Reliability in ML: Final Report

Orr Avrech, 302857065
Github: augmented-distillation

Abstract

Model distillation of complex but accurate model ensembles into fast and simple models can benefit from a data augmentation strategy that aims to reduce the teacher-student error. FAST-DAD [1] suggested a data augmentation scheme, specific for tabular data that uses Gibbs sampling from a self-attention pseudolikelihood estimator. Instead of drawing samples from the full joint distribution, we suggest a simpler and faster approach- sampling directly from the conditionals estimates. In addition, we introduce the use of prediction sets to obtain reliable augmented samples for distillation. We observe that our proposed sampling scheme, is slightly better but is much faster and simpler with the potential of further improving performance by feature-specific augmentations.

1. Background

Knowledge distillation suggests that the knowledge learnt by a complex model (the *teacher*) can be compressed into a simpler model (the *student*) with reduced inference time and memory consumption, which is therefore much easier to deploy. In addition, this strategy can promote interpretability, by choosing a simple understandable model as the student model over accurate, but incomprehensible, complex model ensembles.

Existing work on distillation and model compression mainly focuses on architecture-specific techniques, where both student and teacher models are neural networks, and can be primarily found in the domains of vision, language and speech, where training data is usually plentiful (either labeled or not). Therefore, a general framework that is both model-agnostic and task-agnostic (classification, regression, etc.) can potentially help machine-learning practitioners to deploy simple, fast and accurate models across all domains.

2. Base method: FAST-DAD

2.1 Motivation

The model distillation procedure introduces an approximation error between the teacher and the student. The generalization error of the student fitted on the original training data may be of similar magnitude as the teacher-student approximation error. When this occurs, distillation is not expected to improve generalization accuracy. A key observation in [1], is that increasing the amount of available data for distillation can improve the student-teacher approximation, and hence the generalization error of the student might be reduced. Assuming that the teacher generalizes well enough, the extra data may be labeled by the teacher. The problem therefore boils down to drawing approximate samples from the training data distribution. These samples of unlabeled data will be fed to the teacher for labeling, and will be used to augment the available training data for the student.

2.2 Distillation with Augmented Data

FAST-DAD [1] is a suggested technique to produce **Fast**-and-accurate models via **D**istillation with **A**ugmented **D**ata. Formally, consider a dataset (X_n, Y_n) where $X_n = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ are observations of features sampled from a distribution p and $Y_n = \{y_i\}_{i=1}^n$ are their corresponding labels, sampled from $p_{y|x}$. Distillation then aims to minimize some distance function between the student and the teacher, averaged over the training set. The distance function is determined by a task-specific loss. For instance, MSE for regression and KL-divergence between class probabilities in classification. The statistical error due to distillation, of finding the optimal student with a minimal distance from the teacher, can be reduces by producing more unlabeled data for distillation. This process can be done by estimating the full joint distribution for all features, but it is a hard to task to do so. Instead, one can estimate the univariate conditional distribution of a feature x^i given all other features x^{-i}, i.e to estimate $p\left(x^i|x^{-i}\right)$. FAST-DAD thus suggests the following stages for distillation with augmented data:

1. Estimate the conditional distribution $p\left(x^i|x^{-i}\right)$ for each feature i.

2. Given the estimated conditionals, use Gibbs sampling [2] to draw approximate sample the joint distribution of the features $\tilde{x} \sim p(x)$. This will generate a set of augmented samples $\tilde{X}_m$.

3. Feed the augmented samples as inputs to the teacher model to obtain predictions $\tilde{Y}_m$. In classification, the class probabilities should be obtained.

4. Fit the student model on the augmented training set $(X_n, Y_n) \cup (\tilde{X}_m, \tilde{Y}_m)$. In classification, the true labels Y_n correspond to the actual predicted classes,

while $\tilde{Y}_m$ are the class probabilities, as predicted by the teacher. Therefore, the loss function for the student should take a valid distance between soft-labels and hard-labels into consideration.

