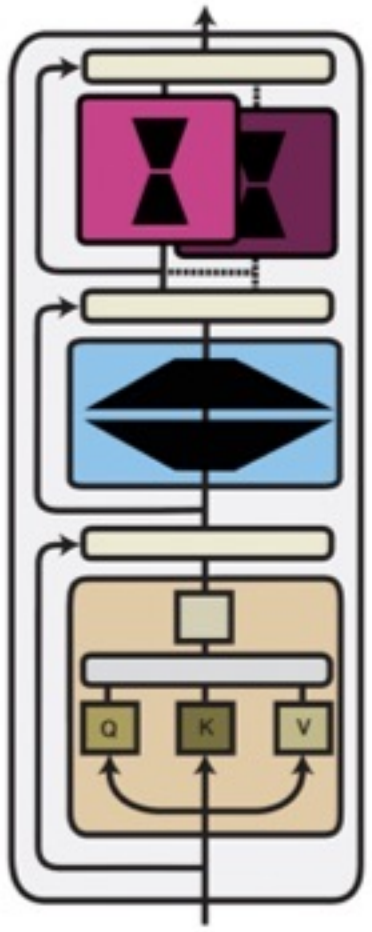
A functional perspective of adaptation

- Function composition augments a model's functions with new task-specific functions:

$$f'_i(\boldsymbol{x}) = f_{\theta_i}(\boldsymbol{x}) \odot f_{\phi_i}(\boldsymbol{x})$$

- Most commonly used in multi-task learning where modules of different tasks are composed.



