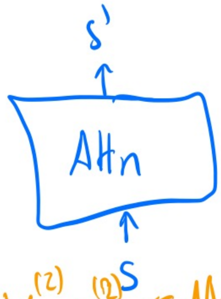
Attn   query   key$_1$ value$_1$

k_2 V_2

k_3 V_3

$\vdots$ $\vdots$

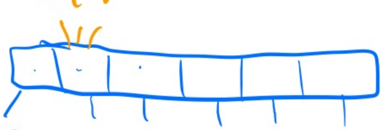


$$\text{out} = \sum_i \text{sim}(q, k_i) V_i$$

$$\text{Out} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

$k^{(2)} q^{(2)} v^{(2)}$ $W_k^{(2)}$ $W_q^{(2)}$ $W_v^{(2)}$ ← Multi-head

self-attn:



W_k, W_q, W_v

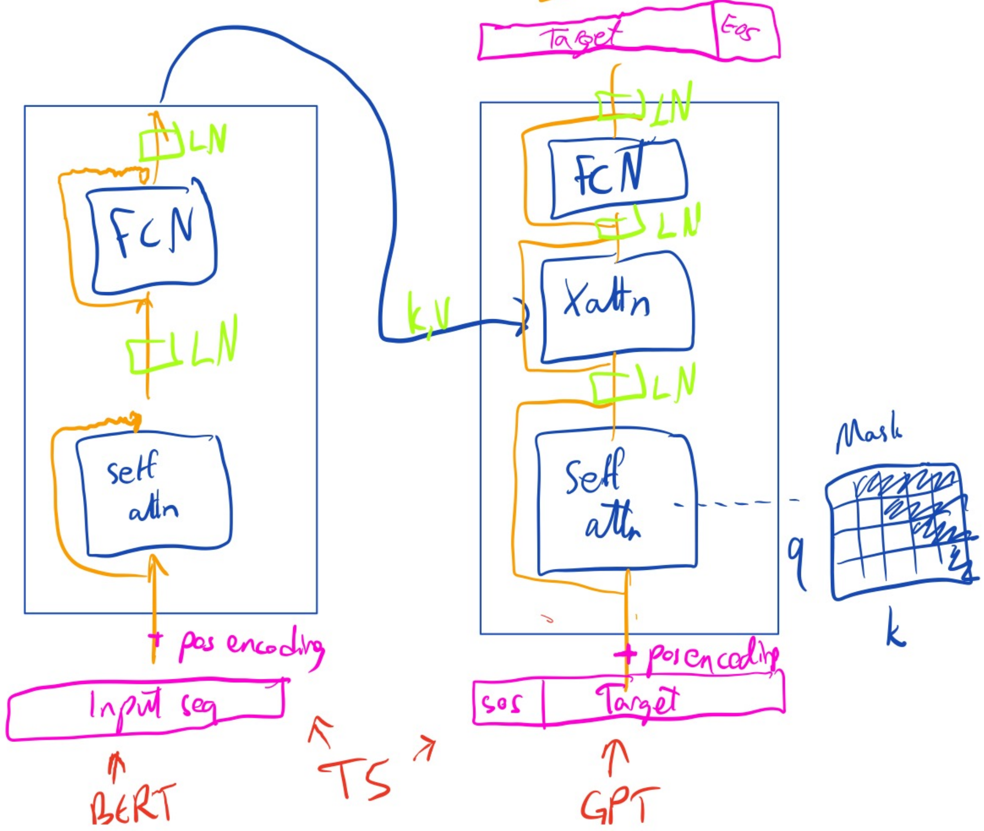
$k_1 = W_k s_1$

$q_1 = W_q s_1$

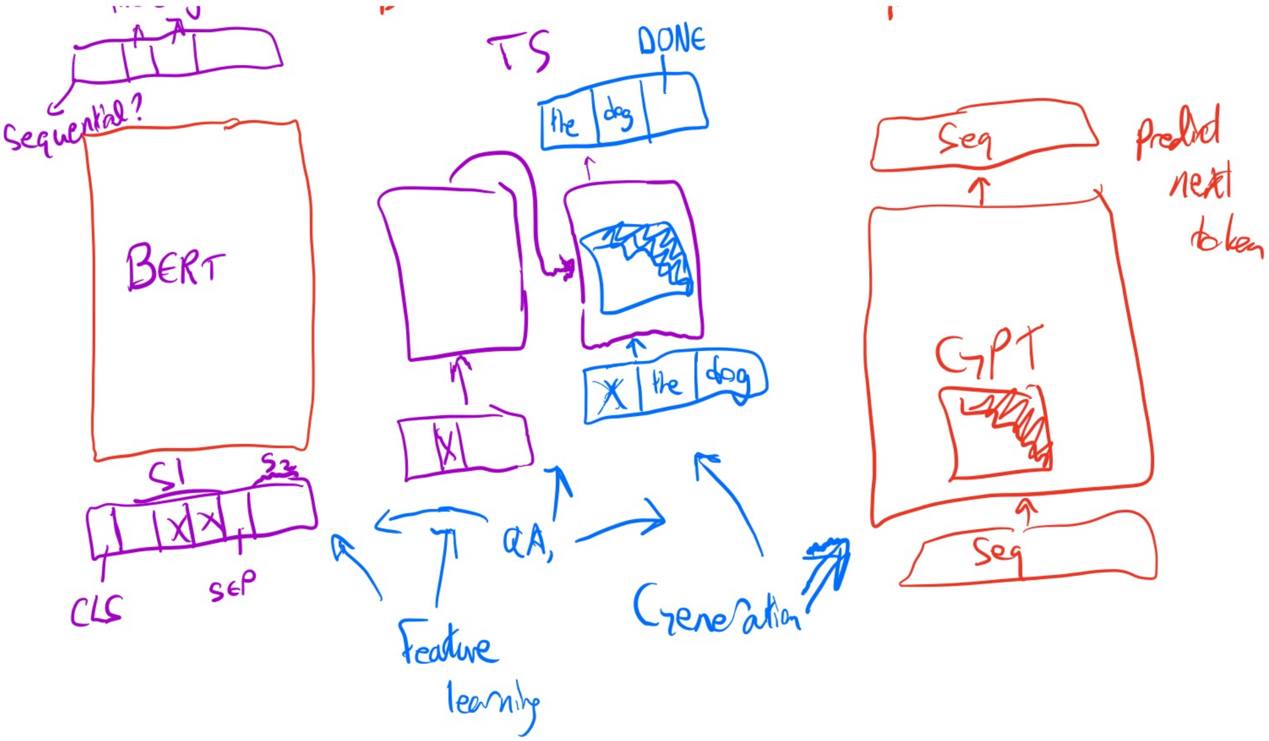
$v_1 = W_v s_1$

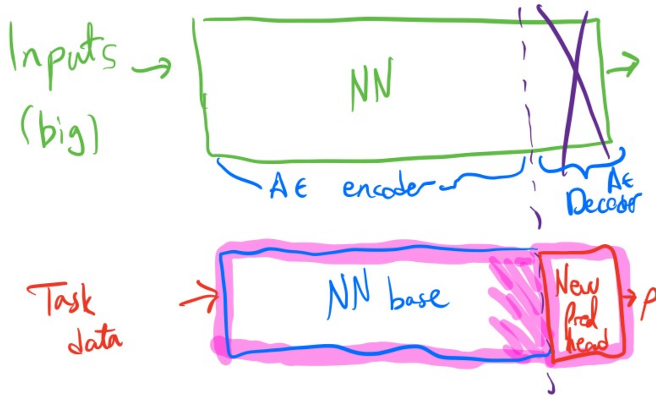
X-attn: q come from a diff. seq.

Transformers   **Encoder**   **Decoder**



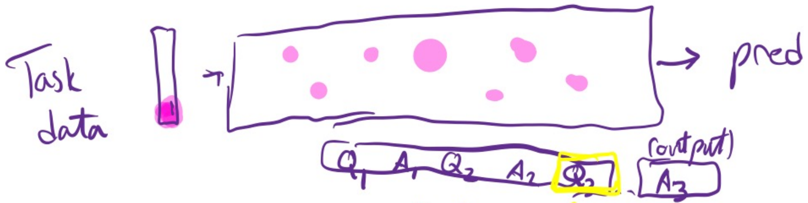
the dog





Proxy task
