Name: Conrad Sachweh
Course: Applied Computer Science (M.Sc.)
Student number: 3082428
Date: June 24, 2018

Seminar "Explainable Machine Learning" 2018

General Data Protection Regulation and Explainable Machine Learning Challenges

Seminar Report

Conrad Sachweh

Contents

1 Introduction

1.1 Motivation

Current machine learning algorithms are broadly used in a black box fashion. The confidence in machine-learning algorithms comes from the fact that they are widely used and we only tune the parameters until the expected behavior is observed. This approach inherently leads to problems with compliance to the General Data Protection Regulation (GDPR), as every algorithm which makes decisions with data of EU-citizens, has to be explainable. First, this report is going to present a paper on GDPR and implications concerning machine learning – Section 2 and Section 3 focuses on a Machine Learning Challenge, which explicitly demands explainability of submitted algorithms.

1.2 Related Work

Explainability is not only becoming a problem now, but building explainable algorithms is an unsolved objective since a long time for artificial intelligence and machine learning processes. Because of this, various machine learning challenges already took place. A platform for this is for example *Codalab.org* [1]. Foremost they are focusing on reproducible research, but there are also challenges adding focus to explainability.

Despite the success of deep-learning algorithms today, explainability of those approaches is explored very poorly. There are explanatory mechanisms for computer vision systems which are not using deep learning [2] and there are also black box approaches [12][11][9][5][6]. It is anticipated that this research topic is growing to be very important in the future, also because of the recently introduced GDPR.

2 General Data Protection Regulation (GDPR)

The paper *European Union regulations on algorithmic decision-making and a right to explanation* [10] is presenting the implications which poses the GDPR to machine learning algorithms.

2.1 Basics

GDPR is, in contrast to its predecessor – the *Data Protection Directive*, a regulation. Therefore, it automatically is prevailing law in every country of the European Union (EU). In consequence, this enables every EU-citizen starting with the 25^{th} May 2018 to

[1] https://competitions.codalab.org/

demand information from every organization which handles their personal data. Against common intuition the law is applicable to every organization regardless of location, only the citizenship of the *data subject* (individual person) is important. In case of nonconformity, fines up to €20 million or 4 % of global revenue, whichever is higher can be enforced. The law does not handle in which way the personal data was acquired, but only how it is processed. So it is assumed that the data was acquired rightful and every data subject is now able to demand the rights, explained in the upcoming section.

2.2 Terms of GDPR

The goals of this whole package of laws is a way to a system with data-protection in mind and privacy by design.

Basis for processing: Every citizen has the right to withdraw organizations their consent. This does not only mean the consent to store the data but can also be used to restrict the use of the data for specific use-cases. Because it is not always possible for individuals to explicitly state what should be allowed or not, they also are allowed to get a log of all processing activities. Consequently every operation, where the "data subjects" data is read, must be logged and also the basis for this access. Additionally, a person responsible for this action, has to be specified. These efforts lead to improvements concerning accountability.

Responsibility and accountability: Creating recommendations for internet platform users is a very common use-case of big amounts of personal data. Every online-shop is dependent on user data to improve its sales. With GDPR it is now mandatory for organizations to be able to give an explanation why a person got a certain recommendation or not. This seems not really important for a recommendation system, but comparable processes are taking place when companies decide the eligibility for a loan or an insurance contract.

Regarding those decisions, the GDPR requests safeguards in a way that at least some human intervention must be possible. Of course the basic premise is, that the algorithm does respect the civil rights and liberties. Another severe requirement is *non-discrimination*. As we know, the world in its current situation is inherently biased towards certain demographic groups. Therefore the results of algorithms working with this big-data will naturally also be prejudiced.

