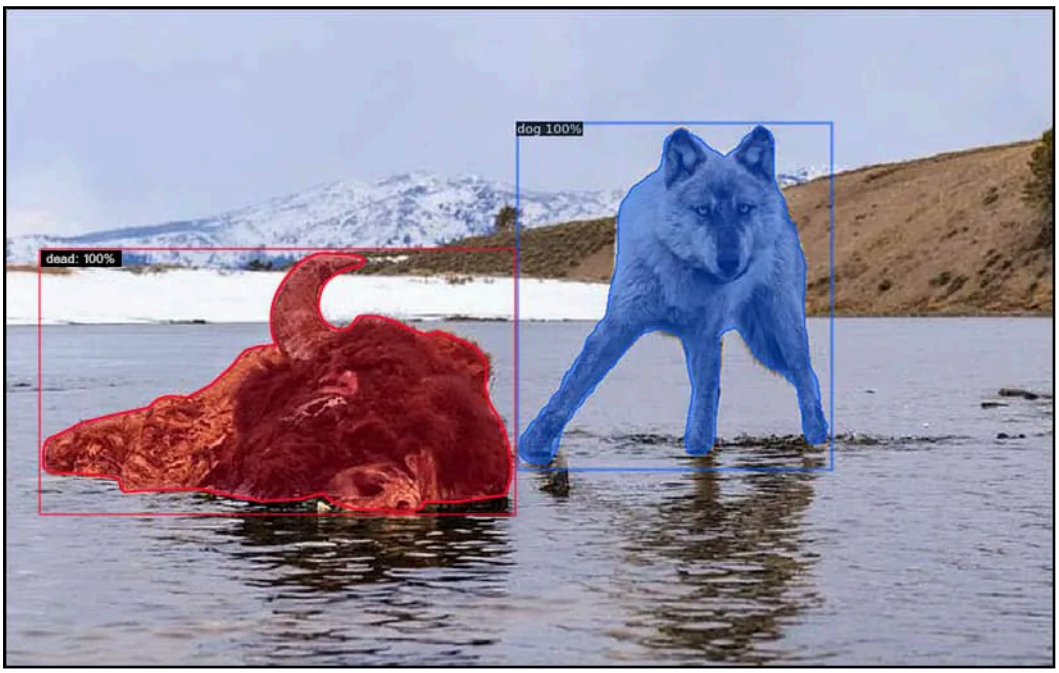
Current AI Practices Could Be Enabling a New Generation of Copyright Trolls

Originally published at https://www.unite.ai/current-ai-practices-could-be-enabling-a-new-generation-of-patent-trolls/



A new research collaboration between Huawei and academia suggests that a great deal of the most important current research in artificial intelligence and machine learning could be exposed to litigation as soon as it becomes commercially prominent, because the datasets that make breakthroughs possible are being distributed with invalid licenses that do not respect the original terms of the public-facing domains from which the data was obtained.

In effect, this has two almost inevitable possible outcomes: that very successful, commercialized AI algorithms that are known to have used such datasets will become the future targets of opportunistic patent trolls whose copyrights were not respected when their data was scraped; and that organizations and individuals will be able to use these same legal vulnerabilities to protest the deployment or diffusion of machine learning technologies that they find objectionable.

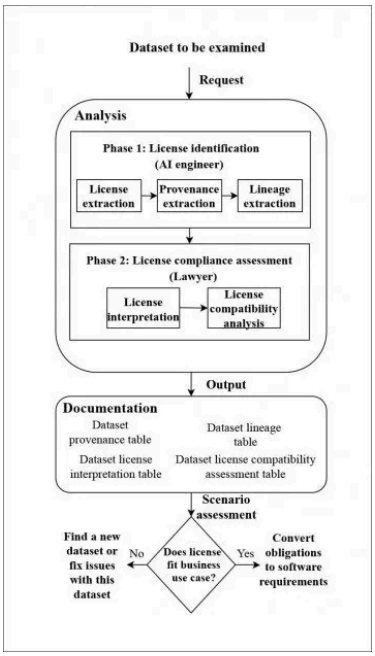
The paper is titled *Can I use this publicly available dataset to build commercial AI software? Most likely not*, and is a collaboration between Huawei Canada and Huawei China, together with York University in the UK and the University of Victoria in Canada.

Five Out of Six (Popular) Open Source Datasets Not Legally Usable

For the research, the authors asked departments at Huawei to select the most desirable open source datasets that they would like to exploit in commercial projects, and selected the six most requested datasets from the responses: CIFAR-10 (a subset of the *80 million tiny images* dataset, since withdrawn

for ‘derogatory terms’ and ‘offensive images’, though its derivatives proliferate); ImageNet; Cityscapes (which contains exclusively original material); FFHQ; VGGFace2, and MSCOCO.

To analyze whether the selected datasets were suitable for legal use in commercial projects, the authors developed a novel pipeline to trace back the chain of licenses as far as was feasible for each set, though they often had to resort to web archive captures in order to locate licenses from now-expired domains, and in certain cases had to ‘guess’ the license status from the nearest available information.



Architecture for the provenance-tracing system developed by the authors. Source: https://arxiv.org/pdf/2111.02374.pdf

The authors found that licenses for five out of the six datasets *‘contain risks associated with at least one commercial usage context’*:

> *‘[We] observe that, except MS COCO, none of the studied licenses allow practitioners the right to commercialize an AI model trained on the data or even the output of the trained AI model. Such a result also effectively prevents practitioners from even using pre-trained models trained on these datasets. Publicly available datasets and AI models that are pre-trained on them are widely being used commercially.’* *

The authors further note that three of the six studied datasets could additionally result in license violation in commercial products if the dataset is modified, since only MS-COCO allows this. Yet data augmentation and sub-sets and super-sets of influential datasets are a common practice.