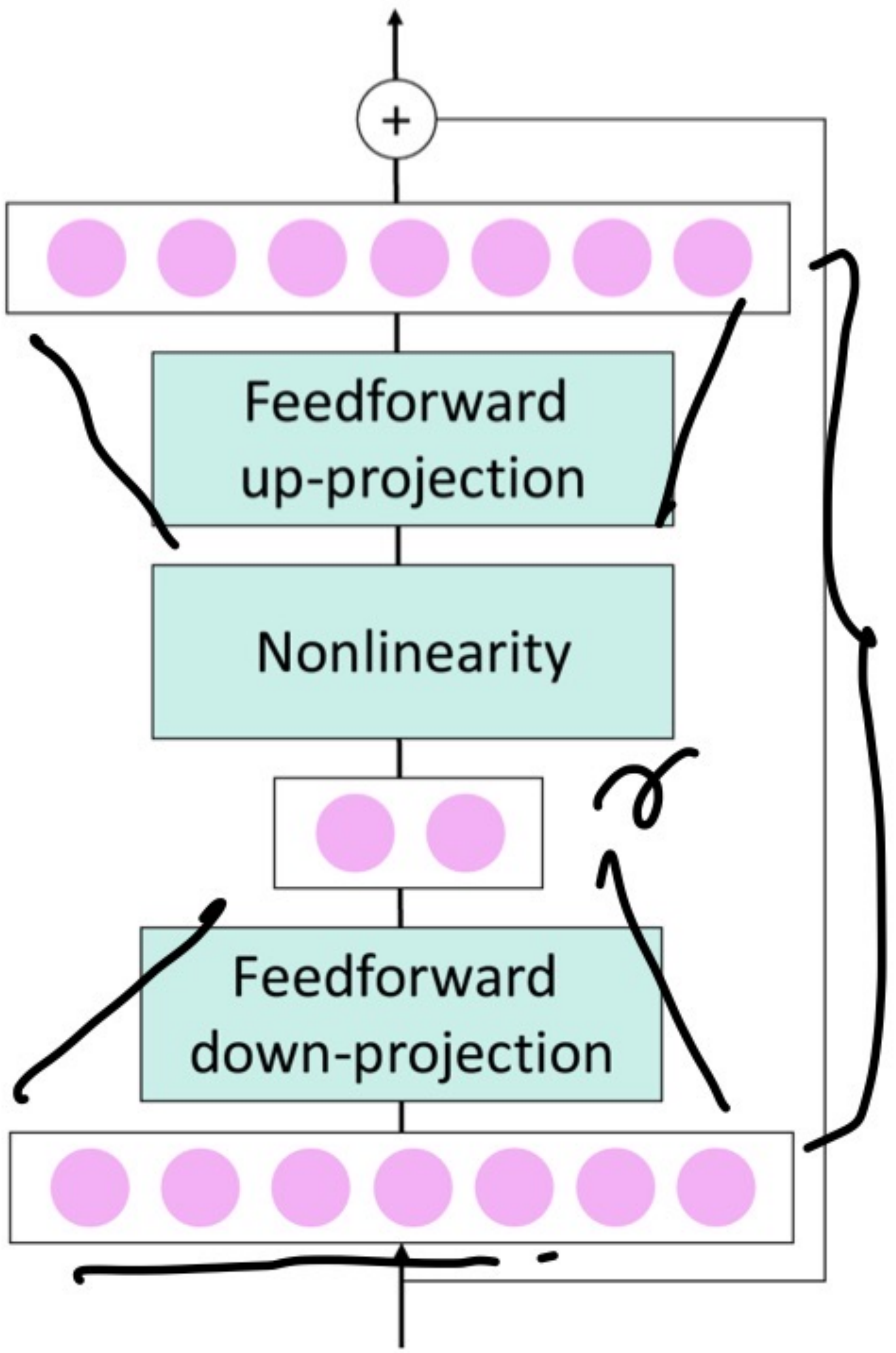
Function Composition

The Adapter Framework

Adapter (Houlsby et al. 2019)

- Insert a new function f_ϕ between layers of a pre-trained model to adapt to a downstream task --- known as "adapters"

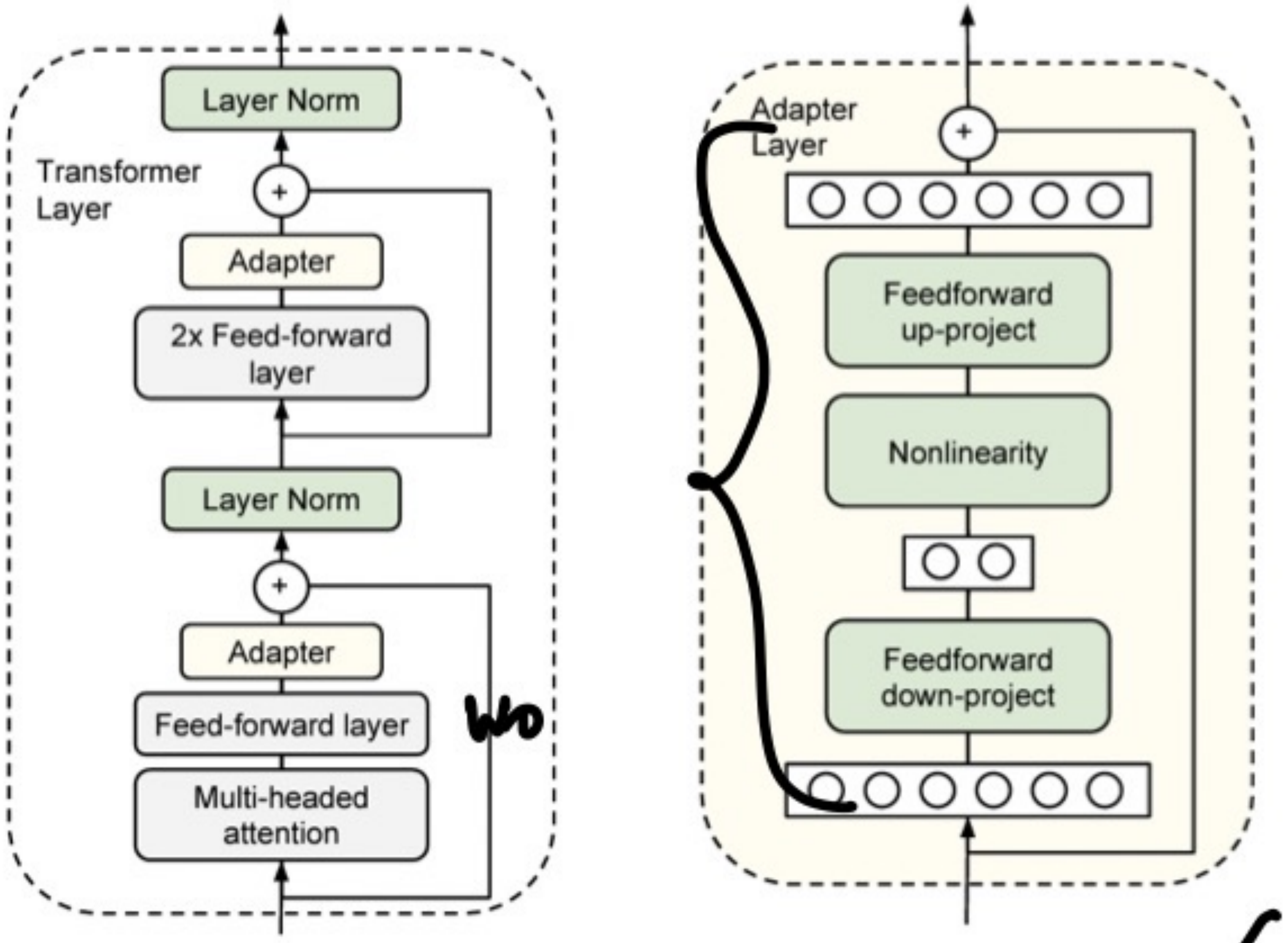
- An adapter in a Transformer layer consists of:
  - A feed-forward down-projection $W^D \in R^{k \times d}$
  - A feed-forward up-projection $W^U \in R^{d \times k}$
  - $f_\phi(\boldsymbol{x}) = W^U(\sigma(W^D \boldsymbol{x}))$



