OxML2024: LLM & Diffusion Model

Wenhan Han

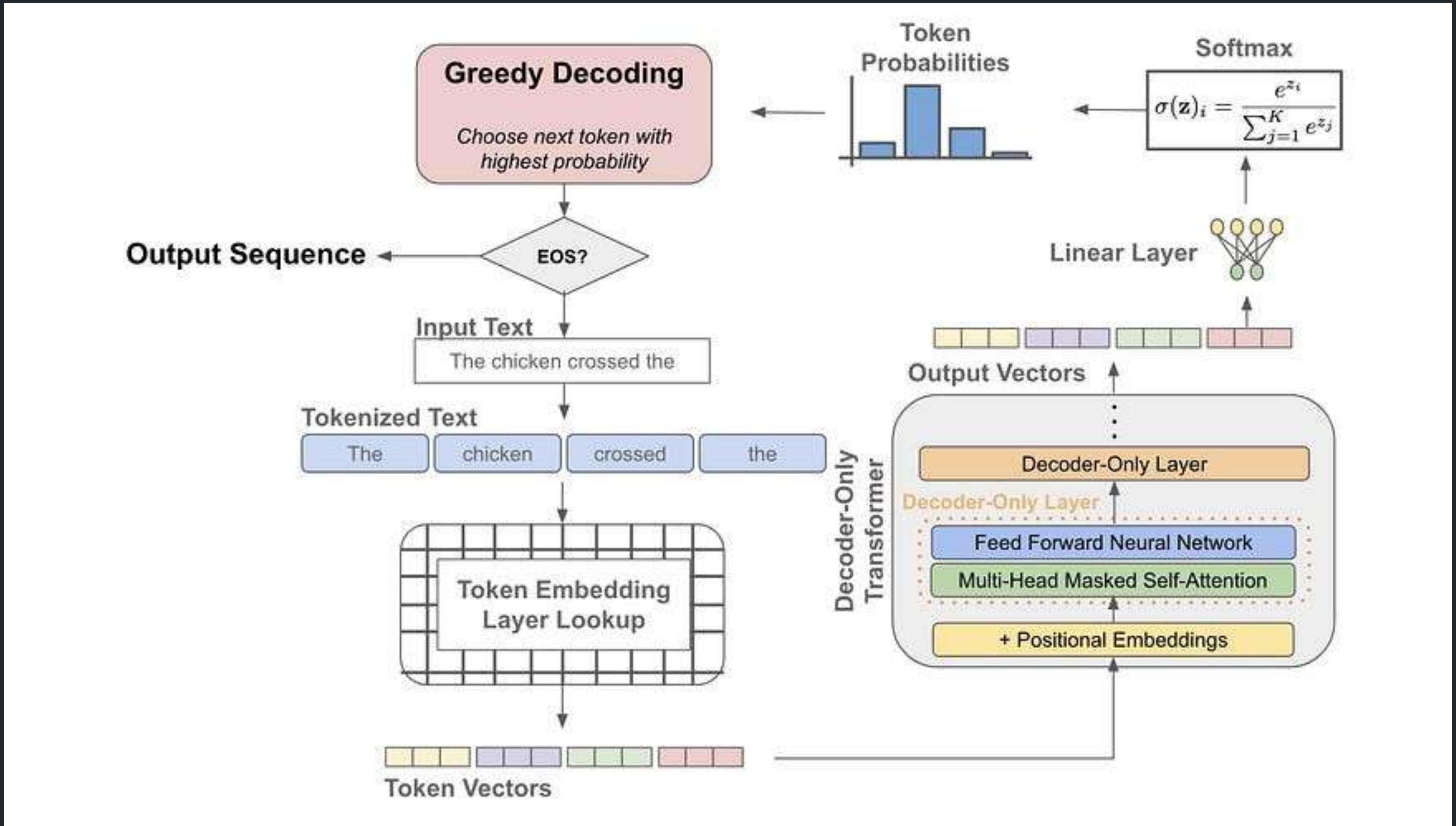
TU Eindhoven

Overview

Colab Notebook: LLM Link Diffusion Link

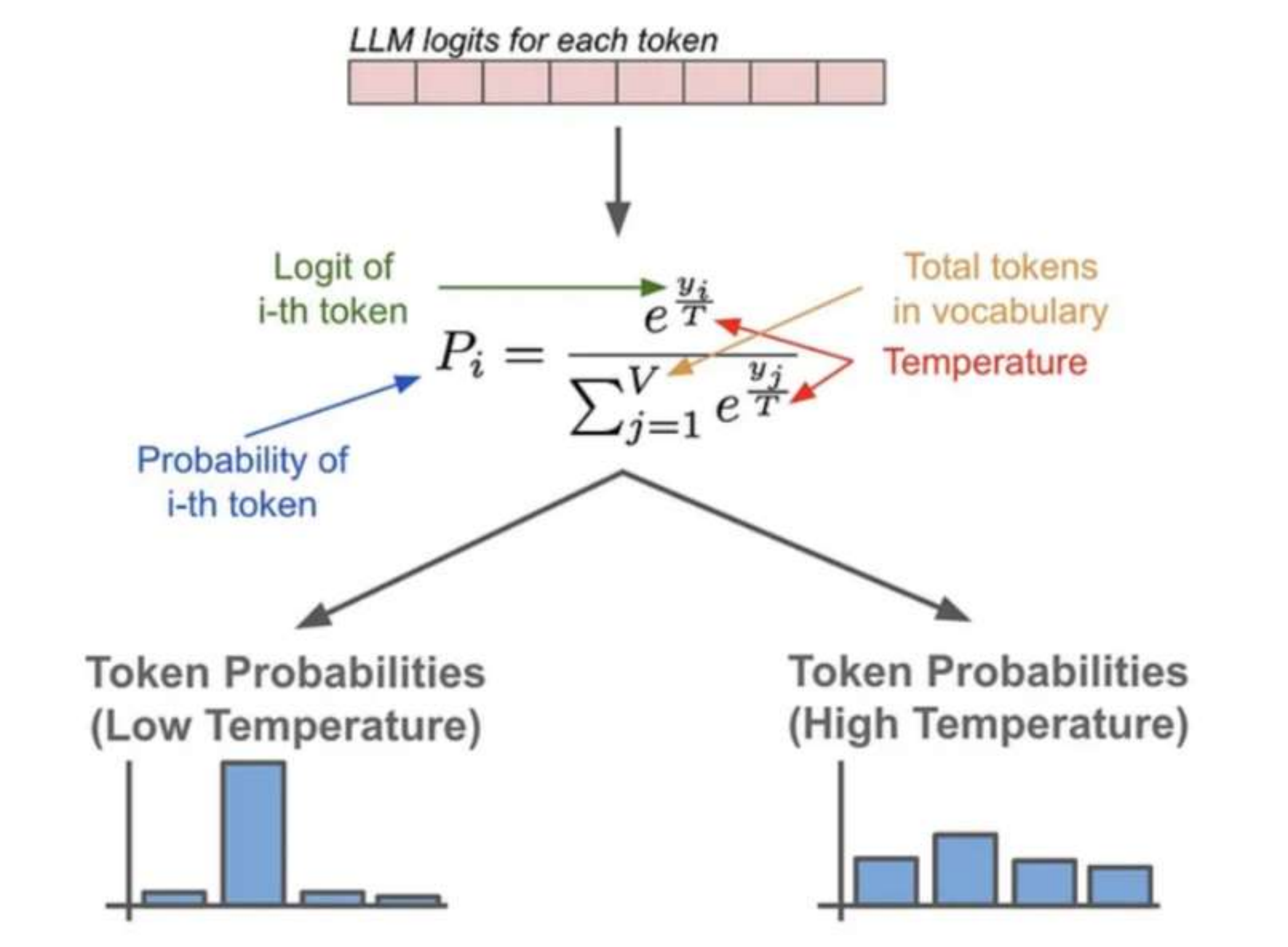
- Access to LLMs through APIs.
- Run LLM inference in local.
- Prompting techniques.
- Finetune LLaMA-3 on single GPU.
- Run Stable Diffusion.
- Train a mini Diffusion Model.

Recall: LLM Inference (Decoder Only)



Recall: LLM Inference (Decoder Only)

