MODEL DISTILLATION ON MULTI-DOMAINED DATASETS

Alexey Orlov
MIPT
orlov.as@hpystech.edu

Andrey Grabovoy
MIPT
grabovoy.av@phystech.edu

ABSTRACT

The paper investigates the problem of reducing the complexity of the approximating model when transferred to new data of lower cardinality. The concepts of a teacher model for large dataset and a student model for a small dataset are introduced. Methods based on the distillation of machine learning models, including the Bayesian approach, are considered. We consider the assumption that solving the optimization problem from the parameters of both models and domains improves the quality of the student's model. A computational experiment is being carried out on real and synthetic datasets for regression, classification and text recognition problems.

Keywords Distillation · Neural network · supervised learning

1 Introduction

One way to improve the quality of a machine learning algorithm is to use a model with a large number of parameters, which answers can be used when training a model with a smaller number of parameters [4]. Models with fewer parameters are more interpretable. The purpose of this work is to reduce the complexity of the machine learning model, as well as the transition to data of lower power. To do this, it is proposed to use two main methods – distillation of models and domain adaptation.

Machine learning model distillation uses model labels with more parameters to train a model with fewer parameters. In paper [4] discusses the method of distillation proposed by G. Hinton, taking into account teacher marks using the function softmax with a temperature parameter, and [8] considers the combination of distillation methods proposed by G. Hinton and privileged information [8] proposed by V. N. Vapnik into a generalized distillation. In probabilistic distillation, a sample generation hypothesis is introduced along with teacher responses. In [2], Bayesian distillation of deep learning models is considered, which uses the posterior distribution of teacher model parameters along with teacher responses. Based on this posterior distribution, the prior distribution of the student model is given. Model distillation is used in a wide class of problems.

Different settings of domain adaptation problems are described in [11], there are settings with partially tagged target domain and not tagged at all. Tasks with an unlabeled target domain are aimed at ensuring that models trained on synthetic data adapt to real data [1]. Thus, domain adaptation uses the tagged data of multiple source domains to perform new tasks in the target domain.

One of the problem definitions of domain adaptation is image style transfer [9, 10]. So, in [9] it is proposed to use selfie images as a source domain and translate them into images of the desired art style based on a selection from the required style. In this way, synthetically generated training data can be used.

In the distillation methods discussed above [4, 8, 2], we consider the case when the teacher and student models approximate samples from different populations. For the distillation problem proposed by Geoffrey Hinton [4], the source and target data sets are the same.

It is proposed to use, in addition to the labels of the teacher on one of the domains, the connection between the domains when training the student model. In this case, close general populations should serve as domains.

Thus, real and generated images can serve as domains of different sizes. As mappings between images, normal noise, convolutional transformations, and image generation using the variance autoencoder model [6] are considered. Here we

study mappings for which the existence of inverses is not considered. It is expected that the quality of the resulting models on one domain will exceed the quality of models that were not trained with teacher marks on another domain.

Real data and a synthetic datasets are used as experimental data. The Fashion-MNIST [12] dataset, consisting of images of clothes, and the MNIST [7] dataset, consisting of images of handwritten numbers, are considered as real data.

2 Distillation problem statement

2.1 Basic distillation problem statement

A dataset is given

$$\mathfrak{D} = \{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^{n}, \quad \mathbf{x_i} \in \mathbb{X}, \quad \mathbf{y_i} \in \{1, ..., R\},$$

where R is the number of classes in the classification problem.

It is assumed that a trained model with a large number of parameters is given — a teacher model. The teacher model $\mathbf{f}$ belongs to the parametric family of functions:

$$\mathfrak{F} = \{\mathbf{f} | \mathbf{f} = \text{softmax}(\mathbf{v}(\mathbf{x})/T), \mathbf{v} : \mathbb{R}^n \to \mathbb{R}^R\}.$$

It is required to train the student model with fewer parameters, taking into account the teacher's answers. The student model $\mathbf{g}$ belongs to the parametric family of functions:

$$\mathfrak{G} = \{\mathbf{g} | \mathbf{g} = \text{softmax}(\mathbf{z}(\mathbf{x})/T), \mathbf{z} : \mathbb{R}^n \to \mathbb{R}^R\},$$

where $\mathbf{v}, \mathbf{z}$ are differentiable parametric functions of a given structure, T is a temperature parameter with the properties:

1. for $T \to 0$ one of the classes has unit probability;
2. for $T \to \infty$ all classes are equally probable.

