Digital, big data and computational forensics

Geradts, Z.

DOI
10.1080/20961790.2018.1500078

Publication date
2018

Document Version
Final published version

Published in
Forensic Science Research

License
CC BY

Link to publication

Citation for published version (APA):
Geradts, Z. (2018). Digital, big data and computational forensics. *Forensic Science Research*, *3*(3), 179-182. https://doi.org/10.1080/20961790.2018.1500078

FORENSIC SCIENCES RESEARCH
2018, VOL. 3, NO. 3, 179–182
https://doi.org/10.1080/20961790.2018.1500078

EDITORIAL

OPEN ACCESS

Digital, big data and computational forensics

Recent years have witnessed significant developments in deep learning and artificial intelligence [1]. For instance, remarkable improvements have been made in automated face comparison systems by using deep learning, compared with the classic approaches [2].

The term "deep learning" is often used to refer to certain kinds of neural networks. The first publications on biological neural networks and the brain date back to the late 1800s [3]. It was not until the rediscovery of the back-propagation algorithm [4] in 1986 that interest in the field was reignited. An artificial neural network is designed following a simple modelling of the brain, and involves a representation of neurons. A neuron receives a specific signal and converts to a different one. Neurons can also be used to learn from examples. They adjust a transfer function. Many neurons are linked together, and are often used in multiple layers. A visual overview is provided in Figure 1.

An example of the application of neural networks is face recognition [5], where these networks examine images of people's faces and find features, such as shapes of nose, ears, and mouth. In such networks, the parameters of thousands or more neurons are adjusted based on training to improve recognition performance. Combined with improved pattern recognition to detect the eyes, mouth, and the position of the face, they yield better results.

This type of network can be used not only for images but also for large amounts of text and other data, including audio. It must be trained for each task on relevant data. In recent years, computation power was been greatly enhanced though Graphical Processing Unites (GPUs) in personal computers (PCs [6]). They have become considerably cheaper over time, and can now be used in parallel.

Neural networks are used for many types of data [7–9]. If the ground truth of a dataset is known, this is good for the training set, since it needs to learn from correct examples. For example, a neural network can be trained to detect with fraudulent as well as benign transactions [10]. This technology can then be applied, and allows for the identification that would not have possible through manual effort.

One problem with neural networks is that they operate like a black box, as the network can be trained on information other than that desired. It is well known that algorithms based on neural networks were used by the US military in the 1980s to recognize enemy artillery tanks hidden in forests [11], and were found to be efficient. In reality, the algorithm did not recognize the tanks, but whether it was sunny or cloudy day, since the training set was made with US tanks in woods during a sunny day and was made with enemy tanks in woods during a cloudy day.

A major advantage of computers is that they can quickly search through large amounts of data. A person becomes tired after manually examining images of faces, and the process is more prone to error [12]. A computer can easily examine several million faces in the same group of facial features. This article discusses definitions of big data relevant to forensics, practical, and ethical considerations of applications and expectations for the future.

Definition of big data

Many definitions of big data have been offered in the literature [13], and most incorporate the volume, variety, and velocity of the data generated as well as the velocity of the analysis needed.

Other factors are important in studying big data [13]. The quality of data is a concern. Some sources are more reliable than others. Moreover, data can be misinterpreted and false conclusions drawn based on an analysis of trends in big data. Consider the example of the service Google Flu Trends that attempted to predict flu activity. Between 2006 and 2010, it could predict well how the flu developed by examining queries for flu on its search engine. Around 2012, this no longer worked, and so the project was terminated [14].

In forensic science, the term "computational forensics" is also used to refer to the automated analysis of forensic traces, and is related to forensic data science [15].

Relevance to forensic sciences

In forensics, the automatic fingerprint identification system (AFIS) was introduced in the 1980s [16]. In the early 1990s, using automated systems of bullets and shells made an important step for comparing striation and impression marks. Similar success has been achieved in automating the recognition of marks and scuffs left by tools and shoes [17, 18], with pattern recognition.

