Битность слоев модели BERT-BASE-UNCASED обеспечивающая сжатие размера в 5.5 раз и сохранение примермлемого качества на наборе данных CoLa

Layer: bert.embeddings.word_embeddings | bit: 5
Layer: bert.embeddings.position_embeddings | bit: 6
Layer: bert.embeddings.token_type_embeddings | bit: 2
Layer: bert.encoder.layer.0.attention.self.query | bit: 2
Layer: bert.encoder.layer.0.attention.self.key | bit: 2
Layer: bert.encoder.layer.0.attention.self.value | bit: 2
Layer: bert.encoder.layer.0.attention.output.dense | bit: 2
Layer: bert.encoder.layer.0.intermediate.dense | bit: 2
Layer: bert.encoder.layer.0.output.dense | bit: 2
Layer: bert.encoder.layer.1.attention.self.query | bit: 2
Layer: bert.encoder.layer.1.attention.self.key | bit: 2
Layer: bert.encoder.layer.1.attention.self.value | bit: 2
Layer: bert.encoder.layer.1.attention.output.dense | bit: 2
Layer: bert.encoder.layer.1.intermediate.dense | bit: 4
Layer: bert.encoder.layer.1.output.dense | bit: 5
Layer: bert.encoder.layer.2.attention.self.query | bit: 3
Layer: bert.encoder.layer.2.attention.self.key | bit: 3
Layer: bert.encoder.layer.2.attention.self.value | bit: 4
Layer: bert.encoder.layer.2.attention.output.dense | bit: 8
Layer: bert.encoder.layer.2.intermediate.dense | bit: 7
Layer: bert.encoder.layer.2.output.dense | bit: 8
Layer: bert.encoder.layer.3.attention.self.query | bit: 3
Layer: bert.encoder.layer.3.attention.self.key | bit: 3
Layer: bert.encoder.layer.3.attention.self.value | bit: 6
Layer: bert.encoder.layer.3.attention.output.dense | bit: 6
Layer: bert.encoder.layer.3.intermediate.dense | bit: 8
Layer: bert.encoder.layer.3.output.dense | bit: 7
Layer: bert.encoder.layer.4.attention.self.query | bit: 8
Layer: bert.encoder.layer.4.attention.self.key | bit: 8
Layer: bert.encoder.layer.4.attention.self.value | bit: 8
Layer: bert.encoder.layer.4.attention.output.dense | bit: 8
Layer: bert.encoder.layer.4.intermediate.dense | bit: 8
Layer: bert.encoder.layer.4.output.dense | bit: 8
Layer: bert.encoder.layer.5.attention.self.query | bit: 8
Layer: bert.encoder.layer.5.attention.self.key | bit: 8

Layer: bert.encoder.layer.5.attention.self.value | bit: 8
Layer: bert.encoder.layer.5.attention.output.dense | bit: 8
Layer: bert.encoder.layer.5.intermediate.dense | bit: 8
Layer: bert.encoder.layer.5.output.dense | bit: 8
Layer: bert.encoder.layer.6.attention.self.query | bit: 8
Layer: bert.encoder.layer.6.attention.self.key | bit: 8
Layer: bert.encoder.layer.6.attention.self.value | bit: 8
Layer: bert.encoder.layer.6.attention.output.dense | bit: 8
Layer: bert.encoder.layer.6.intermediate.dense | bit: 8
Layer: bert.encoder.layer.6.output.dense | bit: 8
Layer: bert.encoder.layer.7.attention.self.query | bit: 6
Layer: bert.encoder.layer.7.attention.self.key | bit: 7
Layer: bert.encoder.layer.7.attention.self.value | bit: 8
Layer: bert.encoder.layer.7.attention.output.dense | bit: 6
Layer: bert.encoder.layer.7.intermediate.dense | bit: 7
Layer: bert.encoder.layer.7.output.dense | bit: 7
Layer: bert.encoder.layer.8.attention.self.query | bit: 4
Layer: bert.encoder.layer.8.attention.self.key | bit: 8
Layer: bert.encoder.layer.8.attention.self.value | bit: 8
Layer: bert.encoder.layer.8.attention.output.dense | bit: 5
Layer: bert.encoder.layer.8.intermediate.dense | bit: 7
Layer: bert.encoder.layer.8.output.dense | bit: 7
Layer: bert.encoder.layer.9.attention.self.query | bit: 7
Layer: bert.encoder.layer.9.attention.self.key | bit: 4
Layer: bert.encoder.layer.9.attention.self.value | bit: 3
Layer: bert.encoder.layer.9.attention.output.dense | bit: 4
Layer: bert.encoder.layer.9.intermediate.dense | bit: 8
Layer: bert.encoder.layer.9.output.dense | bit: 5
Layer: bert.encoder.layer.10.attention.self.query | bit: 3
Layer: bert.encoder.layer.10.attention.self.key | bit: 5
Layer: bert.encoder.layer.10.attention.self.value | bit: 3
Layer: bert.encoder.layer.10.attention.output.dense | bit: 3
Layer: bert.encoder.layer.10.intermediate.dense | bit: 6
Layer: bert.encoder.layer.10.output.dense | bit: 4
Layer: bert.encoder.layer.11.attention.self.query | bit: 4
Layer: bert.encoder.layer.11.attention.self.key | bit: 5
Layer: bert.encoder.layer.11.attention.self.value | bit: 3
Layer: bert.encoder.layer.11.attention.output.dense | bit: 8
Layer: bert.encoder.layer.11.intermediate.dense | bit: 5

Layer: bert.encoder.layer.11.output.dense | bit: 7
Layer: bert.pooler.dense | bit: 7