Alithis

Decentralized traceability for datasets and machine learning models

Eli Vasquez

March 2023

Abstract

This paper proposes a protocol for identifying the datasets used in machine learning models, even in cases where the source of the data is not readily apparent. The protocol leverages blockchain technology and decentralized storage to create a tamper-proof ledger that records the usage of each dataset in machine learning models. The ledger is publicly accessible, providing transparency and promoting openness in machine learning research, while also protecting the privacy of dataset owners. This protocol has the potential to improve trust in machine learning models and promote responsible data usage in AI development

Contents

Introduction

Machine learning models have shown their worth getting better with every iteration and have been improving whenever more data has been loaded to them. Given the recent explosion on generative models and how indistinguishable the generated content is from actual human work a recent concern has been to identify AI generated content and even some governments have prohibited the use of such tools.We take advantage of the improvements in such fields and even though we're privacy concerned, we rarely think about the origin of the data on such models. Since the creation of style transfer learning AI can replicate an artist's work without hassle then diffusion models are capable of generating detail so precise that even watermarks from photo stock were imprinted into the generated content, leaving no doubt on where the training examples were taken.

It's well known and obvious that machine learning models need to ingest data to operate, as that's where they learn patterns. And the quality and power on the model resides on the quality and quantity of the data, this raises the need on some machine learning models to currently ingest new data, and as stated before feeding the model bad data will return not so good results which on itself raises the need of storing a record of what data what was used in what model, yet this information has had the focus on aiding the modeler on having a reference to better tune the model if needed. Providers of such metadata are not entitled to make this information public, and also companies creating it would argue that they want to keep the company or user's privacy. Examples of service providers to the above mentioned are here.

When using a social network, we're consenting to use our data, and for sure the recommendations on them are models created with OUR data, there is no doubt about, but there are cases, such as the one with Facebook and Cambridge Analytica, where data from users has had a questionable impact on society.

There is an intrinsic need for US users of social networks to know WHEN and WHERE our data has been used.

We're proposing a protocol to identify which machine learning models have used what dataset for cases not so obvious as the case mentioned above. We're aiming both for privacy and openness. We rely on smart contracts and persistent decentralized storage to fulfill our goals.

Operation

We need to precisely identify each dataset/data point, a data duplicity operation would not be cost effective or scalable, an efficient manner of marking the data is the hash function. The hash function takes an arbitrary numerical input and returns another fixed length pseudo random numerical data. This is also a one way computation, meaning that we will always get the same hash value for the same image, but we can not get the source data form the hashed value the purpose of this is twofold as we will ensure privacy is kept by not disclosing the source data and this is scalable compared to the original source upload.

We're also aware that there are a lot of data transformations(e.g. image rotation translation) for machine learning models and such transformation of the data would result in a different hash identifier, then we need to rely on the Alithis user to make sure it is using the correct hash. A fist proof of concept is developed in this collab notebook, which we are reading the flower dataset model computing the hash and uploading it to lighthouse.storage our provider for this proof of concept. Look for the example attached here on this link. Here we are creating a manifest JSON file with a hash for each image.Also we computed an image classification model and uploaded a similar manifest with the model metadata. One particularly interesting fact to consider is that the data is shuffled when the model learning occurs, then as the weights are computed differently hashing the binary output will result in different hashes each time the model is created.. We're leaving the verification task to the reader for this version.

Another more interactable demo is found on the alithis mnist DAPP, this version works with the MNIST dataset and also has each of the handwritten images hashed and uploaded to lighthouse, our DAPP lets users login with their wallet address, verify an image hash and write such results in the blockchain.

It is essential that we internet users ask our data processing companies to provide ways for *data usage verification*.

Roadmap

This document fulfills an important milestone in Alithis, as we've established the ground rules and needs for *data usage verification.*

More demos are needed to better validate our protocols, meaning for sure that we will keep on building a non determined amount of DAPPs and other datasets to better prove the functionality described in this paper. This includes APIs that do not require a cryptographic sign to verify data, these APIs would be open to the public.

We're aiming for Filecoin Ehereum Virtual Machine to deploy our smart contracts, yet the main net is not operational on the date this paper is published, we're also aiming in the near to long future for a multichain support Alithis.

Conclusions

It is clear that efficient and scalable identification of datasets and data points is essential for machine learning models. The use of Alithis is a viable solution. The development of a proof of concept with the flower dataset and the MNIST dataset demonstrates the feasibility of this approach.

Moreover, the integration of blockchain technology into the process further enhances the verification and transparency of data usage. It is crucial for data processing companies to provide ways for data usage verification, and the use of hash functions is a step towards achieving this goal.

References

“China Targets AI-Generated Media With New Watermark Requirement -.” *Open Data Science*, 15 December 2022, https://opendatascience.com/china-targets-ai-generated-media-with-new-watermark-requirement/. Accessed 4 March 2023.

“Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content.” *The Verge*, 17 January 2023, https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit. Accessed 4 March 2023.

“Hash function.” *Wikipedia*, https://en.wikipedia.org/wiki/Hash_function. Accessed 6 March 2023.

Heikkilä, Melissa. “A watermark for chatbots can expose text written by an AI.” *MIT Technology Review*, 27 January 2023, https://www.technologyreview.com/2023/01/27/1067338/a-watermark-for-chatbots-can-spot-text-written-by-an-ai/. Accessed 4 March 2023.

“Image classification.” *TensorFlow*, 15 December 2022, https://www.tensorflow.org/tutorials/images/classification. Accessed 6 March 2023.

“Metadata Storage and Management.” *MLOps Community*, https://mlops.community/learn/metadata-storage-and-management/. Accessed 6 March 2023.

“MNIST handwritten digit database.” *Yann LeCun*, http://yann.lecun.com/exdb/mnist/. Accessed 6 March 2023.

“Neural style transfer.” *TensorFlow*, 15 December 2022, https://www.tensorflow.org/tutorials/generative/style_transfer. Accessed 4 March 2023.

"One-way Function | Cryptography." *Crypto-IT*,

http://www.crypto-it.net/eng/theory/one-way-function.html. Accessed 6 March 2023.