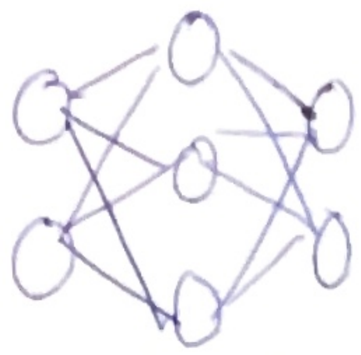
LoRA (Type of - Parameter Efficient Fine Tuning)

trains the model by adding more trainable parameters.



$x \rightarrow h(x)$ → hidden layer.

$h(x) = W_0 x$

W_0 ↓ , x ↘ inputs.

Traincable parameters in a weight matrix (original)

Now say.

$h(x) = W_0 x + \Delta W x$

$W_0 x$ ↓ keep these frozen (no retraining.)

$\Delta W x$ ↓

$\Delta W = BA$ → the no of these parameters are less than W_0 because of low rank decomposition.

↳ new trainable parameter matrix.

LOW RANK DECOMPOSITION

say W_0 is a $d \times k$ matrix

i.e has $d \times k$ trainable parameters

ΔW is introduced by LoRA where

B is a $d \times r$ matrix

A is a $r \times k$ matrix.

*The product BA constructs an ΔW matrix which has same shape as W_0 but with much fewer trainable parameters.

↑

Params in B + Params in A =

$(d \times r) + (r \times k)$

$\Rightarrow r(d+k)$

↙

r (intrinsic rank).

$\therefore r(d+k) << d \times k$.

Example ←

if $d = 1000$

$k = 1000$

using LoRA with rank $= 2$

$2(1000+1000) = 4000$

which is less than original 10^6 parameters.