Modeling Uplift from Observational Time-Series in Continual Scenarios

Anonymous submission

Abstract

As the importance of causality in machine learning grows, we expect the model to learn the correct causal mechanism for robustness even under distribution shifts. Since most of the prior benchmarks focused on vision and language tasks, domain or temporal shifts in causal inference tasks have not been well explored. To this end, we introduce **UpliftCRUD** dataset for modeling uplift in continual learning scenarios. We build a dataset using CRUD log data and construct continual learning tasks under temporal shifts and out-of-domain scenarios.

Introduction

Uplift is defined as Individual Treatment Effect (ITE), but its evaluation metric differs from other causal tasks (Radcliffe and Surry 1999). Following Radclifte and Simpson (2008), individuals can be segmented into four groups along two axes: **received treatment** and **response to it**. *Sure Things* will stay whether or not they receive treatment, and *Lost Causes* will leave in either case. However, *Persuadables* are likely to stay *only if* they receive the treatment, but *Sleeping Dogs* would be annoyed and will eventually leave. Based upon this, the main goal is thus to identify as many *Persuadables* as possible and avoid *Sleeping Dogs* for treatment.

In practice, the bottlenecks of causal models are data availability, scalability, and distribution shifts. In many cases where randomized controlled trials (RCTs) is infeasible, practitioners are given observational data. However, even with a large set of covariates, there may still be unobserved confounders, and the curse of dimensionality would occur. It is problematic, particularly in causal inference, as the chance of violating the positivity assumption increases. In addition, distribution change over time and domains results in wrong validation and, eventually the degradation of the model.

To challenge aforementioned issues with causality even in high-dimensional spaces and bridge the gap between research and practice environments, we constructed **UpliftCRUD** dataset[1], a real-world uplift dataset from mobile game users. The task is to predict uplift to push notifications by recognizing patterns from each user's CRUD history. A model must learn underlying causal mechanisms and continuously adapt itself to distributions varying over time and to other games. To the best of our knowledge, **UpliftCRUD** is the first uplift dataset with time-series under domain shifts.

[1]The dataset will be released to the public in early 2023.

Background

Causal inference and its notations. Potential outcomes framework (Rubin 1974) defines the causal effect as the difference between two potential outcomes $Y(1)-Y(0)$: when receiving treatment ($T=1$) and under control ($T=0$). The fundamental problem of causal inference (Holland 1986) states that either $Y_i(1)$ or $Y_i(0)$ is observable for each unit indexed by $i \in \{1,\dots,n\}$, and the unobserved outcome is called *counterfactual*. To estimate ITE, or uplift, $u_i := Y_i(1) - Y_i(0)$, we, therefore, model Conditional Average Treatment Effect (CATE), *i.e.*, $u(\mathbf{X}) := \mathbb{E}\left[Y(1) - Y(0)|\mathbf{X}\right]$. Among the assumptions needed to identify CATE, two assumptions are crucial and often likely to be violated (Pearl 2010; Neal 2020): unconfoundedness, *i.e.*, $Y \perp\!\!\!\perp T|\mathbf{X}$, and positivity, *i.e.*, $P(T|\mathbf{X}=\mathbf{x}) > 0, \forall \mathbf{x} : P(\mathbf{X}=\mathbf{x}) > 0$.

Time-series modeling. Time-series is a sequence of discrete-time data. Many previous works have dealt with regular time-series, but in this paper, we mainly focus on irregular time-series, where intervals between two consecutive data points are not regular. RNNs (Rumelhart, Hinton, and Williams 1986; Hochreiter and Schmidhuber 1997; Cho et al. 2014), TCNs (Bai, Kolter, and Koltun 2018) with dilated convolutions (Yu and Koltun 2015), and Transformers (Vaswani et al. 2017) have become popular choices for handling irregular time-series data. However, there is no *one-size-fits-all* augmentation strategy in irregular time-series (Yue et al. 2022) except for dropout (Srivastava et al. 2014), or masked autoencoding (Devlin et al. 2018; He et al. 2022).

