The Data Deconflation Problem: Moving from Classical to Emerging Solutions

Roger A. Hallman[1,2] and George Cybenko[1]
[1]*Thayer School of Eningeering, Dartmouth College, Hanover, New Hampshire, U.S.A.*
[2]*Naval Information Warfare Center (NIWC) Pacific, San Diego, California, U.S.A.*

Keywords: Data Deconflation, Deconvolution, Blind Source Separation, Cocktail Party Problem, Simple Data, Complex Data, Deep Learning, Deep Reinforcement Learning (DRL), Generative Adversarial Networks (GANs).

Abstract: Data conflation refers to the superposition data produced by diverse processes resulting in complex, combined data objects. We define the data deconflation problem as the challenge of identifying and separating these complex data objects into their individual, constituent objects. Solutions to classical deconflation problems (e.g., the Cocktail Party Problem) use established linear algebra techniques, but it is not clear that those solutions are extendable to broader classes of conflated data objects. This paper surveys both classical and emerging data deconflation problems, as well as presenting an approach towards a general solution utilizing deep reinforcement learning and generative adversarial networks.

1 INTRODUCTION

The proliferation of Internet-connected devices has led to a flood of complex, conflated data objects from which we can glean a wealth of useful information. For example, distributed sensor networks–critical to large-scale Internet of Things (IoT) systems–continually report real-time data that may be representative of co-located individuals (Wan et al., 2016). Similarly, data reported by medical wearables may be contaminated by patient movements or external influences, or report excess noise due to insufficiently tuned sensors (Tariq et al., 2018). Those conflated data objects must first be separated into their constituent components before any meaningful analysis can be conducted.

Recent advances in deep learning have led to breakthroughs in many classification, recognition, and decision-making tasks; however those results have been limited to tame datasets and performance in relatively benign environments. As a purely motivational example, consider the conflated illustration in Figure 1. While even a human child can identify at least one of the constituent objects (seen individually in Figure 2) in this conflated image, a MATLAB implementation of the well-known Alexnet Object Classifier (Krizhevsky et al., 2012; MathWorks, 2020) is unable to identify any object and returns the following probabilities:

Figure 1: Multiple images have been conflated in such a way that state-of-the-art classifiers cannot identify a single constituent image.

The current known solutions to data deconflation problems rely on well-established linear algebra techniques, but it is not at all clear that these techniques can be generally extended. For instance, behavioral tracking tasks will often generate non-additive superpositions and categorical data that is neither real-valued nor sampled from a uniform spatial or temporal grid. As illustrated by Alexnet's inability to classify any of the constituent images in Figure 1, even current deep learning networks are unlikely to provide

Hallman, R. and Cybenko, G.
The Data Deconflation Problem: Moving from Classical to Emerging Solutions.
DOI: 10.5220/0010530403750380
In *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security (IoTBDS 2021)*, pages 375-380
ISBN: 978-989-758-504-3

Table 1: Alexnet probabilities for Figure 1.

Category	Probability
jigsaw puzzle	0.2270
wreck	0.1252
mud turtle	0.1249
loggerhead	0.0842
terrapin	0.0566

(a) Barn (b) Otter

(c) School Bus (d) Giant Panda

Figure 2: The constituent images that were conflated in Figure 1. Alexnet classifies the subject of these individual images with high probability.

a satisfactory, more generalized solution to the data deconflation problem. (Tangentially, consider Goodfellow et al.'s (Goodfellow et al., 2014b), demonstration that even the addition of seemingly imperceptible noise can lead to misclassifications.)

To that end, we present our vision for a solution to the data deconflation problem which can be be extended to tasks which are beyond currently-known solutions. We believe that a promising approach to the general deconflation problem can be based on iterations between estimating what signal component or element is contributed by what process (accomplished by using a trained deep reinforcement network) and filtering done by using a generative network seeded by small signal samples.

Contribution and Organization. Our primary contribution in this paper is the proposal of what we believe to be a general solution to the data deconflation problem, iteratively using deep reinforcement learning and generative adversarial networks, not only during training but also in the deconflation and classification phase. As far we are aware, there is no general solution for deconflation problems dealing with non-additive superpositions or categorical valued data objects, as often occur in spatial tracking and behavioral deconflation problems.

The remainder of this paper is organized as follows: many deconflation problems, including the cocktail party problem, and their solutions are described in Section 2. Our approach to a general solution to the data deconflation problem is given in Section 3, and concluding remarks are given in Section 4.

