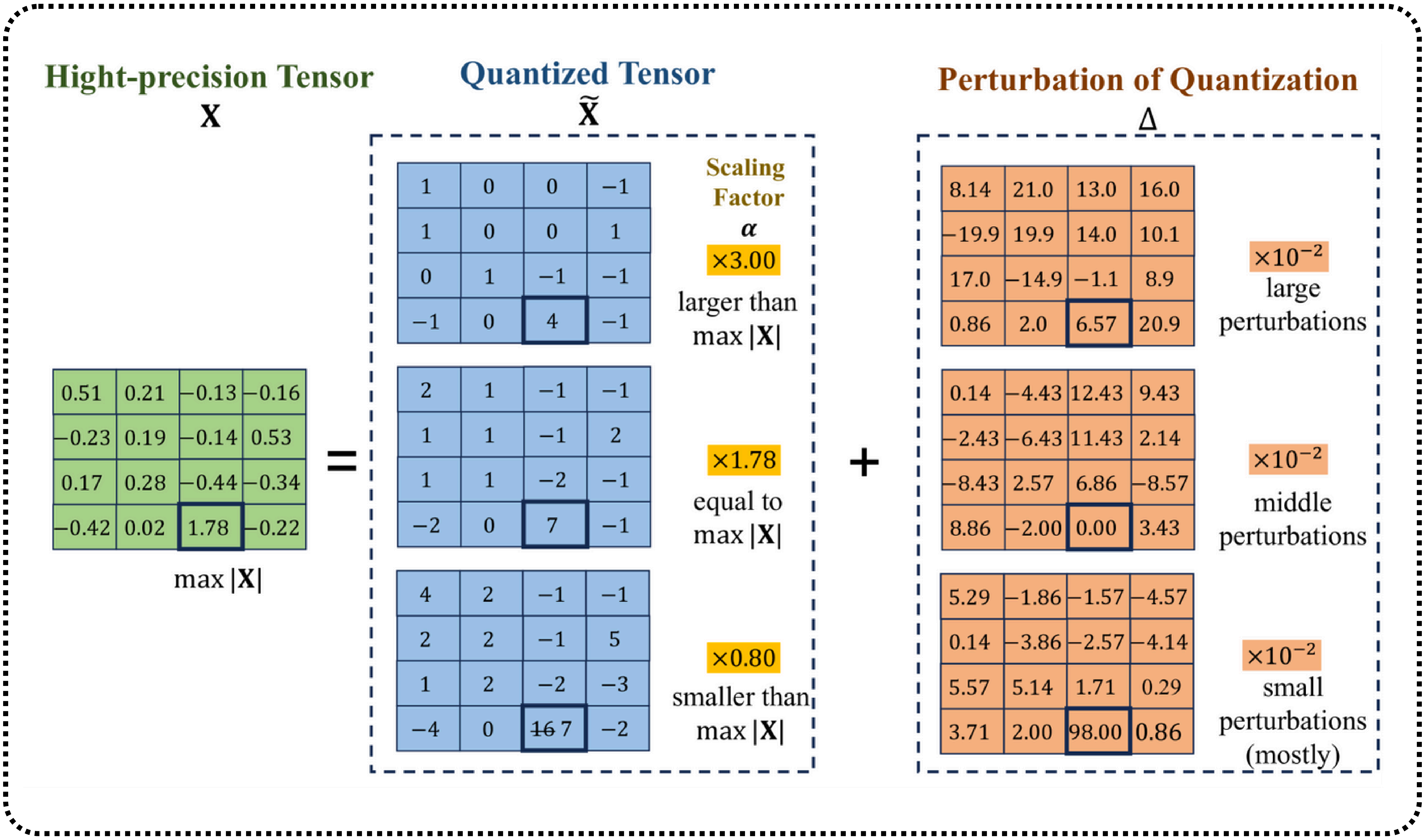
