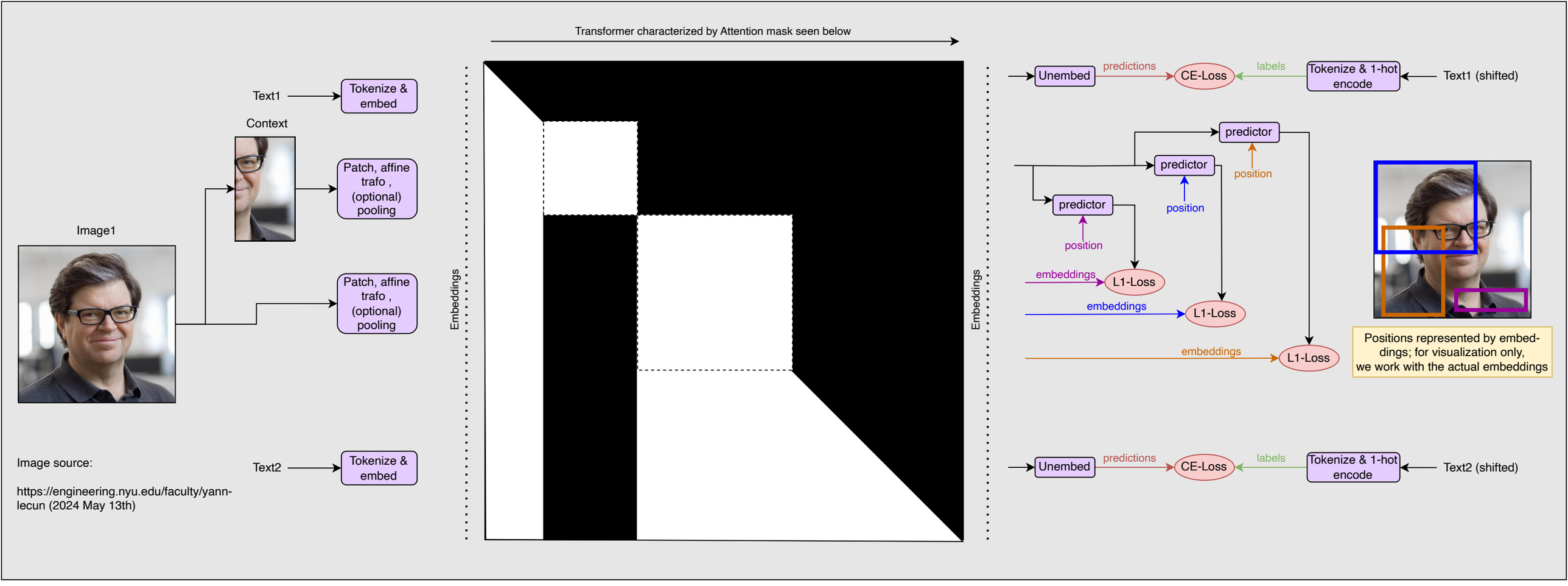
Transformer characterized by Attention mask seen below
Text1
Tokenize & embed
Context
Patch, affine trafo , (optional) pooling
Image1
Patch, affine trafo , (optional) pooling
Image source:
https://engineering.nyu.edu/faculty/yann-lecun (2024 May 13th)
Text2
Tokenize & embed
Embeddings
Embeddings
Unembed
predictions
CE-Loss
labels
Tokenize & 1-hot encode
Text1 (shifted)
predictor
predictor
predictor
position
position
position
embeddings
L1-Loss
embeddings
L1-Loss
embeddings
L1-Loss
Positions represented by embed-dings; for visualization only, we work with the actual embeddings
Unembed
predictions
CE-Loss
labels
Tokenize & 1-hot encode
Text2 (shifted)