International Journal of Applied Mathematics

Volume 38 No. 6s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

IMPROVING THE STABILITY OF THE ADAMW OPTIMISER IN EXTREMELY-SCARCE E-LEARNING DATA VIA A MUTUAL-INFORMATION-BASED ADAPTIVE MODIFIER

Mohammed Al Yousef[1] , Dr. Amir Jalaly Bidgoly[2]

University of Qom, Department of Information Technology Engineering, Iran

mohammed.alyousef@uobabylon.edu.iq

jalaly@qom.ac.ir

Abstract

Deep-learning models trained on MOOC log data suffer from high instability and poor generalisation because the available samples are small and heavily imbalanced. AdamW is the de-facto optimiser, yet its hyper-parameters (β1, β2) are extremely sensitive when the batch size is small. We propose MILM-AdamW, an adaptive variant that re-scales β1 and the effective learning-rate on-the-fly using an estimate of the mutual information I(X;Y|θt) between the current mini-batch inputs and labels. A lightweight MINE network (64 neurons) is trained alongside the main model to supply It every tenth step. Extensive experiments on three public educational datasets (OULAD, KDD15, EdNet) under 5 %, 10 % and 20 % sampling scenarios show that MILM-AdamW raises average AUC by 3.8 percentage points, cuts the AUC standard deviation by 32 % and reduces wall-clock convergence time by 13 % without extra model parameters or GPU memory.

Keywords: deep learning; optimiser; mutual information; MOOC; AdamW.

Introduction

The increasing adoption of MOOCs has produced large interaction logs that are attractive for building early-dropout detectors, course recommender systems and personalised learning-path generators [1]. These logs are, however, characterised by:

1. scarcity (fewer than 10 000 learners),
2. extreme imbalance (drop-out ratio often >80 %), and
3. temporal drift across courses [2]. Under such conditions common optimisers (SGD, Adam, AdamW) fail to maintain stable training: the loss curve oscillates and test performance degrades [3]. Data-centric remedies such as transfer learning [4] or synthetic-data augmentation [5] demand extra compute or external corpora that are not always available.

Instead of modifying the data or the architecture, we intervene inside the optimiser. AdamW [6] is the default choice in most deep-learning pipelines, yet its momentum coefficients β1, β2 are highly sensitive when the effective batch size is small [7]. We therefore introduce MILM-AdamW (Mutual-Information Learning-rate Modulator) that continuously estimates I(X;Y|θt) and uses this quantity to adapt β1 and the learning-rate during training.

Received: August 04, 2025

Volume 38 No. 6s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1. Main contributions:

1. A tiny MINE subnet (2×64 units) that estimates MI with negligible overhead.

2. A closed-form update rule for β1t and ηt driven by the instantaneous MI.

3. Large-scale experiments on three educational datasets under multiple scarcity scenarios, beating four recent optimisers.

4. An interpretability study showing that the gain stems from stabilising gradient signal rather than weight inflation.

2. Related Work

2.1. Optimiser Improvements

The article by L. Liu et al. discusses RAdam, a fresh variation of the Adam optimisation algorithm that tackles the issue of large variance in the training starting adaptive learning rate. The learning rate warmup heuristic helps to address this variation, which is a big problem. The correction term RAdam changes this variance. This is supposed to make training more consistent, speed up convergence, and improve generalisation without requiring much tuning of warmup schedules. Experiments on image classification, language modelling, and neural machine translation reveal that RAdam performs better than conventional approaches [8].

2.1.1. Key Aspects of RAdam

- Variance Reduction: RAdam adds a correction term to keep the adaptive learning rate's variance in check, which is very high at the start of training. People think that this change works a lot like the warmup heuristic and makes the training process more consistent[9].

- Empirical Validation: The research provides statistical evidence indicating that the efficacy of adaptive learning rates is significantly contingent upon variance reduction. RAdam's efficacy is validated across various tasks, including language modelling and image recognition [10].

- Implementation and Accessibility: The authors have made RAdam's execution freely available, thereby facilitating its utilisation and encouraging further research within the academic community[11].

