Humanities-Centered AI: From Machine Learning to Machine Training

Ralf Möller[1]

[1] *Universität zu Lübeck, Institute of Information Systems, Ratzeburger Allee 160, 23562 Lübeck, Germany*

Abstract

In the essay it is argued that machine learning controlled by IT specialists must be replaced with machine training, such that domain experts, e.g., humanities scholars. Furthermore, it is argued that training a machine for a particular task must have a positive impact also on related but different tasks. Only then, one can speak of "true learning" or machine education, an area that is investigated in the emerging field of Humanities-Centred Artificial Intelligence.

Keywords

Humanities, AI, Machine Learning, Machine Training,

1. The Claim

Machine learning is a hot topic these days, for good reasons [1]. Given one half of a large set of training data from a specific application context, computer scientists with a data science background can select model classes and then run algorithms, first, to automatically find appropriate (multiscale) encodings of data, and second, to automatically determine huge sets of model class parameters to derive a specific model or a specific program from training data. With automatic tests to check the performance of learning outcomes on the other half of the training data and possibly learning iterations, the learning result can be further optimized, i.e., so-called hyperparameter values can be suitably fixed or alternate components can be selected for a model class. With the idea of reinforcement learning [2], learning can also be carried out automatically at runtime, albeit within certain limits because the learning space needs to be designed at setup time. The well-known learning approaches dominate artificial intelligence (AI) because – with just training and test data provided by domain experts – computer scientists on their own can successfully build models (or programs) that can be used in specific applications to be used by domain experts. In contrast, the modeling approach to specify domain knowledge [3] requires domain experts to learn and use appropriate formal modeling languages, and usually also intensive cooperation with computer scientists is required to build appropriate models (or programs) that can be used successfully as part of applications.

While the *computer scientists work alone* learning perspective has its merits, it fails if, first,

Humanities-Centred AI (CHAI), Workshop at the 44th German Conference on Artificial Intelligence, September 28, 2021, Berlin, Germany

moeller@ifis.uni-luebeck.de (R. Möller)
https://www.ifis.uni-luebeck.de/index.php (R. Möller)
0000-0002-1174-3323 (R. Möller)

a very useful declarative model for a certain problem context is already well-established in the domain (but possibly unknown to computer scientists) and reconstructing the model by learning will produce subordinate results, or, second, the final goal of the learning process is not known in advance, i.e., a system is to be constructed that is initially configured using machine learning, and then used to carry out certain tasks, with subsequent adaptations required by the application context. Adaptations are usually not foreseeable in the beginning such that it would hardly be possible to set up a respective reinforcement mechanism at system design time. While it might be possible to incrementally specify sets of new training data and then restart machine learning processes, computer scientists remain in the loop all the time this way. Computer science expertise is a scarce resource, however. Thus, we argue that domain experts need to be enabled to control the learning process themselves. Most domain experts will, however, not be IT experts, and it is considered doubtful that often-discussed concepts for ensuring (or improving) data literacy for domain experts will ever work in practice.

The claim is that next-generation AI has the enormous potential to overcome today's computer-science perspective of system design by machine learning (including adaptation by reinforcement) and should move towards a much more powerful paradigm of systems being *trained* by domain experts in a determined way. While initial machine learning, possibly with incremental improvements by reinforcement w.r.t. a spectrum that is foreseen at design time (current AI perspective) is indeed possible and useful, the domain-expert training perspective is required to cope with adaptation requirements occurring in almost all serious applications of intelligent systems in real world contexts.

2. From Machine Learning to Machine Training

Let us consider an information retrieval (IR) scenario in a humanities research context. Current IR systems are not tailored towards a specific domain, which is good on the one hand as the engines are indeed quite versatile then. On the other hand, the catch is that current IR engines can hardly be tailored by its users, e.g., humanities scholars to fulfill specific needs. Next-generation AI should enable scholars, as examples for non-IT-specialists, to train intelligent systems on the job.

