

Heatmaps of attention logits $\frac{1}{\sqrt{d}}\langle Qx_i, Kx_j \rangle$
(wikitext_bert-large-uncased_params-trained_dense-on_layers-24, sample 8)