Continual learning. Continual Learning (CL) aims to effectively learn new tasks and adapt a model to distribution shifts over time while minimizing performance degradation in the learned scenarios, which is called *catastrophic forgetting* (Kirkpatrick et al. 2017). It is also infeasible in practice to fully retrain the model whenever new data is available due to training costs or the unavailability of previous data. Therefore, recent algorithms for this aim to accumulate knowledge and reuse them in future scenarios without forgetting information (iCaRL (Rebuffi et al. 2017), A-GEM (Chaudhry et al. 2019), EWC (Kirkpatrick et al. 2017),

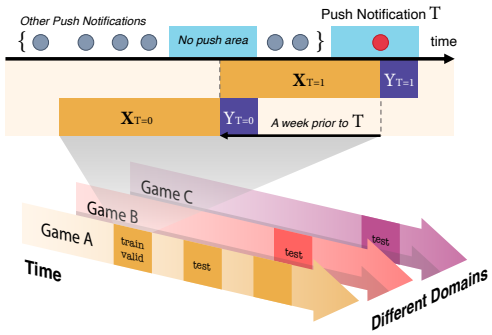

Figure 1: Illustration of dataset construction.

SI (Zenke, Poole, and Ganguli 2017)). Moreover, causal inference tasks require the model to capture the causal mechanism over distributional shifts, on which existing CL algorithms have not focused.

Previous Benchmarks

Benchmarks for uplift modeling. Researchers on uplift have relied on (semi-)synthetic data for testing algorithms since counterfactuals do exist. As of now, the largest observational benchmark is **Criteo dataset** (Eustache et al. 2018) with 12 static features from ∼14M real-world users. Thus far, there has been little motivation to use deep learning, and therefore, related works have been restricted to smaller models (# params $<$ 1K) or other ML algorithms. With regard to time-series, a subset of **MIMIC II/III** (Johnson et al. 2016) has been used for causal discovery or inference. See Moraffah et al. (2021) for a comprehensive review.

Benchmarks for CL. Benchmarks in various fields and tasks with CL scenarios have been introduced, e.g., object recognition in robotics (Fanello et al. 2013; Lomonaco and Maltoni 2017; She et al. 2020), classification tasks in various domains on images (Rebuffi, Bilen, and Vedaldi 2017; Lake, Salakhutdinov, and Tenenbaum 2015; He, Shen, and Cui 2021), videos (Roady et al. 2020), 3D objects (Stojanov et al. 2019), and natural language (Hussain et al. 2021; Srinivasan et al. 2022). However, domain or temporal shifts in causal inference tasks have not been well explored.

UpliftCRUD Dataset

Background. We collected data from a Backend-as-a-Service (BaaS) company specializing in mobile games, allowing game developers to quickly release their apps without backend servers of their own. One of their features is sending **a push notification** to *all users at the same time*.

Construction. In a typical observational dataset, the treatment is not given randomly. In our data, however, the treatment group only exists as the treatment (push message) is given to the whole population simultaneously. To circumvent this problem, for a train set, we sampled a *pseudo*-control group from exactly one week prior to the push to eliminate the time and weekday effect. We also introduced a concept of *no push area*, an -12∼+6 hour window around which no other pushes must exist for preventing interference from the other pushes. Still, some gamers exist in both groups. For a test set, we randomly split those overlapping users into either group to simulate RCT.

Tasks. The dataset consists of three games (A, B, and C) with a total of 16.7M lines of CRUD logs from 5,360 users[2]. Each consists of a triple $(\mathbf{X}_i, T_i, Y_i)$, where $\mathbf{X}_i$ is a sequence of categorical variables along with corresponding timestamps, and $T_i, Y_i \in \{0, 1\}$ are binary indicators of the treatment and whether a user login within three hours. We experimented on uplift modeling in the proposed tasks:

- In-distribution (ID): train with the game A (APR, MAY) and test on 20% random-split holdout set.
- Temporal shift (TS): train with the game A (APR, MAY) and test on the game A (JUN).
- Out-of-distribution (OOD): train with the game A (APR, MAY) and test on the game B with fine-tuning (OOD w/) or on the game C without fine-tuning (OOD w/o).

Experiments

Model	Ckpt	ID	TS	OOD w/	OOD w/o
Dragon	VAL	.091/.056	.006/.003	.118/.038	.037/.023
	MAX		.112/.074	.372/.082	.123/.081
Siamese	VAL	.145/.062	-.036/-.011	.154/.057	-.057/-.030
	MAX		.249/.067	.207/.075	.036/.022
$P(Y=1)$		11.9%	12.2%	5.9%	22.4%

Table 1: Baseline results. VAL denotes the best checkpoint on the holdout set, and MAX denotes the best metric during entire training, showing the discrepancy of the performance.