Many tools for network-based analysis of data are available for law enforcement. The first ones emerged in the 1990s, but were limited in the amount of data that they could analyze.

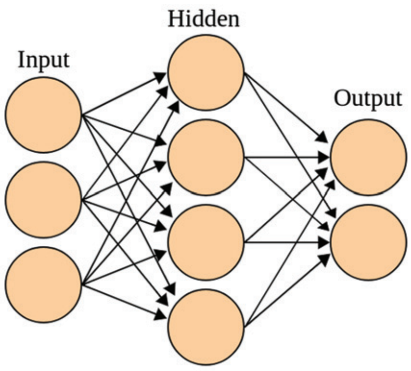

Figure 1. Schematic overview of a neural network (from Wikimedia Commons).

Many companies sell software for the automated analysis of digital data secured from such sources as smartphones, computers, and Internet taps. The need to search through these data has led to the development of increasingly advanced search methods.

Seized data from, for example, a smartphone or a computer, are in part structured and partly unstructured. "Structured" in this context means that a part of the data can be interpreted as it is stored in a standard manner, but another part cannot because it is not stored in a standard or known manner, i.e. it is unstructured. Information from emails and social messaging sites are examples of structured data.

The Netherlands Forensic Institute has developed a system for forensic search for the police [19]. It can be used to search through large amounts of data, e.g. digital traces, including emails, instant messages or images, and video material. Location information from photos and other data can also be plotted on a map using this system. It was developed using an open-source big data-handling software, and can be used in as transparent a manner as possible. Furthermore, the system is continually updated to accommodate file formats used by newly developed apps.

Algorithms are available to determine the make and model of the camera using which a given image was taken, and are used in forensic casework [20]. The small artefacts in sensors of cameras between the different light sensitive elements, is of interest in forensic science. The artefacts between pixels make a pattern, the Photo Response Non-Uniformity (PRNU) pattern.

Based on this pattern, a kind of fingerprint of the sensor can be determined, so evidence that an image has been with a specific camera. The algorithms proposed in [21, 22] can be used to identify the camera that was used to make pornographic images of children.

Another example of the use of computer algorithms is the analysis of large amounts of text. Dujin et al. [23] has researched the use of neural networks in criminal databases and arrived at a number of conclusions. They combined various kinds of information concerning relationships among 22 000 known criminals. This made use of data from social media, police reports, and arrest records. These criminals had been convicted of a variety of crimes, such as drug use and trafficking, extortion, money laundering, and manufacturing synthetic drugs.

From this research and simulations, it appears that the common method of law enforcement to arrest the bosses of the criminal organization appears that the organizations become stronger since leaders are easily replaced, whereas for example toxicologist are much harder to replace. The network is a complex system that adapts rapidly to changing circumstances, and it is impossible to obtain a picture of the activities in a given domain in a classical structure.

Grauss et al. [24] examined the relationship between entities in public datasets from digital data related to Enron, including e-mails and social networking traffic. Emerging entities in a network are also important for detecting abnormal behaviour and networking activities between entities of interest to the investigation. They also showed many predictions concerning behaviour can be made using digital data.

These techniques for predictive data analysis may be important for the prevention of crime and predictive policing. There is a growing recognition in criminology of the opportunities offered by big data [25–27]. The increasing number of sensors in technologies used by people, including smartwatches and medical equipment, has enormous potential if data from them are available for analysis. Location information as well as the fitness trackers can serve as evidence in court [28–30].

Caveats and limitations

In Europe, the General Data Protection Regulation (GDPR) addresses data protection and privacy for all individuals within the European Union (EU). It also addresses the export of personal data outside the EU. Effective in May 2018, the law is expected to have a significant impact on the type and amount of data that can be used within law enforcement, though there are many exceptions. Even in countries with few regulations regarding privacy and ethics, people's perception of the importance of privacy change as they begin sharing less data about themselves [31–33].

