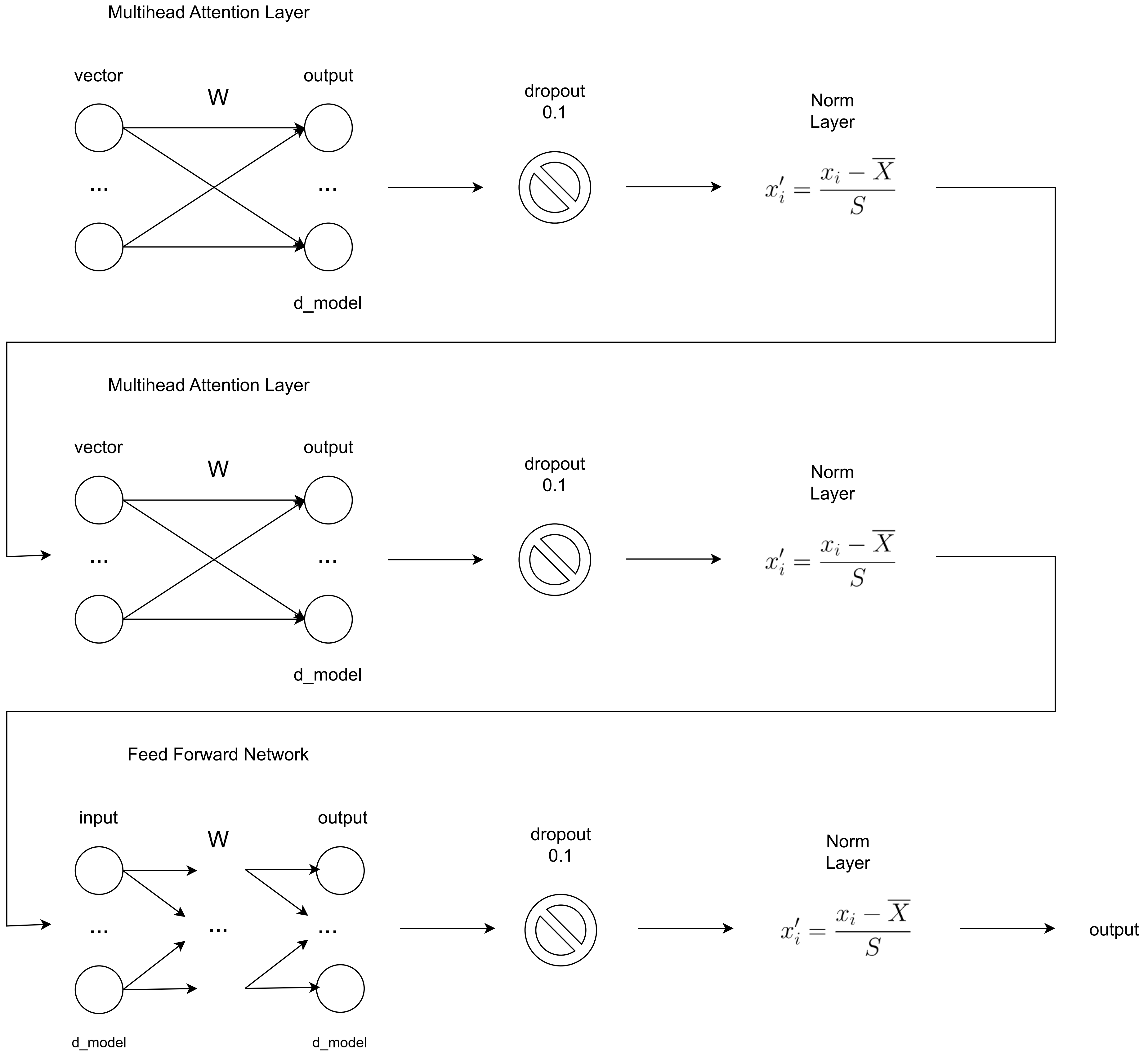
Decoder Layer

Massachusetts
Institute of
Technology


Multihead Attention Layer
vector
W
output
...
...
d_model
dropout
0.1
Norm
Layer
$x'_i = \frac{x_i - \overline{X}}{S}$
Multihead Attention Layer
vector
W
output
...
...
d_model
dropout
0.1
Norm
Layer
$x'_i = \frac{x_i - \overline{X}}{S}$
Feed Forward Network
input
W
output
...
...
...
d_model
d_model
dropout
0.1
Norm
Layer
$x'_i = \frac{x_i - \overline{X}}{S}$
output