Дистилляция знаний с использованием представления выборки в общем латентном пространстве моделей

A Preprint

Седова Анна
sedova.aa@phystech.edu

Горпинич Мария
gorpinich.m@phystech.edu

Бахтеев Олег
bakhteev@phystech.edu

Стрижов Вадим

17 марта 2022 г.

Abstract

В работе исследуется задача дистилляции моделей глубокого обучения. Дистилляция знаний — это задача оптимизации метапараметров, в которой происходит перенос информации модели более сложной структуры, называемой моделью-учителем, в модель более простой структуры, называемой моделью-учеником. В работе рассматривается случай переноса разнородных данных от нескольких моделей-учителей к модели-ученику. В данной работе предлагается построить аналог уже существующих решений с использованием функции оптимизации с триплетными ограничениями, которая показывала более хорошие результаты для классической дистилляции. Вычислительный эксперимент производится на выборке CIFAR-10.

1 Введение

В работе рассматривается разнородная дистилляция знаний с несколькими учителями с использованием промежуточных скрытых представлений учителей и ученика. Оптимизация модели глубокого обучения является вычислительно сложной задачей. Дистилляция знаний является частным случаем оптимизации моделей глубокого обучения.

Дистилляцией знаний называется способ оптимизации модели-ученика, при котором учитывается не только информация, содержащаяся в выборке, но и информация, содержащаяся в модели-учителе, имеющей высокую сложность. Разнородной дистилляцией знаний с несколькими учителями называется обобщение классической дистилляции знаний, в котором есть несколько моделей-учителей, каждый из которых обладает своей собственной, неполной информацией о выборке.

Для выравнивания скрытых пространств предлагается использовать функцию оптимизации с триплетными ограничениями. Функцией оптимизации с триплетными ограничениями называется функция, имеющая следующий вид $E = \sum_{(a,p,n)} max(0, m + ||f(x_a)f(x_p)||_2^2 - ||f(x_a)f(x_n)||_2^2)$. При решении задачи оптимизации с данной функцией потерь расстояние между x_a, называемым якорем, и x_p, называемым положительным входом, минимизируется, а расстояние между якорем и x_n, называемым отрицательным входом, максимизируется.

Сложность дистилляции знаний с помощью скрытого представления о выборке моделей состоит в том, что информация из промежуточных представлений разных моделей может иметь разную размерность. В данной работе предлагается способ сведения этой информации в одно латентное пространство.

Классический случай дистиляяции с одним учителем рассматривается в работе (4). В работах (3; 5) для обучения модели-ученика используется функция оптимизации с триплетными ограничениями. Данное исследование показывает, что эта функция дает более точные результаты на CIFAR 10 и Tiny Image Net.

Это дает основания полагать, что и для разнородной дистилляции с несколькими учителями функция оптимизации с триплетными ограничениями может давать более хорошие результаты.

Вариации дистилляции с несколькими учителями рассматриваются в (1; 7; 6).

Предлагается рассмотреть скрытые представления учителей и ученика получаемые при помощи алгоритмов снижения размерности. Для выравнивания пространств моделей предлагается применять модель автокодировщика с триплетными ограничениями (3). Предложенноые ранее решения (1), multi-teacher-multi-level не использует функцию оптимизации с триплетными ограничениями. (1) также не учитывает, что учителя могут иметь разную информацию о выборке. (5) использует функцию оптимизации с триплетными ограничениями, но не выравнивает информацию в одно латентное пространство.

Вычислительный эксперимент производится на выборке CIFAR-10. (Тут нужно что-то дописать)

Список литературы

[1] Valentin Malykh Irina Piontkovskaya Artur Ilichev, Nikita Sorokin. Multiple teacher distillation for robust and greener models. 2021.

[2] James Philbin Florian Schroff, Dmitry Kalenichenko. Facenet: A unified embedding for face recognition and clustering. 2015.

[3] Jyunichi Miyao Hideki Oki, Motoshi Abe and Takio Kurita. Triplet loss for knowledge distillation. 2020.

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

[5] Radu Tudor Ionescu Mariana-Iuliana Georgescu, Georgian-Emilian Du¸t ˘a. Teacher–student training and triplet loss to reduce the effect of drastic face occlusion. 2015.

[6] Zhongwei Cheng Lin Chen Sumanth Chennupati, Mohammad Mahdi Kamani. Adaptive distillation: Aggregating knowledge from multiple paths for efficient distillation. 2021.

[7] Jun Wang Yuang Liu, Wei Zhang. Adaptive multi-teacher multi-level knowledge distillation. 2021.