The loss function $\mathcal{L}$, which takes into account the teacher's model $\mathbf{f}$ when choosing the student's model $\mathbf{g}$, has the form:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}) = &- \sum_{i=1}^{m} \sum_{r=1}^{R} y_i^r \log g^r(x_i)\big|_{T=1} \\ &- \sum_{i=1}^{m} \sum_{r=1}^{R} f^r(x_i)\big|_{T=T_0} \log g^r(x_i)\big|_{T=T_0},\end{aligned}$$

where $\cdot\big|_{T=t}$ means that the temperature parameter T in the previous function is equal to t.

We get the optimization problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}).$$

2.2 Formulation of the distillation problem for a multidomain sample

Two samples are given:

$$\mathfrak{D}_\text{s} = \{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^{n}, \quad \mathbf{x_i} \in \mathbb{X}_\text{s}, \quad \mathbf{y_i} \in \mathbb{Y}$$
$$\mathfrak{D}_\text{t} = \{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^{m}, \quad \mathbf{x_i} \in \mathbb{X}_\text{t}, \quad \mathbf{y_i} \in \mathbb{Y},$$

where $\mathfrak{D}_\text{s}, \mathfrak{D}_\text{t}$ are the source and destination datasets. In the basic formulation of the distillation problem, it is assumed that $\mathfrak{D}_\text{t} \subset \mathfrak{D}_\text{s}, \mathbb{X}_\text{t} = \mathbb{X}_\text{s}$.

It is assumed that the number of objects in the samples do not match:

$$n \gg m$$

Let the model of the teacher be given on a sample of higher power:

$$\mathbf{f} : \mathbb{X}_\text{s} \to \mathbb{Y}',$$

where $\mathbf{f}$ is the teacher model, $\mathbb{Y}'$ is the evaluation space.

The relationship between the source and target samples is set:

$$\varphi : \mathbb{X}_{\text{t}} \to \mathbb{X}_{\text{s}},$$

where φ is an injective mapping.

It is required to obtain a student model for a low-resource sample:

$$\mathbf{g} : \mathbb{X}_{\text{t}} \to \mathbb{Y}',$$

where $\mathbf{g}$ is the student model.

The paper considers a loss function that takes into account teacher labels and the relationship between domains:

1. for the regression problem:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi) =& \lambda \|\mathbf{y} - \mathbf{g}(\mathbf{x}, \mathbf{w})\|_2^2 \\ &+ (1-\lambda)\|\mathbf{g}(\mathbf{x}, \mathbf{w}) - (\mathbf{f} \circ \varphi)(\mathbf{x})\|_2^2;\end{aligned}$$

2. for classification problem:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi) =& -\lambda \sum_{i=1}^{m} \sum_{r=1}^{R} \mathbb{I}[y_i = r] \log g^r(\mathbf{x}_i, \mathbf{w}) \\ &- (1-\lambda) \sum_{i=1}^{m} \sum_{r=1}^{R} (f \circ \varphi)^r(\mathbf{x}_i) \log g^r(\mathbf{x}_i, \mathbf{w}),\end{aligned}$$

   where λ is a metaparameter specifying the distillation weight, $\mathbb{I}$ is an indicator function.

We get the optimization problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi).$$

3 Computational experiment

The goal of computational experiment is to compare performance of teacher and student models on real and synthetic datasets with or without mapping φ for classification and regression problems.

3.1 Data

1. FMNIST is a dataset of images of clothing items consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes [12].
2. MMIST is a dataset of images of handwritten digits that also consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes [7].
3. The dataset "Bike sharing" is used for the regression task.

3.2 Configuration of algorithm run for classification

A multilayer perceptron with four hidden layers is considered as the teacher model **f**. A multilayer perceptron with one hidden layer is considered as the student model **g**.

Table 1: Description of models

	Teacher	Student
Structure	[784,256,128,64,64,10]	[784,64,10]
Number of parameters	246400	50816

Activation function after each hidden layer – ReLu. We use Adam gradient optimization method [5] to solve the optimization problem.