2 BACKGROUND AND RELATED WORK

We begin by presenting a brief survey of current solutions to deconflation problems, as well as reinforcement learning and generative adversarial networks.

2.1 Blind Source Separation

Blind source separation (BSS) is the process of separating unknown signals that have been mixed in an unknown way (Kofidis, 2016). Specifically, a mixture

$$\mathbf{u}(n) = \mathcal{F}\left(\mathbf{a}(n), \mathbf{v}(n), n\right)$$

mixes N source signals

$$\mathbf{a}(n) = \left[a_1(n), a_2(n), ..., a_N(n)\right]^T,$$

and K noise signals

$$\mathbf{v}(n) = \left[v_1(n), v_2(n), ..., v_K(n)\right]^T,$$

by a mixing system $\mathcal{F}\left(\cdot,\cdot,\cdot\right)$, which yields

$$\mathbf{u}(n) = \left[u_1(n), u_2(n), ..., u_{N\times K}(n)\right]^T.$$

BSS problems have long been an active research topic in both analog and digital signal processing, with numerous demonstrated solutions (O'grady et al., 2005; Comon and Jutten, 2010). BSS problems are vector representative and additive, which means that there are a number of solutions that utilize established linear algebra techniques. Techniques utilized in classical BSS solutions include singular value decompositions, principal component analysis, sparsity enforcement, or other dimensionality reduction methods. For instance, the Joint Approximation Diagonalization of Eigen-matrices algorithm has been implemented to accomplish BSS for both image (Hughes, 2015a) and audio (Hughes, 2015b) samples.

2.1.1 The Cocktail Party Problem

Perhaps the most well-known BSS problem is the Cocktail Party Problem (CPP) (Cherry, 1953), that is the human ability to selectively focus attention on a single voice in a noisy environment. In a typical formulation, an attendee at a cocktail party hears their name spoken by an unknown person outside of their vision and they attempt to identify that person. The CPP has been extended to visual data as well as auditory. Shapiro et al., showed that people have an ability to recognize their own name in otherwise unattended information (Shapiro et al., 1997).

Haykin and Chen (Haykin and Chen, 2005) frame the problem in terms of understanding how the human brain solves this problem and determining whether it is possible to build a machine that can satisfactorily solve it. Their survey of computational approaches detail solutions via (*i*) independent component analysis (ICA) and general BSS approaches, (*ii*) temporal binding and oscillatory correlation, and (*iii*) cortronic networks. They note that while ICA and BSS solutions enjoy decades of support in literature, the approach is not analogous to actual biological solutions. On the other hand, approaches (*ii*) and (*iii*) are inspired by biological processes but rely on the assumption of some prior knowledge (e.g., the language being spoken).

Qian et al. (Qian et al., 2018) survey more recent approaches to the CPP (including deep learning-based solutions). They highlight many impressive results, while pointing out limitations that are analogous to the current solutions' shortcomings mentioned in Section 1. For instance, they highlight greater improvements in recognition for mixed-gender speech than for same-gender speech; inferring that same-gender speech tracing is a more difficult task.

2.2 Process Query Systems

Process Query Systems (PQS) (Cybenko and Berk, 2007) are a more recent solution to deconflation problems that are especially well suited to networked systems, where extracting meaningful information is particularly challenging. By paying attention to process descriptions, PQS are able to solve complex information retrieval tasks within the network. Specifically, PQS take input from arbitrary nodes in a network and build hypotheses about observed events that answer a user's process queries. Multiple hypotheses and models are used to separate observed events, optimally matching them with ongoing processes, and identifying process states.

PQS have been applied to tasks in network administration, including security monitoring (Berk et al., 2003; Berk and Fox, 2005), covert channel detection (Giani et al., 2005), and autonomic server monitoring (Roblee et al., 2005). Additionally, PQS have been used for vehicle tracking using acoustic sensor networks (Berk et al., 2003).

While PQS provide a more general solution to tasks that are beyond the capabilities of BSS and classical deconflation solutions, they are not a general solution. A PQS requires a priori models for underlying processes, as well as heuristics for estimating the number of processes, when those processes begin and end, and track assignments.

2.3 Reinforcement Learning

Reinforcement Learning (RL) is a field of machine learning that seeks to understand, automate, and optimize goal-directed decision making (Sutton and Barto, 2018). Deep Reinforcement Learning (DRL) (François-Lavet et al., 2018) involves harnessing the power of deep neural networks for RL tasks and has led to groundbreaking results, including super-human results in gameplay.