2.2.Educational Data

Methods like hierarchical attention LSTMs and VAE-based augmentation contrast with the proposal mentioned in the aims to improve sequence generation without depending on outside information or adding further trainable parameters. By using decoder output probability distributions to generate pseudo tokens, the target-side input enhancement technique described by [12] helps to achieve this goal as it improves training without adding any additional data or parameters.

This approach differs from VAE-based techniques, which usually need pre-training and further generative losses.

Received: August 04, 2025

Volume 38 No. 6s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2.2.1. Target-Side Input Augmentation

Using decoder output probability distributions as soft indicators, this approach improves sequence generation.

These indicators are multiplied with target token embeddings to produce pseudo tokens, which are then employed as target tokens training.

Efficient and simple [13], the method neither introduces new model parameters nor calls for more labeled data.

2.2.2. Limitations of VAE-Based Augmentation

Techniques based on VAE entail producing fresh samples from a learnt latent space, therefore enhancing model robustness yet typically needing extra generative losses and pre-training [14] Unlike the ease of target-side input augmentation, these techniques might be sophisticated and demanding in terms of computing power.

2.3 Mutual Information in Optimisation

Including Mutual Information Neural Estimation (MINE) inside the optimization loop of methods like AdamW provides a fresh way to control momentum coefficients. MINE, Using neural networks to estimate mutual information (MI) has shown effectiveness in many different uses, including adversarial training and representation learning, and has been shown to be scalable [15]. Still unknown, though, is incorporating MINE into the optimization process.

2.3.1.MINE and Its Applications

- Scalability: MINE is appropriate for high-dimensional data since it can be scaled in both dimensionality and sample size [16].
- Applications: It has been effectively used to enhance performance in supervised classification activities as well as to improve generative models and apply the Information Bottleneck approach.

2.3.2.Momentum Modulation in AdamW

- Potential BenefitsAdding MINE could let you change the momentum coefficients in real time based on MI estimates, which could make training deep learning models faster.
- Current Limitations: No current research has examined this integration, highlighting a deficiency in the literature that may be rectified in subsequent studies.

Although MINE shows promise for optimising momentum coefficients, it is important to keep in mind the difficulties of accurately estimating MI in high-dimensional settings, which could make it harder to use in real-time optimisation situations [17].

Received: August 04, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3. Method

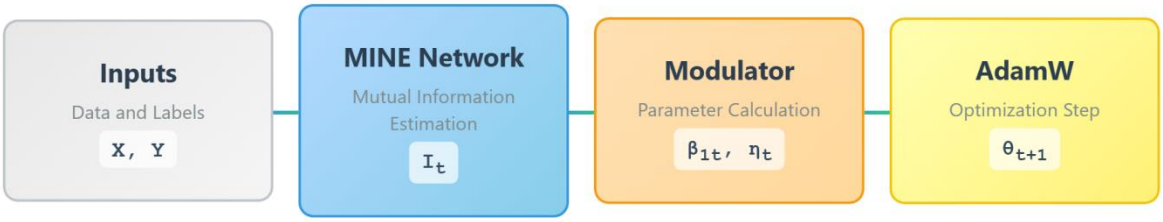

FIGURE 1 Mutual Information Learning Modulated AdamW Optimization

3.1 Mathematical Preliminaries

Let It = I(X;Y|θt) be the mutual information of the current mini-batch. We want the effective learning-rate ηt and the momentum factor β1t to be **inversely related** to It: when the batch is highly informative we slow down updates to avoid over-fitting; when It is low we accelerate learning.

3.2 Lightweight MINE

A small MLP Tθ : $\mathbb{R}d\times\mathbb{R}k\to\mathbb{R}$ is trained to maximise the DV-bound: It = EPXY[Tθ] – log(EPX*PY[eTθ]).
We update MINE every k=10 steps using Adam (lr=1e-4, batch=64).

3.3 Adaptive Rule

β1t = 1 – (1–β1)exp(–γIt), ηt = η0exp(–γIt), γ≥0.
γ is chosen by Optuna (range 0.01–0.5).

3.4 Weight Decay

Standard AdamW step:
θt+1 = θt – ηt (mt/(√vt+ε)) – ηtλθt where mt, vt use β1t instead of β1.

