Hands-on: Model Compression

실습: 모델 압축 파이프라인

PyTorch를 이용한 Knowledge Distillation 및 Quantization 실습

1. Knowledge Distillation 코드

```
# Teacher-Student 증류 손실
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 코드

```
import torch.quantization

# 정적 양자화
model.qconfig = torch.quantization.get_default_qconfig('qnnpack')
model_prepared = torch.quantization.prepare(model)
# 캘리브레이션
model_prepared(calibration_data)
# 양자화 적용
model_quantized = torch.quantization.convert(model_prepared)
```

유용한 도구 및 라이브러리

PyTorch

torch.quantization
torch.nn.utils.prune

TensorFlow

TF-MOT (Model Optimization)
Quantization-Aware Training

ONNX

onnxruntime
Cross-framework

Hugging Face

Optimum
Transformer 최적화

실습 과제:

CIFAR-10 데이터셋으로 ResNet 모델 압축해보기
(증류 + 양자화 파이프라인 구현)