KNOWLEDGE DISTILLATION THROUGH TIME FOR FUTURE EVENT PREDICTION

Skye Gunasekaran, Jason Eshraghian
University of California, Santa Cruz
{akgunase,jeshragh}@ucsc.edu

Ruomin Zhu, Zdenka Kuncic
The University of Sydney

ABSTRACT

Is it possible to learn from the future? Here, we introduce knowledge distillation through time (KDTT). In traditional knowledge distillation (KD), a reliable teacher model is used to train an error-prone student model. The difference between the teacher and student is typically model capacity; the teacher is larger in architecture. In KDTT, the teacher and student models differ in their assigned tasks. The teacher model is tasked with detecting events in sequential data, a simple task compared to the student model, which is challenged with forecasting said events in the future. Through KDTT, the student can use the 'future' logits from a teacher model to extract temporal uncertainty. We show the efficacy of KDTT on seizure prediction, where the student forecaster achieves a 20.0% average increase in the area under the curve of the receiver operating characteristic (AUC-ROC).

1 INTRODUCTION AND RELATED WORK

KD traditionally involves transferring knowledge from a high-capacity teacher model to a more compact representation in a student model (Gou et al., 2021). In predictive modeling, the task of forecasting future events presents a greater challenge than the detection of current events, primarily due to increased stochastic variability. To offset this performance difference, larger capacity models must be used to handle the uncertainty associated with predictive tasks.

Where the tasks of the forecasting model and detection model only differ by the time of detection, we posit that the student model can enhance its predictive capabilities by integrating insights from future data already processed by the teacher model. Our proposed approach, KDTT, can mitigate the elevated uncertainty and stochastic noise characteristics of forecasting tasks. Temporal disparity can ultimately be used to reduce predictive ambiguity and enhance model robustness.

We test KDTT on seizure prediction. Clinicians monitor epilepsy patients by interpreting EEG signals to identify seizures, which has been automated by machine learning (Shoeb & Guttag, 2010). However, the challenge of seizure prediction—determining the confidence of a seizure event within a given time horizon—is a complex task due to patient specificity and seizure stochasticity (Mormann et al., 2007). KDTT is effective as seizure detection and prediction are similar time-series tasks with differing complexity. Prior related work reinforces a prediction model using a binary target as a 'weak label' (Yang et al., 2022). We introduce direct logit matching between the detector and predictor, allowing the forecasting model to be exposed to uncertainty in the detection model.

2 METHODOLOGY

Temporal distillation: Our distillation loss function adheres to those from Hinton et al. (2015) and Kim et al. (2021). Given a teacher model T and a student model S, the total loss $\mathcal{L}$ is defined as a weighted combination of Cross-Entropy (CE) Loss and Kullback Leibler (KL) Divergence Loss:

$$\mathcal{L} = \alpha \times \underbrace{-\sum_i Y_{\text{true},i} \cdot \sigma(\text{logits}_{S,i})}_{\text{Cross-Entropy Loss}} + \beta \times \underbrace{\sum_i \sigma(\text{logits}_{T,i}/t) \cdot \left[\log\left(\frac{\sigma(\text{logits}_{T,i}/t)}{\log \sigma(\text{logits}_{S,i}/t)}\right)\right]}_{\text{Time-Offset KL Divergence Loss}} \quad (1)$$

α and β scale the contribution of each loss function, and $\sigma(\cdot)$ applies the softmax function. Temperature t alters the 'softness' of the softmax for both the teacher and student. Critically, the logits between both models are offset by at least five minutes, such that S aims to forecast whether a seizure will occur five minutes in the future, while T detects seizures at the present step.

Network Architecture: Following other high-performing models for seizure detection (Yang et al., 2023), a pair of convolutional layers encode the pre-processed input (3×3 kernel), a pair of LSTM layers do sequential processing (768 input neurons, 320 hidden neurons), terminated by a dense layer for both teacher and student. The learning rate was $5e-4$ using Adam.

Data Processing Pipeline: EEG signals are classed as either pre-seizure, seizure, or inter-seizure. The teacher model is trained to distinguish between seizure and non-seizure instances. Given that the objective of the student model is to forecast seizures, the student is trained on pre-seizure data and the seizure data itself is concealed. A seizure occurrence period (SOP) of 30 minutes is used, and the seizure prediction horizon (SPH) is set to 5 minutes (Maiwald et al., 2004). Samples are binned into 30 second chunks and pre-processed with a short-time Fourier transform.