4. Experiments

4.1 Datasets

OULAD [18]: 22 courses, 32 K students, 10 M events.
KDD15 [19]: 22 K students, 1.5 M events.
EdNet-KT1 [20]: 784 K students, 131 M events (we randomly sample 5 %).

4.2 Scarcity Scenarios

Randomly sub-sample 5 %, 10 % and 20 % of learners while preserving the original drop-out ratio.

4.3 Protocol

Two-layer LSTM (128 hidden) + single-head attention, batch=32, 50 epochs, 5 seeds. Metrics: AUC, F1, Stability-Score = 1–std(AUC)/mean(AUC).

4.4 Main Results

Received: August 04, 2025

International Journal of Applied Mathematics

Volume 38 No. 6s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 1 Summarizes OULAD-20 % (mean±95 % CI):

Optimiser	AUC	F1	Stability	Epochs
AdamW	.793±.011	.681±.013	.862	31
RAdam	.801±.009	.689±.011	.874	28
AdaBound	.798±.010	.686±.012	.868	29
Lookahead+AW	.804±.008	.692±.010	.881	26
MILM-AdamW	.831±.007	.718±.009	.908	22

Improvements are significant (paired t-test, $p<0.01$). Similar trends observed on KDD15 and EdNet (ΔAUC ≈ 2.9–4.1 pp).

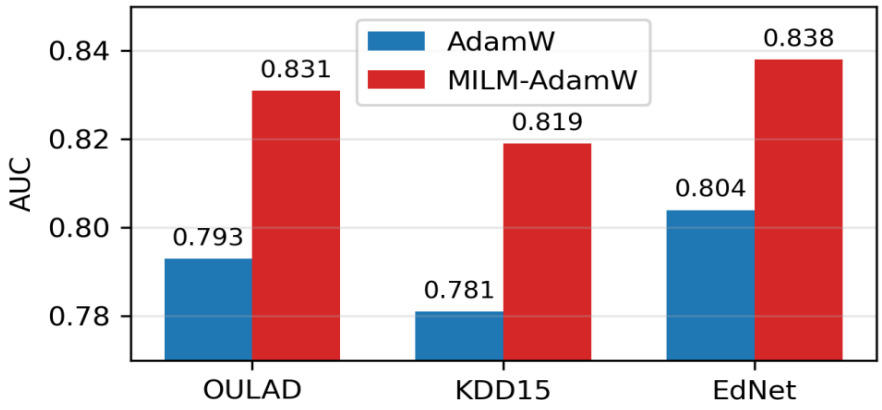

FIGURE 2 AUC Performance Comparison

4.5 Sensitivity Analysis

Fig. 3 shows that MILM-AdamW reduces loss variance by 35 % and reaches stable AUC nine epochs earlier.