By a clever combination of human effort and the use of computers, a considerable amount of time can be saved. Moreover, companies like Microsoft, Google, and Apple are using this idea in large-scale data analysis for the advertising market [34]. All search queries by users in addition to their location information provides invaluable information to these companies about user preferences. Techniques developed in these areas can be used in forensic research, and many companies have rendered their technologies open source, such as PyTorch, Hadoop, TensorFlow, FaceNet, and OpenFace [35–37].

It is expected that anti-forensic techniques will increase as the privacy-awareness increases. Existing anti-forensic tools can be fully wiped or changed easily [38]. It is also conceivable that wrong tracks are added to datasets to corrupt them [39]. The creation of zombie accounts is an instance [40]. Strong encryption is readily available to users.

Most artificial intelligence-based solutions can only function well for one task at a time [41]. Thus, a good algorithm can recognize faces but can perform no other task. According to some studies, algorithms tend to perform better than humans at such games as chess where computers have been known to beat expert human players [42]. However, they are not as good at judging softer information, such as pictures with weaker associated information. Combining a plurality of such algorithms is an important step. We assume here that these algorithms are used in forensic science as tools and their results are critically verified by human experts.

The widespread use of big data raises several ethical and privacy-related issues [43]. An internal solution for privacy and data protection is in the design of systems [44] as well as policies concerning access.

Predictive policing also uses deep learning. Using training on a sample dataset, the system can predict if crimes will occur in a certain neighbourhood. Discussions are ongoing on the ethical aspects of predictive policing as algorithms might make wrong decisions and discriminate against a group of people [33, 45] with the training sets. Validation and human intervention are thus needed depending on use. A feedback mechanism in the software to learn from bad decisions might help improve it. The system should also provide a level of confidence for its results based on the data.

The future perspective

Using artificial intelligence in the criminal justice system can accelerate judicial decisions since they can be automated. Naturally, errors in the decisions found using such techniques should also be considered and validated [46, 47].

Predictive policing can ultimately help prevent crime. It is important to consider the social, ethical, legal, and privacy-related issues involved prior to its acceptance. In the coming years, more research is expected on the combination of large biometric databases and surveillance data [48], e.g. big data projects using biometric features of convicted criminals, which also has ethical implications.

Using artificial intelligence and deep learning, court decisions can be predicted [46]. Another interesting research topic is to examine court decisions and locate misinterpretations of (forensic) evidence, which can be useful in tracking wrongful convictions.

Within this issue of the journal, several insights on new developments are provided in the domain of digital evidence as well as biometric features, along with details on how to use them. Using deep learning to detect the mother tongue of a speaker and traces of chemicals on clothes, and the recognition of the model of a camera from an image taken using it are some developments. The quality assurance and the validation of the results remains an important issue as new algorithms are developed.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Zeno Geradts http://orcid.org/0000-0001-5912-5295

References

[1] Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. JAMA. 2016;316:2366–2367.

[2] Yao M. Understanding the limits of deep learning. VentureBeat [Internet]. 2017 Apr 2 [cited 2017 Apr 9]. Available: https://venturebeat.com/2017/04/02/understanding-the-limits-of-deep-learning/

[3] Bergström RM. Ageing of the brain in the light of neural network models. Geron. 1974;20:4–20.

[4] Zipser D, Andersen RA. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature. 1988;331:679–684.

[5] Kaur M, Kumar D. A review on various techniques for face recognition in digital images. Int J Curr Res Acad Rev. 2016. doi: 10.20546/ijcrar.2016.410.003

[6] Svyatkovskiy A, Kates-Harbeck J, Tang W. Training distributed deep recurrent neural networks with mixed precision on GPU clusters. Proceedings of the Machine Learning on HPC Environments – MLHPC'17; 2017 Nov 12–17; Denver, CO, USA: ACM Press; 2017. p. 1–8.