In spite of the successes in relatively tame and optimized environments, RL and DRL face a multitude of challenges in adoption for real-world tasks (Dulac-Arnold et al., 2019). One such challenge which has recently seen breakthrough results is the credit assignment problem where there are delays between agent actions and rewards (Hung et al., 2019). Specifically, Hung et al. developed an agent memory function that credits past actions and enables them to solve previously intractable problems. Deep Reinforcement Relevance Networks (He et al., 2016) and Dialog State Tracking and Management (Zhao and Eskenazi, 2016) have shown phenomenal success in state tracking and credit assignment in natural language.

2.4 Generative Adversarial Networks

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014a) are a deep learning framework where two deep neural networks, a generator and a discriminator, are simultaneously trained against each other. Specifically, the discriminator is trained to detect real from synthetic data (e.g., differentiating an authentic image versus a synthetic image of a human face (Tariq et al., 2018)) while the generator is trained to generate authentic "looking" synthetic data from a low-dimension seed.

In order to take a low-dimensional data seed and generate synthetic data capable of fooling the discrim-

inator, GANs must effectively impute missing data. Lee et al. (Lee et al., 2019), developed a GAN which converts image imputation into a multi-domain translation task, enabling a single generator and discriminator to successfully estimate missing image data. Following on successes in image data imputation, GANs are being utilized for time series data imputation. Time series data from many sensor networks have an average missing data rate of around 80% and the imputation of that missing data is critical to any analysis efforts. Luo et al. (Luo et al., 2018) implemented a gated recurrent unit (GRU), modified to model temporal irregularity, into their GAN architecture. Furthermore, they developed a loss function that provides a fitness measure for imputed values. Zhang et al. (Zhang et al., 2021) incorporate real data forcing and an encoder network into their GAN architecture to create imputed synthetic data that performs well in numerous downstream tasks.

3 OUR APPROACH TO A GENERAL DATA DECONFLATION SOLUTION

We have now defined BSS and surveyed existing solutions, thus we first propose a generalization of the BSS problem before we present our vision for a general solution.

3.1 From BSS to General Data Deconflation

Data can be conflated in space (e.g., Figure 1), time, and semantics as well as in any combinations of these dimensions. The most common manifestation of the multi-target tracking problem can be both spatial (as arises in occlusion) and temporal (as in track assignment). Pattern of life analyses have to deal with conflated semantics in which, for example, a commuter combines a trip to work with an in-person meeting on the commuter train.

Simple data is data (or a process) coming from a single source. Complex or conflated data consists of interwoven simple data objects coming from multiple sources. Solutions to BSS of complex data require vector respresentable inputs, but it is not apparent that this is broadly possible for general separation tasks. Rather than vector representations, we therefore propose to represent simple data as a state machine (Schneider, 1990) and complex data as state machine synthesis (Ginsburg, 1959).

Our state machine representation for conflated data is presented in Figure 3. We claim that an observed event sequence (i.e., complex data) is the synthesis of an unknown multiplicity of simple data objects. The Data Deconflation Problem is a generalization of the BSS Problem (Section 2.1): given an observed event sequence, which simple data objects are responsible for specific observed events? Furthermore, many separation solutions assume some a priori knowledge–whether a language spoken, some underlying processes, beginning and ending parameters, etc.–so we would like to be able to deconflate complex data without any assumed background knowledge.

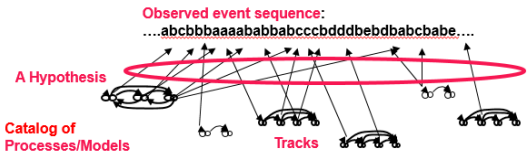

Figure 3: A state machine representation of conflated data objects or processes.

3.2 A General Solution to the Data Deconflation Problem

The approach that we describe below proposes to solve hard deconflation problems by the extension and application of DRL and GANs. We believe that a general solution to the deconvolution problem can be achieved by iterating between estimates of which signal component or element is contributed by which process (accomplished by DRL) and filtering done by using generative networks seeded by small signal samples.

To illustrate this iterative process, refer back to Figure 1. We might be estimating a classification based on a small sample portion of the image and then completing the small portion for that class using a generative model (e.g., sampling a small part of the school bus and using that sample to generate a more complete school bus image). We might then alternately filter in and out the generated constituent image to either isolate it and confirm identification or eliminate it to allow focusing on other objects. Though we are speculating about how a human might solve this particular problem, it is a reasonable starting point for investigating this difficult problem.

