Self-Auxiliary Distillation for Sample Efficient Learning in Google-Scale Recommenders

Yin Zhang
Google DeepMind, USA
yinzh@google.com

Ruoxi Wang
Google DeepMind, USA
ruoxi@google.com

Xiang Li
Google, Inc, USA
lxiange@google.com

Tiansheng Yao
Google, Inc, USA
tyao@google.com

Andrew Evdokimov
Google, Inc, USA
andrewev@google.com

Jonathan Valverde
Google DeepMind, USA
valverdej@google.com

Yuan Gao
Google, Inc, USA
yuangaolearn@google.com

Jerry Zhang
Google, Inc, USA
jerryzh@google.com

Evan Ettinger
Google, Inc, USA
eettinger@google.com

Ed H. Chi
Google DeepMind, USA
edchi@google.com

Derek Zhiyuan Cheng
Google DeepMind, USA
zcheng@google.com

ABSTRACT

Industrial recommendation systems process billions of daily user feedback which are complex and noisy. Efficiently uncovering user preference from these signals becomes crucial for high-quality recommendation. We argue that those signals are not inherently equal in terms of their informative value and training ability, which is particularly salient in industrial applications with multi-stage processes (e.g., augmentation, retrieval, ranking). Considering that, in this work, we propose a novel self-auxiliary distillation framework that prioritizes training on high-quality labels, and improves the resolution of low-quality labels through distillation by adding a bilateral branch-based auxiliary task. This approach enables flexible learning from diverse labels without additional computational costs, making it highly scalable and effective for Google-scale recommenders. Our framework consistently improved both offline and online key business metrics across three Google major products. Notably, self-auxiliary distillation proves to be highly effective in addressing the severe signal loss challenge posed by changes such as Apple iOS policy. It further delivered significant improvements in both offline (+17% AUC) and online metrics for a Google Apps recommendation system. This highlights the opportunities of addressing real-world signal loss problems through self-auxiliary distillation techniques.

ACM Reference Format:
Yin Zhang, Ruoxi Wang, Xiang Li, Tiansheng Yao, Andrew Evdokimov, Jonathan Valverde, Yuan Gao, Jerry Zhang, Evan Ettinger, Ed H. Chi, and Derek Zhiyuan Cheng. 2024. Self-Auxiliary Distillation for Sample Efficient Learning in Google-Scale Recommenders. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3640457.3688041

1 INTRODUCTION

Industrial recommendation systems consider various user feedback signals, such as user clicks and installs, to infer users' nuanced interests. This fundamentally impacts recommendation quality [3, 4, 16]. One common approach is to directly use those signals as labels to train recommenders. However, these labels are not inherently equal in terms of their informative value and training ability [10, 12, 13], especially in industrial contexts. For example, for click-through rate (CTR) models [9], we postulate that positive labels offer more value and reliability compared to negatives in predicting pCTR. This is because negative samples in pCTR models often resemble weak positives, as they have successfully passed through the candidate generation and auction bidding stages. Hence, the distinction between a negative and a positive label is less pronounced than the difference between 0 and 1. In light of this, representing negative labels with an estimated pCTR value, rather than assigning an absolute 0, offers a more granular and informative approach for capturing underlying user preferences. Similar scenarios are also applied in other productions, such as user conversions.

Considering this, in this work, we propose to efficiently uncover user preference by putting more energy on high-quality label training, while improving resolution in low-quality labels through distillation. For example, in the pCTR model, we can train on pseudo-labels $max(y, y')$ that are a mix of original labels y and teacher-generated soft labels y'. This allows the model to focus on learning users' strong preference (represented by valuable positive labels) while obtaining more resolutions among the negatives offered through those soft labels [15, 18]. A key challenge when deploying this method in production is calibration, where many products require that the serving mean prediction need to be the

RecSys '24, October 14–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688041

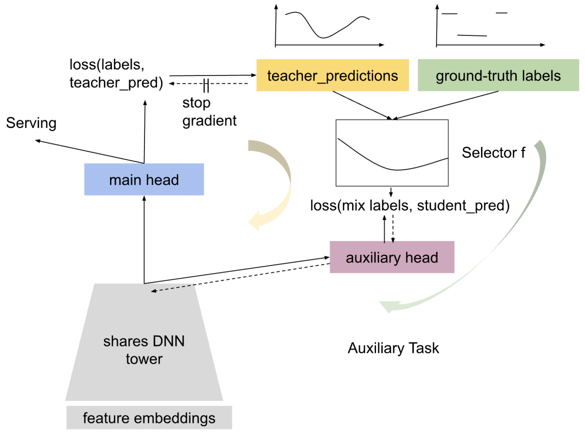

Figure 1: General framework of Self-auxiliary Distillation.

same as the mean of ground-truth labels, to avoid model over/under-predictions. However, directly learning from these pseudo-labels will easily move its mean prediction away from the true-label means. To effectively learn from labels and maintain the calibration, we proposed a novel self-auxiliary distillation framework that (1) introduces the pseudo-label prediction task as an auxiliary task; (2) efficiently generates pseudo-labels through a bilateral-branch architecture and curriculum learning. More importantly, there's no extra serving cost and a negligible training overhead because only an auxiliary head is introduced and the auxiliary task is trained together with main tasks. It makes our method highly scalable for Google-scale recommenders. Self-auxiliary distillation has delivered significant improvements in both offline and online key business metrics, and has been deployed in three major Google products: apps, commerce, and video recommendations.

