TO THE SHORT CONTEXT AND BEYOND: MODELLING LONG SEQUENCES WITH LONGLORA

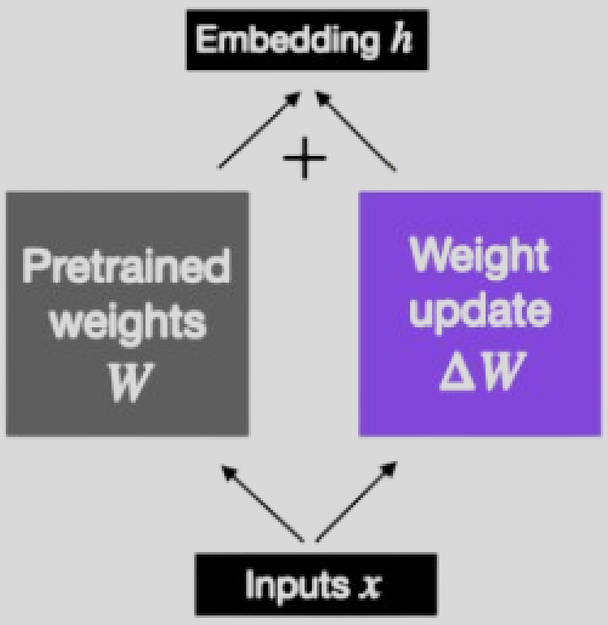
Elisa Forcada Rodríguez
eforcada001@ikasle.ehu.eus

Tutor:
Eneko Agirre

Fine-Tuning

Fine-tuning adjusts a pre-trained model, trained on general tasks, to adapt it to a more specific task

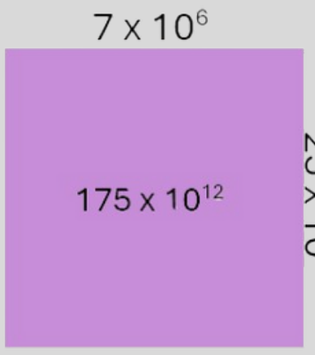
Dense layers of transformers: matrixes with full rank



Even if it is faster and easier than training a model from scratch...

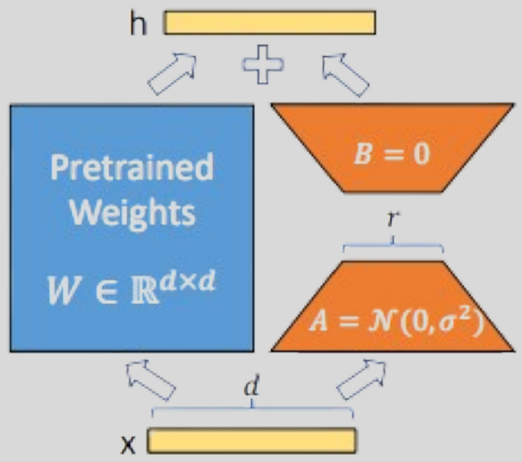
- High computational costs
- High time cost
- Updates all the parameters of the pre-trained models
- Potential problem: I/O bottleneck

GPT-3 175 B VRAM: **1.2TB**



LoRA

Li et al. (2018a); Aghajanyan et al. (2020): Pre-trained language models have a low "instrisic dimension" → still learn efficiently despite a random projection to a smaller subspace

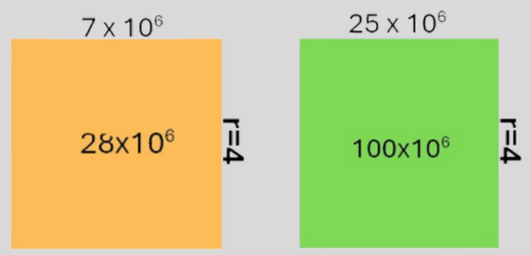


- Reduces computational cost
- Reduces training time
- Reduces the number of parameters to update of the pre-trained models

Main idea: decompose the matrix ΔW into two matrices B and A with much lower rank

