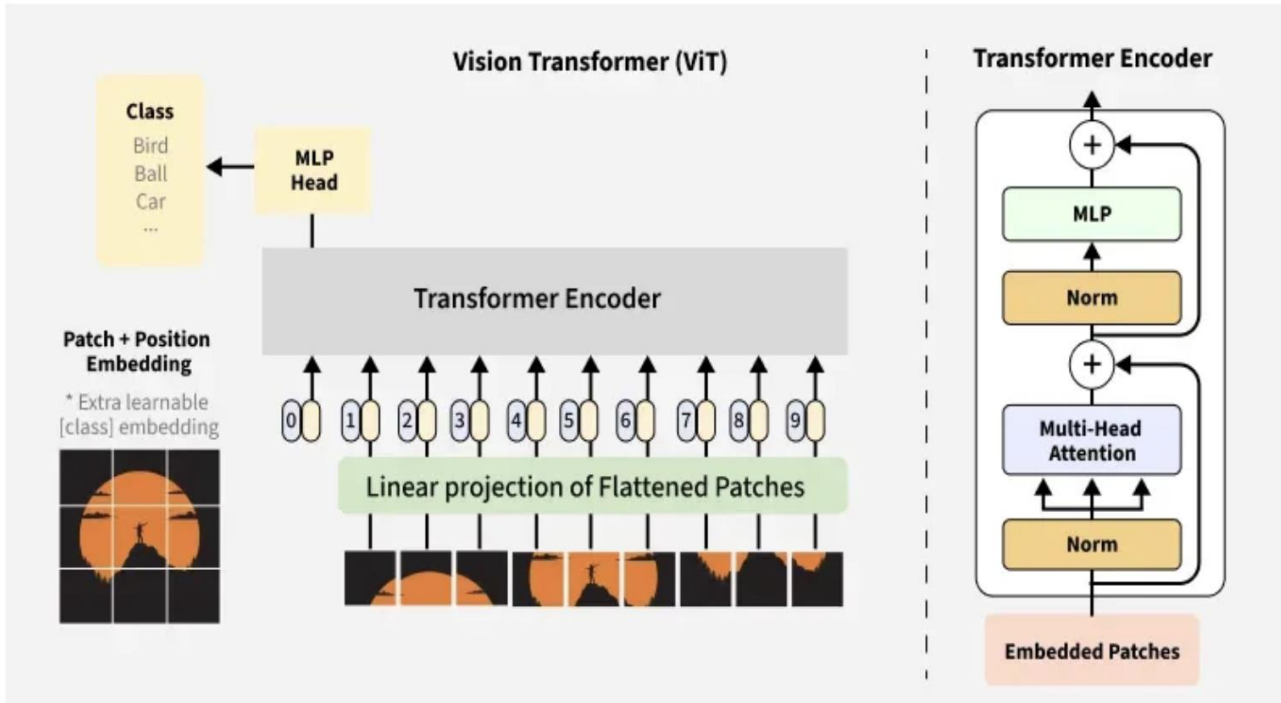



ViT quantization



- ViTForImageClassification
- Fine-tuning for 1 epoch
- Evaluate_model ()
- Measure memory ()
 - ✓ Quantize_vit_like_triton ()
 - ✓ torch.quantization.quantize_dynamic
 - ✓ torchao.quantization.quantize_



	original	pytorch_quant	torchao_quant	our_quant
Accuracy (CIFAR10)	98.08%			98.06%
memory	327.33 MB	2.98 MB	327.33 MB	30. 18 MB

=== RESULTS ===
 Accuracy drop: 0.02%
 Memory reduction: 297.14 MB (90.8%)