Contribution ID: **68** Type: **Poster only**

Licensing, Copyright and AI Training Data: Evidence from 2 Billion Images (Poster Upload)

Monday 23 September 2024 17:14 (1 minute)

We explore the implications of copyright on the composition, quality, and, therefore, inherent bias of AI training datasets. We study the example of LAION5B, a dataset of 5 billion images from the web, which is widely used in research and commercial applications of ML, including generative AI. We document the extent to which images in this dataset have a clear license status and to which extent it may be possible to infer the rightsholder. With this, we can compare the characteristics of a subset of images with either permissive licenses and/or identifiable rightsholders to the full dataset and estimate how the distribution of images and depicted concepts would change if copyright were enforced at scale. This allows us to contribute much-needed empirical evidence to a discussion of the role of copyright in enabling or restricting the availability of unbiased training datasets.

What of the following keywords match your abstract best?

Other

Please tick if you are a PhD student and wish to take part to the poster prize competition!

I am a PhD student

Authors: Prof. PEUKERT, Christian (University of Lausanne, HEC); Mr DALAUD, Irvin (University of Lausanne, HEC); SHCHERBAKOV, Viktor (University of Lausanne, HEC)

Presenters: Prof. PEUKERT, Christian (University of Lausanne, HEC); Mr DALAUD, Irvin (University of Lausanne, HEC); SHCHERBAKOV, Viktor (University of Lausanne, HEC)

Session Classification: Flash talks / poster session