We used Dragonnet (Shi, Blei, and Veitch 2019), Siamese network (Mouloud, Olivier, and Ghaith 2020) with TCN backbones and EWC for CL. Table 1 shows qini coefficients (left) and area under uplift curve (AUUC) (right) (Devriendt et al. 2020), both of which measure incremental gains. The difference between VAL and MAX can be attributed to the difficulty of the proper validation due to distribution shifts. Also, the performance in TS and OOD w/o dropped sharply. We conjecture that it is due to the model being overfitted to the train set without learning causal mechanisms, and a model should perform equally well both in ID and OOD.

Conclusion and Future Work

In this paper, we propose uplift tasks and **UpliftCRUD** dataset, combining causal inference with CL scenarios. We demonstrate that naïvely applying existing methods may fail as uplift modeling tries to predict future behaviors based on historical data. All observational datasets have inherent biases; identifying causal relationship and eliminating undesirable effects would be one of the important follow-up research topics. We believe that learning causal mechanisms which are invariant over time is crucial towards general-level AI, and that the dataset will hopefully contribute to developing these algorithms.

[2]More data will be included in the public release.

References

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2019. Efficient Lifelong Learning with A-GEM. In *International Conference on Learning Representations*.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devriendt, F.; Van Belle, J.; Guns, T.; and Verbeke, W. 2020. Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*.

Eustache, D.; Artem, B.; Renaudin, C.; and Massih-Reza, A. 2018. A Large Scale Benchmark for Uplift Modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London,United Kingdom, August, 20, 2018*. ACM.

Fanello, S. R.; Ciliberto, C.; Santoro, M.; Natale, L.; Metta, G.; Rosasco, L.; and Odone, F. 2013. iCub World: Friendly Robots Help Building Good Vision Data-Sets. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 700–705.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, Y.; Shen, Z.; and Cui, P. 2021. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110: 107383.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960.

Hussain, A.; Holla, N.; Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2021. Towards a robust experimental framework and benchmark for lifelong language learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Lomonaco, V.; and Maltoni, D. 2017. CORe50: a New Dataset and Benchmark for Continuous Object Recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78, 17–26.

Moraffah, R.; Sheth, P.; Karami, M.; Bhattacharya, A.; Wang, Q.; Tahir, A.; Raglin, A.; and Liu, H. 2021. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 1–45.

Mouloud, B.; Olivier, G.; and Ghaith, K. 2020. Adapting neural networks for uplift models. *arXiv preprint arXiv:2011.00041*.

Neal, B. 2020. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*.

Pearl, J. 2010. Causal inference. *Causality: objectives and assessment*, 39–58.

Radcliffe, N.; and Surry, P. 1999. Differential response analysis: Modeling true responses by isolating the effect of a single action. *Credit Scoring and Credit Control IV*.

Radclifte, N. J.; and Simpson, R. 2008. Identifying who can be saved and who will be driven away by retention activity. *Journal of Telecommunications Management*, 1(2).

Rebuffi, S.-A.; Bilen, H.; and Vedaldi, A. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Roady, R.; Hayes, T. L.; Vaidya, H.; and Kanan, C. 2020. Stream-51: Streaming Classification and Novelty Detection From Videos. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *nature*, 323(6088): 533–536.

She, Q.; Feng, F.; Hao, X.; Yang, Q.; Lan, C.; Lomonaco, V.; Shi, X.; Wang, Z.; Guo, Y.; Zhang, Y.; Qiao, F.; and Chan, R. H. M. 2020. OpenLORIS-Object: A Robotic Vision Dataset and Benchmark for Lifelong Deep Learning. In *2020 International Conference on Robotics and Automation (ICRA)*, 4767–4773.

Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Srinivasan, T.; Chang, T.-Y.; Alva, L. L. P.; Chochlakis, G.; Rostami, M.; and Thomason, J. 2022. CLiMB: A Continual Learning Benchmark for Vision-and-Language Tasks. *arXiv preprint arXiv:2206.09059*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958.

Stojanov, S.; Mishra, S.; Thai, N. A.; Dhanda, N.; Humayun, A.; Yu, C.; Smith, L. B.; and Rehg, J. M. 2019. Incremental Object Learning From Contiguous Views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8980–8987.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, 3987–3995. PMLR.