2.2.1 Estimating Conditionals via Self-Attention

As mentioned above, the first stage of the proposed strategy is to estimate the conditional distributions for all features in the data. Instead of modelling each conditional separately, the paper takes a novel approach and suggests a simultaneous estimation of all conditionals using a self-attention-based encoder [3] trained with a pseudolikelihood objective,

$$\hat{\theta} = -\arg\min_{\theta} \frac{1}{n} \sum_{x} \sum_{i=1}^{d} \log p \left(x^i | x^{-i}; \theta \right)$$

where each conditional is parameterized as a mixture of Gaussians: $p\left(x^i|x^{-i};\theta\right) = \sum_k \lambda_k N(x^i; \mu_k, \sigma_k^2)$. The parameters to estimate (θ) are $\lambda_k, \mu_k, \sigma_k$ which depend only on x^{-i} and are the output of the attention-based encoder. The masking mechanism of the self-attention applied on a given dimension i , enables the conditioning on x^{-i}.

2.2.2 Gibbs Sampling from the Learnt Conditionals

Assuming that our learnt conditionals of the features from the previous step are accurate, we can initialize the sampler at some random sample from the original training set $x \in X_n$ and cycle through the features with a random ordering of the features. In each step (j), we replace the value of one feature x^i using a sample from its conditional distribution $p\left(x^i|x^{-i}\right)$ and use this value $x^i_{(j)}$ when conditioning on that feature in the next step $(j+1)$. After every feature has been resampled, a round of Gibbs sampling has been completed.

2.3 Experiments

Multiple tabular datasets were examined, including regression tasks and binary/multi-class classification tasks. Several teacher models were tested, using AutoML frameworks such as AutoGluon [4], AutoSklearn [5] and H2O. Four types of student models: (1) Neural Network, (2) CatBoost, (3) LightGBM and (4) Random Forest. In Figure 1, the main results of the paper are presented. One can notice that there's a major improvement in speed between the teacher (TEACHER dot) and the distilled version of the best student on average (GIB-1 dot), with a minor drop in accuracy between the two. Moreover, each student model is outperformed by its distilled counterpart, with similar latency measurements.

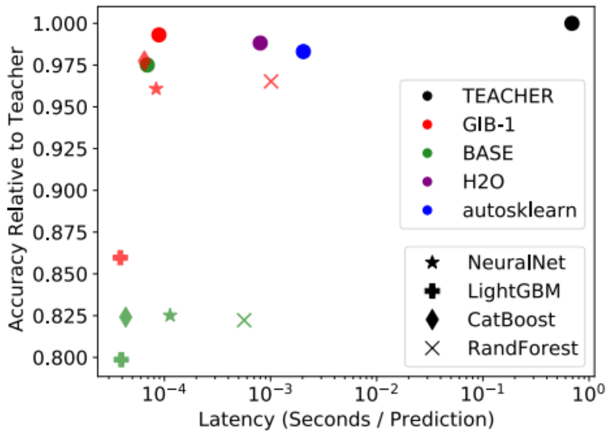


Figure 1: Test accuracy vs. latency of several student and teacher AutoML models averaged over 30 datasets. TEACHER denotes the performance of AutoGluon, while H2O and autosklearn are competing AutoML frameworks. GIB-1 indicates the results of FAST-DAD after 1 round of Gibbs sampling which yielded the best results according to the paper. BASE denotes the student model fit on the original data. The dots of GIB-1 and BASE are represent the best model, out of 4 student models (1)-(4) with averaged accuracy over all the datasets.

2.4 Discussion

FAST-DAD presents a novel and generic framework for building fast and accurate models via distillation of large, ensemble-based AutoML models. This strategy adopt several new techniques to improve distillation, with the focus on tabular data: *(i)* Instead of directly trying to estimate the multivariate distribution of the features for generating augmented samples, the suggested framework only involves an estimation of the conditional univariate distribution for each feature, dependent on all other features. *(ii)* All conditionals estimations are learnt once via a single self-attention-based encoder. *(iii)* Gibbs sampling based on original data initialization with the conditional estimates, can quite efficiently augment the dataset for distillation.
However, the suggested strategy has a few clear limitations:

1. One major limitation, is that the augmented features are fed to the teacher for labeling (either class-probability in classification or scalar value in the case of regression). Therefore, the performance of the student after distillation is significantly affected by the reliability of the teacher's predictions.

2. Although the efficiency of the Gibbs sampling is improved due to the conditional estimates, the task of drawing approximate samples from the full joint distribution is much harder and prone to approximation errors than simply sampling from the conditional estimates. Furthermore, the choice of augmenting minority groups within a given feature, is not possible when

drawing samples from $p(x)$. The fact that the best results for distillation were obtained after one round of Gibbs sampling is casts a shadow on its relevance.

