The Ultimate Compression Pipeline ~ Ajay Maheshwari

In [6]:
```
! pip install -q tensorflow-model-optimization
```

In [7]:
```
import tensorflow as tf
import tf_keras as keras

import numpy as np
import tempfile
import zipfile
import os
```

In [8]:
```
def get_gzipped_model_size(model):
  with tempfile.NamedTemporaryFile(suffix=".h5") as temp_file:
    model.save(temp_file.name)

    _, zipped_file = tempfile.mkstemp('.zip')
    with zipfile.ZipFile(zipped_file, 'w', compression=zipfile.ZIP_DEFLATED) as f:
      f.write(temp_file.name)

    # print(f"Zipped model is saved at: {zipped_file}")

    x = os.path.getsize(zipped_file)

    os.remove(zipped_file)
    # print(f"Temporary zip file removed: {zipped_file}")

    return x / 1000


def print_model_weights_sparsity(model):
    for layer in model.layers:
        if isinstance(layer, keras.layers.Wrapper):
            weights = layer.trainable_weights
        else:
            weights = layer.weights
        for weight in weights:
            if "kernel" not in weight.name or "centroid" in weight.name:
                continue
            weight_size = weight.numpy().size
            zero_num = np.count_nonzero(weight == 0)
            print(
                f"{weight.name}: {zero_num/weight_size:.2%} sparsity ",
                f"({zero_num}/{weight_size})",
            )


def get_gzipped_model_size2(file):

  _, zipped_file = tempfile.mkstemp('.zip')
  with zipfile.ZipFile(zipped_file, 'w', compression=zipfile.ZIP_DEFLATED) as f:
    f.write(file)

  return os.path.getsize(zipped_file)/1000


def eval_model(interpreter):
  input_index = interpreter.get_input_details()[0]["index"]
  output_index = interpreter.get_output_details()[0]["index"]

  prediction_digits = []
  for i, test_image in enumerate(test_images):
    test_image = np.expand_dims(test_image, axis=0).astype(np.float32)
    interpreter.set_tensor(input_index, test_image)

    interpreter.invoke()

    output = interpreter.tensor(output_index)
    digit = np.argmax(output()[0])
    prediction_digits.append(digit)

  prediction_digits = np.array(prediction_digits)
  accuracy = (prediction_digits == test_labels).mean()
  return accuracy

model_acc = []
model_sz = []
```

Creating a Base Model - 1.0

In [9]:
```
# Load MNIST dataset
mnist = keras.datasets.mnist
(train_images, train_labels), (test_images, test_labels) = mnist.load_data()

# Normalize the input image so that each pixel value is between 0 to 1.
train_images = train_images / 255.0
test_images  = test_images / 255.0

model = keras.Sequential([
  keras.layers.InputLayer(input_shape=(28, 28)),
```

```
  keras.layers.Reshape(target_shape=(28, 28, 1)),
  keras.layers.Conv2D(filters=12, kernel_size=(3, 3),
                      activation=tf.nn.relu),
  keras.layers.MaxPooling2D(pool_size=(2, 2)),
  keras.layers.Flatten(),
  keras.layers.Dense(10)
])

opt = keras.optimizers.Adam(learning_rate=1e-3)

# Train the digit classification model
model.compile(optimizer=opt,
              loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
              metrics=['accuracy'])

model.fit(
    train_images,
    train_labels,
    validation_split=0.1,
    epochs=10
)
```

```
WARNING:absl:At this time, the v2.11+ optimizer `tf.keras.optimizers.Adam` runs slowly on M1/M2 Macs, please use th
e legacy TF-Keras optimizer instead, located at `tf.keras.optimizers.legacy.Adam`.
```

```
Epoch 1/10
1688/1688 [==============================] - 5s 3ms/step - loss: 0.3176 - accuracy: 0.9106 - val_loss: 0.1256 - val
_accuracy: 0.9647
Epoch 2/10
1688/1688 [==============================] - 4s 3ms/step - loss: 0.1250 - accuracy: 0.9652 - val_loss: 0.0815 - val
_accuracy: 0.9805
Epoch 3/10
1688/1688 [==============================] - 4s 3ms/step - loss: 0.0860 - accuracy: 0.9758 - val_loss: 0.0702 - val
_accuracy: 0.9820
Epoch 4/10
1688/1688 [==============================] - 4s 3ms/step - loss: 0.0701 - accuracy: 0.9794 - val_loss: 0.0633 - val
_accuracy: 0.9833
Epoch 5/10
1688/1688 [==============================] - 5s 3ms/step - loss: 0.0604 - accuracy: 0.9826 - val_loss: 0.0691 - val
_accuracy: 0.9815
Epoch 6/10
1688/1688 [==============================] - 6s 3ms/step - loss: 0.0538 - accuracy: 0.9840 - val_loss: 0.0601 - val
_accuracy: 0.9842
Epoch 7/10
1688/1688 [==============================] - 6s 3ms/step - loss: 0.0482 - accuracy: 0.9858 - val_loss: 0.0628 - val
_accuracy: 0.9853
Epoch 8/10
1688/1688 [==============================] - 6s 3ms/step - loss: 0.0441 - accuracy: 0.9865 - val_loss: 0.0546 - val
_accuracy: 0.9872
Epoch 9/10
1688/1688 [==============================] - 6s 3ms/step - loss: 0.0402 - accuracy: 0.9882 - val_loss: 0.0558 - val
_accuracy: 0.9858
Epoch 10/10
1688/1688 [==============================] - 6s 3ms/step - loss: 0.0368 - accuracy: 0.9887 - val_loss: 0.0550 - val
_accuracy: 0.9873
```