Each of the samples consists of a training and a test part, while the training part is divided into multi-resource and low-resource parts. The training part contains 60,000 objects, the multi-resource part contains 59,000 objects, the low-resource part contains 1,000 objects, and the test part contains 10,000 objects.

To analyze the quality of distillation, we use integral quality criterion [3].

Table 2: Datasets

Dataset	Explanation	Dataset size
FashionMNIST-Train	Training part	60000
FashionMNIST-Big	Multi-resource part	59000
FashionMNIST-Small	Low-resource part	1000
FashionMNIST-Test	Test part	10000
MNIST-Train	Training part	60000
MNIST-Big	Multi-resource part	59000
MNIST-Small	Low-resource part	1000
MNIST-Test	Test part	10000

3.3 Configuration of algorithm run for regression

Data from "Bike sharing" dataset. In this case, the training part is divided into multi-resource and low-resource parts. The training part contains 9000 objects, the multi-resource part contains 8700 objects, the low-resource part contains 300 objects, and the test part contains 1000 objects.

Table 3: Datasets

Datasets	Explanation	Dataset size
Reg-Train	Training part	9000
Reg-Big	Multi-resource part	8700
Reg-Small	Low-resource part	300
Reg-Test	Test part	1000

A multilayer perceptron with four hidden layers is considered as the teacher model **f**. A multilayer perceptron with one hidden layer is considered as the student model **g**.

Table 4: Description of models

	Teacher	Student
Structure	[15,256,128,64,64,1]	[15,32,1]
Number of parameters	48960	512

4 Conclusion

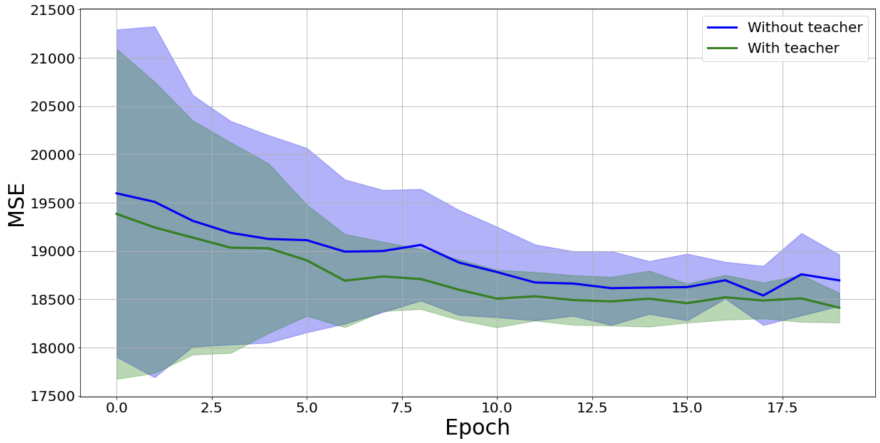

Figure 1: The mean squared error between the true and the student's predicted values on the test set. All results are averaged over 5 runs.

Figure 1 shows a plot of the deferred test sample mean squared error between the true values of the objects and the values predicted by the student model.

It can be seen from the graph that the model using the teacher's responses shows the best value of MSE.

Table 5: Model quality

Student	Teacher	Mapping φ	MSE
Reg-Train	—	—	$18694{,}9 \pm 266$
Reg-Train	Reg-Train	—	$\mathbf{18411{,}4 \pm 153{,}3}$

References

[1] Hongruixuan Chen, Chen Wu, Yonghao Xu, and Bo Du. Unsupervised domain adaptation for semantic segmentation via low-level edge information transfer, 2021.

[2] A.V. Grabovoy and V.V. Strijov. Bayesian distillation of deep-learning models, 2021.

[3] A.V. Grabovoy and V.V. Strijov. Probabilistic interpretation of the distillation problem, 2022.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[7] Cortes C. LeCun Y. Mnist handwritten digit database., 2010.

[8] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information, 2016.

[9] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications, 2021.

[10] Srikanth Tammina. Transfer learning using vgg-16 with deep convolutional neural network for classifying images, 2019.

[11] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey, 2018.

[12] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.