УДК 004.85

Разработка метода слияния экспертов в архитектурах генеративных моделей на основе mixture of experts

Пакульневич К.М. (ИТМО), **Бессонницын Е.С.** (ИТМО)

Научный руководитель – кандидат технических наук Муравьев С.Б. (ИТМО)

Введение. Многие современные генеративные модели используют смеси экспертов (mixture of experts) внутри своей архитектуры для увеличения качества, при этом на работу экспертов приходится большая доля вычислений. Использование методов слияния моделей для них может позволить сократить вычисления, производимые различными экспертами. Несмотря на то, что существует много алгоритмов слияния моделей [1], данные методы зачастую слишком требовательны к моделям и не учитывают специфику экспертов в ансамбле. Целью данной работы является создание алгоритма эффективного слияния экспертов, который позволит значительно ускорить большие модели MoE-архитектуры.

Основная часть. В смеси при каждом срабатывание выбирается фиксированное количество экспертов, при этом среди них можно найти группы экспертов, которые часто выбираются вместе. Предложенный алгоритм использует данную особенность для того, чтобы создать сжатую модель из часто встречающихся вместе экспертов, где для слияния выбираются эксперты на основе частоты совместной активации экспертов.

Сам процесс сжатия базируется на предложенном в [2] методе, его особенностью является то, что он не опирается на функциональные свойства экспертов, вместо этого он функционально объединяет их вычисления в одно векторное пространство. Базовый метод, предложенный в указанной статье позволяет не просто построить новую модель, но и дает возможность из выхода новой модели получить приближенную оценку для значений исходных моделей.

Мы реализовали алгоритм, который находит пары экспертов на каждом MLP слое модели и объединяет их. Нами были поставлены эксперименты на нескольких открытых больших языковых MoE моделях.

Выводы. Предложен метод сжатия модели, состоящей из смеси экспертов. Использование алгоритма для открытых моделей показывает значимую способность к оптимизации вычислений. Предварительное тестирование предложенного метода показывает ускорение в 5-10 %, в зависимости от количества экспертов на слой и особенностей архитектуры.

Список использованных источников:

1. Enneng Yang et al. Model Merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities //arXiv preprint arXiv:2408.07666. – 2024.
2. George Stoica, Daniel Bolya et al. ZipIt! Merging Models from Different Tasks without Training //arXiv preprint arXiv:2305.03053. – 2023.