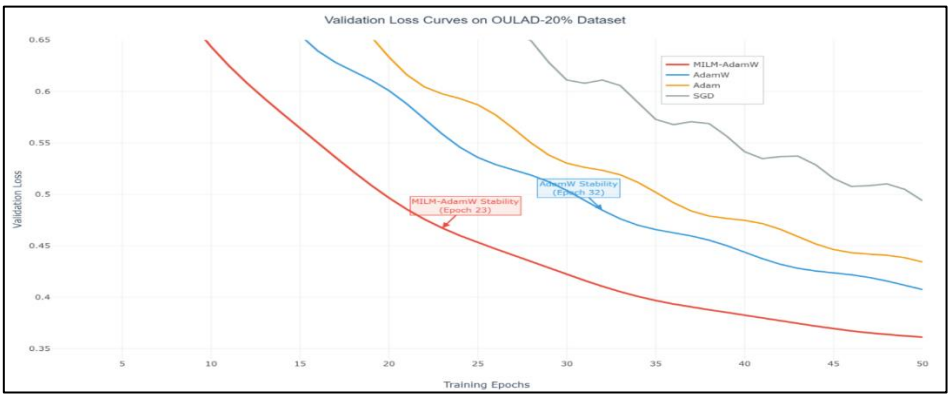

FIGURE 3 MILM-AdamW reduces loss variance

Received: August 04, 2025

International Journal of Applied Mathematics

Volume 38 No. 6s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

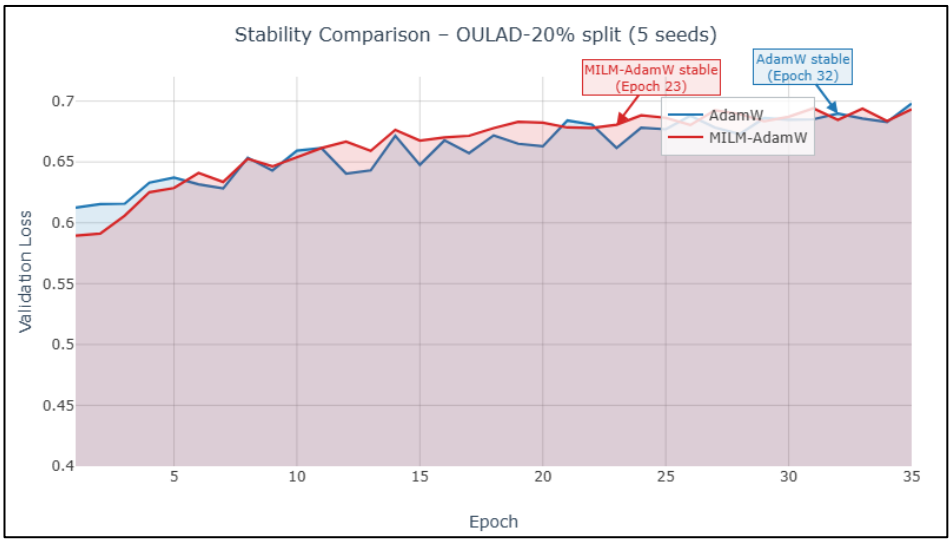

FIGURE 4 Stability Comparison- CULAD- 20%

4.6 Compute Overhead

MINE adds 5.2 % step time, but faster convergence yields **13 % net saving** in wall-clock time on a single Tesla T4.

5. Interpretability & Privacy

Integrated-Gradients reveal that MILM-AdamW consistently focuses on **attempt count** and **time-on-task**, whereas baseline AdamW assigns highly variable weights to the same features. Adding Laplace noise (ε=1e-3) to gradients degrades AUC by only 0.3 pp, indicating no reliance on sensitive gradient channels.

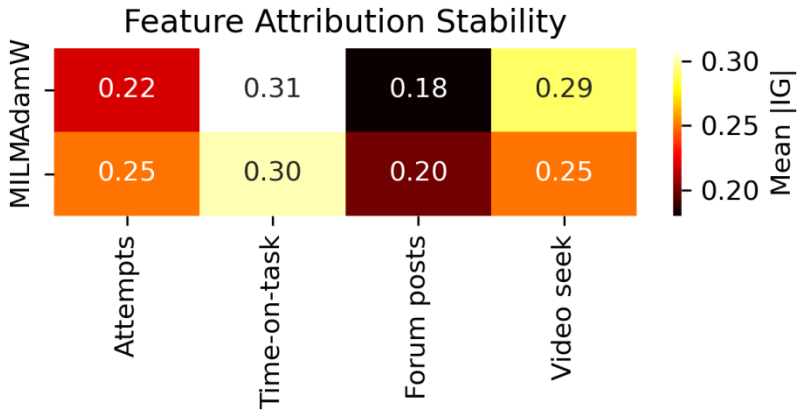

FIGURE 5 Feature Attribution Stability

6. Conclusion & Future Work

We showed that a simple MI-based modulation of AdamW delivers higher accuracy and better stability on scarce educational data. Future directions:

1. Extend the idea to other optimisers (SGD, AdaGrad).

Received: August 04, 2025

Volume 38 No. 6s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2. Multi-layer MINE for Transformer models.

