Universidad Alfonso X el Sabio

Bases de la cuantización y las diferencias entre PTQ y QAT

Germán Llorente y Carlos Puigserver

October 16, 2023

1 Introducción

La cuantización es un proceso esencial en el campo del aprendizaje automático y la inteligencia artificial, donde los modelos complejos se simplifican para su implementación en dispositivos con recursos limitados. En este artículo, se explorarán las bases de la cuantización y se compararán dos enfoques populares: la Cuantización Posterior al Entrenamiento (PTQ) y el Entrenamiento Consciente de la Cuantización (QAT).

2 Cuantización: Fundamentos

La cuantización es el proceso de convertir modelos de punto flotante en modelos de punto fijo, reduciendo así el consumo de memoria y energía durante la inferencia. En un modelo cuantizado, los valores de los parámetros y las activaciones se representan con menos bits, lo que ahorra espacio de almacenamiento y mejora la eficiencia computacional. La cuantización se utiliza ampliamente en dispositivos móviles, sistemas embebidos y hardware especializado para acelerar las inferencias de modelos de aprendizaje automático.

3 Cuantización Posterior al Entrenamiento (PTQ)

PTQ es un enfoque de cuantización donde un modelo pre-entrenado se convierte a formato cuantizado después del entrenamiento. Durante este proceso, se determinan los rangos de los valores de los parámetros y activaciones en el modelo pre-entrenado. Luego, estos valores se cuantizan a un número fijo de bits para la implementación en hardware específico. PTQ es rápido y fácil de implementar, pero puede perder precisión debido a la falta de ajuste fino durante el entrenamiento. Esto puede ser aceptable en ciertas aplicaciones donde la pérdida de precisión es tolerable en comparación con los beneficios de la reducción del tamaño del modelo y la mejora de la eficiencia computacional.

4 Entrenamiento Consciente de la Cuantización (QAT)

QAT, por otro lado, es un enfoque donde el modelo se entrena directamente con cuantización en mente. Durante el entrenamiento, se aplican operaciones de cuantización simulada en el modelo, lo que permite que los parámetros se ajusten para adaptarse mejor a las limitaciones de la cuantización. Esto lleva a modelos cuantizados que son más precisos en comparación con PTQ. Sin embargo, el entrenamiento consciente de la cuantización puede ser más complejo y computacionalmente intensivo, ya que requiere modificar los algoritmos de entrenamiento y puede aumentar el tiempo de entrenamiento.

5 Comparación y Conclusiones

En resumen, tanto la Cuantización Posterior al Entrenamiento como el Entrenamiento Consciente de la Cuantización son técnicas importantes para implementar modelos de aprendizaje automático en dispositivos con recursos limitados. PTQ es más rápido y fácil, pero puede sacrificar la precisión, especialmente en modelos complejos donde la información detallada es crucial. QAT, por otro lado, ofrece una mayor precisión al permitir que el modelo se ajuste específicamente para la cuantización, pero a costa de una mayor complejidad y tiempo de entrenamiento.

La elección entre estos enfoques depende de las necesidades específicas de la aplicación y los recursos disponibles. En situaciones donde la precisión es fundamental y se dispone de recursos computacionales suficientes, QAT puede ser la mejor opción. Por otro lado, en aplicaciones donde la eficiencia computacional y el tamaño del modelo son críticos y la pérdida de cierta precisión es aceptable, PTQ puede ser una elección adecuada. En última instancia, la selección del método de cuantización debe equilibrar la precisión, la eficiencia y los recursos disponibles para satisfacer los requisitos del caso de uso específico.