3. Tabular data typically contains mixed data types (e.g. numerical, categorical, text, etc). In order to create a model that is independent of the feature types and orders, the paper suggested a generic modelling for the data in the form of mixture of Gaussians, where the same number of mixture components is used for all features. It does represents categorical features numerically with dequantization [6], but the correctness of this form of modelling needs to be further justified.

4. The proposed solution is said to be model-agnostic. However, although the actual process of generating augmented samples and soft-labeling them with the teacher, is agnostic to the student's model, the last stage of distillation is actually model-dependent. For example, in classification, neural network student are trained using cross-entropy applied on soft-labels, while random forest students are used as an ensemble of regressors with hard-labels for the original data and soft-labels for the teacher-predicted class labels. Therefore, distillation is not completely generic, or can be said to seamlessly applied on any black-box algorithm.

3. Our Work

In this work, we wanted to address some of the observed limitations in [1].

1. We present a comparison between two sampling methods. The first is the Gibbs sampler, which is presented in the paper. The second sampling scheme, is sampling from the conditional distributions estimates, which can be done in several ways. One way is to run all samples (or batches of samples), and for each batch randomly pick a feature i and estimate its conditional distribution. Then, we can use this estimate to sample a 1-dim value of the feature i given all other features in that sample. This kind of sampling technique should have a few advantages over Gibbs. First, given that the pseudolikelihood estimated are accurate, then the approximated samples should be very accurate as well, since it is merely sampling from a univariate distribution. Second, conditional sampling should faster than Gibbs sampling since it does not involve an iterative process of encoder predictions. Third, this strategy allows us to choose specific features to sample from, which can be helpful for augmenting certain features of interest or for finding hidden information that is not represented in the original data.

2. One major limitation in FAST-DAD, and in knowledge distillation in general, is the fact the student's performance is bounded by the accuracy of teacher predictions. Being able to provide reliable predictions by the teacher, could potentially improve the distillation error. Therefore, we suggest using conformal prediction by the teacher. Conformal prediction can be used for any black-box algorithm and can be applied on several tasks, such as classification [7] and regression [8]. Therefore, is sound appealing to use this technique as a part of the generic framework of FAST-DAD.

4. Results

4.1 Setup

We evaluate various methods for distillation on two datasets: Adult Census Income (Kaggle) and Phoneme (OpenML), both considered as binary classification tasks. The Adult Census Income consists of mixed data features, as well as a variety of unrepresented features (e.g., race). The Phoneme datasets is composed of continuous features only. Both datasets are centered and scaled, for all training and prediction purposes. Label encoding is applied on the relevant categorical features. We adopt AutoGluon [4] as our teacher model. AutoGluon is fit for each training data for up to 10 minutes and includes and auto-stack that allows extensive model ensembling. Random Forest is picked as the student model.

4.2 Prediction Sets

In the binary classification case, the following prediction sets are possible:

$$\{\{0\},\{1\},\{0,1\},\emptyset\}$$

Therefore, if a teacher prediction is either an empty set or a full set (both classes are in), then we assume that the prediction is not reliable and discard it prior to distillation. We produce the prediction sets for the teacher by using split-conformal with a probability conformity score, as in [7].

4.3 Distillation Strategies

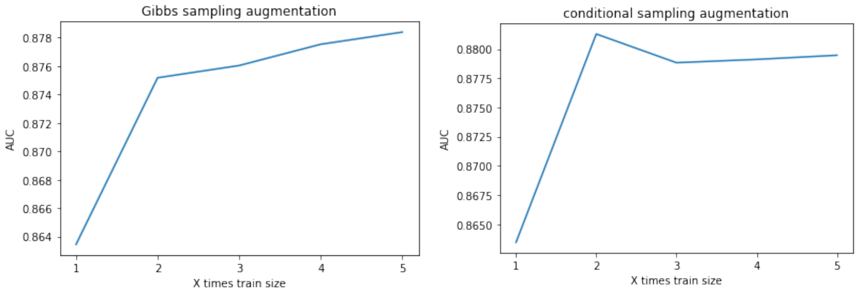
We compare between the following strategies: **TEACHER**: The complex model produced by AutoGluon on the training data. **GIB-1**: the FAST-DAD approach with one round of Gibbs sampling. **COND**: unviariate sampling for a random feature given all other features, for each sample of the original training data. **GIB-1-SC**: one round of Gibbs sampling for generating the augmented samples. Soft-labeling only the most reliable predictions after the teacher's fit using split-conformal. Prediction set of a single class are kept for distillation. **COND-SC**: conditional sampling + split conformal. From Table 1, we can observe that both the Gibbs sampling and the conditional sampling are quite similar in results, but both demonstrate a successful distillation since the student's performance is inferior. In addition, the use of prediction sets had little to no effect on the accuracy of the distilled models.