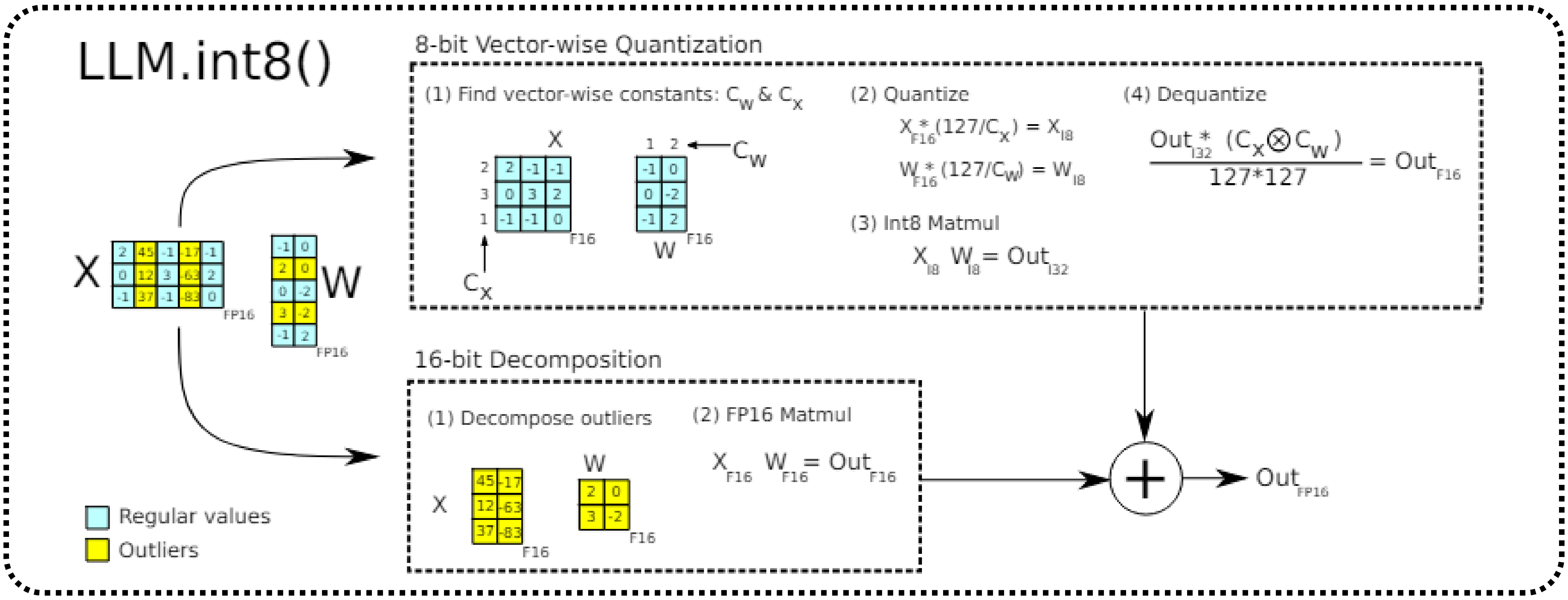
Mastering LLMs