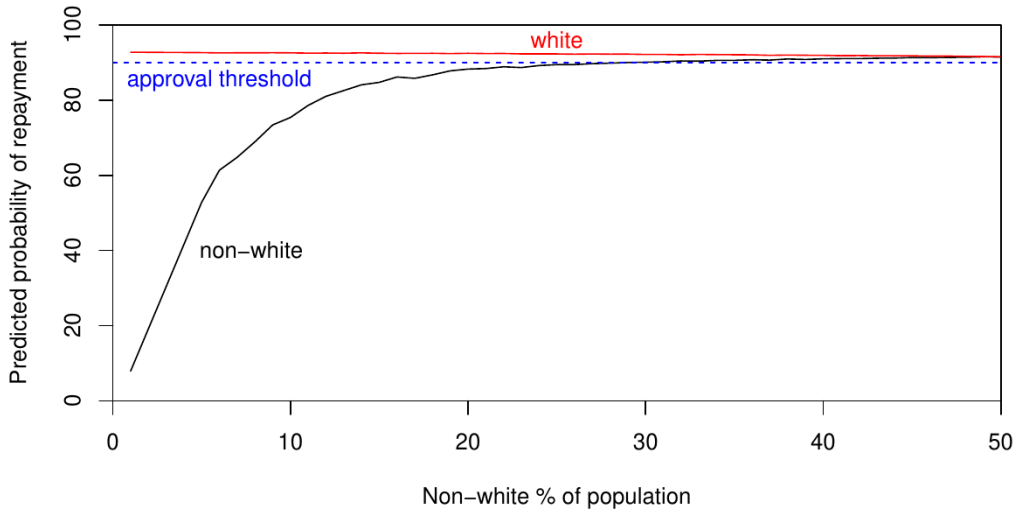

Figure 1: Discriminating underrepresented groups in training set with a risk avers logistic regression classifier. Source: [7]

Further demands of GDPR: Additionally, to the introduced rights, GDPR tries to regulate other important parts of data processing overall. Only to mention briefly, there has to be a data protection officer in every organization, which is responsible to lecture people and monitoring compliance. At the same time handling of data breaches are also standardized and the well-known *right to be forgotten* is going to be continued as the *right to erasure*.

2.3 Consequences

Whereas the promise as a system with "data protection and privacy by design" obviously is favorable for every user, GDPR certainly has some deficiencies in terms of real-world issues. For example blockchain as a contemporary hype topic is in its principles not build to be GDPR compliant, as there is no way to remove old data.

A major concern has to be made regarding the used data-sets. They are inherently biased and will lead to (unintended) discrimination. Because machine learning algorithms are based on big amounts of data, the very first step for developers to think of, has to be the selection of training data.

The simple example depicted in Figure 1 is showing the approval rate for a loan extension for the non-white population group. The training set consists of 500 people and on the X-axis the ratio is increased. Per default the initial approval value of a person is 95 %

and the approval threshold is 90 %. Because this is a risk avers logistic regression classifier underrepresented groups are penalized with a deduction of approval score which leads to discrimination solely because the group is underrepresented in the training data.

3 Explainable Machine Learning Challenge

In order to handle the legal requirements and also because it is common knowledge that algorithms which can be explained are beneficial, machine learning algorithm competitions were started. One particular interesting competition is described in the paper *Design of an explainable machine learning challenge for video interviews* [7] by Escalante et. al.

3.1 Employed Data Set

For finding large amounts of video data, the organizers used the nearly infinite amount of video blogs (vlog) stream-able on YouTube. Each of the 10 000 selected videos was split up into snippets of 15 s to represent a job interview video. Vlogs typically consist of one person talking and facing the camera directly, therefore this data-set has a high resemblance to actual interview videos.

This manufactured data-set was labeled by the service "Amazon Mechanical Turk" (human workers) for personality traits and the newly introduced *job-interview variable*. Additionally to the labels every video was transcribed by a human transcription service. Hence, the training and test data should be of satisfactory quality.

3.2 Challenge Objectives

How this, so called *first impression model* was implemented was totally up to the participants. A visualization of the pipeline which was expected to be build is visualized in Figure 2.

First impressions For a human being, first impressions are evaluated normally within 100 ms after first sight [14]. Inferences are usually based on stereotypical knowledge and are very important for our social behavior. Widely acknowledged is the segmentation into the *Big Five Personality Traits:* Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism. Recognizing these traits with algorithms is a emerging topic in image recognition and language processing.

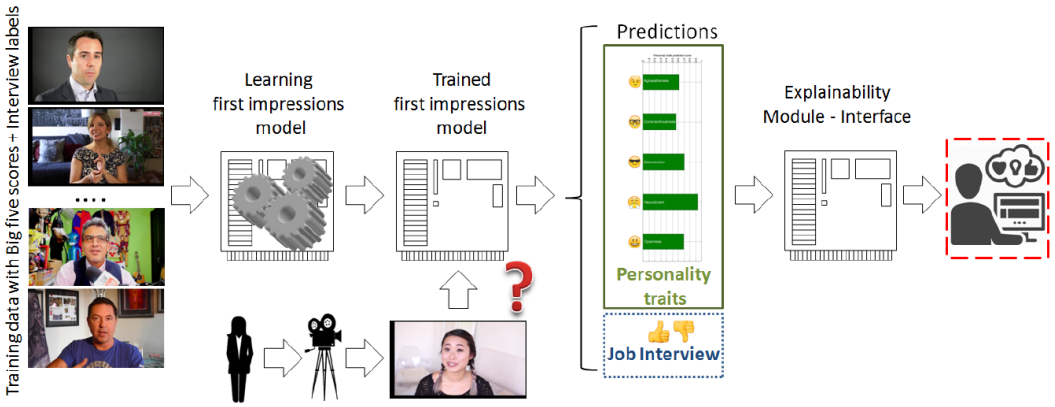

Figure 2: Challenge pipeline. Source: [7]

In addition to the first impression-traits, the *job-interview variable* had to be predicted. It must be mentioned that the ground truth decisions were also made by Amazon Mechanical Turk people, concerning videos which absolutely have nothing to do with job interviews.

