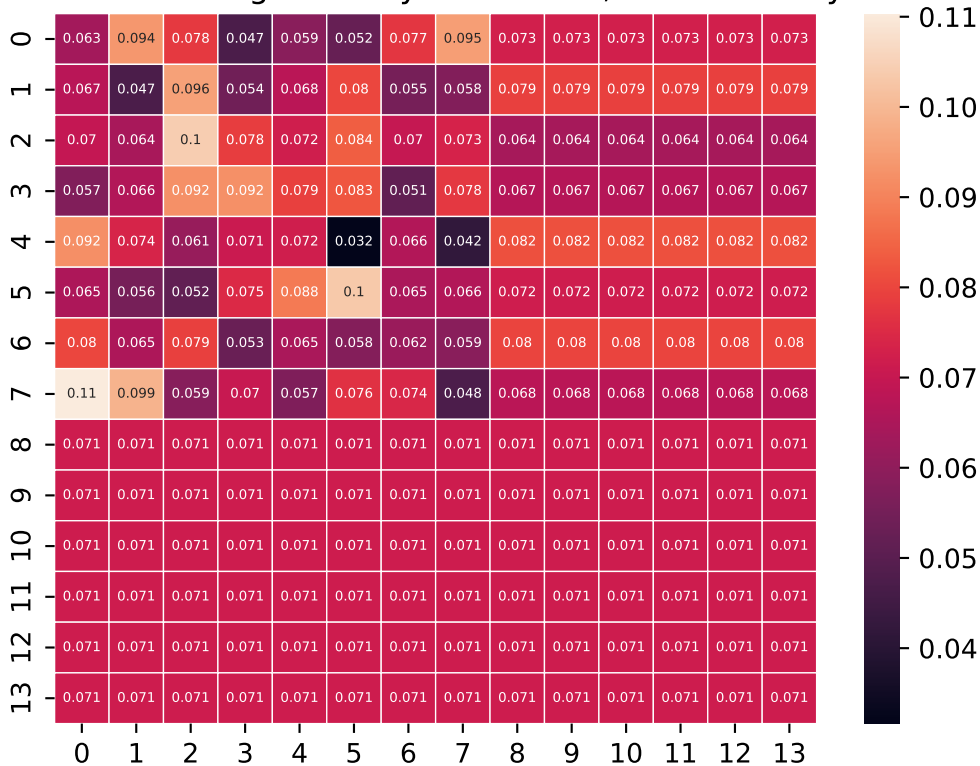Step 0 attention weights at layer 0 head 0, with accuracy 0.180