Training Pipeline: First, the teacher is pre-trained on a specific patient for seizure detection. During training of the student model, logits are obtained from the teacher model at the future time step, i.e., the student model lags behind the teacher by 5 minutes. This is because prediction takes place prior to detection. The student model strives to make the same conclusion as the teacher, while using EEG signals that occur earlier in time. The loss in 1 is used, and both models are trained for 100 epochs.

3 RESULTS

We evaluate KDTT on 11 patients from the CHB-MIT database (Shoeb, 2009). We used the hyperparameters $\alpha = 0.7$, $\beta = 0.5$, and $t = 4.3$. We additionally test KDTT performance where KL Divergence from 1 is replaced with the mean-square error (MSE) loss for direct logit matching. These are compared with a base student model trained without the KL Divergence term from 1. Note the vast degradation of the base student model relative to the teacher models. This is unsurprising, given the stochastic nature of seizure forecasting, and it highlights the difficulty of future time prediction, even where the time window is 30 s. Table 1 shows an average improvement of 15.0% when using KL, and 20.0% when using MSE. All results shown are an average of 5 trials per model.

ID	T (MSE)	T (KL)	S (Base)	S (MSE)	S (KL)
1	0.93	1.0	0.93 ± 0.05	1.0 ± 0.01	0.99 ± 0.00
2	1.0	1.0	0.16 ± 0.09	0.73 ± 0.25	0.65 ± 0.36
3	0.99	0.99	0.1 ± 0.09	0.86 ± 0.17	0.81 ± 0.11
5	1.0	1.0	0.50 ± 0.23	0.81 ± 0.36	0.73 ± 0.14
10	1.0	1.0	0.38 ± 0.07	0.43 ± 0.26	0.39 ± 0.22
13	0.87	0.88	0.91 ± 0.13	0.95 ± 0.06	0.94 ± 0.03
18	0.98	0.97	0.45 ± 0.24	0.45 ± 0.21	0.56 ± 0.17
19	1.0	1.0	0.99 ± 0.01	0.58 ± 0.16	0.64 ± 0.26
20	1.0	1.0	0.80 ± 0.15	0.88 ± 0.29	0.68 ± 0.08
21	0.72	0.54	0.85 ± 0.02	0.39 ± 0.09	0.35 ± 0.04
23	0.63	0.71	0.56 ± 0.04	0.87 ± 0.13	0.84 ± 0.10
Avg	0.93	0.92	0.60 ± 0.04	0.72 ± 0.02	0.69 ± 0.04

Table 1: Results: Mean AUC-ROC w/std. T: Teacher model. S: Student model. S (Base): no KDTT.

4 CONCLUSION

To summarize, we propose KDTT and verify its effectiveness using seizure forecasting as an application. The improvement of AUC-ROC highlights the importance of using latent uncertainty in present-time data to train forecasting models. Future work can span label-free forecasting, to developing models of predictive coding. Code URL: `github.com/SkyeGunasekaran/KDTT`

URM STATEMENT

We acknowledge that the first author meets multiple URM criteria as outlined by the ICLR 2024 TinyPaper URM criteria.

REFERENCES

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Taehyeon Kim, Jaehoon Oh, NakYil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. *arXiv preprint arXiv:2105.08919*, 2021.

Thomas Maiwald, Matthias Winterhalder, Richard Aschenbrenner-Scheibe, Henning U Voss, Andreas Schulze-Bonhage, and Jens Timmer. Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic. *Physica D: nonlinear phenomena*, 194 (3-4):357–368, 2004.

Florian Mormann, Ralph G Andrzejak, Christian E Elger, and Klaus Lehnertz. Seizure prediction: the long and winding road. *Brain*, 130(2):314–333, 2007.

Ali H Shoeb and John V Guttag. Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 975–982, 2010.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Yikai Yang, Nhan Duy Truong, Jason K Eshraghian, Armin Nikpour, and Omid Kavehei. Weak self-supervised learning for seizure forecasting: a feasibility study. *Royal Society Open Science*, 9(8):220374, 2022.

Yikai Yang, Jason K Eshraghian, Nhan Duy Truong, Armin Nikpour, and Omid Kavehei. Neuromorphic deep spiking neural networks for seizure detection. *Neuromorphic Computing and Engineering*, 3(1):014010, 2023.