[7] Sameer VU, Naskar R, Musthyala N, et al. Deep learning based counter–forensic image classification for camera model identification. In: Digital forensics and watermarking. Berlin: Springer International Publishing; 2017. p. 52–64.

[8] Geradts ZJ, Franke K. Editorial for big data issue. Digit Investig. 2015;15:18–19.

[9] What Is Big Data? | SAS US [Internet]. Cary, North Carolina, US: SAS Institute Inc. [cited 29 Mar 2017]. Available: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html.

[10] Sokolov V. Discussion of 'Deep learning for finance: deep portfolios': discussion of 'Deep learning for finance: deep portfolios. Appl Stoch Model Bus Ind. 2017;33:16–18.

[11] Dreyfus HL, Dreyfus SE. What artificial experts can and cannot do. AI & Soc. 1992;6:18–26.

[12] Simoni DA, Motter BC. Human search performance is a threshold function of cortical image separation. J Vis. 2010;3:228.

[13] Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manag. 2015;35:137–144.

[14] Lohr S. Google flu trends: the limits of big data. New York Times. 2015;12–15.

[15] Franke K, Srihari SN. Computational forensics: an overview. In: Computational forensics. Berlin, Heidelberg: Springer; 2008. p. 1–10.

[16] Geradts Z, Bijhold J. New developments in forensic image processing and pattern recognition. Sci Justice. 2001;41:159–166.

[17] Geradts Z, Keijzer J. The image-database REBEZO for shoeprints with developments on automatic classification of shoe outsole designs. Forensic Sci Int. 1996;82:21–31.

[18] Geradts ZJ, Keijzer J, Keereweer I. Automatic comparison of striation marks and automatic classification of shoe prints. In: Rudin LI, Bramble SK, editors. Proceeding of SPIE 1995 Sept 1; San Diego, CA.SPIE; 1995. p. 151.

[19] van Baar RB, van Beek HMA, van Eijk EJ. Digital forensics as a service: a game changer. Digit Investig. 2014;11:S54–S62.

[20] Cooper AJ. Improved photo response non-uniformity (PRNU) based source camera identification. Forensic Sci Int. 2013;226:132–141.

[21] Eindrapport PRNU NCTV. Available: https://www.nctv.nl/binaries/eindrapport-prnu_tcm31-30331.pdf.

[22] Gisolf F, Malgoezar A, Baar T, et al. Improving source camera identification using a simplified total variation based noise removal algorithm. Sci Rep. 2013;10:207–214.

[23] Duijn PAC, Kashirin V, Sloot PMA. The relative ineffectiveness of criminal network disruption. Sci Rep. 2014;4:4238.

[24] Graus D, Tsagkias M, Weerkamp W, et al. Dynamic collective entity representations for entity ranking. Proceedings of the Ninth ACM International Conference on Web Search and Data Mining – WSDM '16; 2016 Feb 22-25; Francisco, CA:ACM Press; 2016. p. 595–604.

[25] Williams ML, Burnap P. Cyberhate on social media in the aftermath of Woolwich: a case study in computational criminology and big data. Br J Criminol. 2016;56:211–238.

[26] Chan J, Bennett Moses L. Is big data challenging criminology? Theor Criminol. 2016;20:21–39.

[27] Smith GJD, Bennett Moses L, Chan J. The challenges of doing criminology in the big data era: towards a digital and data-driven approach. Br J Criminol. 2017;57:259–274.

[28] Bodin WK, Jaramillo D, Marimekala SK. Security challenges and data implications by using smartwatch devices in the enterprise. 2015 12th International Conference & Expo on Emerging Technologies for a Smarter World (CEWIT); 2015 Oct 19-20; Melville, NY. Piscataway, NJ: IEEE Press; 2015. p. 1–5.

[29] Ichikawa F, Chipchase J, Grignani R. Where's the phone? A study of mobile phone location in public spaces. IEE Mobility Conference 2005. The Second International Conference on Mobile Technology, Applications and Systems; 2005 Nov 15-17; Guangzhou, China. Piscataway, NJ: IEEE Press; 2005. p. 142–142.