Step 0 attention weights at layer 0 head 1, with accuracy 0.180

Step 0 attention weights at layer 0 head 2, with accuracy 0.180

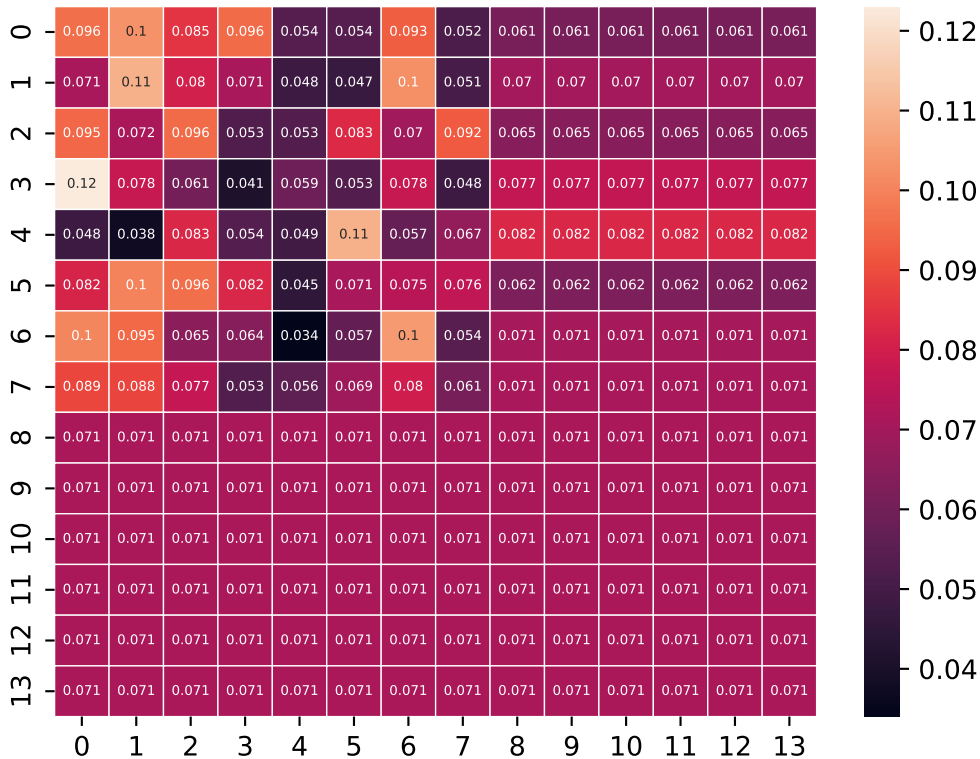Step 0 attention weights at layer 0 head 3, with accuracy 0.180

Step 0 attention weights at layer 1 head 0, with accuracy 0.180
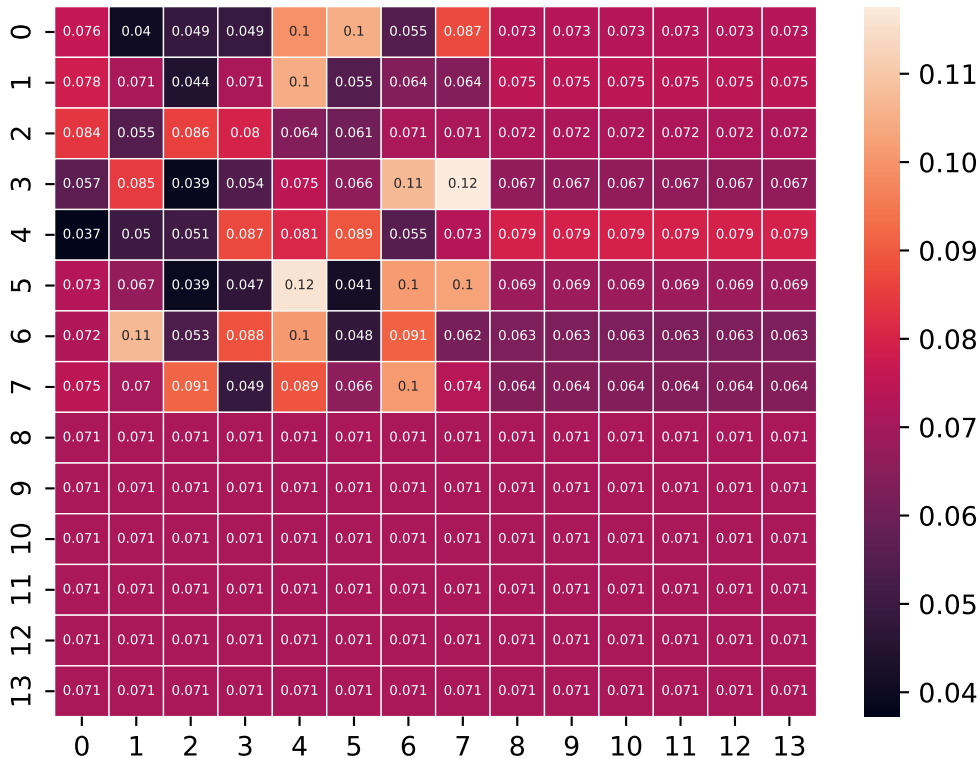
Step 0 attention weights at layer 1 head 1, with accuracy 0.180