Providing a set of documents as a reference library and, in addition, a query string, will allow the intelligent agent [4] behind a search engine, first, to focus on documents in the respective repository that match the search string and, second, to relate documents to similar or complementary reference library documents by exploiting, e.g., topic models [5]. Given reference library documents selected in beforehand by a scholar, the agent is trained w.r.t. topic models and, possibly, w.r.t. interesting datasets that are referred in the documents of the reference library. Thus, not only documents can be found, but also related datasets, be they similar or complementary. If the reference library is extended (or adapted) the agent can explain the corresponding changes of result sets w.r.t. changes in the result of previous queries. This way, the trainer, namely the humanities scholar, can directly see the effect of providing a reference library. In the case of humanities research, this kind of training to build topic models will enable the search agent underneath to also compare heterogeneous datasets based on the association of the datasets with the papers that discuss or describe them.

3. Long-Term Perspective: From Machine Training to Machine Education and True Learning

In general, we assume that the agent manages to exploit the directives of humanities scholars by updating an internal model $\mathcal{M}$, a search string and a topic model in the example above. With new documents provided as an extension to the reference library, the model $\mathcal{M}$ is converted into a new model $\mathcal{M}'$. With the updated $\mathcal{M}'$, the notion of relatedness between documents is adapted. While this is an important effect, it is merely a local effect, namely an effect on the effectiveness of topic-based information retrieval. Local effects on a single model define the standard learning mode on which many, if not all, artificial intelligence learning scenarios are based. Transforming $\mathcal{M}$ into $\mathcal{M}'$ via machine training can, however, hardly be seen as *education* because the learning effect is indeed visible only for the model of a very specific task that the agent carries out. The question is: How can updates to the reference library and the corresponding improvements in IR via an updated model $\mathcal{M}'$ be exploited to positively affect the fulfillment of other goals the agent might be given, that is, other goals alongside IR goals? On the other hand, if other models not related to information retrieval are updated, this should indeed also have effects on information retrieval. Only if (significant) effects of the update of a goal-specific model $\mathcal{M}$ to $\mathcal{M}'$ on other goals were achieved, machine training would be effective in the long run, and only then we can talk about *education*. Thus, being *educated* requires that training w.r.t. a specific task (or goal) also has some (positive) impact on other tasks (or goals) of the agent, which is the intrinsic idea of *education*. Learning resulting in education is called true learning. The question is: how can this be achieved?

To solve the tasks assigned to it an agent uses algorithms for solving problems defined on a model to be used for a task. To be more precise, usually the problems are defined w.r.t. the interpretation $\mathcal{I}(\mathcal{M})$ of the model $\mathcal{M}$. The function $\mathcal{I}$ defines the semantics of the model, and we write $\mathcal{M}^{\mathcal{I}}$ instead of $\mathcal{I}(\mathcal{M})$ for brevity. Please note that the semantics given by $\mathcal{I}$ is just used to define computational problems, there is no need to compute the value $\mathcal{I}(\mathcal{M})$, which might even be infinite. A small example is appropriate here.

Let us assume that $\mathcal{M}_{\tau_1}$ is a search string plus a topic model as indicated above. Both are used to specify an information need in an information retrieval task τ_1. In this case, $\mathcal{M}_{\tau_1}^{\mathcal{I}}$ is the set of documents from a repository that match the search string $\mathcal{M}_{\tau_1}$ and the topic model. An (inference) problem is to check whether a given document d found in the repository is in $\mathcal{M}_{\tau_1}^{\mathcal{I}}$ (relevance decision problem $d \in \mathcal{M}_{\tau_1}^{\mathcal{I}}$). Checking the relevance decision problem for all documents d in a repository and returning all d for which this is the case, is called the information retrieval problem. The information retrieval problem is used to formalize task τ_1[1].

Let us further assume that for (or while) carrying out another task τ_2, new knowledge about synonyms for words is made available to the agent via training, i.e., another model $\mathcal{M}_{\tau_2}$ is extended (or adapted) to obtain $\mathcal{M}'_{\tau_2}$. If the effect of training an agent w.r.t. a task τ_2 by transforming $\mathcal{M}_{\tau_2}$ into $\mathcal{M}'_{\tau_2}$ is to be called education, we must make sure that the change to $\mathcal{M}_{\tau_2}$ by learning is also effective for problem solving used for carrying out other tasks, information retrieval τ_1 say. We now assume that formal problems used to solve τ_1 are defined w.r.t. $\mathcal{M}_{\tau_1}^{\mathcal{I}}$ as indicated above. Since $\mathcal{M}_{\tau_1}^{\mathcal{I}}$ is not changed, or the change of $\mathcal{M}_{\tau_2}$ into $\mathcal{M}'_{\tau_2}$ to be