Out[9]: `<tf_keras.src.callbacks.History at 0x29f3ae610>`

Evaluate the baseline model and save it for later usage

In [10]:
```
_, baseline_model_accuracy = model.evaluate(
    test_images, test_labels, verbose=0)

print('Baseline test accuracy:', baseline_model_accuracy)

sz = get_gzipped_model_size(model)
print("Base model size: ",  sz , ' KB' )

model_acc.append(baseline_model_accuracy)
model_sz.append(sz)
```

```
Baseline test accuracy: 0.982699990272522
Base model size:  235.287  KB
```

---------- Checkpoint Point 1 ---------

Pruning and then fine-tuning the model - 2.0

In [11]:
```
import tensorflow_model_optimization as tfmot

pruning_params = {
          'pruning_schedule': tfmot.sparsity.keras.ConstantSparsity(0.5, begin_step=0, frequency=100)
}

callbacks = [
      tfmot.sparsity.keras.UpdatePruningStep()
]


pruned_model = model

for i in range(1):
    prune_low_magnitude = tfmot.sparsity.keras.prune_low_magnitude

    pruned_model = prune_low_magnitude(pruned_model, **pruning_params)

    # learning rate for fine-tuning
    opt = keras.optimizers.Adam(learning_rate=1e-5)
```

```
    pruned_model.compile(
      loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
      optimizer=opt,
      metrics=['accuracy'])

    # Fine-tune model
    pruned_model.fit(
      train_images,
      train_labels,
      epochs=10,
      validation_split=0.1,
      callbacks=callbacks)

    stripped_pruned_model = tfmot.sparsity.keras.strip_pruning(pruned_model)

    print_model_weights_sparsity(stripped_pruned_model)

    pruned_model = stripped_pruned_model
```

```
WARNING:absl:At this time, the v2.11+ optimizer `tf.keras.optimizers.Adam` runs slowly on M1/M2 Macs, please use th
e legacy TF-Keras optimizer instead, located at `tf.keras.optimizers.legacy.Adam`.
```

```
Epoch 1/10
1688/1688 [==============================] - 7s 4ms/step - loss: 0.0664 - accuracy: 0.9786 - val_loss: 0.0738 - val
_accuracy: 0.9795
Epoch 2/10
1688/1688 [==============================] - 6s 3ms/step - loss: 0.0497 - accuracy: 0.9852 - val_loss: 0.0677 - val
_accuracy: 0.9822
Epoch 3/10
1688/1688 [==============================] - 6s 4ms/step - loss: 0.0461 - accuracy: 0.9866 - val_loss: 0.0661 - val
_accuracy: 0.9828
Epoch 4/10
1688/1688 [==============================] - 7s 4ms/step - loss: 0.0445 - accuracy: 0.9871 - val_loss: 0.0652 - val
_accuracy: 0.9832
Epoch 5/10
1688/1688 [==============================] - 9s 5ms/step - loss: 0.0433 - accuracy: 0.9873 - val_loss: 0.0644 - val
_accuracy: 0.9832
Epoch 6/10
1688/1688 [==============================] - 9s 5ms/step - loss: 0.0422 - accuracy: 0.9878 - val_loss: 0.0637 - val
_accuracy: 0.9833
Epoch 7/10
1688/1688 [==============================] - 9s 5ms/step - loss: 0.0413 - accuracy: 0.9882 - val_loss: 0.0630 - val
_accuracy: 0.9835
Epoch 8/10
1688/1688 [==============================] - 9s 5ms/step - loss: 0.0406 - accuracy: 0.9886 - val_loss: 0.0624 - val
_accuracy: 0.9835
Epoch 9/10
1688/1688 [==============================] - 9s 5ms/step - loss: 0.0399 - accuracy: 0.9889 - val_loss: 0.0619 - val
_accuracy: 0.9835
Epoch 10/10
1688/1688 [==============================] - 9s 5ms/step - loss: 0.0392 - accuracy: 0.9890 - val_loss: 0.0616 - val
_accuracy: 0.9835
conv2d_1/kernel:0: 50.00% sparsity  (54/108)
dense_1/kernel:0: 50.00% sparsity  (10140/20280)
```