dmodel

r×D

Where is the adapter placed?



$K=2$

Figure 2. Architecture of the adapter module and its integration with the Transformer. **Left:** We add the adapter module twice to each Transformer layer: after the projection following multi-headed attention and after the two feed-forward layers. **Right:** The adapter consists of a bottleneck which contains few parameters relative to the attention and feedforward layers in the original model.

GPT-3

$+ K + d_{model}$

$2 \times (d_{model} \times K + K \times d_{model}) \times L$

How many parameters?

Consider one adapter

- Feed-forward down-projection: $r \times d_{model} + r$
- Feed-forward up-projection: $d_{model} \times r + d_{model}$

- For L layers in the decoder, there would be $2L$ adapters
- Number of parameters: $2L \times (2 \times d_{model} \times r + d_{model} + r)$

Comparison of various PEFT methods

Method	Hyperparameters	# Trainable Parameters	WikiSQL	MNLI-m
Fine-Tune	-	175B	73.8	89.5
PrefixEmbed	$l_p = 32, l_i = 8$	0.4 M	55.9	84.9
	$l_p = 64, l_i = 8$	0.9 M	58.7	88.1
	$l_p = 128, l_i = 8$	1.7 M	60.6	88.0
	$l_p = 256, l_i = 8$	3.2 M	63.1	88.6
	$l_p = 512, l_i = 8$	6.4 M	55.9	85.8
PrefixLayer	$l_p = 2, l_i = 2$	5.1 M	68.5	89.2
	$l_p = 8, l_i = 0$	10.1 M	69.8	88.2
	$l_p = 8, l_i = 8$	20.2 M	70.1	89.5
	$l_p = 32, l_i = 4$	44.1 M	66.4	89.6
	$l_p = 64, l_i = 0$	76.1 M	64.9	87.9
Adapter^H	$r = 1$	7.1 M	71.9	89.8
	$r = 4$	21.2 M	73.2	91.0
	$r = 8$	40.1 M	73.2	91.5
	$r = 16$	77.9 M	73.2	91.5
	$r = 64$	304.4 M	72.6	91.5
LoRA	$r_v = 2$	4.7 M	73.4	**91.7**
	$r_q = r_v = 1$	4.7 M	73.4	91.3
	$r_q = r_v = 2$	9.4 M	73.3	91.4
	$r_q = r_k = r_v = r_o = 1$	9.4 M	74.1	91.2
	$r_q = r_v = 4$	18.8 M	73.7	91.3
	$r_q = r_k = r_v = r_o = 2$	18.8 M	73.7	**91.7**
	$r_q = r_v = 8$	37.7 M	73.8	**91.6**
	$r_q = r_k = r_v = r_o = 4$	37.7 M	74.0	**91.7**
	$r_q = r_v = 64$	301.9 M	73.6	91.4
	$r_q = r_k = r_v = r_o = 64$	603.8 M	73.9	91.4

40×12288

4×12288×16

3×12288×2×96