Our approach to the deconflation problem takes place over two phases. In the first phase we use GANs to model potential simple data objects based on observed complex data. Once simple data models have been generated, we will use DRL to approximate labeled complex data training sets by processes of inter-

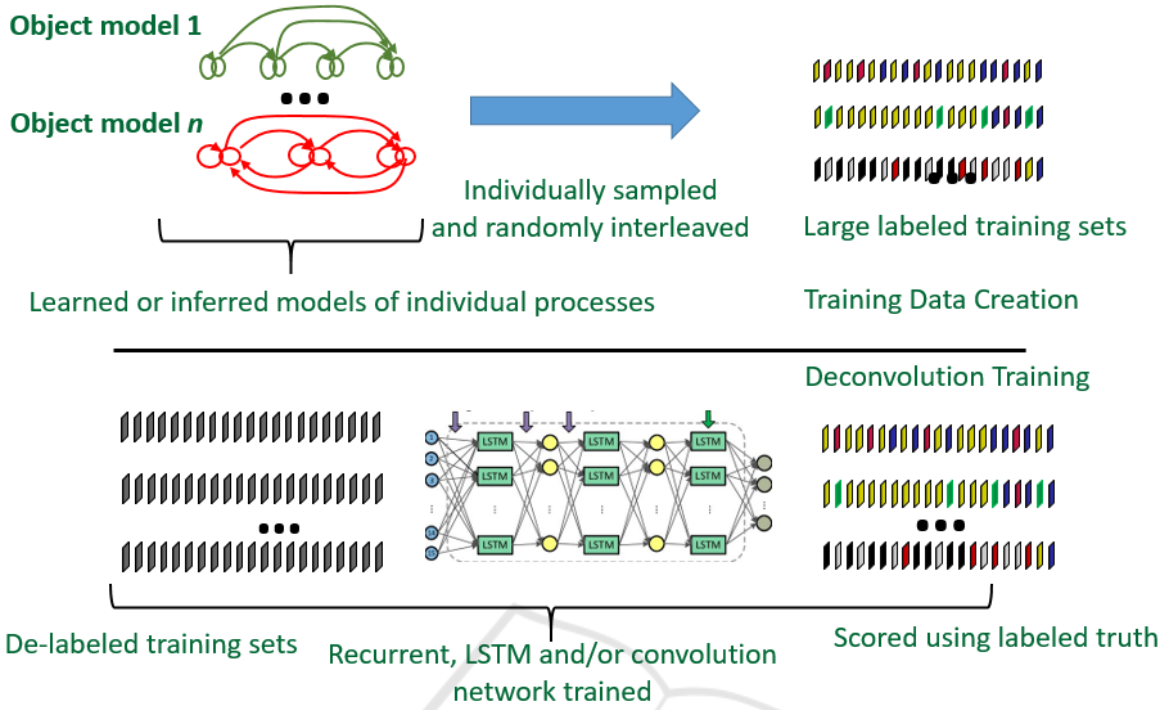

Figure 4: Our training process for deconflating complex data.

leaving and repetition of models. Once we have created an approximation of observed complex data, we use deep neural networks that have been trained to deconvolve complex data. Both phases of this approach are illustrated in Figure 4. This process is analogous to learning to play a game–the observed complex data sequence is the game state and label assignments are the player moves.

4 CONCLUSION

The adoption and emerging ubiquity of Internet-connected devices is leading us to a digital environment that is full of complex data streams that must be correctly deconflated in order to conduct meaningful analysis. While much of this data can be adequately separated through traditional BSS solutions, a non-trivial amount of this complex data is not vector representable and thus requires new deconflation solutions. In this paper we have described complex data objects that cannot be deconflated by current BSS solutions, and for which we have proposed a more general data deconflation problem. Furthermore, we have presented our vision for a general solution to the data deconflation problem that extends recent advances in DRL and GANs.

We are currently working on an initial proof-of-concept implementation. Other ongoing work on this effort includes a rigorous generalization of the data conflation process from vector representations to state machine representation (Section 3.1). We are also designing experiments to determine the appropriate structures for recurrent and/or convolutional neural networks to learn minimal simple data object models. Once we have demonstrated results with established with complex spatio-temporal data, we will extend our approach to non-spatio-temporal data, such as semantic conflations that might appear in pattern of life tracking.