Step 0 attention weights at layer 1 head 2, with accuracy 0.180

Step 0 attention weights at layer 1 head 3, with accuracy 0.180

Step 500 attention weights at layer 0 head 0, with accuracy 0.604

Step 500 attention weights at layer 0 head 1, with accuracy 0.604

Step 500 attention weights at layer 0 head 2, with accuracy 0.604

Step 500 attention weights at layer 0 head 3, with accuracy 0.604

Step 500 attention weights at layer 1 head 0, with accuracy 0.604
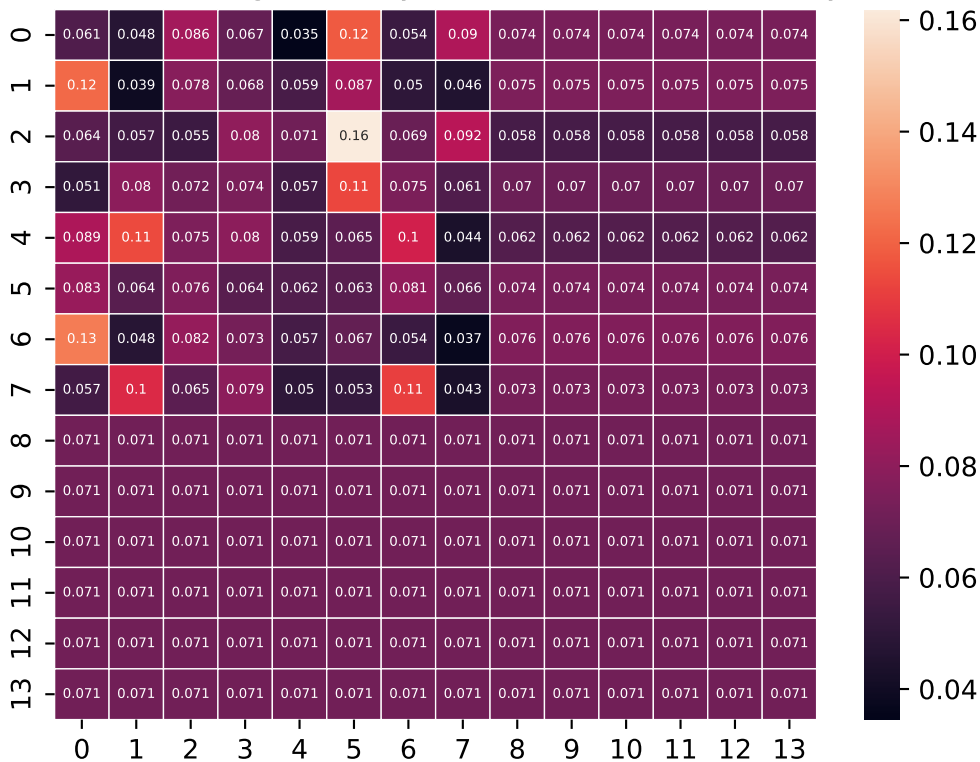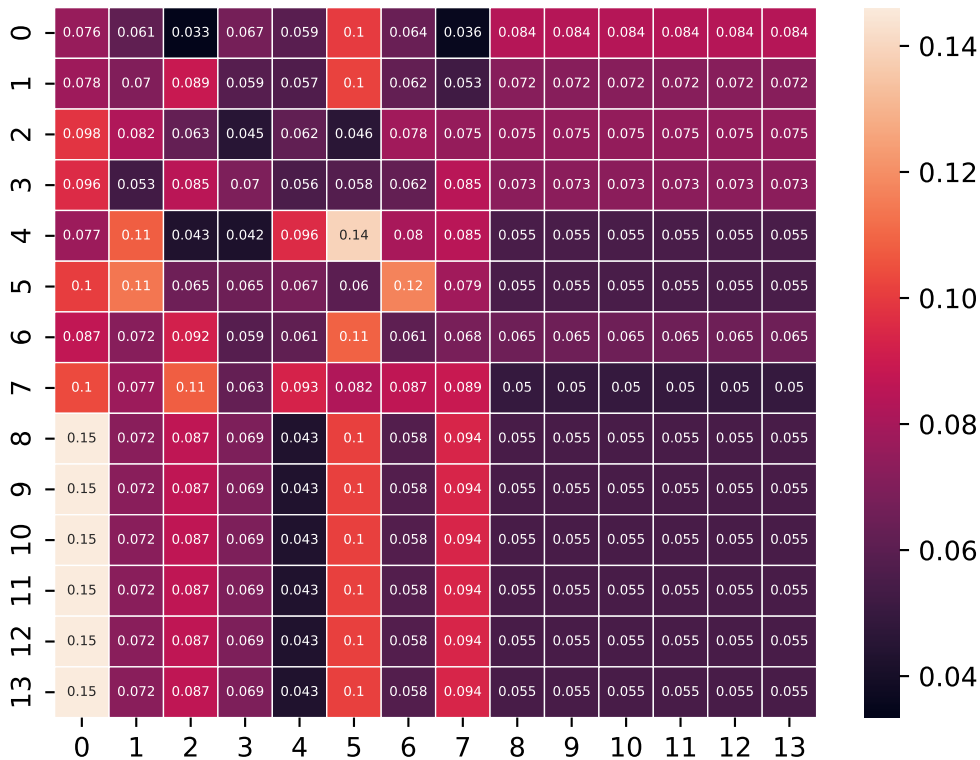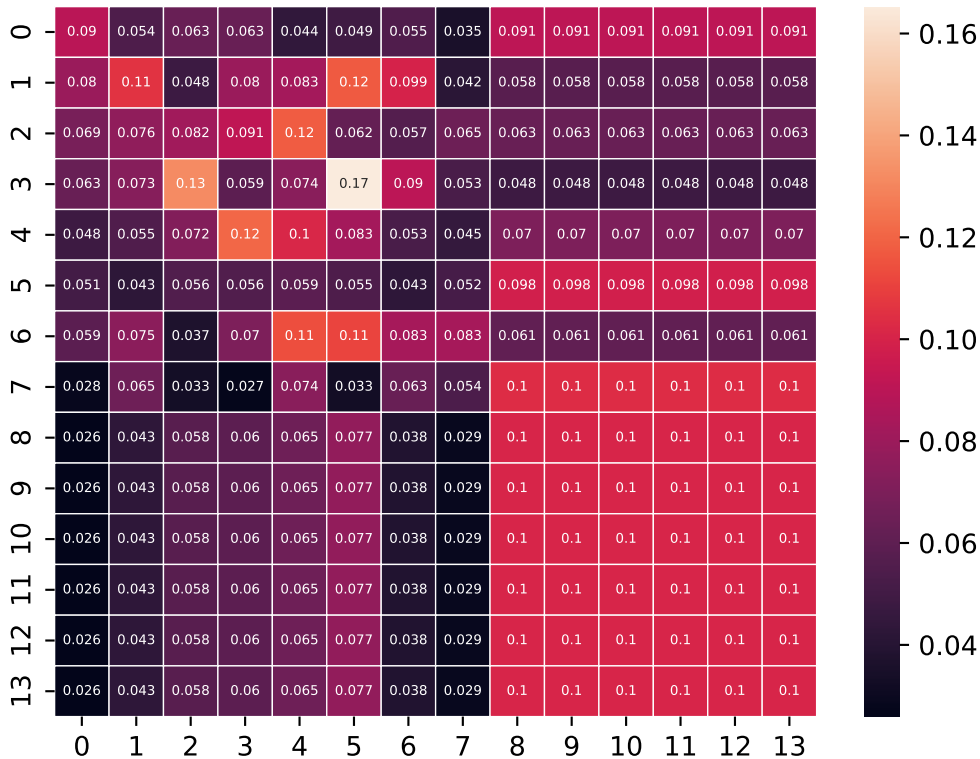
Step 500 attention weights at layer 1 head 1, with accuracy 0.604