[30] Zhang D, Vasilakos AV, Xiong H. Predicting location using mobile phone calls. ACM SIGCOMM Comput Commun Rev. 2012;42:295–296.

[31] Unsworth K. Questioning trust in the era of big (and small) data: questioning trust in the era of big (and small) data. Bull Assoc Inf Sci Technol. 2014;41:12–14.

[32] Wang Z, Yu Q. Privacy trust crisis of personal data in China in the era of Big Data: the survey and countermeasures. Comput Law Secur Rev. 2015;31:782–792.

[33] Haberman CP, Ratcliffe JH. The predictive policing challenges of near repeat armed street robberies. Policing. 2012;6:151–166.

[34] Ross S. Apple Vs. Microsoft Vs. Google: How Their Business Models Compare (AAPL, MSFT) [Internet]. NY, US: Investopedia, LLC. [cited 29 Mar 2017]. Available: http://www.investopedia.com/articles/markets/111015/apple-vs-microsoft-vs-google-how-their-business-models-compare.asp

[35] Gao D, Zhao Y, Gao J, et al. Comparison and analysis of the open-source frameworks for deep learning. DEStech Trans Comput Sci Eng. 2017. doi: 10.12783/dtcse/mcsse2016/10975

[36] OpenFace: Models and Accuracies [Internet]. Openface; 2017 Jan 22 [cited 2017 Mar 29]. Available: https://cmusatyalab.github.io/openface/models-and-accuracies/#pre-trained-models.

[37] Baltrusaitis T, Robinson P, Morency LP. OpenFace: an open source facial behavior analysis toolkit. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV); 2016 Mar 7-10; Lake Placid, NY. Piscataway, NJ: IEEE Press; 2016. p. 1–10.

[38] Singh G, Singh K. Improved JPEG anti-forensics with better image visual quality and forensic undetectability. Forensic Sci Int. 2017;277:133–147.

[39] Moon PJ. The development of anti-forensic tools for android smartphones. J Korea Inst Electron Commun Sci. 2015;10:95–101.

[40] Li S, Li X, Yang H, et al. A zombie account detection method in microblog based on the PageRank. In: 2017 IEEE International Conference on Software Quality. Reliability and Security Companion (QRS-C); 2017 Jul 25-29; Prague, Czech Republic. Piscataway, NJ: IEEE Press; 2017. p. 267–270.

[41] Ruocco K. Related Articles to Artificial Intelligence: The Advantages and Disadvantages. Arrk. [cited 23 Sept 2017]. Author's manuscript available at: https://www.arrkgroup.com/thought-leadership/artificial-intelligence-the-advantages-and-disadvantages/.

[42] Levene M. Chess metaphors, artificial intelligence and the human mind. Comput J. 2011;54:1560.

[43] Keenan B. Mireille Hildebrandt. Smart Technologies and the End(s) of Law [bookreview]. Mod Law Rev. 2016;79:305–318.

[44] Hansen M. Data protection by design and by Default à la European General Data Protection Regulation. In: Lehmann A, Whitehouse D, FH Simone, et al., editors. Privacy and identity management. Facing up to next steps. New York: Springer International Publishing; 2016. p. 27–38.

[45] Jefferson BJ. Predictable policing: predictive crime mapping and geographies of policing and race. Ann Am Assoc Geogr. 2018;108:1–16.

[46] Hutson M. Artificial intelligence prevails at predicting Supreme Court decisions. Science. 2017.

[47] Rzevski G. A new direction of research into artificial intelligence. Artif Intell. 2008:1–16.

[48] Shen Y, Shen M, Chen Q. Measurement of the new economy in China: big data approach. China Econ J. 2016;9:304–316.

Zeno Geradts

Netherlands Forensic Institute, Den Haag, The Netherlands University of Amsterdam, Institute of Informatics, Amsterdam, The Netherlands

geradts@uva.nl