transformer compression

attention ❄

MLP layer ❄

ReLU

$W^\top$ W $W^\top$ W

input W $W^\top$ output

residual stream