There are three main categories of model compression techniques.

The first technique is pruning. With pruning, unimportant parameters are removed from the model based upon an importance function. Pruning can be done globally or locally, and with different granularities.

The second technique is knowledge distillation. With knowledge distillation, a teacher model is used to help train a student model. Typically, the teacher model will be larger than the student model. With some knowledge distillation schemes, the teacher model does not need to be pre-trained and can learning with the student.

The third technique is quantization. With quantization, less bits are used to store the model parameters, thus reducing the size of the model. However, using less bits results in a loss in precision, which can affect the model’s accuracy. Some techniques such as quantization-aware training (QAT) can be used to help address this.