[1] We neglect ranking here.

effective on τ_1 the change must indeed have an influence on the interpretation $\mathcal{I}$ used to define the information retrieval problem. In contrast to standard learning processes that transform $\mathcal{M}_{\tau_2}$ into $\mathcal{M}'_{\tau_2}$, education processes should not only transform $\mathcal{M}_{\tau_2}$ into $\mathcal{M}'_{\tau_2}$ but should also transform $\mathcal{I}$ into $\mathcal{I}'$. Now, if $\mathcal{I}$ was transformed into $\mathcal{I}'$, this would mean that knowledge about synonyms acquired for task τ_2 is used for the information retrieval task τ_1. Indeed, when now $\mathcal{I}'$ is used in problems for τ_1, we still have $\mathcal{M}_{\tau_1}$ but need to deal with $\mathcal{M}_{\tau_1}^{\mathcal{I}'}$ that possibly denotes a larger (or adapted) set of documents. Changing $\mathcal{M}_{\tau_2}$ into $\mathcal{M}'_{\tau_2}$ can be seen as *education* when, as a by-product, $\mathcal{I}$ is transformed into $\mathcal{I}'$ and then $\mathcal{I}'$ is used further on to adapt the semantics of the unchanged model $\mathcal{M}_{\tau_1}$ used in the IR task τ_1. This is what we have in mind when talking about *educating by training* and *true learning*. Algorithms for solving the information retrieval problem problems based on $\mathcal{I}'$ need to be automatically adapted to *realize* $\mathcal{I}'$, and it is all but clear how this can be accomplished.

4. True Learning in the Humanities

In the use of digital libraries, where humanities scholars often have to choose from a variety of search criteria, not only do problems arise in users' choice of effective criteria, but users are also characterized by different personas [6]. That is, the same information need exists for the different personas, but the model interpretation $\mathcal{M}^{\mathcal{I}}$ varies and leads to different results for the classical approaches such as retrieving the goal-specific model $\mathcal{M}'$ from $\mathcal{M}$, although the results should be the same here. The problem of identifying the interpretation needs of humanities scholars via personas has so far been attempted to be solved via approaches such as information seeking [7]. However, as argued above, it is a form of machine training which has the previously identified personas as input.

True learning is, for example, to identify the personas based on $\mathcal{I}$, and then transform $\mathcal{I}$ to $\mathcal{I}'$ to identify the information need of the humanities scholars. Another approach for true learning in the Humanities is, for example, to identify the humanities scholars' context, which could be represented as $\mathcal{I}$ initially. In addition, $\mathcal{I}$ and the search string $\mathcal{M}'_{\tau}$ must be treated differently in the field of the Humanities than in classical information retrieval approaches because the human understanding of a term and the treatment of the same term by the computer is different [7].

We argue, however, that artificial intelligence must evolve into the direction illustrated above to support true learning while still being beneficial [8], a direction that is investigated in the new field of *Humanities-Centred Artificial Intelligence* (CHAI).

References

[1] A. V. Joshi, Machine Learning and Artificial Intelligence, Springer, 2020.

[2] R. Sutton, A. Barto, Reinforcement Learning: An Introduction (2nd Ed.), MIT Press, 2018.

[3] C. M. Bishop, Model-based machine learning (2012). doi:10.1098/rsta.2012.0222, 371(1984).

[4] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach (4th Edition), Pearson, 2020.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[6] M. Al-Shboul, A. Abrizah, Information Needs: Developing Personas of Humanities Scholars, in: The Journal of Academic Librarianship, 2014. doi:10.1016/j.acalib.2014.05.016.

[7] G. Buchanan, S. J. Cunningham, A. Blandford, J. Rimmer, C. Warwick, Information seeking by humanities scholars, in: A. Rauber, S. Christodoulakis, A. M. Tjoa (Eds.), Research and Advanced Technology for Digital Libraries, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 218–229.

[8] S. Russell, Human Compatible: AI and the Problem of Control, USA: Viking Press, 2019.