ACKNOWLEDGEMENTS

Roger A. Hallman is partially supported by the United States Department of Defense SMART Scholarship for Service Program, funded by USD/R&E (The Under Secretary of Defense-Research and Engineering), National Defense Education Program (NDEP) / BA-1, Basic Research.

REFERENCES

Berk, V., Chung, W., Crespi, V., Cybenko, G., Gray, R., Hernando, D., Jiang, G., Li, H., and Sheng, Y. (2003). Process query systems for surveillance and awareness. In *In Proc. System. Cyber. Infor.(SCI2003*. Citeseer.

Berk, V. and Fox, N. (2005). Process query systems for network security monitoring. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense IV*, volume 5778, pages 520–530. International Society for Optics and Photonics.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Cybenko, G. and Berk, V. H. (2007). Process query systems. *Computer*, 40(1):62–70.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J. (2018). An introduction to deep reinforcement learning.

Giani, A., Berk, V., Cybenko, G., and Hanover, N. (2005). Covert channel detection using process query systems. In *proceedings of: FLoCon*.

Ginsburg, S. (1959). Synthesis of minimal-state machines. *IRE Transactions on Electronic Computers*, (4):441–449.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Haykin, S. and Chen, Z. (2005). The cocktail party problem. *Neural computation*, 17(9):1875–1902.

He, J., Chen, J., He, X., Gao, J., Li, L., Deng, L., and Ostendorf, M. (2016). Deep reinforcement learning with a natural language action space. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1621–1630.

Hughes, K. (2015a). Blind source separation on images with shogun. (Accessed via Internet Web Archive) http://shogun-toolbox.org/static/notebook/current/bss_image.html.

Hughes, K. (2015b). Blind source separation with the shogun machine learning toolbox. https://nbviewer.jupyter.org/github/kevinhughes27/bss_jade/blob/master/bss_jade.ipynb.

Hung, C.-C., Lillicrap, T., Abramson, J., Wu, Y., Mirza, M., Carnevale, F., Ahuja, A., and Wayne, G. (2019). Optimizing agent behavior over long time scales by transporting value. *Nature communications*, 10(1):1–12.

Kofidis, E. (2016). Blind source separation: Fundamentals and recent advances (a tutorial overview presented at sbrt-2001). *arXiv preprint arXiv:1603.03089*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Lee, D., Kim, J., Moon, W.-J., and Ye, J. C. (2019). Collagan: Collaborative gan for missing image data imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2487–2496.

Luo, Y., Cai, X., Zhang, Y., Xu, J., and Yuan, X. (2018). Multivariate time series imputation with generative adversarial networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1603–1614.

MathWorks (2020). Alexnet convolutional neural network. https://www.mathworks.com/help/deeplearning/ref/alexnet.html.

O'grady, P. D., Pearlmutter, B. A., and Rickard, S. T. (2005). Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, 15(1):18–33.

Qian, Y.-m., Weng, C., Chang, X.-k., Wang, S., and Yu, D. (2018). Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering*, 19(1):40–63.

Roblee, C., Berk, V., and Cybenko, G. (2005). Implementing large-scale autonomic server monitoring using process query systems. In *Second International Conference on Autonomic Computing (ICAC'05)*, pages 123–133. IEEE.

Schneider, F. B. (1990). The state machine approach: A tutorial. *Fault-tolerant distributed computing*, pages 18–41.

Shapiro, K. L., Caldwell, J., and Sorensen, R. E. (1997). Personal names and the attentional blink: A visual" cocktail party" effect. *Journal of Experimental Psychology: Human Perception and Performance*, 23(2):504.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tariq, S., Lee, S., Kim, H., Shin, Y., and Woo, S. S. (2018). Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd international workshop on multimedia privacy and security*, pages 81–87.

Wan, P., Hao, B., Li, Z., Zhou, L., and Zhang, M. (2016). Time differences of arrival estimation of mixed interference signals using blind source separation based on wireless sensor networks. *IET Signal Processing*, 10(8):924–929.

Zhang, Y., Zhou, B., Cai, X., Guo, W., Ding, X., and Yuan, X. (2021). Missing value imputation in multivariate time series with end-to-end generative adversarial networks. *Information Sciences*, 551:67–82.

Zhao, T. and Eskenazi, M. (2016). Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10.