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 datasets for computer vision and text processing tasks.

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 [5]. 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 [5] discusses the method of distillation proposed by G. Hinton, taking into account teacher marks using the function softmax with a temperature parameter, and [9] considers the combination of distillation methods proposed by G. Hinton and privileged information [9] proposed by V. N. Vapnik into a generalized distillation. In probabilistic distillation, a sample generation hypothesis is introduced along with teacher responses. In [3], 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 [14], 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 [10, 11]. So, in [10] 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 [5, 9, 3], we consider the case when the teacher and student models approximate samples from different populations. For the distillation problem proposed by Geoffrey Hinton [5], 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 [7] 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 [15] dataset, consisting of images of clothes, and the MNIST [8] 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 classification problem:

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

   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 datasets with or without mapping φ for computer vision and natural language processing tasks. To analyze the quality of distillation an integral quality criterion is proposed [4].

3.1 Data

1. We use a subset of ImageNet — a dataset of images, for which we need to solve the classification problem into 10 classes. The dataset consists of a training part and a test part, and the training part is divided into multi-resource and low-resource parts.[2].
2. OPUS-100 is English-centric, meaning that all training pairs include English on either the source or target side. The article used fr-en and de-en datasets. [16]

3.2 Configuration of algorithm run for CV

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

We present information about the structure and number of parameters for the student and teacher models in Table 1.

Activation function after each hidden layer – ReLu. We use Adam gradient optimization method [6] 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.

The Table 2 describes the datasets on which the CV computational experiment took place.

Table 2: ImageNet dataset

Dataset	Explanation	Dataset size
ImageNet-Train	Training part	9469
ImageNet-Big	Multi-resource part	8469
ImageNet-Small	Low-resource part	1000
ImageNet-Test	Test part	3925

3.3 Configuration of algorithm run for NLP

The OPUS100 dataset was divided into a training part for the teacher, consisting of German-English sentences, and a training dataset and a test dataset for the student, consisting of French-English sentences. The teacher's training dataset contained 5000 sentences, the student's training dataset contained 2000 sentences, and the test dataset contained 500 sentences.

We used the student **g** and teacher **f** models as the transformer model based on article "Attention Is All You Need" [13] and Adam gradient optimization method [6] to solve the optimization problem. The NLLB model[12] was used as the φ mapping. This model translated French sentences into German sentences.

The Table 3 describes the datasets on which the NLP computational experiment took place.

Table 3: OPUS100 dataset

Dataset	Explanation	Language	Dataset size
Teacher-Train	Teacher training part	de-en	5000
Student-Train	Student training part	fr-en	2000
Student-Test	Test part	fr-en	500

4 Conclusion

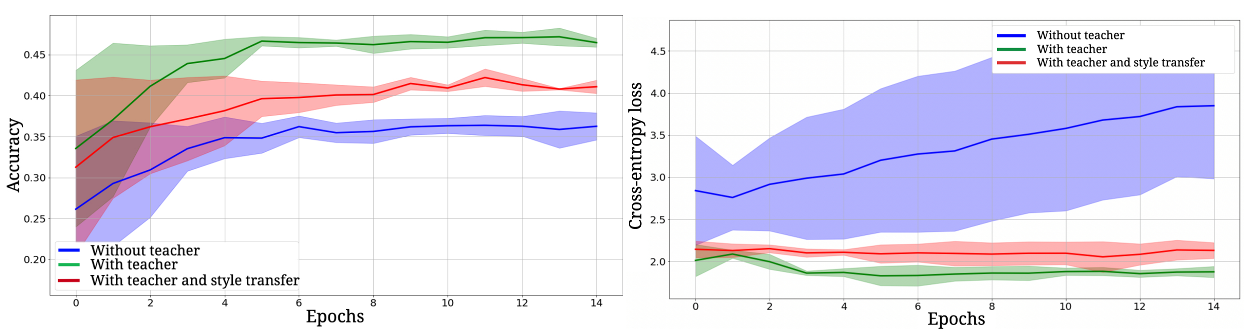

Figure 1: Quality of approximation on the test sample. All results are averaged over 5 runs. a) accuracy; b) cross-entropy error between true and predicted student labels

As can be seen from Figure 1, models trained with the use of a teacher achieve better quality and accuracy.

Table 4: Model quality for CV

Student	Teacher	Mapping φ	Accuracy	Cross-entropy loss	Integral criterion
ImageNet-Small	—	—	$0{,}363 \pm 0{,}017$	$3{,}849 \pm 0{,}866$	$46{,}615 \pm 11{,}498$
ImageNet-Small	ImageNet-Big	—	$\mathbf{0{,}465 \pm 0{,}005}$	$\mathbf{1{,}876 \pm 0{,}066}$	$\mathbf{26{,}488 \pm 0{,}996}$
ImageNet-Small	ImageNet-Big	StyleTransfer	$0{,}411 \pm 0{,}008$	$2{,}131 \pm 0{,}093$	$29{,}476 \pm 1{,}495$

The Table 4 shows the results of a comparison of student models obtained with and without the use of distillation.

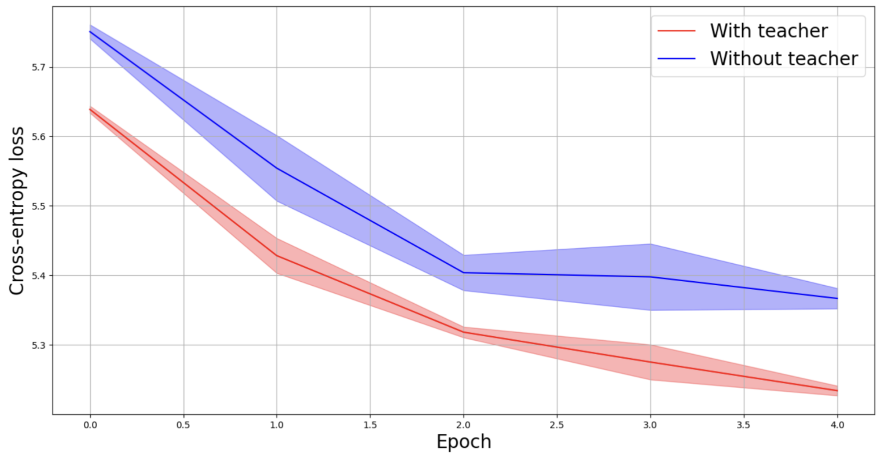

Figure 2: Cross-entropy error on the test dataset. All results are averaged over 3 runs.

As can be seen from Figure 2, models trained with the use of a teacher achieve better quality.

The Table 5 shows the results of a comparison of student models obtained with and without the use of distillation.

Table 5: Model quality for NLP

Student	Teacher	Mapping φ	Cross-entropy loss	BLEU
Student-Train	—	—	$5{,}367 \pm 0{,}015$	$0{,}0282$
Student-Train	Teacher-Train	NLLB	**5,233 ± 0,007**	**0,0572**

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] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database, 2009.

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

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

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

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

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

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

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

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

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

[12] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

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

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

[16] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation, 2020.