Sampling Parameter: Temperature

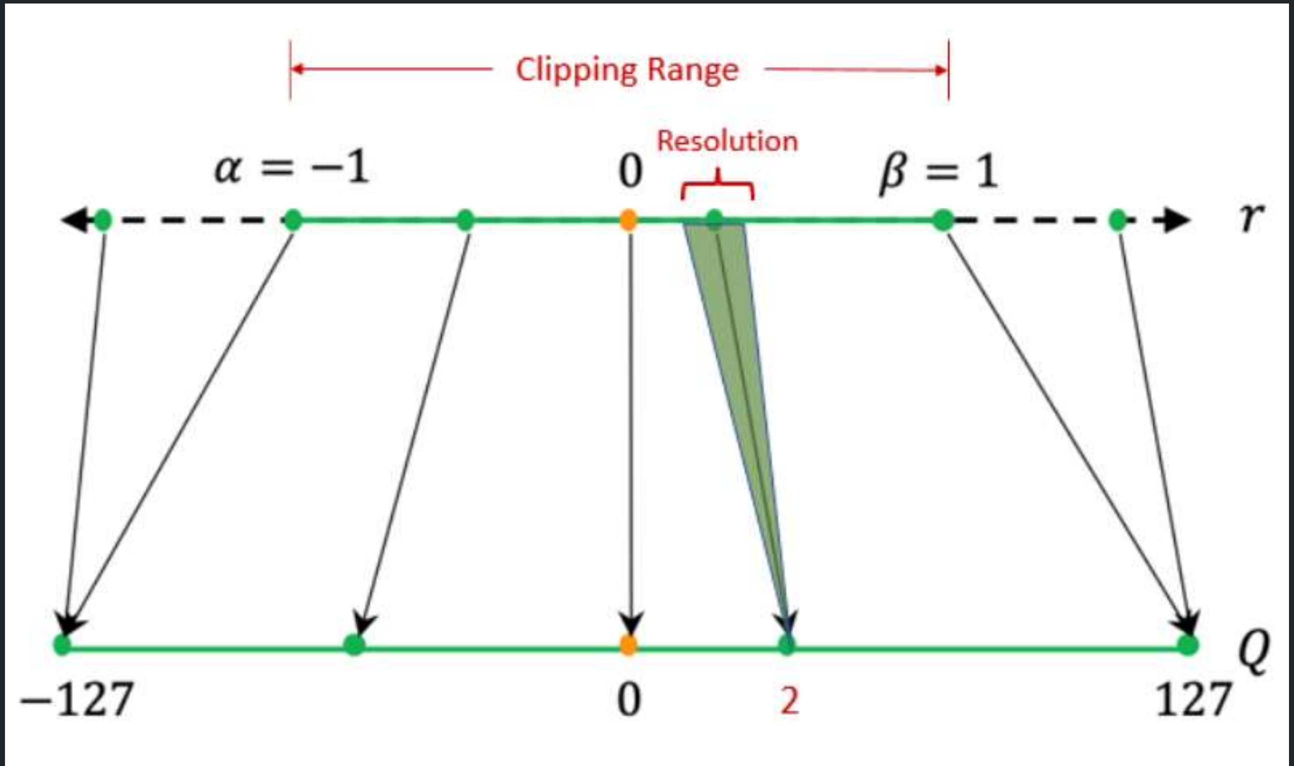


In the context of a language model, the probability of the next token (word or sub-word unit) is given by a distribution P, which is a function of the model's internal parameters and the input context.

High Temperature: The scaling effectively "flattens" the distribution, making less probable tokens more likely to be sampled.

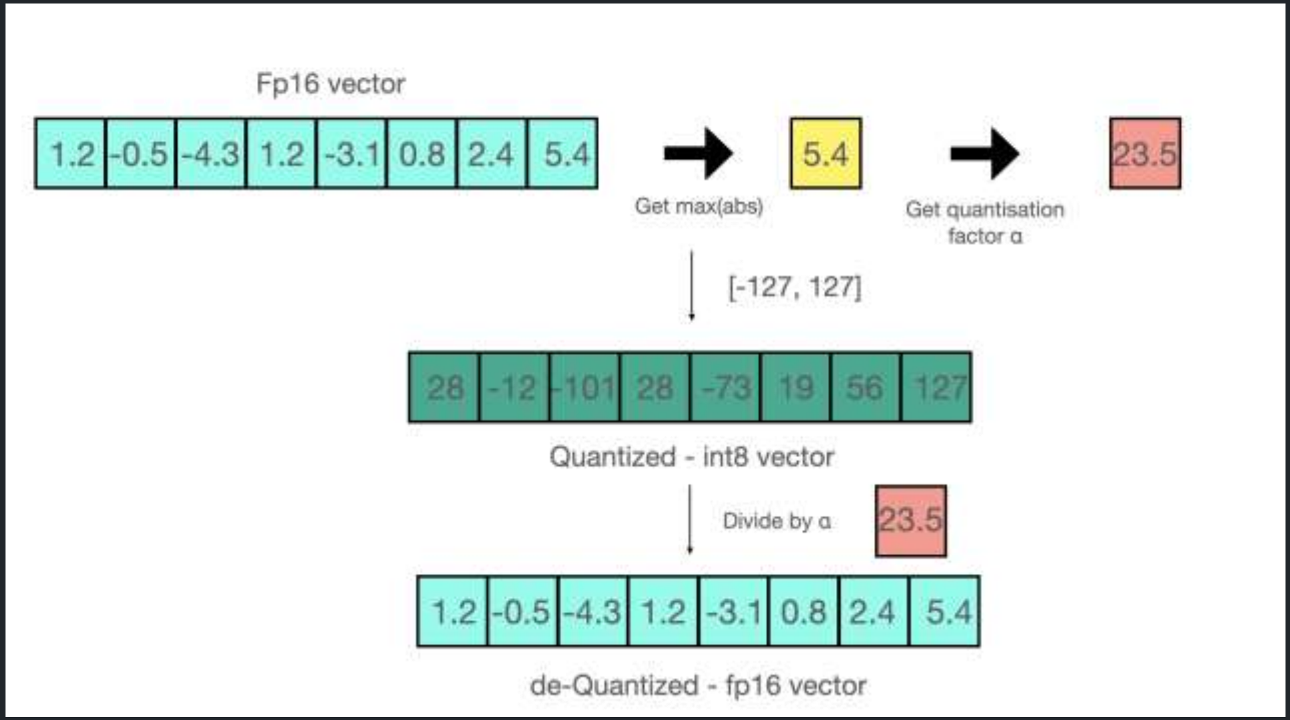
Low Temperature: The distribution becomes "sharper," concentrating the probability mass on the most likely tokens.

