**Why is this an important area of study? [0.5 points]**

This area of study involves reducing the model size (fewer parameters), which can significantly reduce the computing requirements.

**Describe two different techniques/approaches discussed [0.5 points]**

1. **Pruning**: Given an existing model, we can remove some of the model’s parameters without losing too much performance. To optimize the model for compressibility, we can use Sharpness-Aware Minimization
2. Knowledge Distillation: Instead of removing parts from existing models, we train a new but smaller model using the existing model as the teacher.

**Discuss relative strengths [0.5 points] and weaknesses [0.5 points] of the two techniques described above. [1 point total]**

* Pruning can substantially reduce the number of parameters, but it requires hardware support for sparse operations (not calculating part of a neural network layer).
* Knowledge distillation enables us to perform adaptive computation: use the smaller model for easier problems and full model for more difficult ones, and the parameters of two models can be shared. However, this requires additional control flow logic.