In the case of CIFAR-10, the original compilers did not create any conventional form of license at all, only requiring that projects using the dataset include a citation to the original paper that accompanied the release of the dataset, presenting a further obstruction to establishing the legal status of the data.

Further, only the CityScapes dataset contains material which is exclusively generated by the originators of the dataset, rather than being ‘curated’ (scraped) from network sources, with CIFAR-10 and ImageNet using multiple sources, each of which would need to be investigated and traced back in order to establish any kind of copyright mechanism (or even a meaningful disclaimer).

No Way Out

There are three factors that commercial AI companies seem to be relying upon to protect them from litigation around products that have used copyrighted content from datasets freely and without permission, to train AI algorithms. None of these afford much (or any) reliable long-term protection:

1: Laissez Faire National Laws

Though governments around the world are compelled to relax laws around data-scraping in an effort not to fall back in the race towards performant AI (which relies on high volumes of real world data for which regular copyright compliance and licensing would be unrealistic), only the United States offers full-fledged immunity in this respect, under the Fair Use Doctrine – a policy that was ratified in 2015 with the conclusion of Authors Guild v. Google, Inc., which affirmed that the search giant could freely ingest copyrighted material for its Google Books project without being accused of infringement.

If the Fair Use Doctrine policy ever changes (i.e. in response to another landmark case involving sufficiently high-powered organizations or corporations), it would likely be considered an *a priori* state in terms of exploiting current copyright-infringing databases, protecting former use; but not *ongoing* use and development of systems that were enabled through copyrighted material without agreement.

This puts the current protection of the Fair Use Doctrine on a very provisional basis, and could potentially, in that scenario, require established, commercialized machine learning algorithms to cease operation in cases where their origins were enabled by copyrighted material – even in cases where the model's weights now deal exclusively with permitted content, but were trained on (and made useful by) illegally copied content.

Outside the US, as the authors note in the new paper, policies are generally less lenient. The UK and Canada only indemnifies the use of copyrighted data for non-commercial purposes, while the EU's Text and Data Mining Law (which has not been entirely overridden by the recent proposals for more formal AI regulation) also excludes commercial exploitation for AI systems that do not comply with the copyright requirements of the original data.

These latter arrangements mean that an organization can achieve great things with other people's data, up to – but not including – the point of making any money out of it. At that stage, the product would either become legally exposed, or arrangements would need to be drawn up with literally millions of copyright holders, many of whom are now untraceable due to the shifting nature of the internet – an impossible and unaffordable prospect.

2: Caveat Emptor

In cases where infringing organizations are hoping to defer blame, the new paper also observes that many licenses for the most popular open source datasets auto-indemnify themselves against any claims of copyright abuse:

> *'For instance, ImageNet's license explicitly requires practitioners to indemnify the ImageNet team against any claims arising from use of the dataset. FFHQ, VGGFace2and MS COCO datasets require the dataset, if distributed or modified, to be presented under the same license.'*

Effectively, this forces those using FOSS datasets to absorb culpability for the use of copyrighted material, in the face of eventual litigation (though it does not necessarily protect the original compilers in a case where the current climate of 'safe harbor' is comprised).

3: Indemnity Through Obscurity

The collaborative nature of the machine learning community makes it fairly difficult to use corporate occultism to obscure the presence of algorithms that have benefited from copyright-infringing datasets. Long-term commercial projects often begin in open FOSS environments where the use of datasets is a matter of record, at GitHub and other publicly-accessible forums, or where the origins of the project have been published in preprint or peer-reviewed papers.

Even where this is not the case, model inversion is increasingly capable of revealing the typical characteristics of datasets (or even explicitly outputting some of the source material), either providing proof in itself, or enough suspicion of infringement to enable court-ordered access to the history of the algorithm's development, and details of the datasets used in that development.

Conclusion

The paper depicts a chaotic and ad hoc use of copyrighted material obtained without permission, and of a series of license chains which, followed logically as far back as the original sourcing of the data, would require negotiations with thousands of copyright holders whose work was presented under the aegis of sites with a wide variety of licensing terms, many precluding derivative commercial works.

The authors conclude:

> *'Publicly available datasets are being widely used to build commercial AI software. One can do so if [and] only if the license associated with the publicly available dataset provides the right to do so. However, it is not easy to verify the rights and obligations pro-vided in the license associated with the publicly available datasets. Because, at times the license is either unclear or potentially invalid.'*

Another new work, entitled *Building Legal Datasets*, released on November 2nd from the Centre for Computational Law at Singapore Management University, also emphasizes the need for data scientists to recognize that the 'wild west' era of ad hoc data gathering is coming to a close, and mirrors the recommendations of the Huawei paper to adopt more stringent habits and methodologies in order to ensure that dataset usage does not expose a project to legal ramifications as the culture changes in time, and as the current global academic activity in the machine learning sector seeks a commercial return on years of investment. The author observes*:

> *'[The] corpus of legislation affecting ML datasets is set to grow, amid concerns that current laws offer insufficient safeguards. The draft AIA* [EU Artificial Intelligence Act]*, if and when passed, would significantly alter the AI and data governance landscape;other jurisdictions may follow suit with their own Acts. '*

* *My conversion of inline citations to hyperlinks*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai