

Heatmaps of attention logits  $\frac{1}{\sqrt{d}}\langle Qx_i, Kx_j \rangle$   
(imdb\_albert-xlarge-v2\_params-trained\_dense-on\_layers-48, sample 8)

