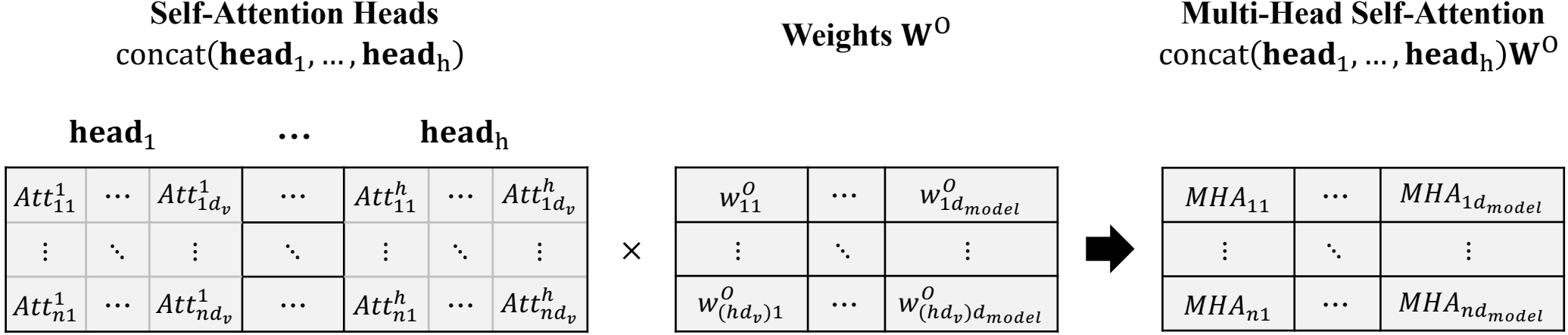
Self-Attention Heads
concat(head$_1$, … , head$_h$)
Weights W^O
Multi-Head Self-Attention
concat(head$_1$, … , head$_h$)W^O
head$_1$
…
head$_h$
Att_{11}^1 … $Att_{1d_v}^1$ … Att_{11}^h … $Att_{1d_v}^h$
Att_{n1}^1 … $Att_{nd_v}^1$ … Att_{n1}^h … $Att_{nd_v}^h$
×
w_{11}^O … $w_{1d_{model}}^O$
$w_{(hd_v)1}^O$ … $w_{(hd_v)d_{model}}^O$
MHA_{11} … $MHA_{1d_{model}}$
MHA_{n1} … $MHA_{nd_{model}}$