

Transformer characterized by Attention mask seen below

Embeddings

Embeddings



Image1

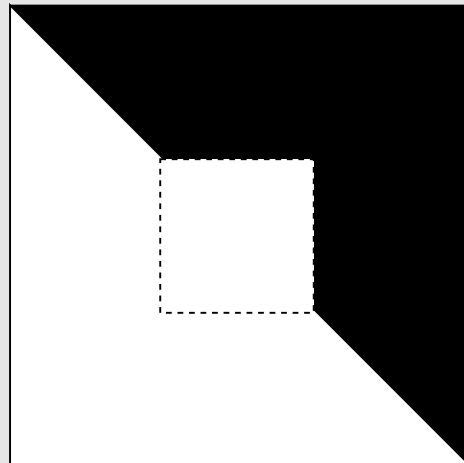
Text1

Tokenize & embed

Patch, affine trafo ,  
(optional)  
pooling

Text2

Tokenize & embed



Text1 (shifted)

Tokenize & 1-hot  
encode

CE-Loss

Unembed

Text2 (shifted)

Tokenize & 1-hot  
encode

CE-Loss

Unembed

Image source:

<https://engineering.nyu.edu/faculty/yann-lecun> (2024 May 13th)