At last, the "Explainability Module" should be build to enable interpretation also for non-technical people, preferred in a textual way.

3.3 Evaluation Process

All participants were given the same development data with the available ground truth. In the last week of participation the submissions were run on the competition-evaluation system against the final evaluation data in order to allow very last tweaks. The result from this *technical evaluation* is the first cornerstone for the evaluation process along with *model interpretability* and *creativity* of the approach.

As this challenge explicitly is focused on *explainability* it demands reasoning why a decision was preferred over all other possibilities. Another important property therefore is how confident the algorithm is with its decision. But also the human configuration side, why some parameters for the algorithm were chosen, has to be explained. In the end all these statements together should be condensed into a description of the decision-making of the algorithm [13].

3.4 Approaches

As this work is a successor to another challenge, the previous work of the winning submission was used as a baseline for this competition [8]. Roughly speaking the first round

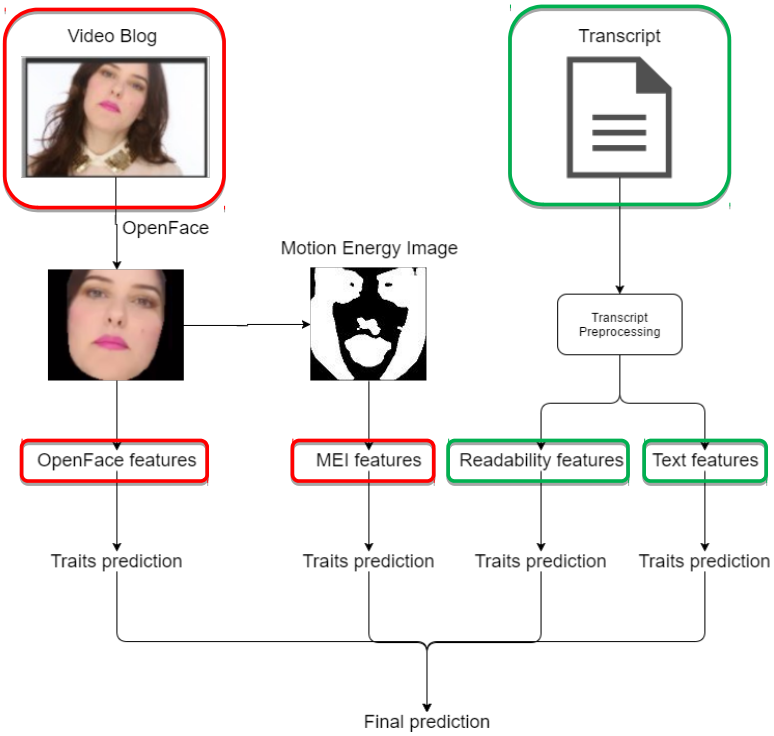

Figure 3: Winning system pipeline. Source: [13]

of this challenge were all done with a combination of well-known algorithms and the winning submission of the second round was no exception. All valid submissions performed equally good compared to the baseline.

3.5 Winning Solution

The winning approach used a pipeline which is displayed in Figure 3. They used solely the video and transcript of the video for classification. For feature extraction from the video part, Openface [1] was used [2]

3.5.1 Video Features

Alongside the different traits, Openface also extracts the face-part of the image, which can be used for additional processing. For this approach they extracted the Motion-

[2]Openface: This open-source software implements face-recognition using a deep neural network. It is written in python and uses Torch [4] as neural-network backend.

Energy Image (MEI) from the face which minimizes eventual background noise from the video, for example if it was recorded in a public space. To be more precise, the paper describes using the *weighted* MEI which is a normalized version of MEI and was introduced in the work of Biel et al. [3]. The white areas, shown in the example picture in Figure 3 are moving a lot whereas black areas are showing little movement.

The authors mention, that the approach only works for videos with a scene showing only one face, as Openface would otherwise extract all face-areas and features for any person in the scene. Nevertheless, this should not be a limitation in regards to the employed data-set.

3.5.2 Transcript Features

The transcript of the video was searched for their definition of "linguistic sophistication". This was done with 8 implementations from the Natural Language Toolkit (NLTK) [3]. The authors clearly state that the used measures were originally designed for written language and arenormally needing a lot more textual input than available with 15 s clips. Therefore they also used the measure of total word count compared to unique words to reflect complexity which thereby qualifies as "linguistic sophistication".