Beyond general scenarios, our method also effectively addresses the industry-wide challenge of signal loss problems. Recent changes in Apple iOS policy [5] has led to a substantial decrease in crucial conversion labels, significantly hurting model performance. For example, Apple's iOS App Tracking Transparency policy [6–8] requires user's explicit consent for conversion tracking. This resulted in biased per example label, with few positives and a large amount of false negatives, as non-consented user conversions cannot be distinguished from true negatives. To address the signal loss problem, we applied a variant of our method, which has delivered significant offline and online improvements in a Google Apps recommendation system.

2 SELF-AUXILIARY DISTILLATION

We propose the Self-auxiliary Distillation, a novel self-distillation framework to facilitate efficient model learning through mixed labels. It has a shared bottom tower with two heads (depicted in Figure 1). (1) The main head trains on the ground-truth labels, and serves both the inference head and a teacher to generate soft labels. To ensure well calibration at serving time, its mean prediction need to be the same as mean label. (2) The auxiliary head works as student, and learns from a combination of teacher-generated soft labels from main head and the original ground-truth labels.

Specifically, the auxiliary head leverages a bilateral branch [17] to flexibly learn from the two types of labels. One branch (distillation branch shown in yellow), distills knowledge from the soft labels produced by the main head. This allows the auxiliary head to simultaneously receive feedback from the main head to align their learning objectives, including information on the current training status and teacher logit distribution. The other branch (scratch training branch shown in green), learns from ground-truth labels. This branch is specifically designed to further train examples where the main head hasn't fully extracted all the predictive information, such as those with high-quality labels or rich information. Then, a selector merges the two types of labels from each branch as $f(y, y')$ for the auxiliary head training. Different adaptors [14], curriculum learning strategies [2] and other functions (e.g. max) can be used as f. The structure enables the auxiliary head to train on diverse label combinations without affecting the serving mean prediction in main head. Consequently, our method can flexibly prioritize and allocate more resources to high-quality labels and maintain calibration.

It is also worth highlighting that the remaining model components (bottom tower and feature embeddings) are shared between teacher and student, and updated jointly during training (shown in Figure 1). This makes self-auxiliary distillation a more cost-efficient and scalable framework.

3 EXPERIMENTS AND ONLINE STUDY

The mixed labels in auxiliary head play a pivotal role in boosting model performance. We conduct a comprehensive study of self-auxiliary distillation (SEAD) against related SOTA and variations in productions. Table 1 shows their offline performances on three Google production models. Each model is continuously trained and updated in a streaming fashion [1]. For SEAD, $max(y, y')$ is applied in the selector, where $y = 1$ means the user clicks or installs, otherwise $y = 0$; y' is the soft labels produced from the main head.

As shown in Table 1, self-auxiliary distillation delivered significant improvements across different production models. In auxiliary head, training only on original labels (OGL) showed inconsistent results, often causing model instabilities; training only on distillation labels (DTL) did not bring extra improvements. These confirm that in the self-auxiliary distillation, the scratch training branch primarily drives model performance improvements, while the distillation branch stabilizes the model and ensures consistency with the main head. Training solely on positive labels in the auxiliary task showed inferior performance, highlighting the importance of distilling negative examples. Additionally, stopping the gradient updates from auxiliary head on shared components (SGA) showed neutral performance, verifying that incorporating SEAD auxiliary task improves the model performance rather than the infra trick. Doubling the learning rate (LR) triggered instability issue in some cases, and we find optimizing the learning rate complemented the gains from self-auxiliary distillation.

For the selectors, utilizing positive labels and distilling on negatives $max(y, y')$ outperforms the vice versa $min(y, y')$, which is consistent with our assumptions. If curriculum learning is used in the selector, emphasizing positive labels (e.g. by assigning them higher weights) in early training stages tends to enhance the model

Table 1: Self-auxiliary distillation applied to different production models. Note that +0.1% is considered significant due to the large traffic in online A/B experiments.

AUC	SEAD	OGL	DTL	LR	SGA	$min(y, y')$
pCTR Model1	+0.35%	+0.35%	neutral	+0.03%	neutral	+0.20%
pCTR Model2	+0.20%	blow-up	neutral	blow-up	neutral	-0.03%
pCVR Model	+0.26%	+0.26%	neutral	+0.22%	neutral	-0.05%

performance. Besides offline improvements, online A/B testing demonstrated that self-auxiliary distillation yields significant improvements in key business metrics in the three Google products.