In [12]:
```
pruned_model.compile(
      loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
      optimizer=opt,
      metrics=['accuracy'])
```

In [ ]:
```
```

In [13]:
```
_, pruned_model_accuracy = pruned_model.evaluate(
    test_images, test_labels, verbose=0)

print('Pruned Model test accuracy:', pruned_model_accuracy)

sz = get_gzipped_model_size(pruned_model)
print("Stripped model size: ",  sz , ' KB' )

model_acc.append(pruned_model_accuracy)
model_sz.append(sz)
```

```
Pruned Model test accuracy: 0.9807999730110168
Stripped model size:  203.828  KB
```

---------- Checkpoint Point 2 ---------

Knowledge Distillation - 3.0

In [14]:
```
from keras import layers
from keras import ops
import numpy as np

class Distiller(keras.Model):
    def __init__(self, student, teacher):
        super().__init__()
        self.teacher = teacher
        self.student = student

    def compile(
        self,
        optimizer,
```

```
        metrics,
        student_loss_fn,
        distillation_loss_fn,
        alpha=0.1,
        temperature=3,
    ):
        super().compile(optimizer=optimizer, metrics=metrics)
        self.student_loss_fn = student_loss_fn
        self.distillation_loss_fn = distillation_loss_fn
        self.alpha = alpha
        self.temperature = temperature

    def compute_loss(
        self, x=None, y=None, y_pred=None, sample_weight=None, allow_empty=False
    ):
        teacher_pred = self.teacher(x, training=False)
        student_loss = self.student_loss_fn(y, y_pred)

        distillation_loss = self.distillation_loss_fn(
            ops.softmax(teacher_pred / self.temperature, axis=1),
            ops.softmax(y_pred / self.temperature, axis=1),
        ) * (self.temperature**2)

        loss = self.alpha * student_loss + (1 - self.alpha) * distillation_loss
        return loss

    def call(self, x):
        return self.student(x)
```

In [15]:
```
teacher = keras.Sequential(
    [
        keras.layers.InputLayer(input_shape=(28, 28)),
        keras.layers.Reshape(target_shape=(28, 28, 1)),
        keras.layers.Conv2D(filters=16, kernel_size=(3, 3), activation=tf.nn.relu),  # Increase filters
        keras.layers.MaxPooling2D(pool_size=(2, 2)),
        keras.layers.Conv2D(filters=24, kernel_size=(3, 3), activation=tf.nn.relu),  # Add another layer
        keras.layers.MaxPooling2D(pool_size=(2, 2)),
        keras.layers.Flatten(),
        keras.layers.Dense(units=128, activation=tf.nn.relu),  # Add a hidden layer
        keras.layers.Dense(10)
    ],
    name="teacher",
)


# teacher = keras.Sequential(
#     [
#         keras.layers.InputLayer(input_shape=(28, 28)),
#         keras.layers.Reshape(target_shape=(28, 28, 1)),
#         keras.layers.Conv2D(filters=12, kernel_size=(3, 3),
#                             activation=tf.nn.relu),
#         keras.layers.MaxPooling2D(pool_size=(2, 2)),
#         keras.layers.Flatten(),
#         keras.layers.Dense(10)
#     ],
#     name="teacher",
# )


teacher.compile(
    optimizer=keras.optimizers.Adam(),
    loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
    metrics=[keras.metrics.SparseCategoricalAccuracy()],
)
```

```
WARNING:absl:At this time, the v2.11+ optimizer `tf.keras.optimizers.Adam` runs slowly on M1/M2 Macs, please use th
e legacy TF-Keras optimizer instead, located at `tf.keras.optimizers.legacy.Adam`.
```