Quantization

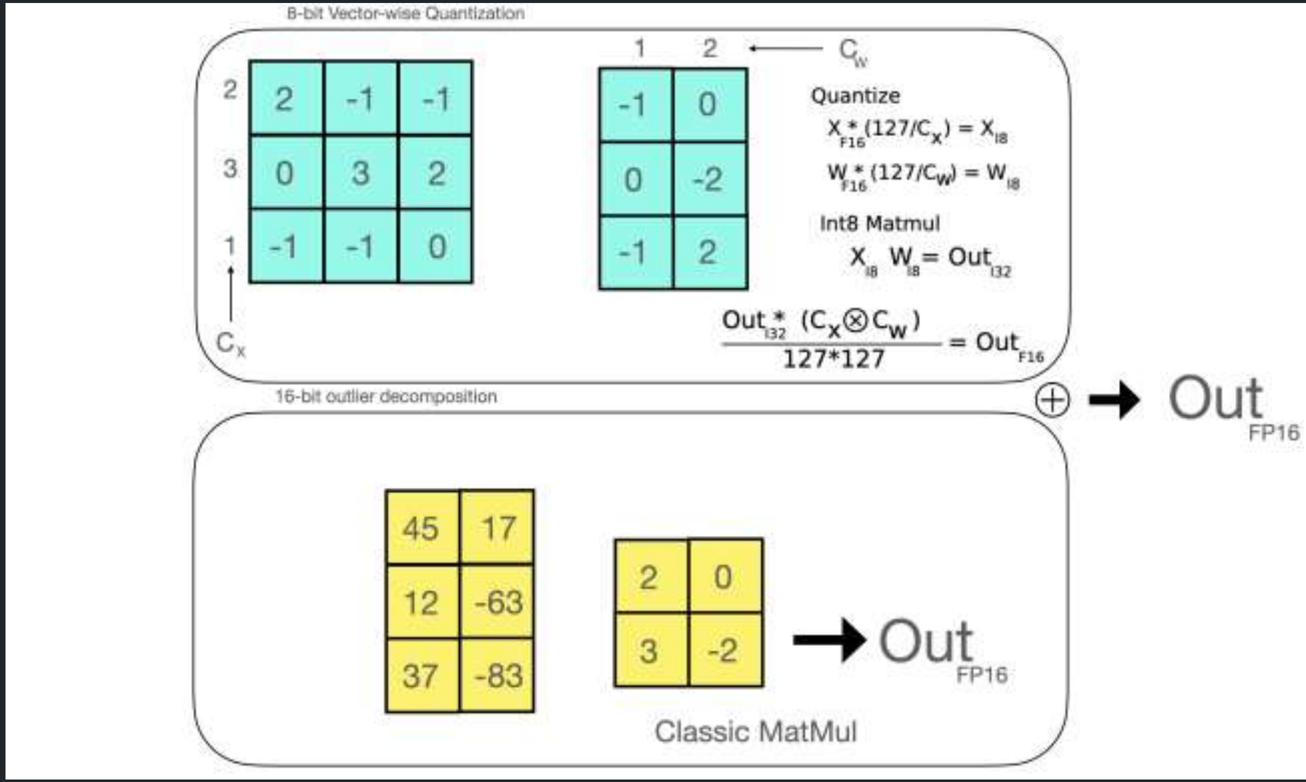


The two most common 8-bit quantization techniques are zero-point quantization and absolute maximum (absmax) quantization. Zero-point quantization and absmax quantization map the floating point values into more compact int8 (1 byte) values.

