Hands-on: Model Compression

Practice: Model Compression Pipeline

Knowledge Distillation and Quantization Practice using PyTorch

1. Knowledge Distillation Code

```
# Teacher-Student distillation loss
def distillation_loss(student_logits, teacher_logits, labels, T=3):
    # Soft targets
    soft_loss = nn.KLDivLoss()(F.log_softmax(student_logits/T, dim=1),
                                F.softmax(teacher_logits/T, dim=1)) * T*T
    # Hard targets
    hard_loss = nn.CrossEntropyLoss()(student_logits, labels)
    return 0.7 * soft_loss + 0.3 * hard_loss
```

2. INT8 Quantization Code

```
import torch.quantization

# Static quantization
model.qconfig = torch.quantization.get_default_qconfig('qnnpack')
model_prepared = torch.quantization.prepare(model)
# Calibration
model_prepared(calibration_data)
# Apply quantization
model_quantized = torch.quantization.convert(model_prepared)
```

Useful Tools and Libraries

PyTorch

torch.quantization
torch.nn.utils.prune

TensorFlow

TF-MOT (Model Optimization)
Quantization-Aware Training

ONNX

onnxruntime
Cross-framework

Hugging Face

Optimum
Transformer Optimization

Practice Assignment:

Compress ResNet model with CIFAR-10 dataset
(Implement distillation + quantization pipeline)