不对称的 b-bit 量化技术

不对称的 b-bit 量化技术 是一种针对数据分布进行优化的量化方法，与传统的对称量化不同，其特点在于：

---

1. 什么是量化?

量化是指将连续值（如浮点数）映射到离散值（整数）的过程，常用于压缩数据、减少存储需求和加速计算。对于 b-bit 量化：

- 数值范围被划分为 2^b 个离散值（或称量化级别）。
- 量化后数据占用的比特数为 b 比特。

---

2. 对称量化 vs 不对称量化

- **对称量化**：量化范围以零为中心，分布在 $[-\max, \max]$ 内。公式：

$$q = \text{round}\left(\frac{x}{s}\right) \tag{1}$$

  - s 是缩放因子。
  - 适合数据分布对称的场景。

- **不对称量化**：量化范围并非以零为中心，分布在 $[\beta, \beta + \gamma \times (2^b - 1)]$。
  - **量化公式**：

$$\hat{l}_i = \text{clamp}\left(\frac{l_i - \beta_i}{\gamma_i}, 0, 2^b - 1\right) \tag{2}$$

    - β_i：偏移因子，用于调整数据分布的起点。
    - γ_i：缩放因子，用于确定量化步长。
  - **反量化公式**：

$$l_i = \hat{l}_i \times \gamma_i + \beta_i \tag{3}$$

---

3. 不对称量化的优点

- **适应非对称分布**：适合数据分布不以零为中心的情况，例如所有数据都偏向正值或某个固定范围。
- **提高量化效率**：通过引入偏移因子 β，可以更紧密地覆盖数据分布，有效减少量化误差。
- **减少动态范围需求**：相比对称量化，它不需要为了覆盖负值范围而浪费表示能力。

---

4. 具体实现中的细节

- **b-bit 的作用**：
  - b 表示量化位宽，例如 $b = 8$ 时，数据被映射到 $2^8 = 256$ 个离散值。
  - 量化后只需存储 b-bit 的整数，相比浮点数（通常占 32 位或 16 位），显著减少存储空间。
- **关键参数学习**：
  - **偏移因子** β 和 **缩放因子** γ 是在训练或微调过程中通过优化学习得到的。
  - 它们确保量化后的整数值能有效表示原始数据分布。
- **clamp 操作**：用于限制量化值在 $[0, 2^b - 1]$ 范围内，避免量化值溢出。

---

5. 应用场景

- **深度学习模型压缩**：用于压缩神经网络权重或激活值，特别是在分布范围不对称时。
- **图像处理与压缩**：如文中提到的高斯参数量化，通过学习偏移和缩放因子，减少存储数据所需的比特数。

总结

不对称的 b-bit 量化通过引入偏移因子 β 和缩放因子 γ，能够更高效地适配数据分布，尤其是非对称分布，最终达到压缩存储和降低计算成本的目的。