Quantization

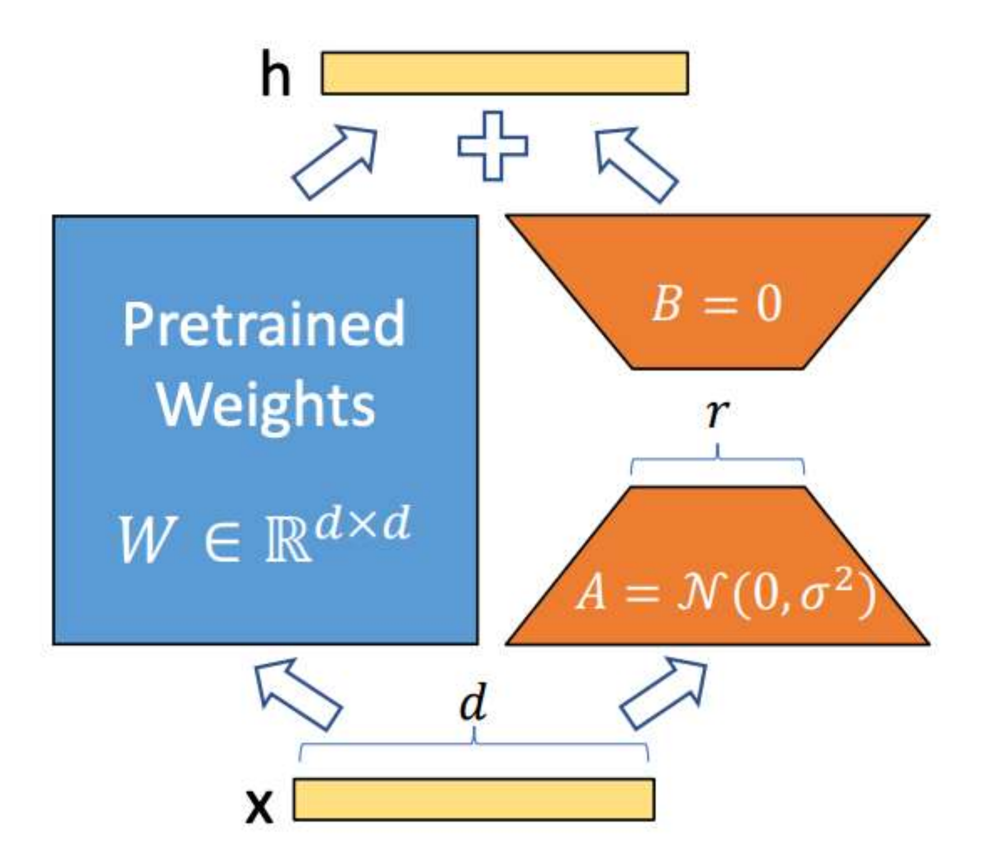


Quantization
Handle Outliers



- From the input hidden states, extract the outliers (i.e. values that are larger than a certain threshold) by column.
- Perform the matrix multiplication of the outliers in FP16 and the non-outliers in int8.
- Dequantize the non-outlier results and add both outlier and non-outlier results together to receive the full result in FP16.