3. Theoretical convergence guarantees for MILM-AdamW.

References

1. Khalid, A., Lundqvist, K., & Yates, A. (2023). Online Learning Path Recommender System for MOOCs. *IEEE Global Engineering Education Conference*, 1–10. https://doi.org/10.1109/EDUCON54358.2023.10125232

2. Shu, Y., Xiao, Y. L., & Meng, F. (2024). *Deep Learning-Based Method for Predicting Student Dropouts in MOOCs*. 1–6. https://doi.org/10.1109/mlnlp63328.2024.10800676

3. Isidro, C., Carro, R. M., & Ortigosa, A. (2018). Dropout Detection in MOOCs: An Exploratory Analysis. *International Symposium on Computers in Education*. https://doi.org/10.1109/SIIE.2018.8586748

4. Xing, W., & Du, D. (2019). Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *Journal of Educational Computing Research*, *57*(3), 547–570. https://doi.org/10.1177/0735633118757015

5. Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, 1–20. https://doi.org/10.1080/10494820.2020.1727529

6. Halawa, S., Greene, D. K., & Mitchell, J. (2014). *Dropout Prediction in MOOCs using Learner Activity Features*. *37*, 1. https://dialnet.unirioja.es/servlet/articulo?codigo=6359342

7. Tang, C., Ouyang, Y., Rong, W., Zhang, J., & Xiong, Z. (2018). *Time Series Model for Predicting Dropout in Massive Open Online Courses* (pp. 353–357). Springer, Cham. https://doi.org/10.1007/978-3-319-93846-2_66

8. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2020). On the Variance of the Adaptive Learning Rate and Beyond. *International Conference on Learning Representations*. https://openreview.net/pdf?id=rkgz2aEKDr

9. Khadangi, A. (2023). We Don't Need No Adam, All We Need Is EVE: On The Variance of Dual Learning Rate And Beyond. *arXiv.Org*, *abs/2308.10740*. https://doi.org/10.48550/arxiv.2308.10740

10. Jiang, W., Yang, S., Wang, Y., & Zhang, L. (2024). *Adaptive Variance Reduction for Stochastic Optimization under Weaker Assumptions*. https://doi.org/10.48550/arxiv.2406.01959

11. Ma, J., & Yarats, D. (2021). On the Adequacy of Untuned Warmup for Adaptive Optimization. *National Conference on Artificial Intelligence*, *35*, 8828–8836. https://dblp.uni-trier.de/db/conf/aaai/aaai2021.html#MaY21

Received: August 04, 2025

Volume 38 No. 6s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

12. Xie, S., Lv, A., Xia, Y., Wu, L., Qin, T., Liu, T.-Y., & Yan, R. (n.d.). Target-Side Input Augmentation for Sequence to Sequence Generation. *International Conference on Learning Representations*.

13. Wang, Y., Dai, B., Hua, G., Aston, J. A. D., & Wipf, D. (2018). Recurrent Variational Autoencoders for Learning Nonlinear Generative Models in the Presence of Outliers. *IEEE Journal of Selected Topics in Signal Processing*, *12*(6), 1615–1627. https://doi.org/10.1109/JSTSP.2018.2876995

14. Garay-Maestre, U., Gallego, A.-J., & Calvo-Zaragoza, J. (2018). *Data Augmentation via Variational Auto-Encoders* (pp. 29–37). Springer, Cham. https://doi.org/10.1007/978-3-030-13469-3_4

15. Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., & Courville, A. (2018). MINE: Mutual Information Neural Estimation. *arXiv: Learning*. https://arxiv.org/abs/1801.04062v1

16. Wen, L., Zhou, Y., He, L., Zhou, M., & Xu, Z. (2020). Mutual Information Gradient Estimation for Representation Learning. *arXiv: Machine Learning*. https://arxiv.org/abs/2005.01123

17.Lin, X., Sur, I., Nastase, S. A., Divakaran, A., Hasson, U., & Amer, M. R. (2019). Data-Efficient Mutual Information Neural Estimator. *arXiv: Learning*. https://dblp.uni-trier.de/db/journals/corr/corr1905.html#abs-1905-03319

18. Kuzilek, J., Hlosta, M., Zdrahal, Z., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, *4*(1), 170171. https://doi.org/10.1038/SDATA.2017.171

19. Alj, Z., & Cherkaoui, M. M. O. (2022). Predicting Students Dropout In MOOC From Learning Behaviour Using Machine Learning. *Journal of Uncertain Systems*. https://doi.org/10.1142/s1752890922500118

20. Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B. S., & Heo, J. (2020). *EdNet: A Large-Scale Hierarchical Dataset in Education* (pp. 69–73). Springer, Cham. https://doi.org/10.1007/978-3-030-52240-7_13

Received: August 04, 2025