Method	*Phoneme*	*Adult Census Income*
STUDENT	86.3	88.0
GIB-1	87.8	89.4
COND	**87.9**	**89.5**
GIB-1-SC	**87.9**	89.4
COND-SC	**87.9**	89.4
TEACHER	97.6	92.6

Table 1: Evaluation of the several distillation techniques on both datasets. **ROC-AUC** is taken as the measure of success for these tasks.

4.4 Analysis of the Sampling Techniques

Both sampling techniques demonstrate a desired behavior when increasing the number of samples for distillation. This phenomenon can be observed in Figure 2, where one can observe that an increasing number of augmented samples generally leads to better accuracy. However, there is an evident saturation after some amount of samples. In Figure 3, we can see the time differences between the two sampling techniques. Conditional sampling is unsurprisingly, much faster than the Gibbs sampler. Since all conditional estimates are predicted at once by the self-attention encoder, there is a plateau with an increasing number of features.



(a)

(b)

Figure 2: ROC-AUC vs. the number of augmented samples used for distillation (as a multiplication of the original training set size). (a) GIB-1 (b) COND. The results are on Phoneme. We expect an increasing accuracy with the increasing number of augmented samples.

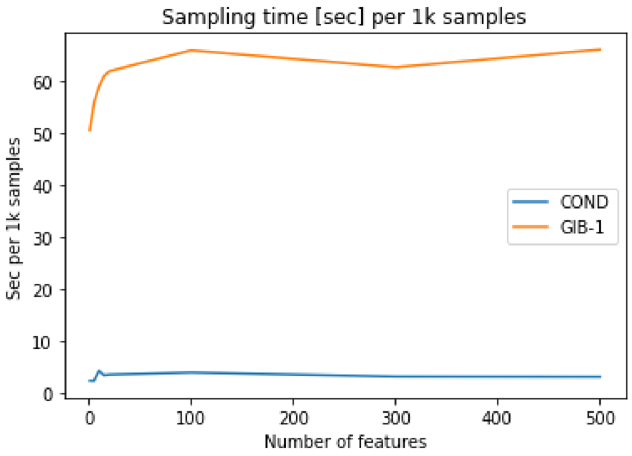


Figure 3: Sampling time in seconds per 1k random samples **(not related to any dataset)** vs. the number of features. The evident plateau is due to a simultaneous estimation of all conditionals by the encoder. In the Gibbs sampling case, we do however expect a slight increase with the number of features because of its iterative nature.

5. Conclusions

In this work, we tried to improve existing methods of model distillation on tabular data in order to obtain fast, accurate and deployable models. We used the proposed attention-based-encoder of FAST-DAD and used its conditional estimates to draw samples from univariate distributions instead of trying to draw samples from the full joint distributions. This sampling technique is much simpler, faster and has the potential of improving performance by treating features of interest more carefully. We also introduced the use of prediction sets, that are supposed to filter-out unreliable predictions by the teacher. In the examined binary classification setup, we have not seen any improvement with this method. The binary classification is rather degenerated, and we expect to see better results by applying this method on different tasks.

References

[1] Rasool Fakoor, Jonas Mueller, Nick Erickson, Pratik Chaudhari, and Alexander J. Smola. *Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation*. 2020. arXiv: 2006.14284 [cs.LG].

[2] Stuart Geman and Donald Geman. “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), pp. 721–741.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. 2017.

[4] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data*. 2020. arXiv: 2003.06505 [stat.ML].

[5] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. “Auto-sklearn: Efficient and Robust Automated Machine Learning”. In: *Automated Machine Learning: Methods, Systems, Challenges*. Cham: Springer International Publishing, 2019, pp. 113–134.

[6] Benigno Uria, Iain Murray, and Hugo Larochelle. *RNADE: The real-valued neural autoregressive density-estimator*. 2014. arXiv: 1306.0186 [stat.ML].

[7] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. “Classification with Valid and Adaptive Coverage”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 3581–3591.

[8] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. *Conformalized Quantile Regression*. 2019. arXiv: 1905.03222 [stat.ME].