

Heatmaps of attention logits  $\frac{1}{\sqrt{d}}\langle Qx_i, Kx_j \rangle$   
(wikitext\_albert-xlarge-v2\_params-trained\_dense-off\_layers-48, sample 9)

