Training Foundation Models as Data Compression: On Information, Model Weights and Copyright Law

Anonymous Authors[1]

1. Introduction

Generative models training with self-supervised learning, i.e., by maximizing the likelihood of training data, as commonly done for foundation models (Bommasani et al., 2021) such as large language models (LLMs) (Bubeck et al., 2023; Gemini Team et al., 2023; Touvron et al., 2023) and diffusion models (Ho et al., 2020; Rombach et al., 2022), can be seen as a form of *(lossy or lossless) compression* (MacKay, 2003). From this perspective, the training algorithm plays the role of the compression algorithm; the inference (feed-forward) algorithm is the de-compression algorithm (with the input passed to the model working as a decoding key); and the model's weights represent the compressed version of the training set.

Deletang et al. (2024) discuss how a language model can implement a lossless compression process in an offline setting, i.e., through a fixed set of model parameters derived from training. We move a step further and claim the self-supervised training to be a lossy or lossless compression process, during which the whole training set is encrypted into the model's weights. This is demonstrated by the fact that the model can reproduce certain portions of training samples (Carlini et al., 2023; Kandpal et al., 2022), which can cause issues at the copyright level (Henderson et al., 2023). It is almost impossible to decompress each possible training sample perfectly and in its entirety, i.e., without any loss of information. Nonetheless, certain training samples can indeed be retrieved (Carlini et al., 2021), and more advanced techniques might lead to an even higher degree of "retrievability". This suggests the idea that backpropagation and stochastic gradient descent make the model's weights the best possible compressed version of the training set or, more correctly, batches of it at a time. The training-as-compressing analogy is summarized in Fig. 1.

However, such models are commonly used as a starting point for additional training rather than to generate outputs directly: the pre-trained model can be fine-tuned for downstream tasks (Tunstall et al., 2022) or inducing desired behaviors, e.g., to align it with human preferences (Leike et al., 2018). The analogy depicted above in the figure can be extended to fine-tuned models as well, with the only difference being that the source data would be both the training set and the pre-trained model's weights.

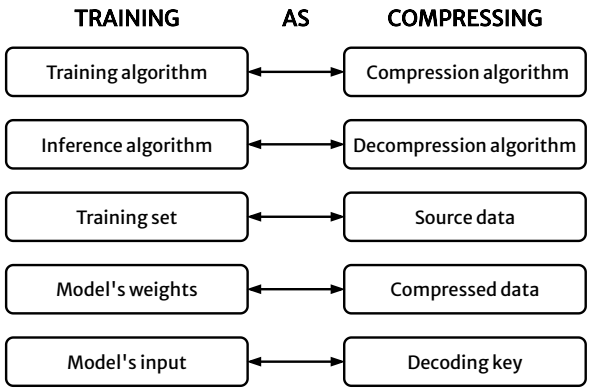

Figure 1. The *training-as-compressing* analogy. The training set is compressed into the model's weights via a training algorithm; the source data can be retrieved using the appropriate model's input.

2. Training-as-Compressing and Copyright

Though perhaps not particularly relevant for developing such models, the *training-as-compressing* perspective can shed new light on the open copyright issues related to generative modeling. While the software code responsible for the training and inference of a generative model can fall under copyright law as a computer program and the algorithmic method is a mathematical model and thus not protected (World Trade Organization, 1994), whether the model's weights can be protected or not is an open question. Indeed, if the model's weights represent a compressed version of the training set, and the training set is protected by copyright laws, then the weights are also subject to them. Assuming that the training set is protected in some ways (we will discuss it later), the weights can thus be seen as either a) a lossy or lossless compressed copy of it, or b) (a lossy compressed version of) a derivative work of it.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Seeing the weights as a mere compressed copy of the training set (not different from a zipped file) is seducing since the weights are meant to contain all the information necessary to reconstruct the original samples given a certain input (i.e., the decoding key). However, the final result is usually lossy, and the common scenario is that what we obtain after decompression is similar, but not exactly equal, to the original work. If the differences are not substantial, then it can still be considered a copy; however, it can also lead to a non-negligible modification or transformation of the training data. This second option seems to match the definition of derivative works. This opens up a different perspective: what the weights are compressing might not be the original training set, but a new, derivative work (substantially different from, but still based on, the original) whose creation happens concurrently with weights' learning and whose only existence is due to the weights themselves. Nonetheless, a derivative work must still satisfy the originality requirement to be protected by copyright. Whether or not the trainers' role in choosing data, algorithms, and parameters is sufficient for claiming authorship (and thus protection) of the model's weights is still an open question.