In [16]:
```
teacher.fit(train_images, train_labels, epochs=12)
teacher.evaluate(test_images, test_labels)
```

```
Epoch 1/12
1875/1875 [==============================] - 15s 8ms/step - loss: 0.1590 - sparse_categorical_accuracy: 0.9517
Epoch 2/12
1875/1875 [==============================] - 11s 6ms/step - loss: 0.0520 - sparse_categorical_accuracy: 0.9844
Epoch 3/12
1875/1875 [==============================] - 11s 6ms/step - loss: 0.0368 - sparse_categorical_accuracy: 0.9886
Epoch 4/12
1875/1875 [==============================] - 10s 6ms/step - loss: 0.0285 - sparse_categorical_accuracy: 0.9911
Epoch 5/12
1875/1875 [==============================] - 10s 5ms/step - loss: 0.0211 - sparse_categorical_accuracy: 0.9929
Epoch 6/12
1875/1875 [==============================] - 11s 6ms/step - loss: 0.0158 - sparse_categorical_accuracy: 0.9947
Epoch 7/12
1875/1875 [==============================] - 12s 6ms/step - loss: 0.0140 - sparse_categorical_accuracy: 0.9952
Epoch 8/12
1875/1875 [==============================] - 11s 6ms/step - loss: 0.0114 - sparse_categorical_accuracy: 0.9962
Epoch 9/12
1875/1875 [==============================] - 11s 6ms/step - loss: 0.0091 - sparse_categorical_accuracy: 0.9970
Epoch 10/12
1875/1875 [==============================] - 10s 6ms/step - loss: 0.0077 - sparse_categorical_accuracy: 0.9973
Epoch 11/12
1875/1875 [==============================] - 11s 6ms/step - loss: 0.0080 - sparse_categorical_accuracy: 0.9973
Epoch 12/12
1875/1875 [==============================] - 10s 6ms/step - loss: 0.0061 - sparse_categorical_accuracy: 0.9980
313/313 [==============================] - 1s 3ms/step - loss: 0.0535 - sparse_categorical_accuracy: 0.9884
```

Out[16]:
```
[0.053500182926654816, 0.9883999824523926]
```

In [17]:
```
distiller = Distiller(student=pruned_model, teacher=teacher)
distiller.compile(
    optimizer=keras.optimizers.Adam(),
    metrics=[keras.metrics.SparseCategoricalAccuracy()],
    student_loss_fn=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
```

```
    distillation_loss_fn=keras.losses.KLDivergence(),
    alpha=0.1,
    temperature=10,
)
```

```
WARNING:absl:At this time, the v2.11+ optimizer `tf.keras.optimizers.Adam` runs slowly on M1/M2 Macs, please use th
e legacy TF-Keras optimizer instead, located at `tf.keras.optimizers.legacy.Adam`.
```