LoRA

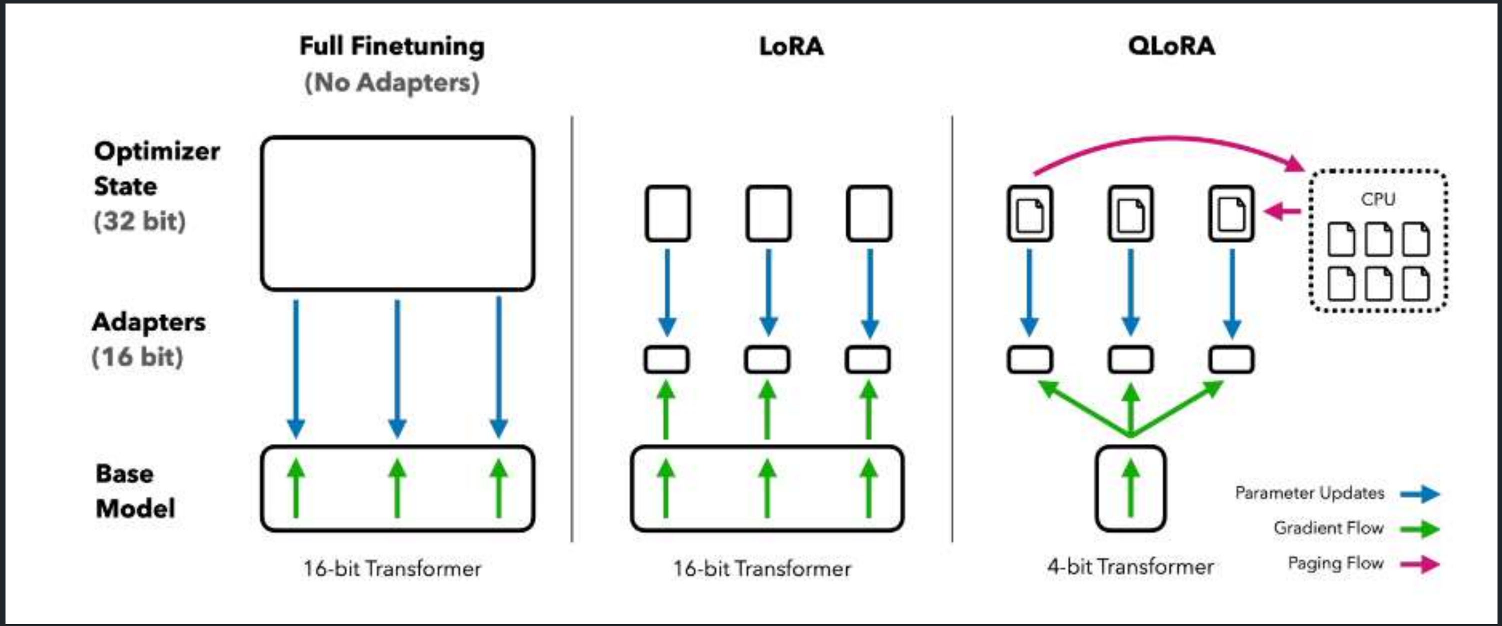


We only train A and B. The finetuned weight becomes

$$W = W_o + \Delta W = W_o + BA$$

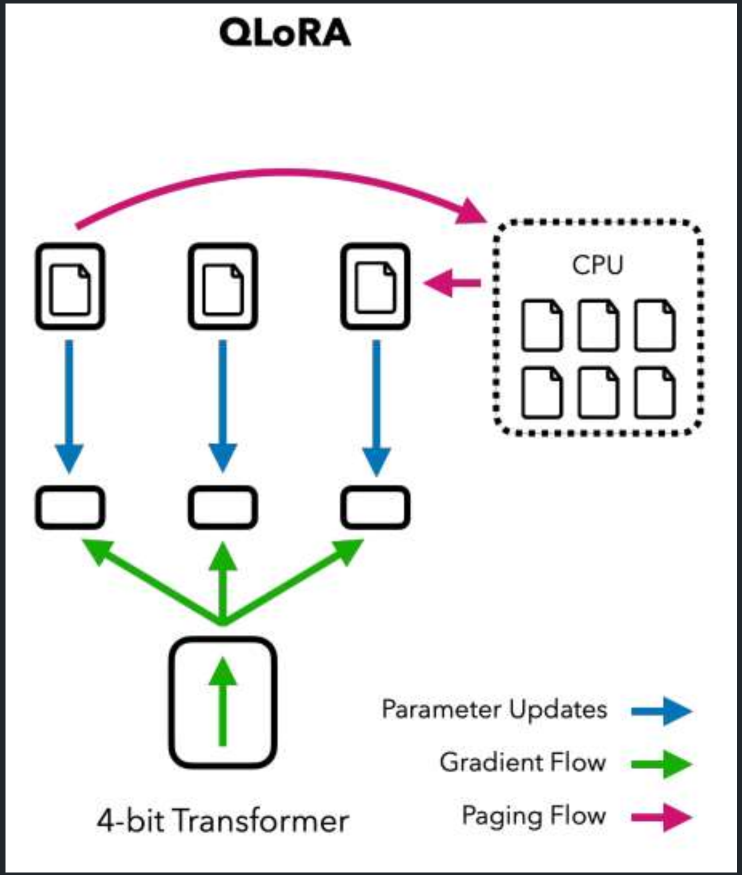
W_o denotes the pre-trained parameter weights.

QLoRA



QLoRA is the extended version of LoRA which works by quantizing the precision of the weight parameters in the pre trained LLM to 4-bit precision.

QLoRA



QLORA introduces multiple innovations designed to reduce memory use without sacrificing performance:

- 4-bit NormalFloat, an information theoretically optimal quantization data type for normally distributed data that yields better empirical results than 4-bit Integers and 4-bit Floats.
- Double Quantization, a method that quantizes the quantization constants, saving an average of about 0.37 bits per parameter (approximately 3 GB for a 65B model).
- Paged Optimizers, using NVIDIA unified memory to avoid the gradient checkpointing memory spikes that occur when processing a mini-batch with a long sequence length.