Until now, we have assumed that the training set is protected under copyright law. The whole training set can be protected as a database or a collective work, i.e., a collection of separate and independent works (Lee et al., 2024). However, the collective work must constitute an intellectual creation because of the selection and arrangement of its content; the same criteria also apply to databases. The current trend for training foundation models seems to go in the opposite direction. Although a certain degree of data pre-processing is always present, the common choice is to collect as much data as possible for example from the Web, thus threatening the requirement of making a careful and original selection or arrangement. Fine-tuning data sets or training sets for more specific domains are more likely to be eligible for protection as collective works. Still, this interpretation does not seem to cover the usual foundation models' training sets. On the other hand, single training samples are often protected under copyright law (Bandy & Vincent, 2021). Even though the training goal aims to compress batches of samples at a time, thus potentially leading to a compression that is optimal for a subset of works when considered together but not when considered separately, the single works can still be decompressed from the resulting model, at least in principle. This suggests that the model's weights can be interpreted as a copy (or a derivative work) of all the independent training samples, and not only of the training set as a whole.

3. Implications

Understanding the model's weights as a copy or a derivative work of protected works leads to two crucial implications. First, it provides a legal framework to understand them, removing the veil of uncertainty surrounding this issue. While it is hard to claim that they can be protected as a derivative work under copyright law due to a lack of valid authorship (Otero, 2021), protection can still be applied by considering the file containing the model's weights as a database. Indeed, they are a collection of floating point numbers, which can be retrieved independently; the significant investments required for obtaining them make the model's weights eligible for the sui generis right (thus providing certain rights to those who have invested in the database constitution independently from its copyright protection) (Sousa e Silva, 2024). In other words, the *sui generis right* can protect the investment; our copyright perspective can link the model's weights back to the training data, providing a new perspective over one of the several issues concerning the generative-AI supply chain (Lee et al., 2024). The same considerations still hold in the case of a fine-tuned model. According to Lee et al. (2024), it would be considered as a derivative work of the pre-trained model (and also of the fine-tuning data), thus simply making fine-tuning an additional link in the chain.

Second, it provides a potential framework for works generated by the model. Indeed, decompressing the information from the model might be seen as producing a derivative work of the weights, thus a derivative work of a copy of a protected work or a derivative work of a derivative work of (a copy of) a protected work. Either way, this link between the output and the training data may help enforce their copyrights. It is worth noting that the EU text and data mining (TDM) exceptions (European Union, 2019) apply for TDM purposes such as training the model, therefore for making that copy or derivative work; however, they do not apply for further derivative works from the model. The main consequence is that authorization from the training set's rightsholders would be required (or else the reproduction or adaptation right would be triggered), allowing for agreements for compensation of training data authors. In addition, generated works would need to respect the moral rights of training data, even when their economic rights have expired. Whether or not such a new derivative work is protected by copyright will depend on the human contribution (i.e., the input to the model), in particular, on its substantiality and its being the main contribution of the originality (Guadamuz, 2017; Franceschelli & Musolesi, 2022).

We believe that a multi-disciplinary analysis of these problems and implications from the point of view of Information Theory is of fundamental importance for practitioners and researchers from both the technological and legal points of view. Our research agenda also includes a rigorous formalization of the problem as a basis for rigorous legal analysis of this complex yet fascinating area.

References

Bandy, J. and Vincent, N. Addressing "documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Proc. of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *Proc. of the 30th USENIX Security Symposium (USENIX Security 21)*, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *Proc. of the 11th International Conference on Learning Representations (ICLR'23)*, 2023.

Deletang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, M., Orseau, L., Hutter, M., and Veness, J. Language modeling is compression. In *Proc. of the 12th International Conference on Learning Representations (ICLR'24)*, 2024.

European Union. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, 2019.

Franceschelli, G. and Musolesi, M. Copyright in generative deep learning. *Data & Policy*, 4:e17, 2022.

Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Guadamuz, A. Do androids dream of electric copyright? comparative analysis of originality in artificial intelligence generated works. *Intellectual Property Quarterly*, 2:1–24, 2017.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS'20)*, 33:6840–6851, 2020.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *Proc. of the 39th International Conference on Machine Learning (ICML'22)*, 2022.

Lee, K., Cooper, A. F., and Grimmelmann, J. Talkin' 'bout AI generation: Copyright and the generative-AI supply chain, 2024. arXiv:2309.08133 [cs.CY].

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction, 2018. arXiv:1811.07871 [cs.LG].

MacKay, D. J. *Information Theory, Inference and Learning algorithms*. Cambridge University Press, 2003.

Otero, B. G. Machine Learning models under the Copyright microscope: Is EU Copyright fit for purpose? *GRUR International*, 70(11):1043–1055, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, pp. 10684–10695, 2022.

Sousa e Silva, N. Are AI models' weights protected databases?, 2024. https://copyrightblog.kluweriplaw.com/2024/01/18/are-ai-models-weights-protected-databases/ [last access: May 27, 2024].

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tunstall, L., von Werra, L., and Wolf, T. *Natural Language Processing with Transformers*. O'Reilly, 2022.

World Trade Organization. Agreement on Trade-Related Aspects of Intellectual Property Rights, 1994.