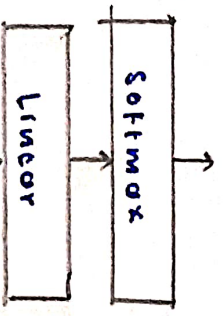


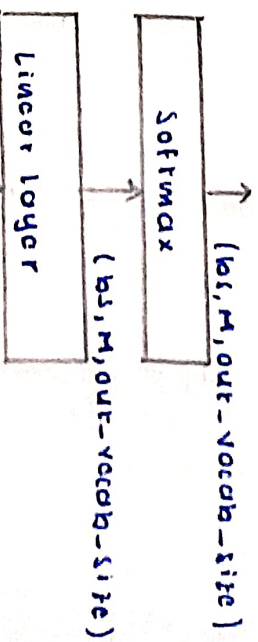
## Output Probabilities



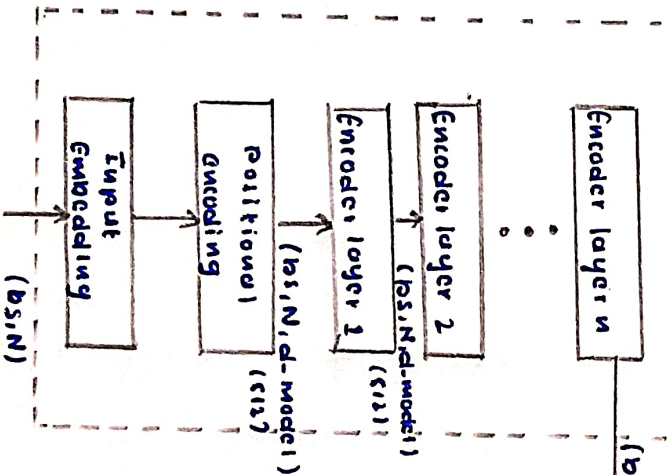
## Transformer Model Architecture

Notation	Explanation
$b, l$	Batch size
$d$ -model	Dimension of model embedding
$N$	Length of input sequence (# of tokens)
$M$	Length of output sequence (# of tokens)
out-vocab-size	Size of output vocabulary
$S$	number of attention head.

## Output Probabilities



## Encoder



## Decoder