- reconstruction
- Masked prediction
- contrastive (similarity, clustering)
- classification
⋮

What to Finetune



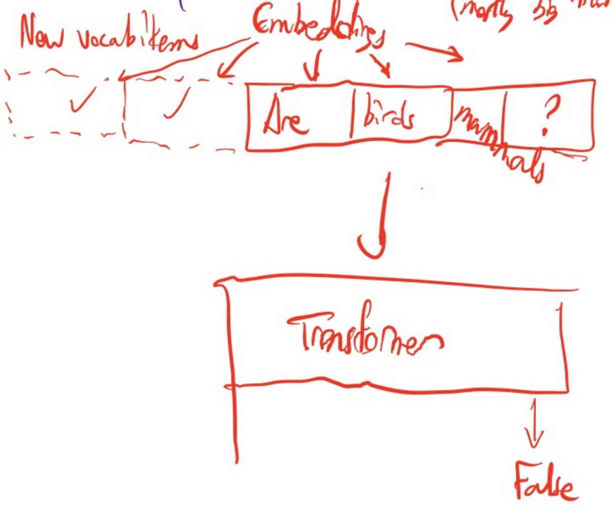
Feature Extraction (Linear probing	Full finetuning (partial OK too)	Hard Prompting	Soft Prompting	Everything Else
	Bet perf	No (or minimal) training data	Doesn't mess up good pretrained models	
← Almost every model ↗		Capped performance	Good for batches w mixed tasks	
Moderate amount of data	-LOTS of data -one model/task = memory	Mostly for LLM	Needs moderate amount	

Not the best perf

memory

of data

Doesn't work for all models (mostly big transformers)


New vocab items
Embeddings
✓
✓
Are
birds
mammals
?
Transformer
False