attention_logit t-SNE embeddings (wikitext bert-large-uncased params-trained dense-on layers-24) after layer 2 after layer 1 after layer 3 after layer 4 100 100 100 100 50 50 50 50 0 0 -50-50 -50-50-100-100-100-10050 50 100 0 50 50 100 0 100 -100 0 -50100 -100 -50-50-100-100 -50after layer 5 after layer 8 after layer 6 after layer 7 100 100 100 100 50 50 50 50 0 0 0 -50-50 -50 -50-100-100-100-50 50 100 50 100 -5050 100 50 100 -100 -100-50-100-100-50after layer 9 after layer 10 after layer 11 after layer 12 100 100 100 100 50 50 50 50 0 0 0 -50-50-50-50-100-100-100-10050 100 50 100 0 50 100 50 100 -100-500 -100-500 -100-50-100-500 after layer 13 after layer 14 after layer 15 after layer 16 100 100 100 100 50 50 50 50 0 0 -50-50 -50-50-100-100-100 -100 50 -50 50 0 50 -100-500 100 -100-500 50 100 0 -50100 -100100 -100after layer 17 after layer 18 after layer 19 after layer 20 100 100 100 100 50 50 50 50 0 0 0 0 -50 -50 -50 -50 -100 | -100 -100-100 -100 -50 0 50 100 -100 -50 0 50 100 -50 0 50 100 -100 -50 0 50 100 -100 after layer 23 after layer 21 after layer 22 after layer 24 100 100 100 100 50 50 50 50 0 0 A T 0 0 -50-50 -50 -50-100 -100 -100 -100-100-50 0 50 100 -100-50 50 100 -100-50 0 50 100 -100 -50 50 100