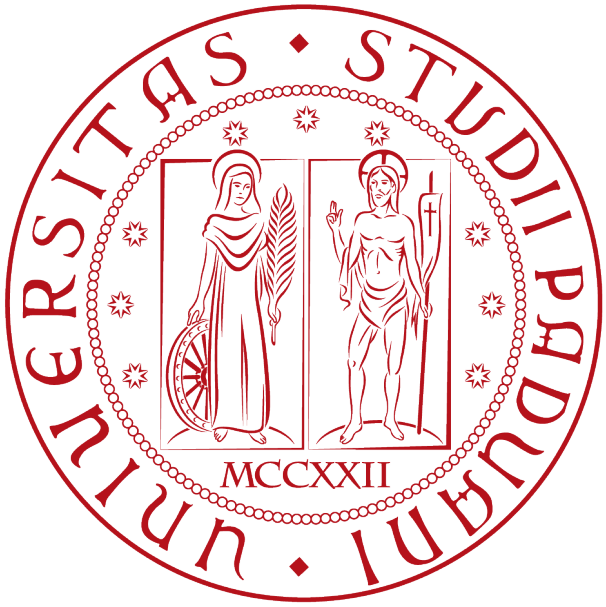

Advanced Topics in Computer Science
Trustworthy AI report

Alberto Lazari - 2089120

Computer Science – Other training activities

June 2023

1. Introduction

The aim of this report is to give a simple summary of the 2020 paper "Interpretable and Differentially Private Predictions", from F. Harder, M. Bauer and M. Park, and later make some considerations on the broader trustworthy AI context of the ideas proposed in the paper. Finally, some personal critical thoughts on the topic are left.

2. Paper

2.1. Trade-offs in ML models

The first topic addressed by the paper are trade-offs that arise when modeling machine learning systems.
There are three key features of an ML model that struggle to coexist:

1. **Interpretability**: is it possible to explain why an ML system gave a specific classification to a specific input?
2. **Privacy**: is it possible to retrieve information about the data the system was trained on, using only the system's output?
3. **Accuracy**: does the system actually work? I.e., are predictions accurate most of the times, with a suitably high probability?

Trying to increase one of these aspects, naturally lowers one of the others:

- increasing privacy may lead to more complex models, with less regular predictions that are more difficult to explain (interpretability) and are less accurate;
- increasing accuracy may reduce privacy, because better performances can probably reveal more about training data, but also make the model much more difficult to explain (neural networks, for instance).

Various works already addressed the trade-off between privacy and accuracy, but no one ever considered interpretability in the variables. Since EU's GDPR (*General Data Protection Regulation*) states that algorithms have to be interpretable and private by design it's important to find a balance between the three variables, otherwise accuracy would have to be sacrificed.

The main topic addressed by the paper is exactly that of finding a model to better equilibrate interpretability, privacy and accuracy in ML systems predictions.

2.2. LLM

To achieve the goal of balancing the trade-off the paper suggests to use a simple and analytical model, satisfying:

A. Local *and* global explanations
B. Low number of parameters, to have a good privacy – accuracy trade-off
C. Good level of expressiveness to be useful even with complex data

The proposal is to use *Locally Linear Maps* (LLMs) for such a model.

2.2.A. Local and global explanation

Using LLMs allows to provide both local and global predictions explanations.

Various methods are explainable, but only on a local scope: it is possible to explain the classification of a particular input, but not that of an entire class of inputs. With these methods the only possible way to understand the system as a whole it is necessary to test various inputs to get an idea of how it works on specific classes, but it is not possible only looking at the model itself.

LLMs allow this property instead: when analyzing the output of a specific input it is possible to quickly have a grasp of what lead to a particular classification.

When classifying an input, the model assigns weights to every single linear map. The wights themselves are a great indicator of the explanation of predictions about a single input, but also about a class as a whole. In fact, judging by the weight of a linear map it's possible to understand what class that specific linear map is sensible to.

2.2.B. Privacy vs accuracy

As stated before, there are various existing methods to get a good trade-off between privacy and accuracy. As seen in the previous section using LLM it is possible to have a good level of interpretability (both local and global). Finding a good trade-off between privacy and accuracy means that LLM provides a good trade-off between all the three variables.

Differential Privacy It is important to note that the paper considers privacy in the context of *Differential Privacy* (DP), which defines the quantity of privacy loss between two datasets that differ in the presence of one element.
The privacy loss is higher the higher is the probability of revealing the presence of that one element in the dataset, through the system's output.

The *composability* property of DP states that the composition of DP values degrades the total privacy. For this reason LLM has to reduce the number of parameters and perturb them to increase privacy, which is done combining two methods:

- **Random projections**: used to reduce the dimension of the input. It allows to have a lower number of linear maps
- **Perturbation**: noise is added to each linear map, at each step of the training

Only adding noise to the output would lead to a great loss in accuracy, this is the reason to also use random projections.

