Marco
			Friday	15 - 16	SP G3.13	
Fairness	Identifying and Reducing Gender Bias in Word-level Language Models	https://arxiv.org/abs/1904.03035	Tuesday	16 - 17	SP C4.203 GIS-Studio	Weeks 1&2: Phillip, Weeks 3&4: Andreas
			Friday	10 - 11	SP B0.201	
Fairness	Mitigating Unwanted Biases with Adversarial Learning	https://arxiv.org/abs/1801.07593	Tuesday	18 - 19	SP C4.203 GIS-Studio	Leon
			Wednesday	18 - 19	SP C4.203 GIS-Studio	
Fairness	Man is to Programmer as Woman is to Homemaker?	https://arxiv.org/abs/1607.06520	Tuesday	13 - 14	SP C4.203 GIS-Studio	Morris
			Friday	12 - 13	SP B0.201	
Fairness	Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure	https://lmrt.mit.edu/sites/default/fil	Thursday	12 - 13	SP C4.203 GIS-Studio	Simon
			Friday	14 - 15	SP B0.201	
Fairness	Equity of Attention: Amortizing Individual Fairness in Rankings	https://arxiv.org/abs/1805.01788	Wednesday	16 - 17	SP C4.203 GIS-Studio	Marco
			Friday	16 - 17	SP B0.201	

FACT-AI Course Assignment: Grading Matrix

Grade		<= 5 (fail)	6 (sufficient)	7 (satisfactory)	8 (good)	9 (very good)	10 (excellent)
Project (40%)							
	Project Design	Unsystematic and/or no validated use of research and design methodologies. Insufficient explanation. How are the results tested and/or verified?	Adequate use of research and design methodologies. Limited explanation.	Adequate use of research and design methodologies. Explained and justified.	Use of the right research and design methodologies. Well-explained and well justified.	Profound and critical use of research and design methodologies. Very clear and validated design.	Excellent demonstration of research and design methodologies.
	Positioning of project	Project not positioned w.r.t. new literature, the FACT-field and reproducibility papers.	Project is somewhat positioned.	Project is sufficiently positioned in literature.	Project is correctly positioned in literature.	Project is well positioned within literature.	Project is integrated within literature, even from different fields/sources.
	Creativity	The project does not make an original contribution. E.g. the picked paper is just reproducible or not is reproducible without any extra insights.	Project does not really make any original contribution. The results are reproducible, with limited efford or not reproducible with limited insights (why is this not working?).	Project team had at least one original contribution to reproduce the work and/or go beyond the original results of the paper. .	Project team came up with several original ideas to reproduce the paper and/or go beyond the original results, design options and/or concepts not initiated or thought of by the supervisor.	Project team came up with many original ideas, design options and/or concepts to reproduce the work and/or go beyond the originial results. Not initiated or thought of by the supervisor.	Project team surprised us all with some brilliant new ideas, design options and/or concepts, both in breadth and depth.
Code base (20%)							
	Technical quality	Insufficient	Sufficient	Satisfactory	Good	Very Good	Excellent
	Reproducability of your results by the TA's.	Not reproducible. The project results should be reproducible by the TA's	N/A	With some effort the results are reproducible by the TA's.	N/A	Without any effort the results are reproducible by the TA's	N/A
Paper (30%)							
	Content	Report shows no coherence of content.	Report shows sufficient coherence of content.	Report fulfils all requirements in terms of content.	Good report in terms of content.	Very good report in terms of content.	Excellent report in terms of content.
	Form	Structure needs considerable improvement. General presentation of the content (text and figures) not very effective.	Structure need some improvement. General presentation of the content (text and figures) is sufficient.	Structure is acceptable. General presentation of the content (text and figures) is satisfactory.	Clear structure. Good presentation of the content (text and figures).	Well-structured document. General presentation of the content (text and figures) is effective.	Very well-structured document. General presentation of the content (text and figures) is very effective.
	Quality of writing	Poorly expressed. Document contains serious spelling and grammatical errors.	Reasonably expressed argumentation. Document contains some spelling and grammatical errors.	Sufficiently expressed argumentation. The document contains little spelling and grammatical errors.	Expressed and formulated well. Document has a nice flow. Document contains only minor spelling and grammatical errors.	Expressed and formulated very well. Document has a smooth flow with sufficient transitions. Document is without any spelling and grammatical errors.	Excellent expressed and formulated report. Document has a smooth flow with effective transitions. Spelling and grammatically error free.
Presentation (10%)							
	Content	Presentation lacks detail and does not support conclusions. Irrelevant information presented.	Presentation lacks detail, and is just enough to support conclusions.	Presentation has sufficient detail to support conclusions.	Presentation has a good level of detail to support conclusions.	Presentation has the right level of detail to support the conclusions and to understand the recommendations.	Presentation has the perfect level of detail to support the conclusions and to understand the recommendations.
	Form	Presentation is unstructured and chaotic. No (proper) use of visual aids.	Logical structure of presentation is poor. Improvements to the structure should be made. Use of visual aids can be improved.	Logical structure of presentation is reasonable but needs some improvement. Sufficient use of visual aids.	Presentation has good logical structure, the essentials are separated from the ancillary. Good use of visual aids.	Presentation has very good logical structure, the essentials are clearly separated from the ancillary. Good use of visual aids.	Presentation has excellent logical structure, the essentials are very well separated from the ancillary. Perfect use of visual aids.
	Performance	Poorly expressed and formulated. Unclearly presented. Audience was ineffectively addressed.	Expression and formulation can be improved. Not always clearly presented.	Expressed and formulated adequately. Most of the time clearly presented. Audience was sufficiently addressed.	Well expressed and formulated. Clearly presented. Audience was well addressed.	Very well expressed, formulated and clearly presented.	Expressed, formulated and presented with great style, clarity and effectiveness. Audience was very well addressed and engaged.