# Дистилляция данных

***Медведев Дмитрий Владимирович***

*Студент*

*Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия*

*E-mail:* dm.medvedev97@gmail.com

***Научный руководитель*** *— Дьяконов Александр Геннадьевич*

В машинном обучении задача дистилляции данных [1] представляет собой агрегирование всей возможной информации из оригинальной, как правило, довольно большой тренировочной выборки в более компактную. При этом обычно требуется, чтобы обучение моделей на дистиллированных данных позволяло достигать приемлемого качества за меньшее число шагов оптимизации (например, шагов градиентного спуска). Также необходимо, чтобы у новой выборки была обобщающая способность, позволяющая обучать даже те модели, которые не участвовали в процедуре дистилляции. Все это позволяет применять дистиллированные данные в важных практических задачах. Одной из такой задач является поиск архитектуры нейронной сети. Дистиллированные данные позволяют значительно ускорить обучение кандидатов, а соответственно и процедуру их отбора. Отбор обученных кандидатов осуществляется на основе качества их решений на реальных валидационных данных.

Обычно алгоритм дистилляции сводится к оптимизации самих объектов [1] выборки. Сначала объекты (в данной работе объектами выборки являются изображения) инициализируются случайным шумом. Затем с помощью этих синтетических объектов идет продедура обучения ученика (случайно выбранной модели машинного обучения). После этого считается ошибка решения учеником задачи (в данной работе рассматривается задача классификации) на реальных данных. Обновление синтетичесих объектов происходит с помощью градиентого спуска. При этом градиенты можно посчитать обратным распространением ошибки через процедуру обучения модели ученика на этих же объектах. Можно отметить, что шаг такой процедуры требует больших затрат как по времени так и по памяти, поэтому стали появляться альтернативные алгоритмы дистилляции. Так, в [2] было предложено использовать теорему о неявной функции и эффективную аппроксимацию обратного гессиана. В [3] авторы предложили поменять функцию ошибки на расстояние между векторами градиентов этой ошибки по параметрам ученика, которые получаются при обучении на обычных и дистиллированных данных.

Альтернативой дистилляции некоторого заранее заданного числа синтетических объектов является обучение модели учителя. Так в [4] вводится генеративная модель (Generative Teaching Network), способная из шума и меток класса создавать необходимые для обучения ученика синтетические изображения. Такой подход, по словам авторов, позволяет обучающимся на сгенерированных данных моделям получать лучшее по сравнению с [1] качество. В то же время авторы статьи для дистилляции рассматривали только обратное распространение ошибки через процедуру обучения. Это приводит к значительному ограничению данного метода в практическом использовании: обучение учителя может требовать гораздо больше вычислительных ресурсов чем задача поиска архитектуры, для которой учителя и могут обучать.

Чтобы решить эту проблему в данной работе исследуется возможность сокращения вычислительных ресурсов обучения Generative Teaching Network с помощью методов предложенных в [2] и [3]. А также исследуется влияние этих процедур дистилляции, эффективных с точки зрения ресурсов, на итоговое качество моделей учеников. Для верификации результатов рассматривается стандартная задача классификации рукописных цифр [5]. В качестве результата можно отметить, что оба варианта значительно уменьшают требования к использованию памяти и позволяют обучить учителя генерировать данные на которых модели ученика достигают качества сравнимого с тем, что достигается при использовании более тяжелой процедуры дистилляции.

**Литература**

1. Wang T., Zhu J., Torralba A., Efros A. A. Dataset Distillation. CoRR; abs/1811.10959 (2018)
2. Lorraine J., Vicol P., Duvenaud D. Optimizing Millions of Hyperparameters by Implicit Differentiation. CoRR; abs/1911.02590 (2019)
3. Zhao B., Mopuri K. R., Bilen H. Dataset Condensation with Gradient Matching. CoRR; abs/2006.05929 (2020)
4. Such F. P., Rawal A., Lehman J., Stanley K. O., Clune J. Generative Teaching Networks: Accelerating Neural Architecture Search by Learning to Generate Synthetic Training Data. CoRR; abs/1912.07768 (2019)
5. MNIST Handwritten Digit Database: `http://yann.lecun.com/exdb/mnist/`