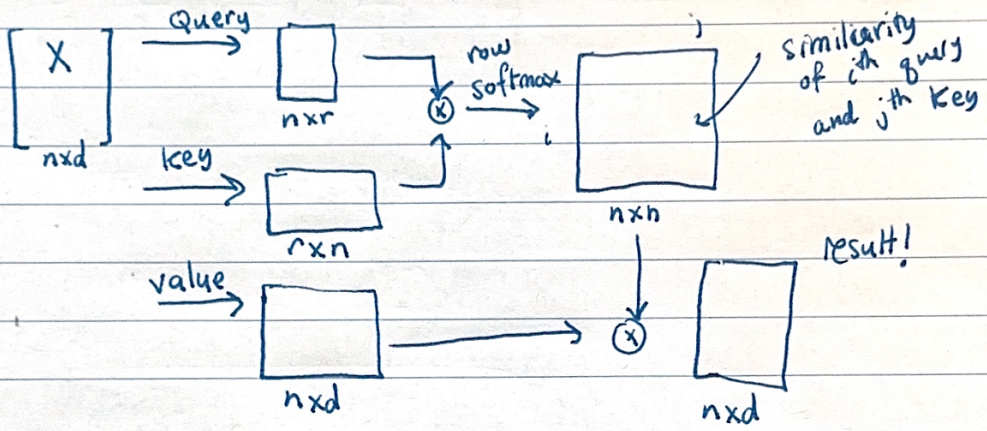
LoRA and Sparse Auto encoders

Plan
Review
LoRA
Sparse Auto encoders

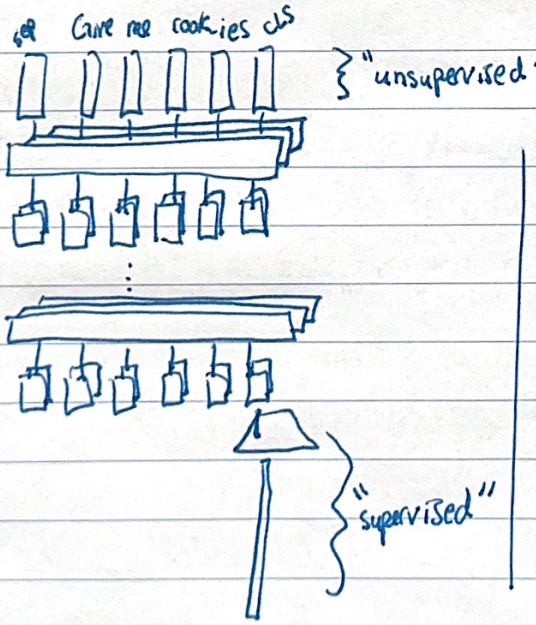
Logistics
Scribe
check in
Zoom
Project!!

Review

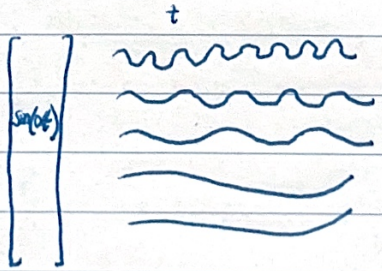


+ cross-attention

Complexity: $(n \cdot d \cdot r) + (n \cdot d \cdot d) + (n \cdot r \cdot n) + (n \cdot n \cdot d)$



1,065,066,880 vs
11:24am Tues 14 Jan, 2025



Low Rank Adaptation

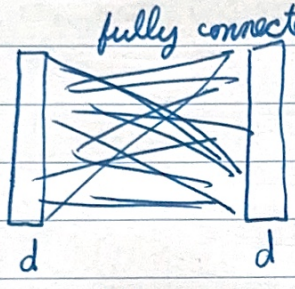
Motivation: Training Large Models = $$$

↳ Train more efficiently?

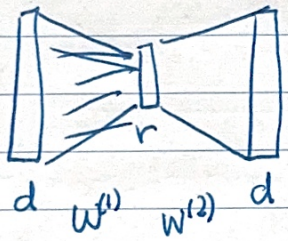
↳ Finetune on our own dataset?

Qs:

1. What takes the most time?
2. How can we speed up?



complexity: $O(d^2)$



Complexity: $O(dr)$

frozen ↙ ↙ update wrt these

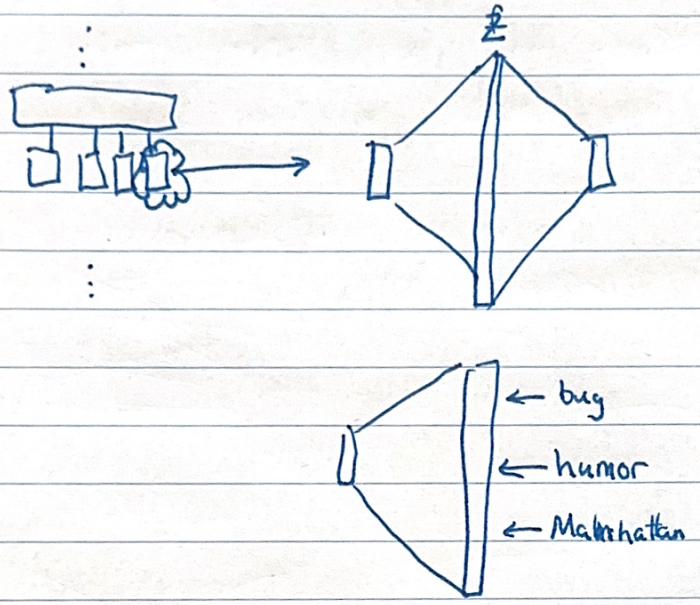
$$W = W^{orig} + W^{(1)} W^{(2)}$$

Sparse Autoencoders

Motivation: What is going in model?

↳ Attention heat maps

↳ visualize with interpretation



- trivial except sparsity penalty
  e.g. dist(z, 0)

Hard (for me) to train ∴