Multimedia datasets: challenges and future possibilities

Thu Nguyen[1], Andrea M. Storås[1,2], Vajira Thambawita[1],
Steven A. Hicks[1], Pål Halvorsen[1,2], and Michael A. Riegler[1]

[1] SimulaMet, Oslo, Norway
[2] OsloMet, Oslo, Norway

Abstract. Public multimedia datasets can enhance knowledge discovery and model development as more researchers have the opportunity to contribute to exploring them. However, as these datasets become larger and more multimodal, besides analysis, efficient storage and sharing can become a challenge. Furthermore, there are inherent privacy risks when publishing any data containing sensitive information about the participants, especially when combining different data sources leading to unknown discoveries. Proposed solutions include standard methods for anonymization and new approaches that use generative models to produce fake data that can be used in place of real data. However, there are many open questions regarding whether these generative models hold information about the data used to train them and if this information could be retrieved, making them not as privacy-preserving as one may think. This paper reviews some important milestones that the research community has reached so far in important challenges in multimedia data analysis. In addition, we discuss the long-term and short-term challenges associated with publishing open multimedia datasets, including questions regarding efficient sharing, data modeling, and ensuring that the data is appropriately anonymized.

Keywords: Datasets · Privacy · Modelling · Multimedia.

1 Introduction

Recent technological advancements in computation have allowed us to use complex models to perform tasks like object detection or data generation. These models often require thousands, if not millions, of data samples to perform well and can take considerable amounts of time to train. Simultaneously, open datasets are becoming increasingly popular in the interest of free and transparent research. These datasets are often published in association with machine learning (ML) benchmarks and challenges or simply made public in the interest of open science. However, these datasets are becoming larger, especially multimedia datasets, complicating efficient data storage and sharing and opening up more privacy-related issues. As one of the primary venues for publishing open datasets are benchmarks and ML challenges, finding ways to distribute data to the participants can be challenging, potentially limiting the number of people that can participate.

Making a dataset that contains sensitive information publicly available can pose certain risks in terms of privacy retention for those who participated in the data collection. Moreover, datasets that do not have any sensitive information can also become sensitive when combined with other open datasets [6].

Traditional data anonymization techniques that obfuscate sensitive data by modifying specific values (masking) [4], such as removing the last part of the personal identification number, are well-established and can usually ensure privacy if performed correctly. Besides the traditional methods of privatizing data, researchers have recently also explored the use of synthetic data as an option to preserve participants' privacy. Nevertheless, there is still a lot of research to be done to ensure that the models that create fake data cannot be reverse-engineered or that the fake data does not contain identifiable information in and of itself. Also, as more datasets are released to the public, a problem may arise where persons participate in multiple datasets and can be identified through cross-referencing.

While the challenges mentioned above are important for dealing with data modeling, they are usually not addressed in research articles. This motivates us to analyze them in unison, identify important open challenges that need to be addressed, and discuss future possibilities. More specifically, we raise the following questions:(i) What are the challenges to making large multimodal datasets more accessible to everyone; (ii) How can privacy be enhanced while keeping data open; (iii) How can domain knowledge supplement available data for the development of multimedia ML models, increase explainability, and contribute to privacy protection?

The rest of the paper is organized as follows: In Section 2, we discuss existing problems in multimedia data storage and sharing. Next, in Section 3, we consider the modeling problem, which arises after the data has been stored (and shared), by mainly combining information from multiple datasets and incorporating domain knowledge for improved model performance. In Section 4, we examine data privacy, which is relevant for both data storage, sharing, and modeling, and highlight some future research directions. We conclude the paper in Section 5.

2 Data storage and sharing

Collecting data can be a strenuous task depending on the application and domain. Not only does it require the collection of the data itself, but it also needs to be turned into something that can be used. In general, more data is always better, as one may not always know what is helpful before it is applied to a given task. Therefore, compression and removal of seemingly uninteresting data could limit the potential of a given approach, which again could lead to unwanted downstream effects like poor prediction results or highly constrained statistical models. For example, suppose we were to collect data from a medical device and compress this using a certain codec. In that case, we may limit our models to only consuming data that has been compressed with this codec. Compression is often lossy, meaning that details that might be important are lost. Collecting raw data, however, worsen the issue of storage as it often takes up an immense amount of space.

Several challenges exist regarding data storage. High-quality data usually requires a lot of storage space and should be quickly accessible in order to not slow down the model training process. This is especially relevant for high-quality image and video data. Consequently, the costs associated with storing this data may become too large and pose limitations on who can use it, especially in areas with less developed infrastructure. Furthermore, uploading and downloading these massive datasets requires adequate bandwidth, which is not a certainty in large parts of the world. Besides accessing the data, processing and analyzing large amounts of data is computationally demanding and expensive.