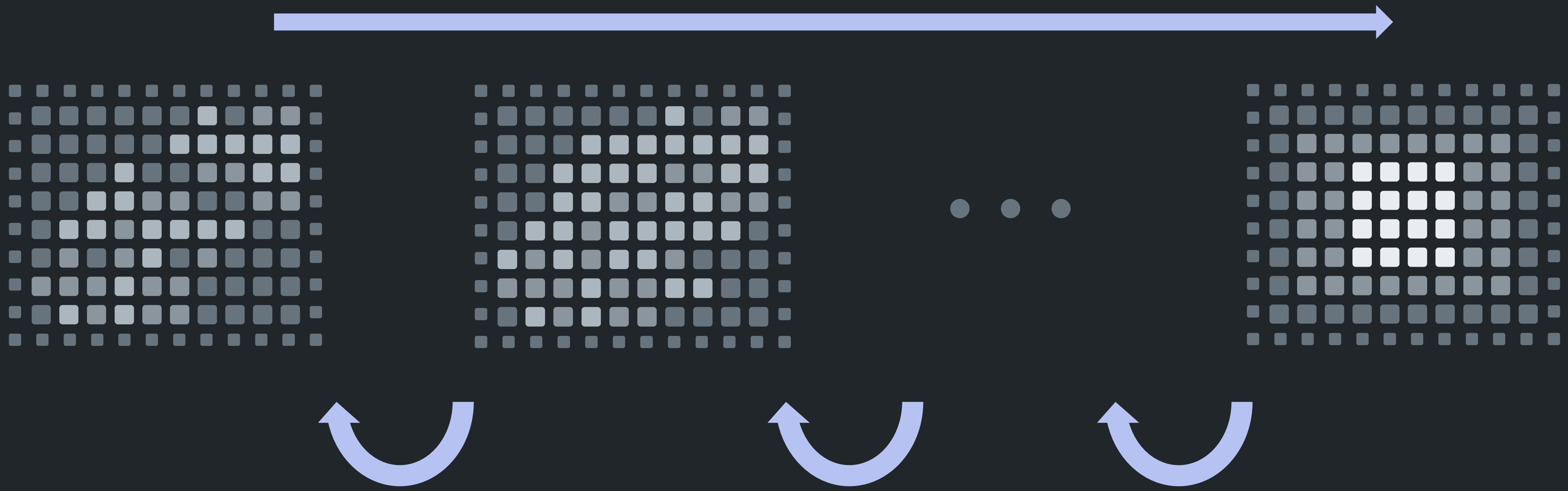
Recall: Diffusion Model

A Diffusion Model or Score-based Generative Model perturbs the input data into standard Gaussian noise and learns the reverse process. By sampling from the Gaussian distribution, the Diffisuion Model can iteratively reconstruct the samples.

Recall: Diffusion Model

To be general, the forward process follows a Stochastic Differential Equation (SDE) that has the form as

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$$

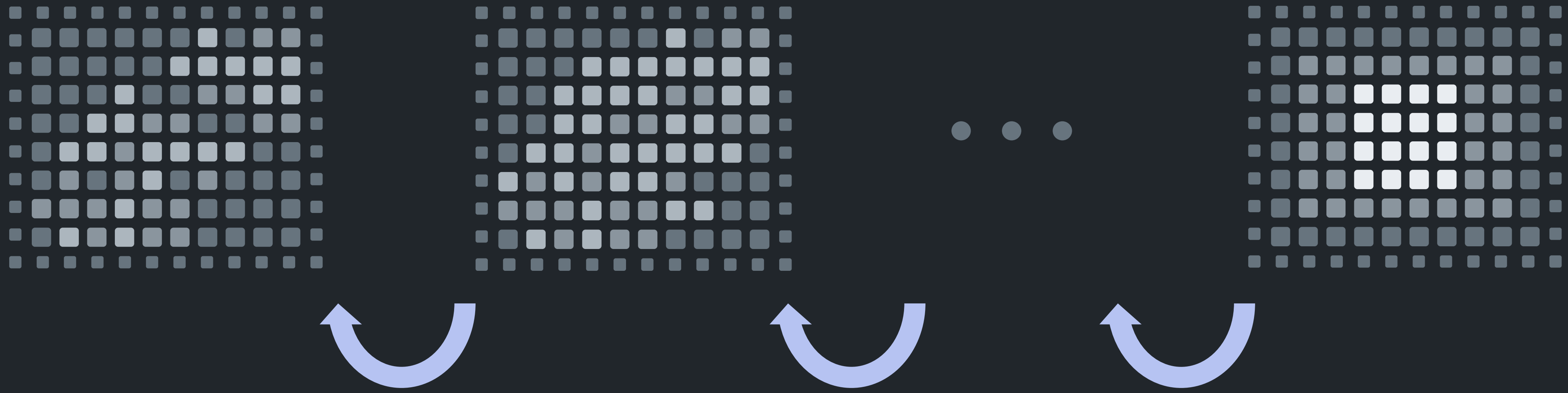


There is a corresponding reverse SDE

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log_{p_\sigma}(\mathbf{x})}]dt + g(t)d\mathbf{w}$$