2.2.C. Expressiveness

The LLM method is comparable to other works, that reflect flexibility and expressiveness. An example, reported in the paper, is the *Mixture of Experts* (ME) model.

MEs are very similar to LLMs and both models can be directly translated into one another: local expert models represent a specialized model over a specific part of the input, just like linear maps are specific of one class of the input.

Linear maps, though, can be even more flexible when multiple linear maps are associated with a class. This allows for multiple combinations to better fit the possible inputs.

2.3. Test results

2.3.A. MNIST

When compared with different, accurate models, the LLM model gives very similar results. In a test on the MNIST database the LLM model has a mean accuracy which is less than 1% inferior than other privatized models (with the same privacy loss of $\varepsilon = 2$ and $\varepsilon = 0.5$).
This result is impressive, considering that the other existing privatized models are far less explainable.

2.3.B. Fashion-MNIST

Another test done on the Fashion-MNIST database gives less impressive results, however with a still decent accuracy (about 10% less, around 80-83% of accuracy).

2.3.C. Medical dataset

In the paper yet another test is reported, consisting in a disease classification over medical data. When compared to a non-privatized model, here LLM manages to keep an accuracy of about 2% less for the non-private and private ($\varepsilon = 1.5$) models and 4% for strongly private training ($\varepsilon = 0.2$).

Results for this test also highlight the fact that the LLM model, in every configuration (non-private, private with high and low privacy loss) has very consistent linear maps: they are very similar in each case, though incrementing the privacy makes them slightly more homogeneous (because of the added noise).

2.4. Conclusions

The actual, experimental tests prove the fact that a good trade-off between interpretability, privacy and accuracy is indeed possible, using the LLM model.

The authors, though, leave some open questions and topics for the future:

- What happens when LLM is trained on more complex data sets? Complexity limits of this model are still to be analyzed and there is much to study about how it behaves when the number of linear maps grows significantly
- Is it possible to make the model interact with more large and complex models (such as neural networks) in a privacy-preserving way?

3. About broader trustworthy AI context

When put into the broader trustworthy AI horizon, the paper contributions have some interesting improvements over some of the main topics of the subject.

When considering the *Assessment List for Trustworthy AI* (ALTAI) the greatest improvements can be seen mostly on:

- **Requirement #1**: Human Agency and Oversight
- **Requirement #3**: Privacy and Data Governance
- **Requirement #4**: Transparency

3.1. Privacy and Data Governance

This is another rock-solid assumption in the paper: the model has to be private. As with interpretability, it proved that a model can have comparable security results, but with great levels of accuracy and still keeping the interpretability.

In this case LLM is not the first model to consider this aspect, nor the best, but a virtuous example that privacy can be treated seriously, without giving up on performance.

3.2. Transparency

This is probably the most important point the paper revolves around. It starts pointing out that no other model ever considered to find a trade-off between privacy and accuracy, but with the premise that the system should be interpretable and explainable.

When compared with other, already existing models, LLM proved to reach similar results, but with the added feature of being simple and explainable, by design.
This is a great starting point for modern, next-generation models, that could incorporate such a model, or the philosophy behind it, allowing all new models to be easily understandable.

3.3. Human Agency and Oversight

Being transparent and easily explainable, the LLM model is easier than more obscure alternatives to be managed and supervised by a human. When talking about oversight especially, being able to

explain clearly the reasoning behind specific choices of the AI is important and a great advantage to have.

It is not that easy to intervene and change the AI system manually, though probably possible by tuning the linear maps weights to avoid particular answers (maybe toward bias) or make some classes more precisely identifiable.

3.4. Other requirements

Regarding the other requirements it's difficult to say whether it improves or not. It mostly depends on actual implementations and real applications.

Topics like diversity, bias, safety or accountability were not addressed by the paper, because they were outside its scope. The proposal is a generic model, but a simple and clear one, thus it may be able to improve some aspects, given it is used and implemented well.

4. Personal evaluation

I think the paper provides a very useful and interesting improvement in the theory of ML systems. What I find particularly interesting is the fact that it is not an incremental step over an already established concept. The authors really tried to provide an alternative to other models, but thinking outside of the box to give an answer to modern constraints and requirements, like GDPR.

I particularly enjoy the simplicity of the model: it's minimal and intuitive, which can help to better shape future extensions and improvements.

I think that starting to build the model with privacy and interpretability in mind should be the standard, for legal and requirements reasons, but also because from there it's possible to build more efficient and accurate implementations, while it's more difficult to increment privacy and interpretability aspects, once a model is well established and implemented in various forms.

The only disadvantage I found in the LLM model is that it seems to be *too simplistic*. As the authors allude in the conclusion, it is yet to be seen whether the model is capable of keeping the accuracy high, even with more complex data.