Another aspect related to data sharing is that it is currently not possible to detect what data was used to train a model. Open datasets are usually published with a license, like those that fall under the Creative Commons licenses. These licenses may, for example, restrict the usage of the data only to be used for non-commercial purposes. Still, one cannot prove whether a specific entity has broken the license. Furthermore, open benchmarks and challenges often prohibit the usage of training data outside the provided development datasets, but there is no way of knowing whether a participant used more data than allowed. Possible solutions could be to develop some watermarking method for ML models that state which data the model was trained on or freeze an image of the training environment. To our knowledge, none of these solutions exist yet.

3 Modeling

3.1 Incoporating information from already available datasets

Data can be hard to obtain in many cases, such as for rare or neglected diseases. Moreover, when a dataset is collected from a limited population, e.g., from one hospital, there is a risk of overfitting the ML model. This can make the model fail or perform substantially worse when applied in other populations. Consequently, reusing related datasets or combining them can be an efficient way to provide more information to the model and make it more generalizable.

There are several potential ways to incorporate knowledge from other datasets into a current dataset or model [8]: Multi-task learning learn multiple tasks together by sharing information across tasks, while transfer learning takes the weights that a model learned when solving one task and uses this as a basis for training a model to solve another task. Moreover, datasets can be combined before training the model, and future research should investigate how to perform such combinations more efficiently.

Although data fusion exists for analyzing data from multiple sources, more research is needed to investigate sophisticated ways to learn from different data types. Rather than independently extracting and learning the features from different data modalities, efforts should be made to extract and analyze those modalities together, especially when the data types differ greatly from each other. For example, one can jointly model magnetic resonance imaging (MRI) data and tabular data. This way of multi-modal modeling could also be combined with multi-task learning.

3.2 Incorporating prior knowledge into data modeling

Domain knowledge, such as knowing that benign cancers often are circular [1], can aid in the modeling process and lead to better model performance. Including domain knowledge as mathematical equations can also increase the interpretability of the model, making it easier for end-users to trust it [16].

There are several ways to incorporate prior knowledge into data modeling. In [11], the authors propose a method that relies on a loss function to optimize linear constraints on the output space for weakly supervised segmentation. Ulas and colleagues incorporate domain knowledge by including a cerebral blood flow (CBF) model as a part of the loss function of a neural network [18]. The CBF is highly relevant for the task, and the authors find the neural network with the modified loss function to outperform other methods. Alternatively, special layers can be introduced to the network. The deep learning method Varmole [7] adds a biological drop-connect layer to the neural network. This layer includes a matrix indicating the association between small differences in the DNA called single nucleotide polymorphisms (SNPs) and specific genes. Bochare et al. use domain knowledge to design an algorithm for creating virtual data instances. When the virtual instances are added to the original training data, the accuracy of the ML model improves compared to when the model is only trained on the original data [2].

With the recent advances in physics-inspired deep learning methods, prior knowledge can be included by adding a penalty term to the loss function of the neural network. This way, we can constrain the data to satisfy the given dynamic. The penalty term could be, for example, a partial differential equation that is usually used to model the dynamic of interest. Moreover, the coefficients of the equation can be optimized while training the neural network. The solution of the differential equations could be obtained using a deep learning model for differential equations [17]. To our knowledge, there is no work that goes in this direction for multimedia data yet.

Prior knowledge can help to identify important features that should be collected for a particular dataset, or the knowledge can be incorporated directly into the models, as earlier exemplified. Therefore, studying whether prior knowledge can reduce the number of features that need to be collected, stored, and shared can be an interesting research topic. Furthermore, fewer data samples could lead to improved privacy protection. However, there is still possibly some trade-off between these benefits and the performance of a prediction model.

Despite its potential for reducing the required amount of training data, improving privacy protection, enhancing learning, and increasing explainability of ML systems, domain knowledge is seldom applied for ML modeling on multimedia data. We believe this is an important and exciting direction for future work. Further on, attention should be paid to explainability aspects when incorporating domain knowledge into ML models.

4 Data privacy

Masking data for training or testing of ML models is an established method for protecting the privacy of the individuals in the dataset. Data masking replaces original data with data that looks and acts realistic but is actually not. Moreover, it should be impossible to restore the original data from the masked data without having access to an encryption key or similar types of extra information [4].

When the original data cannot be made public, it can sometimes be useful to publish aggregated data instead. For tabular data, dimensionality reduction methods such as Principle Component Analysis (PCA), singular value decomposition (SVD), tensor decompositions, and Auto-encoders (AE) can be applied, and the dimension-reduced version of the original data can be made public instead. If a study is conducted on volunteers in such a manner, it may attract more volunteers as well because they know that their data will not be made public directly. Sometimes, ML models can solve tasks at acceptable levels without access to private data. For example, swarm learning was used to predict the mutational status of cancer from histopathology images without including personal data in the analysis [12]. Careful considerations should be made regarding whether or not it is necessary to include all available information. A potential increase in model performance due to the inclusion of sensitive data might be outweighed by reduced privacy.

Sometimes, a person's identity and private information in a dataset can be retrieved through some minor details. Indeed, according to OpenAIRE[3], sensitive data also includes datasets that can be combined into personal or sensitive data. Therefore, it's important to prevent unmasking by using data fusion, prior knowledge, or collecting extra data (such as web scraping) to identify individuals from the data. In addition, it is worth investigating if incorporating domain knowledge can help preserve privacy better and improve privacy for synthetic data.

