УДК 004.89

ПРИМЕНЕНИЕ СОВРМЕННЫХ ПОДХОДОВ К ЗАДАЧЕ ДИСТИЛЛЯЦИИ ЗНАНИЙ ДЛЯ ДИАЛОГОВЫХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ НА БАЗЕ АРХИТЕКТУРЫ ТРАНСФОРМЕР

Кузьмин А.Д. (Университет ИТМО)
Научный руководитель — к.т.н. Шуранов Е.В. (Университет ИТМО)

Введение. В данной работе на примере популярных подходов рассмотрена специфика такой задачи как дистилляция знания. Кроме того, проведён анализ последних наиболее успешных работ, показавших качественную передачу знаний на соответствующих им прикладных задачах, а также воспроизведены эксперименты связанные с некоторыми из них. Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 622282 «Разработка русскоязычного персонифицированного диалогового агента с динамической долгосрочной памятью».

Основная часть. Несмотря на внушительный рост доступных ресурсов большинства вычислительных устройств, компактные процессоры не способны рассчитать результат работы крупных искусственных нейронных сетей за допустимое во многих прикладных областях время. Одним из решений для запусков подобных вычислений является использование сетей с меньшим числом параметров, но при этом обеспечивающих качество близкое к качеству исходной модели. Создание подобных моделей путём передачи “знаний” от более крупной модели к меньшей называют дистилляцией знаний. Первым решением задачи дистилляция знаний принято считать работу “Model compression” [1]. Авторами был предложен алгоритм для “сжатия” ансамбля классификаторов в один единственный классификатор для повышения скорости работы модели. В дальнейшем такой подход к “ускорению” моделей стал интересен многим исследователям и получил широкое приложение во множестве прикладных областей.

На текущем уровне развития всё большую популярность набирают модели на базе архитектуры Трансформер. Однако, именно данные модели зачастую очень требовательны к ресурсам вычислительной машины. Применение классических подходов для дистилляции знаний к данным моделям затруднительно из-за их блока внимания. Ввиду этой проблемы данная работа предлагает применить подход, основанный на кросс-модальной дистилляции, предлагающий использовать Трансформер в качестве модели-учителя для свёрточной нейросетевой модели [2].

Выводы. В данной работе приведены современные методы дистилляции знаний, учитывающие особенности данной архитектуры. Кроме того, на основе прикладных метрик, высчитываемых для диалоговых моделей, сделан вывод о пользе тех или иных методов в будущих исследованиях текущего проекта.

Список использованных источников:

1. C. Bucilua, R. Caruana, and A. Niculescu-Mizil, “Model compression” in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2006, pp. 535–541
2. Gong Y., Khurana S., Rouditchenko A., Glass J. CMKD: CNN/Transformer-Based CrossModel Knowledge Distillation for Audio Classification // 2022