In [18]:
```
# Distill teacher se student
distiller.fit(train_images, train_labels, epochs=15)

distiller.evaluate(test_images, test_labels)
```

```
Epoch 1/15
1875/1875 [==============================] - 14s 7ms/step - sparse_categorical_accuracy: 0.9704
Epoch 2/15
1875/1875 [==============================] - 14s 7ms/step - sparse_categorical_accuracy: 0.9712
Epoch 3/15
1875/1875 [==============================] - 13s 7ms/step - sparse_categorical_accuracy: 0.9702
Epoch 4/15
1875/1875 [==============================] - 13s 7ms/step - sparse_categorical_accuracy: 0.9690
Epoch 5/15
1875/1875 [==============================] - 14s 7ms/step - sparse_categorical_accuracy: 0.9689
Epoch 6/15
1875/1875 [==============================] - 13s 7ms/step - sparse_categorical_accuracy: 0.9701
Epoch 7/15
1875/1875 [==============================] - 13s 7ms/step - sparse_categorical_accuracy: 0.9712
Epoch 8/15
1875/1875 [==============================] - 14s 7ms/step - sparse_categorical_accuracy: 0.9723
Epoch 9/15
1875/1875 [==============================] - 13s 7ms/step - sparse_categorical_accuracy: 0.9730
Epoch 10/15
1875/1875 [==============================] - 14s 7ms/step - sparse_categorical_accuracy: 0.9738
Epoch 11/15
1875/1875 [==============================] - 14s 7ms/step - sparse_categorical_accuracy: 0.9742
Epoch 12/15
1875/1875 [==============================] - 13s 7ms/step - sparse_categorical_accuracy: 0.9752
Epoch 13/15
1875/1875 [==============================] - 10s 6ms/step - sparse_categorical_accuracy: 0.9752
Epoch 14/15
1875/1875 [==============================] - 10s 6ms/step - sparse_categorical_accuracy: 0.9759
Epoch 15/15
1875/1875 [==============================] - 11s 6ms/step - sparse_categorical_accuracy: 0.9761
313/313 [==============================] - 1s 951us/step - sparse_categorical_accuracy: 0.9794
```

Out[18]:
```
0.9793999791145325
```

In [19]:
```
_, acc = distiller.student.evaluate(
    test_images, test_labels, verbose=0)

print('Distilled Model test accuracy:', acc)

sz = get_gzipped_model_size(distiller.student)
print("Distilled model size: ",  sz , ' KB' )

model_acc.append(acc)
model_sz.append(sz)
```

```
Distilled Model test accuracy: 0.9793999791145325
Distilled model size:  227.846  KB
```

---------- Checkpoint Point 3 ---------

Weight Clustering - 4.0

In [20]:
```
def print_model_weight_clusters(model):
    for layer in model.layers:
        if isinstance(layer, keras.layers.Wrapper):
            weights = layer.trainable_weights
        else:
            weights = layer.weights
        for weight in weights:
            # ignore auxiliary quantization weights
            if "quantize_layer" in weight.name:
                continue
            if "kernel" in weight.name:
                unique_count = len(np.unique(weight))
                print(
                    f"{layer.name}/{weight.name}: {unique_count} clusters "
                )
```

In [21]:
```
import tensorflow_model_optimization as tfmot
from tensorflow_model_optimization.python.core.clustering.keras.experimental import (
    cluster,
)

cluster_weights = tfmot.clustering.keras.cluster_weights
CentroidInitialization = tfmot.clustering.keras.CentroidInitialization

cluster_weights = cluster.cluster_weights

clustering_params = {
  'number_of_clusters': 8,
  'cluster_centroids_init': CentroidInitialization.KMEANS_PLUS_PLUS,
  'preserve_sparsity': True
}

sparsity_clustered_model = cluster_weights(distiller.student, **clustering_params)
```

```
sparsity_clustered_model.compile(optimizer='adam',
              loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
              metrics=['accuracy'])

print('Train sparsity preserving clustering model:')
sparsity_clustered_model.fit(train_images, train_labels,epochs=6, validation_split=0.1)
```

```
Train sparsity preserving clustering model:
Epoch 1/6
1688/1688 [==============================] - 9s 4ms/step - loss: 0.0841 - accuracy: 0.9755 - val_loss: 0.0567 - val
_accuracy: 0.9835
Epoch 2/6
1688/1688 [==============================] - 6s 4ms/step - loss: 0.0686 - accuracy: 0.9794 - val_loss: 0.0638 - val
_accuracy: 0.9813
Epoch 3/6
1688/1688 [==============================] - 6s 4ms/step - loss: 0.0630 - accuracy: 0.9798 - val_loss: 0.0603 - val
_accuracy: 0.9843
Epoch 4/6
1688/1688 [==============================] - 6s 4ms/step - loss: 0.0624 - accuracy: 0.9800 - val_loss: 0.0655 - val
_accuracy: 0.9813
Epoch 5/6
1688/1688 [==============================] - 6s 4ms/step - loss: 0.0589 - accuracy: 0.9815 - val_loss: 0.0634 - val
_accuracy: 0.9832
Epoch 6/6
1688/1688 [==============================] - 6s 4ms/step - loss: 0.0528 - accuracy: 0.9832 - val_loss: 0.0730 - val
_accuracy: 0.9780
```