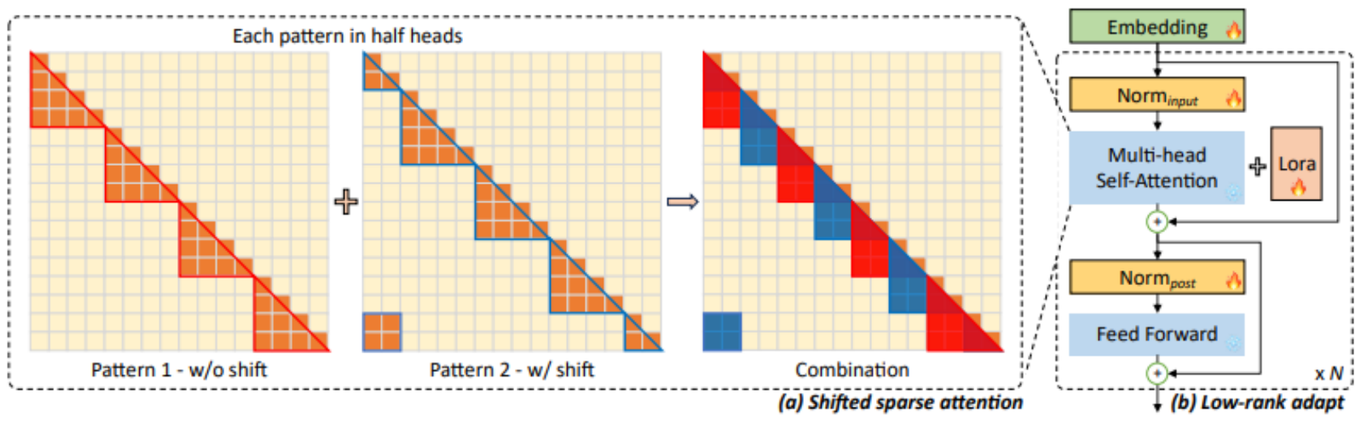
$$h = W_0 x + \Delta W x = W_0 x + BAx$$

GPT-3 175 B VRAM: **35MB** with **r=4**



LongLoRA

If we want to increase the pre-trained context lenght → LoRA loses its good performance, even significantly enlarging r

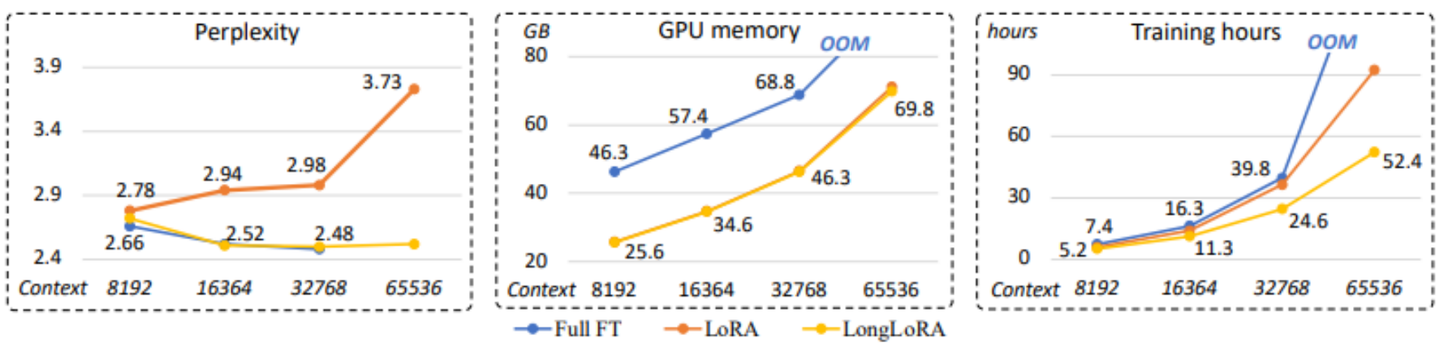


Normal attention: divide the lengthened context lenght into equal groups (Pattern 1)
Problem: no relationships between groups

MAIN IDEAS

1. Shifted sparse attention (S2 -Attn)
   a. Pattern 1: partition in equal groups
   b. Pattern 2: shifting the partition by half group size in half attention heads
   c. Combination of Pattern 1 and Pattern 2
2. LoRA for embedding and normalization layers (original LoRA only adapts attention weights)
3. Easy implementation (2 lines of code!)

Results



Effectiveness of S2 -Attn under different context lengths

Setting	Position Embedding	Training Attention	Training Shift	Target Context Length 8192	16384	32768
Full Attn	PI (Chen et al., 2023)	Long	-	8.02	8.05	8.04
Short Attn		Short	✗	8.29	8.83	9.47
S^2-Attn		Short	✓	8.04	8.03	8.08

References