Addressing Signal loss. We then discuss how self-auxiliary distillation effectively mitigate the signal loss problems. Data and models are affected by the Apple iOS policy [7, 11], where there are only two sources of conversion labels: per example label for consented users, and coarse aggregated labels (i.e. SKAdNetwork [1]). This resulted in biased per example label with few positives and a large amount of false negatives, as non-consented user conversions cannot be distinguished from true negatives. Thus, we applied the self-auxiliary distillation idea with some modifications. The new model contains a teacher tower and a student tower. For teacher, the main head directly learns from an unbiased data traffic, and the auxiliary head learns from policy-affected biased data. Student learns from $max(y, y')$ and then uses SKAdNetwork for calibration.

Table 2: Our proposed model performances on signal loss settings. +0.1% is considered significant.

Model	AUC	SKAdNetworkAUC	SimulationAUC	Online
SEAD	+17%	+3%	+2%	positive

For evaluating signal loss problems with no ground-truth labels, we consider difference aspects: the model's AUC on predicting user's event-level feedback, SKAdNetwork label AUC, simulated performance where we have full ground-truth labels, and online A/B experiments. Table 2 shows that our method significantly improves all the offline and online evaluations metrics. Notably, the improvement in SKAdNetwork labels demonstrated our method's strong generalization ability, as no SKAdNetwork label is used in distillation. Our model has been deployed in a Google App recommender, delivering significant improvements for signal loss problems.

4 CONCLUSION

Sample efficient learning from labels is a crucial problem for recommenders. Our proposed self-auxiliary distillation is a well-tested solution in Google-scale systems and also shows effectively mitigating the real-word signal loss problems in productions.

5 SPEAKER BIO

Yin Zhang: Yin is a senior engineer and tech lead. She focuses on privacy-aware learning, model generalization and LLM for recommendation. She got her PhD from TAMU working on Recommender Research.

Ruoxi Wang: Ruoxi is a senior staff research engineer and tech lead specializing in model and data efficiency. Her expertise spans efficient feature cross learning, transfer learning, and sample efficient learning, as well as LLM for Recommenders. She holds a PhD in applied math from Stanford, and enjoys her time hanging out with her golden retriever Meimei.

REFERENCES

[1] Rohan Anil, Sandra Gadanho, Da Huang, Nijith Jacob, Zhuoshu Li, Dong Lin, Todd Phillips, Cristina Pop, Kevin Regan, Gil I Shamir, et al. 2022. On the factory floor: ML engineering for industrial-scale ads recommendation models. *arXiv preprint arXiv:2209.05310* (2022).

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[3] Chi Chen, Hui Chen, Kangzhi Zhao, Junsheng Zhou, Li He, Hongbo Deng, Jian Xu, Bo Zheng, Yong Zhang, and Chunxiao Xing. 2022. Extr: click-through rate prediction with externalities in e-commerce sponsored search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2732–2740.

[4] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 935–944.

[5] Garrett Johnson, Julian Runge, and Eric Seufert. 2022. Privacy-centric digital advertising: Implications for research. *Customer Needs and Solutions* 9, 1 (2022), 49–54.

[6] Reinhold Kesler. 2022. The Impact of Apple's App Tracking Transparency on App Monetization. *Available at SSRN 4090786* (2022).

[7] Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. 2022. Goodbye tracking? Impact of iOS app tracking transparency and privacy labels. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 508–520.

[8] Lennart Kraft, Bernd Skiera, and Tim Koschella. 2023. Economic impact of opt-in versus opt-out requirements for personal data usage: The case of apple's app tracking transparency (ATT). *Available at SSRN 4598472* (2023).

[9] Junwei Pan, Wei Xue, Ximei Wang, Haibin Yu, Xun Liu, Shijie Quan, Xueming Qiu, Dapeng Liu, Lei Xiao, and Jie Jiang. 2024. Ad Recommendation in a Collapsed and Entangled World. *arXiv preprint arXiv:2403.00793* (2024).

[10] Andreas Pfadler, Huan Zhao, Jizhe Wang, Lifeng Wang, Pipei Huang, and Dik Lun Lee. 2020. Billion-scale recommendation with heterogeneous side information at taobao. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1667–1676.

[11] Ian Thomas. 2021. Planning for a cookie-less future: How browser and mobile privacy changes will impact marketing, targeting and analytics. *Applied marketing analytics* 7, 1 (2021), 6–16.

[12] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 269–277.

[13] Abdou Youssef. 1999. Image downsampling and upsampling methods. *National Institute of Standards and Technology* (1999).

[14] Yin Zhang, Ruoxi Wang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Lichan Hong, James Caverlee, and Ed H Chi. 2023. Empowering Long-tail Item Recommendation through Cross Decoupling Network (CDN). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5608–5617.

[15] Zhilu Zhang and Mert Sabuncu. 2020. Self-distillation as instance-specific label smoothing. *Advances in Neural Information Processing Systems* 33 (2020), 2184–2195.

[16] Jiawei Zheng, Hao Gu, Chonggang Song, Dandan Lin, Lingling Yi, and Chuan Chen. 2023. Dual Interests-Aligned Graph Auto-Encoders for Cross-domain Recommendation in WeChat. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4988–4994.

[17] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9719–9728.

[18] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. 2021. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *arXiv preprint arXiv:2102.00650* (2021).

[1] https://developer.apple.com/documentation/storekit/skadnetwork/