Out[21]:
```
<tf_keras.src.callbacks.History at 0x2c7e6c410>
```

In [22]:
```
stripped_clustered_model = tfmot.clustering.keras.strip_clustering(sparsity_clustered_model)

print("Model sparsity:\n")
print_model_weights_sparsity(stripped_clustered_model)

print("\nModel clusters:\n")
print_model_weight_clusters(stripped_clustered_model)
```

```
Model sparsity:

kernel:0: 41.67% sparsity  (45/108)
kernel:0: 62.65% sparsity  (12706/20280)

Model clusters:

conv2d_1/kernel:0: 8 clusters
dense_1/kernel:0: 8 clusters
```

In [23]:
```
stripped_clustered_model.compile(optimizer=opt,
              loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
              metrics=['accuracy'])
```

In [24]:
```
_, stripped_clustered_model_accuracy = stripped_clustered_model.evaluate(
    test_images, test_labels, verbose=0)

print('Clustered Model test accuracy:', stripped_clustered_model_accuracy)

sz = get_gzipped_model_size(stripped_clustered_model)
print("Clustered model size: ",  sz , ' KB' )

model_acc.append(stripped_clustered_model_accuracy)
model_sz.append(sz)
```

```
Clustered Model test accuracy: 0.9782000184059143
Clustered model size:  165.533  KB
```

---------- Checkpoint Point 4 ---------

Quantization - 5.0

In [25]:
```
quant_aware_annotate_model = tfmot.quantization.keras.quantize_annotate_model(
              stripped_clustered_model)
quant_model = tfmot.quantization.keras.quantize_apply(
              quant_aware_annotate_model,
              tfmot.experimental.combine.Default8BitClusterPreserveQuantizeScheme(preserve_sparsity=True))

quant_model.compile(optimizer='adam',
              loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
              metrics=['accuracy'])
print('Training after quantization model:')
quant_model.fit(train_images, train_labels, batch_size=128, epochs=3, validation_split=0.1)
```

```
Training after quantization model:
Epoch 1/3
WARNING:tensorflow:Gradients do not exist for variables ['conv2d_1/kernel:0', 'dense_1/kernel:0'] when minimizing t
he loss. If you're using `model.compile()`, did you forget to provide a `loss` argument?

WARNING:tensorflow:Gradients do not exist for variables ['conv2d_1/kernel:0', 'dense_1/kernel:0'] when minimizing t
he loss. If you're using `model.compile()`, did you forget to provide a `loss` argument?

WARNING:tensorflow:Gradients do not exist for variables ['conv2d_1/kernel:0', 'dense_1/kernel:0'] when minimizing t
he loss. If you're using `model.compile()`, did you forget to provide a `loss` argument?

WARNING:tensorflow:Gradients do not exist for variables ['conv2d_1/kernel:0', 'dense_1/kernel:0'] when minimizing t
he loss. If you're using `model.compile()`, did you forget to provide a `loss` argument?

422/422 [==============================] - 4s 7ms/step - loss: 0.0423 - accuracy: 0.9870 - val_loss: 0.0619 - val_a
ccuracy: 0.9817
Epoch 2/3
422/422 [==============================] - 3s 7ms/step - loss: 0.0389 - accuracy: 0.9879 - val_loss: 0.0605 - val_a
```

```
ccuracy: 0.9832
Epoch 3/3
422/422 [==============================] - 3s 7ms/step - loss: 0.0379 - accuracy: 0.9882 - val_loss: 0.0589 - val_a
ccuracy: 0.9825
```

Out[25]:
```
<tf_keras.src.callbacks.History at 0x2c840a890>
```

In [26]:
```
print("Final Model clusters:")
print_model_weight_clusters(quant_model)
print("\nFinal Model sparsity:")
print_model_weights_sparsity(quant_model)
```

```
Final Model clusters:
quant_conv2d_1/conv2d_1/kernel:0: 8 clusters
quant_dense_1/dense_1/kernel:0: 8 clusters

Final Model sparsity:
conv2d_1/kernel:0: 42.59% sparsity  (46/108)
dense_1/kernel:0: 63.03% sparsity  (12782/20280)
```

