Поиск низкобитовых матриц кодирования в сотовых системах связи пятого поколения

Тышова Ольга Александровна

Студент

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

E-mail: s02200227@gse.cs.msu.ru

Научный руководитель *— Кропотов Дмитрий Александрович*

Одной из основополагающих технологий в системах связи пятого поколения является многопользовательская система передачи данных с множественным входом и выходом (Multi-user (MU) MIMO), которая основана на многоантенной приёмопередаче и обеспечивает значительное увеличение пропускной способности между абонентами и базовой станцией. Канал передачи данных в Multi-user MIMO описывается линейной моделью $y = G(HWx + n)$, где x ndash; данные, передаваемые базовой станцией абоненту, W ndash; матрица кодирования, n ndash; аддитивный белый шум, H – канальная матрица, G – матрица приема, вычисляемая абонентом. При больших размерах антенной решетки в десятки или сотни элементов матричное умножение в формуле выше становится затратной операцией. Вычисление этого произведения может быть сильно ускорено, если использовать низкобитовое представление матрицы W и вектора x. В данной работе рассматривается подход, когда матрица кодирования есть решение задачи оптимизации с ограничениями

$$\begin{cases} \sum_{k=1}^{K} L_k \log_2(1 + SINR_k^{eff}(W, H_k, G_k, \sigma, P)) \to \max_W \\ ||w_i||^2 \leq \frac{P}{T}. \end{cases}$$

Для нахождения низкобитовой (или квантизованной) W исследуются методы, используемые для квантизации нейронных сетей, поскольку дискретная задача оптимизации поиска квантизованной матрицы через оптимизацию среднеквадратичного отклонения от вещественной является NP-трудной задачей. Рассматриваются два метода квантизации ndash; квантизация после обучения (Post-Training Quantization) и квантизация во время обучения (Quantization Aware Training). Post-Training Quantization представляет собой перевод уже найденной оптимальной матрицы W в целые числа выбранной битности: производится отображение диапазона значений вещественной матрицы на интервал значений выбранной битности, затем идет округление. В Quantization Aware Training производится оптимиза-

ция с учетом квантизации ndash; перевод в целые числа симулируется значениями во float32 за счет специальных функций имитации квантизации (Fake Quantization), которые позволяют вычислять аппроксимацию градиента через операцию округления. В данной работе были применены оба метода квантизации для нахождения целочисленной матрицы кодирования для нисходящей передачи. Моделирование проводилось на наборе канальных матриц, сгенерированных в симуляторе QuadRiGa. Сравнение низкобитовых матриц W для разных битностей, полученных методом Post-Training Quantization, с вещественной матрицей представлены на рисунке.

.png .pdf .jpg .mps .jpeg .jbig2 .jb2 .PNG .PDF .JPG .JPEG .JBIG2 .JB2 .eps

Пропускная способность, рассчитанная для вещественной матрицы W (сплошная линия) и ее низкобитовых аппроксимаций (пунктир), полученных методом Post-Training Quantization

Литература

1. Wu H., Integer quantization for deep learning inference: Principles and empirical evaluation. arXiv preprint arXiv:2004.09602, 2020.
2. Krishnamoorthi R. Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342, 2018