databooster

an initiative of

doi

data innovation alliance

Databooster Call for participation: ML model and dataset licensing

Description

In many applications of machine learning (ML), foundation models are becoming the starting point for downstream analyses and tasks. For example, any text-bound task will often begin with a large language model, which has been pre-trained on vast text corpora scraped from the web. Subsequently, these models are tuned on a few annotated samples for the specific task at hand. This unsupervised pre-training ensures that the model has learned good representations, such that the concrete task can be solved with only a few annotated samples.

Recently, we have observed a trend of big tech companies limiting the licensing of foundation models to non-commercial uses. While this gatekeeping is understandable from a business perspective, it poses a significant challenge for innovative small to mid-sized companies, given that they rely heavily on the results of state-of-the-art research, which is dominated by these very companies.

At the same time, it is unclear how licensing of public datasets impacts the usability of models that used these datasets, e.g., is a model that was trained on restrictive data still subject to the same license restrictions, even if it was trained on other (more permissive) data?

Given these observations, we foresee a need to clarify the impact of ML model and dataset license restrictions and the extent to which companies within the data innovation alliance are facing similar challenges and how they are addressing them.

Objectives

1. To determine the extent to which small to mid-sized companies are encountering the challenge of restrictive licenses on ML models and datasets.

2. To understand the implications of combining various license texts in a given ML project, e.g., training a *MIT* licensed model on a non-commercial creative commons license (such as *CC BY-NC-SA 4.0*).

3. To identify amendments to and/or modifications of original licensing texts after, e.g., finetuning pre-trained models on internal data, making architectural changes, etc.

4. To explore the possibility of forming a collaboration to curate data and/or pre-train models on public datasets, if there is sufficient interest among the participants.

Conclusion:

The ML model and data landscape is less and less open posing grave challenges to small- and mid-size companies with limited computational resources. We consider it a timely question to address at this point and hope that by leveraging the data booster program, a future bottleneck due to restrictive licensing can be evaded.

Requested support

Companies might find themselves in a similar situation, which could pave the way to a collaborative effort in the future.

Research partners could provide the necessary legal expertise.