Recent advancements in deep generative models, such as Generative Adversarial Networks (GANs) [5] and diffusion models [13], show promising results of using synthetic data [15, 14] to mimic the original data distributions without sharing privacy-sensitive real datasets. However, research is still being done to find leakage between real and synthetic data to ensure that data generated by generative models is privacy-preserving. While effective methods exist for masking sensitive data, more research should be directed toward potential privacy issues when synthetic data generation models are made public, also in the context of multimodal datasets.

Differential privacy [3] is another well-known technique often used to retain the privacy of large datasets. However, differential privacy works by introducing noise, which could affect the distribution of small datasets. In this regard, using this technique in fields that often have small datasets, like the medical domain, is challenging and needs more research.

The best way to protect privacy is to delete sensitive entries in the dataset. In addition, the development of missing data imputation techniques so far allows

[3] https://www.openaire.eu/sensitive-data-guide

Table 1. Overview of future challenges related to public sharing and modeling of multimedia datasets and suggestions for future research.

Topics	Future challenges
Data storage	Solutions to upload and download data more efficiently Methods to store and quickly analyze large amounts of data
Check the actual training data	Develop watermark for models Freeze image of training environment
Information from other datasets	Investigate how to efficiently combine datasets Develop sophisticated methods for joint extracting and modeling of multiple data modalities
Incorporating domain knowledge	Methods for multimedia modeling that incorporate domain knowledge Investigate if domain knowledge can improve explainability of multimedia models Explore domain knowledge to reduce amount of data needed to collect, store and share
Data privacy	Develop and improve methods for synthetic data generation Improved privacy preserving metrics Improved methods for differential privacy on small datasets

dealing with various types of missing data [9], data types, and sizes [10]. Consequently, these techniques can be applied when certain entries in the dataset cannot be included due to privacy issues.

5 Conclusion and future research directions

In this paper, we highlighted some important trends in multimedia data storage, sharing, usage, and privacy protection and identified significant challenges that need to be addressed. A summary of the challenges is listed in Table 1. In conclusion, multimedia datasets and their respective analyses have challenges that are still not addressed, and it is necessary to use all the available resources, including domain knowledge, to enhance privacy protection and model training. In addition, adding constraints to the model, especially the ones that are intuitive and obvious, not only can aid the performance but also the explainability of the model, which leads to a better understanding of which data sources are relevant. In the long run, this can help to enhance privacy and sharing even more. The challenges discussed in this work are all essential for the multimedia domain and should be further explored in the future.

References

1. Barata, C., et al.: A survey of feature extraction in dermoscopy image analysis of skin cancer. IEEE journal of biomedical and health informatics (2018)
2. Bochare, A., et al.: Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer. International Journal of Medical Engineering and Informatics (2014). https://doi.org/10.1504/IJMEI.2014.060245
3. Dwork, C., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science (2014)
4. Goyal, C.: Data Masking: Need, Techniques & Solutions. International Research Journal of Management Science & Technology (IRJMST) (2015)
5. Gui, J., et al.: A review on generative adversarial networks: Algorithms, theory, and applications. IEEE Transactions on Knowledge and Data Engineering (2021)
6. Narayanan, A., et al.: Robust de-anonymization of large sparse datasets (2008). https://doi.org/10.1109/SP.2008.33
7. Nguyen, N.D., et al.: Varmole: a biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. Bioinformatics (2021)
8. Nguyen, T., et al.: Combining datasets to increase the number of samples and improve model fitting (2022). https://doi.org/10.48550/ARXIV.2210.05165
9. Nguyen, T., et al.: DPER: Direct Parameter Estimation for Randomly missing data. Knowledge-Based Systems (2022)
10. Nguyen, T., et al.: Principle Components Analysis based frameworks for efficient missing data imputation algorithms. arXiv preprint arXiv:2205.15150 (2022)
11. Pathak, D., et al.: Constrained convolutional neural networks for weakly supervised segmentation (2015)
12. Saldanha, O.L., et al.: Swarm learning for decentralized artificial intelligence in cancer histopathology. Nature Medicine (2022). https://doi.org/10.1038/s41591-022-01768-5
13. Sohl-Dickstein, J., et al.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics (2015)
14. Thambawita, V., et al.: DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. Scientific reports (2021)
15. Thambawita, V., et al.: DeepSynthBody: the beginning of the end for data deficiency in medicine (2021). https://doi.org/10.1109/ICAPAI49758.2021.9462062
16. Tjoa, E., et al.: A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems (2021). https://doi.org/10.1109/TNNLS.2020.3027314
17. Tu, S.N.T., et al.: FinNet: Solving Time-Independent Differential Equations with Finite Difference Neural Network. arXiv:2202.09282
18. Ulas, C., et al.: DeepASL: Kinetic Model Incorporated Loss for Denoising Arterial Spin Labeled MRI via Deep Residual Learning (2018). https://doi.org/10.1007/978-3-030-00928-1_4