

Probabilities

Softmax



Linear



Add & Norm



Feed-Forward



Add & Norm



Masked Multi-Head Attention



Block

Add Position
Embeddings



Embeddings

Decoder Inputs

*Repeat for number of
encoder blocks*



Transformer Decoder