In [27]:
```
converter = tf.lite.TFLiteConverter.from_keras_model(quant_model)
converter.optimizations = [tf.lite.Optimize.DEFAULT]
final_tflite_model = converter.convert()
final_model_file = 'final_model.tflite'
# Save the model.
with open(final_model_file, 'wb') as f:
    f.write(final_tflite_model)


sz = get_gzipped_model_size2(final_model_file)
print("Final model size: ", sz, ' KB')

model_sz.append(sz)
```

```
INFO:tensorflow:Assets written to: /var/folders/px/z8lb6znd6q95tq6vlznyb0s40000gn/T/tmpn8r4jvz3/assets
```

```
INFO:tensorflow:Assets written to: /var/folders/px/z8lb6znd6q95tq6vlznyb0s40000gn/T/tmpn8r4jvz3/assets
```

```
Final model size:  6.928  KB
```

```
/Users/ajaymaheshwari/anaconda3/lib/python3.11/site-packages/tensorflow/lite/python/convert.py:964: UserWarning: St
atistics for quantized inputs were expected, but not specified; continuing anyway.
  warnings.warn(
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
W0000 00:00:1716870668.282998   24405 tf_tfl_flatbuffer_helpers.cc:390] Ignored output_format.
W0000 00:00:1716870668.283500   24405 tf_tfl_flatbuffer_helpers.cc:393] Ignored drop_control_dependency.
2024-05-28 10:01:08.285833: I tensorflow/cc/saved_model/reader.cc:83] Reading SavedModel from: /var/folders/px/z8lb
6znd6q95tq6vlznyb0s40000gn/T/tmpn8r4jvz3
2024-05-28 10:01:08.287096: I tensorflow/cc/saved_model/reader.cc:51] Reading meta graph with tags { serve }
2024-05-28 10:01:08.287101: I tensorflow/cc/saved_model/reader.cc:146] Reading SavedModel debug info (if present) f
rom: /var/folders/px/z8lb6znd6q95tq6vlznyb0s40000gn/T/tmpn8r4jvz3
2024-05-28 10:01:08.297178: I tensorflow/compiler/mlir/mlir_graph_optimization_pass.cc:388] MLIR V1 optimization pa
ss is not enabled
2024-05-28 10:01:08.298196: I tensorflow/cc/saved_model/loader.cc:234] Restoring SavedModel bundle.
2024-05-28 10:01:08.334799: I tensorflow/cc/saved_model/loader.cc:218] Running initialization op on SavedModel bund
le at path: /var/folders/px/z8lb6znd6q95tq6vlznyb0s40000gn/T/tmpn8r4jvz3
2024-05-28 10:01:08.343882: I tensorflow/cc/saved_model/loader.cc:317] SavedModel load for tags { serve }; Status:
success: OK. Took 58052 microseconds.
2024-05-28 10:01:08.392379: I tensorflow/compiler/mlir/tensorflow/utils/dump_mlir_util.cc:268] disabling MLIR crash
reproducer, set env var `MLIR_CRASH_REPRODUCER_DIRECTORY` to enable.
```

In [28]:
```
interpreter = tf.lite.Interpreter(final_model_file)
interpreter.allocate_tensors()

final_test_accuracy = eval_model(interpreter)

print('Final test accuracy:', final_test_accuracy)

model_acc.append(final_test_accuracy)
```

```
INFO: Created TensorFlow Lite XNNPACK delegate for CPU.
WARNING: Attempting to use a delegate that only supports static-sized tensors with a graph that has dynamic-sized t
ensors (tensor#12 is a dynamic-sized tensor).
```

```
Final test accuracy: 0.9825
```

---------- Checkpoint Point 5 ---------

In [29]:
```
for i in range(len(model_acc)):
    print(f"Accuracy = { round(model_acc[i]*100,2)} with size = {model_sz[i]} KB ")
```

```
Accuracy = 98.27 with size = 235.287 KB
Accuracy = 98.08 with size = 203.828 KB
Accuracy = 97.94 with size = 227.846 KB
Accuracy = 97.82 with size = 165.533 KB
Accuracy = 98.25 with size = 6.928 KB
```

------- Final Comparison Summary -------