3.5.3 Traits Prediction

All the features extracted via the mentioned methods are now combined with a linear model to retain interpretability of this prediction step. First a Principal Component Analysis (PCA) is applied to all features and then the final prediction is made with a linear regression model.

3.5.4 Explainability

As the explainability of the approach is a needed qualification, a report for every prediction was compiled. In Figure 4 the language report is shown. Another textual report is supplied individually for the visual features and also a combined assessment (see Figure 5) was created.

4 Conclusion & Discussion

The newly created GDPR is posing a challenge to all machine learning application. But on the other hand it give us the chance to understand algorithms better, which we are

[3] https://www.nltk.org/

```
******************
* USE OF LANGUAGE *
******************

Here is the report on the person's language use:


** FEATURES OBTAINED FROM SIMPLE TEXT ANALYSIS **
Cognitive capability may be important for the job.  I looked at a few very simple
text statistics first.

*** Amount of spoken words ***
This feature typically ranges between 0.000000 and 90.000000.  The score for this
video is 47.000000 (percentile:  62).
In our model, a higher score on this feature typically leads to a higher overall
assessment score.
```

Figure 4: Description fragment. Source: [13]

```
"**************************************************
* ASSESSMENT REPORT FOR VIDEO 2c42A4Z7qPE.001.mp4: *
**************************************************

On a scale from 0.0 to 1.0, I would rate this person's interviewability
as 0.497947.
Below, I will report on linguistic and visual assessment of the person.
Percentiles are obtained by comparing the person against scores of
6000 earlier assessed people.
```

Figure 5: Final assessment. Source: [13]

already using as a black box. Of course at first this will introduce a high overhead into developing new algorithms and porting old code, but in the end a better understanding can emerge.

Needless to say that whereas the GDPR had a long development phase it only manages the data protection challenges of the internet very poorly. A lot of questions are still not fully governed and big parts are in need to be interpreted where they apply. Especially for the use of data in order to advertise products and user-tracking, lawsuits have to clarify when *consent* is given.

Concerning the approaches to building algorithms with explainability in mind, the presented approach in Section 3.5 is not outstanding and satisfactory from the standpoint of a non-technical person. But challenges like this are going to increase the likelihood of algorithms really to be able to explained. Unfortunately looking at today's proposals, we still have a long road ahead of us.

References

1. B. Amos, B. Ludwiczuk, and M. Satyanarayanan. *OpenFace: A general-purpose face recognition library with mobile applications*. Technical report. CMU-CS-16-118, CMU School of Computer Science, 2016.
2. T. Berg and P. N. Belhumeur. "How do you tell a blackbird from a crow?" In: *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE. 2013, pp. 9–16.
3. J.-I. Biel, O. Aran, and D. Gatica-Perez. "You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube." In: *ICWSM*. 2011.
4. R. Collobert, K. Kavukcuoglu, and C. Farabet. "Torch7: A Matlab-like Environment for Machine Learning". In: *BigLearn, NIPS Workshop*. 2011.
5. P. Dabkowski and Y. Gal. "Real time image saliency for black box classifiers". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6970–6979.
6. E. Elenberg, A. G. Dimakis, M. Feldman, and A. Karbasi. "Streaming weak submodularity: Interpreting neural networks on the fly". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4047–4057.
7. H. J. Escalante, I. Guyon, S. Escalera, J. Jacques, M. Madadi, X. Baró, S. Ayache, E. Viegas, Y. Güçlütürk, U. Güçlü, et al. "Design of an explainable machine learning challenge for video interviews". In: *International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 3688–3695.
8. H. J. Escalante, V. Ponce-López, J. Wan, M. A. Riegler, B. Chen, A. Clapés, S. Escalera, I. Guyon, X. Baró, P. Halvorsen, et al. "Chalearn joint contest on multimedia challenges beyond visual analysis: An overview". In: *23rd International Conference on Pattern Recognition (ICPR), 2016*. IEEE. 2016, pp. 67–73.
9. R. C. Fong and A. Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". *arXiv preprint arXiv:1704.03296*, 2017.
10. B. Goodman and S. Flaxman. "European Union regulations on algorithmic decision-making and a right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.
11. S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4768–4777.

12. M. T. Ribeiro, S. Singh, and C. Guestrin. “Why should i trust you?: Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 1135–1144.
13. A. S. Wicaksana and C. C. Liem. “Human-Explainable Features for Job Candidate Screening Prediction”. In: *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE*. IEEE. 2017, pp. 1664–1669.
14. J. Willis and A. Todorov. “First impressions: Making up your mind after a 100-ms exposure to a face”. *Psychological science* 17:7, 2006, pp. 592–598.