Analytics Vidhya

Day 31: Quantization 101

Making LLMs Lighter

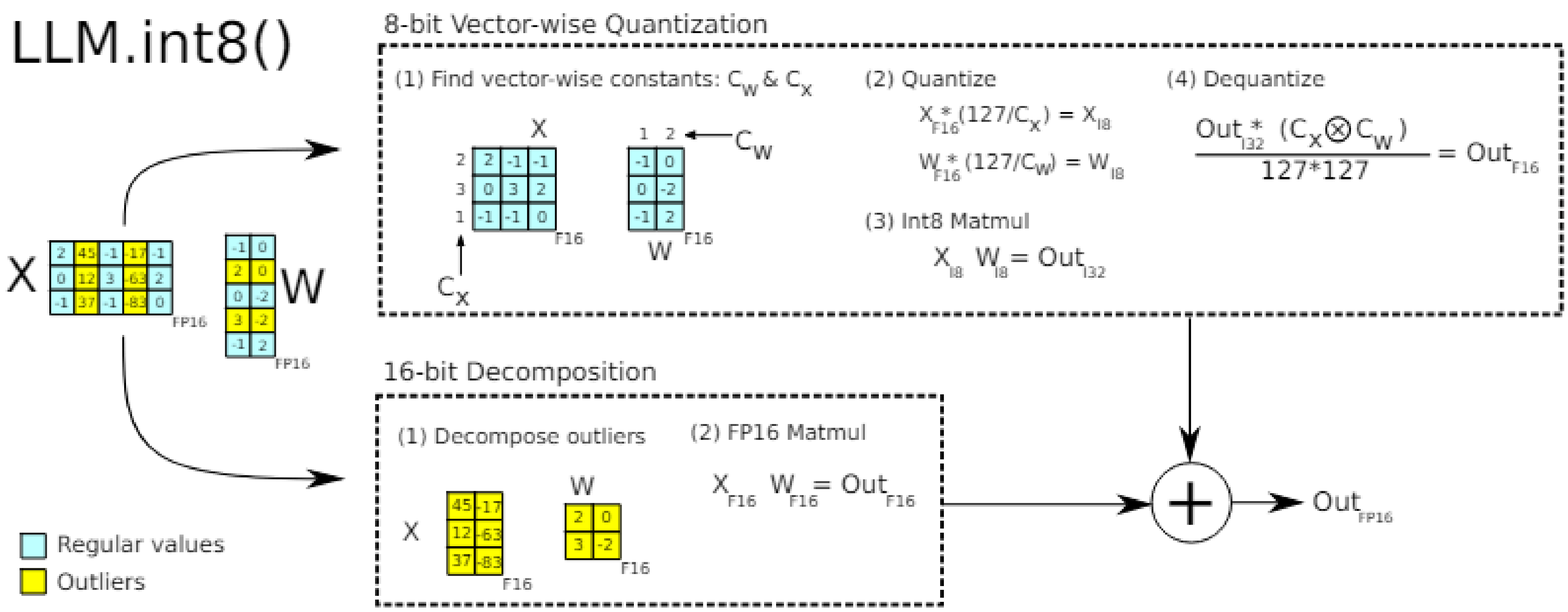




- Large Language Models (LLMs) are powerful but resource-hungry. Running them efficiently, especially on consumer GPUs, requires a game-changer: Quantization.

What is Quantization?

- Quantization is a technique that reduces a model’s memory footprint and computational requirements by representing weights and activations with lower precision (e.g., 4-bit or 8-bit instead of 16-bit or 32-bit).



Why Quantization?

- Reduced Memory Usage – Load larger models on smaller hardware

- Faster Inference – Less computation = quicker responses

- Lower Power Consumption – Crucial for edge AI & mobile deployment

- Cheaper Deployment – Enables cost-efficient scaling

Types of Quantization for LLMs

Post-Training Quantization (PTQ)

- Convert a trained FP16/FP32 model to a lower precision format
- Fast & easy, but may lead to slight accuracy loss

Quantization-Aware Training (QAT)

- Train the model while simulating quantization effects
- Better accuracy but requires retraining

Popular Quantization Methods

- **GPTQ** (Generalized Post-Training Quantization) – Optimized for minimal performance loss
- **AWQ** (Activation-aware Weight Quantization) – Preserves accuracy in low-bit models
- **Bitsandbytes** (8-bit / 4-bit Quantization) – Hugging Face integration for easy model loading

How to Use Quantization in LLMs?

- Run models with 4-bit quantization using bitsandbytes

```
from transformers import AutoModelForCausalLM, AutoTokenizer
import torch
import bitsandbytes as bnb

model_name = "mistralai/Mistral-7B-v0.1"

model = AutoModelForCausalLM.from_pretrained(
    model_name,
    device_map="auto",
    load_in_4bit=True,  # Enables 4-bit quantization
    bnb_4bit_compute_dtype=torch.float16
)

tokenizer = AutoTokenizer.from_pretrained(model_name)
```

Speed up inference with GPTQ quantized models

```
from auto_gptq import AutoGPTQForCausalLM

quantized_model =
AutoGPTQForCausalLM.from_quantized("TheBloke/Llama-2-
13B-GPTQ")
```

Does Quantization Affect Accuracy?

- Yes, but modern quantization methods (GPTQ, AWQ) minimize loss while offering huge efficiency gains.

Who Benefits from Quantization?

✅ Researchers needing faster experiments
✅ Startups with limited cloud budgets
✅ Developers running LLMs locally on consumer GPUs
✅ AI on mobile & edge devices

Analytics Vidhya

Stay Tuned for **Day 32** of

Mastering LLMs