Score function estimated by a score model

Recall: Diffusion Model



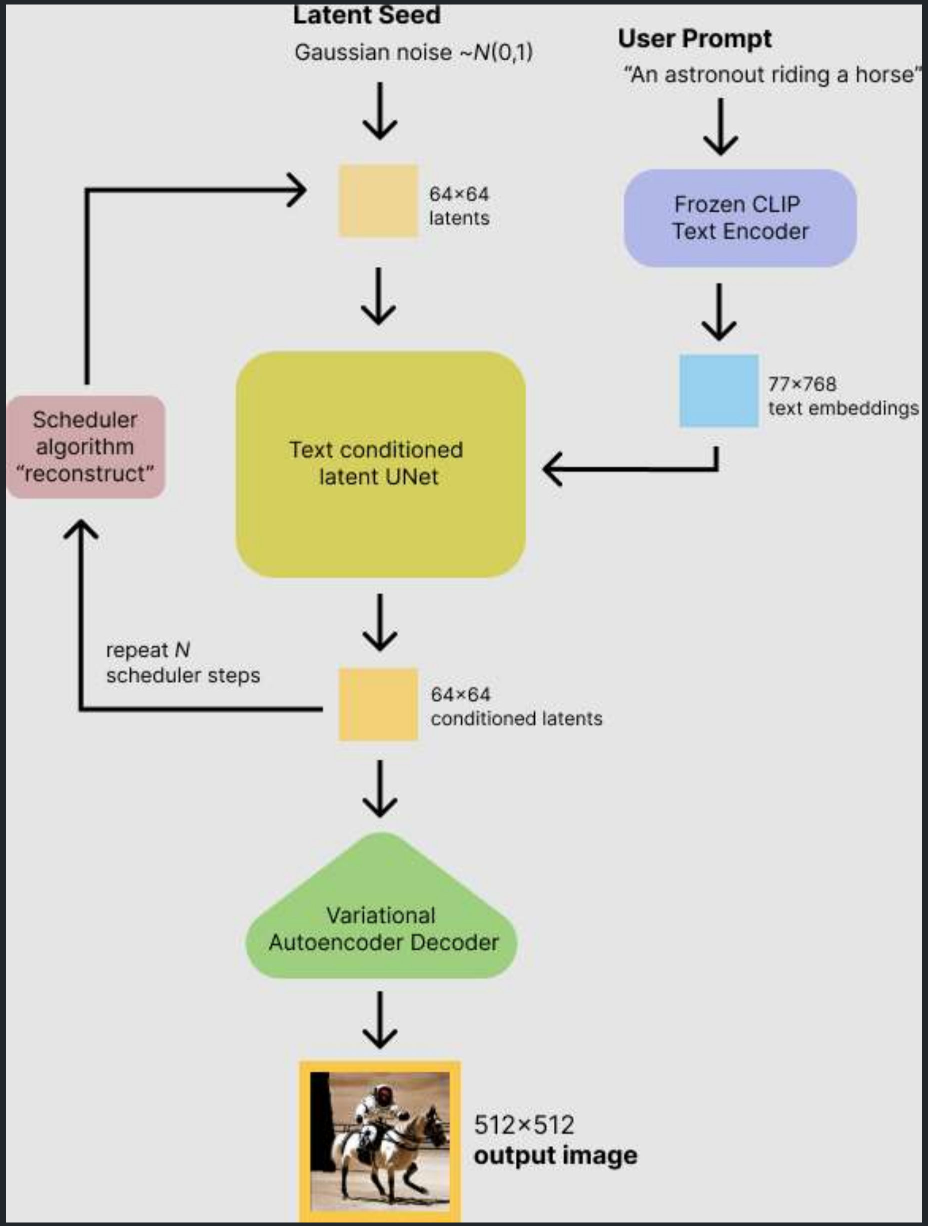
The sampling can follow a simple numerical SDE solver
Euler-Maruyama Method.

$$\Delta \mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t) s_\theta(\mathbf{x}, t)]\Delta t + g(t)\sqrt{|\Delta t|}\mathbf{z}_t$$

$$\mathbf{x} = \mathbf{x} + \Delta \mathbf{x}$$

$$t = t - \Delta t$$

Stable Diffusion



Stable Diffusion is a large text to image diffusion model trained on billions of images.

Thanks