Step 500 attention weights at layer 1 head 2, with accuracy 0.604

Step 500 attention weights at layer 1 head 3, with accuracy 0.604

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.0083	0.033	0.018	0.016	0.56	0.0059	0.31	0.01	0.0058	0.0057	0.014	0.0038	0.011	0.0045
1	0.027	0.057	0.045	0.088	0.18	0.061	0.17	0.053	0.041	0.051	0.067	0.053	0.08	0.03
2	0.065	0.095	0.049	0.045	0.15	0.045	0.15	0.074	0.052	0.057	0.085	0.048	0.035	0.049
3	0.065	0.11	0.056	0.082	0.11	0.052	0.18	0.055	0.033	0.049	0.067	0.039	0.057	0.052
4	0.065	0.067	0.073	0.085	0.065	0.054	0.16	0.068	0.047	0.069	0.045	0.075	0.057	0.07
5	0.066	0.061	0.063	0.054	0.066	0.099	0.11	0.063	0.049	0.067	0.1	0.099	0.037	0.065
6	0.083	0.045	0.051	0.055	0.049	0.094	0.077	0.072	0.054	0.083	0.1	0.081	0.065	0.088
7	0.042	0.052	0.044	0.087	0.14	0.059	0.15	0.061	0.046	0.07	0.066	0.062	0.072	0.052
8	0.056	0.12	0.048	0.043	0.14	0.046	0.12	0.056	0.047	0.049	0.085	0.094	0.035	0.06
9	0.059	0.076	0.063	0.047	0.17	0.055	0.1	0.049	0.041	0.052	0.065	0.092	0.043	0.085
10	0.057	0.12	0.066	0.068	0.075	0.052	0.1	0.079	0.05	0.051	0.064	0.083	0.059	0.08
11	0.071	0.05	0.072	0.08	0.16	0.048	0.1	0.044	0.055	0.081	0.075	0.045	0.068	0.047
12	0.078	0.083	0.04	0.092	0.1	0.048	0.13	0.055	0.027	0.081	0.076	0.049	0.055	0.084
13	0.044	0.082	0.064	0.068	0.12	0.059	0.13	0.045	0.041	0.072	0.062	0.093	0.053	0.067

Step 3000 attention weights at layer 0 head 0, with accuracy 0.596

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.0039	0.013	0.12	0.003	0.041	0.0012	0.046	0.003	0.053	0.0018	0.71	0.00053	0.00017	7.8e-05
1	0.02	0.64	0.025	0.0096	0.018	0.013	0.02	0.022	0.027	0.023	0.11	0.021	0.022	0.028
2	0.0097	0.023	0.18	0.019	0.077	0.0098	0.11	0.026	0.19	0.067	0.24	0.036	0.0075	0.0013
3	0.02	0.035	0.13	0.067	0.057	0.022	0.067	0.054	0.2	0.16	0.13	0.038	0.024	0.008
4	0.018	0.0075	0.19	0.0096	0.038	0.012	0.069	0.041	0.14	0.073	0.34	0.023	0.012	0.027
5	0.044	0.014	0.098	0.03	0.03	0.032	0.03	0.046	0.11	0.065	0.14	0.063	0.027	0.27
6	0.0093	0.0014	0.017	0.0025	0.0046	0.0054	0.008	0.016	0.014	0.01	0.048	0.0085	0.0082	0.85
7	0.0044	0.0011	0.0057	0.0018	0.0023	0.0023	0.0027	0.0042	0.0038	0.005	0.013	0.0058	0.0029	0.95
8	0.003	0.00053	0.0032	0.0018	0.0022	0.0049	0.0045	0.013	0.029	0.013	0.01	0.0072	0.0067	0.9
9	0.0015	0.00038	0.003	0.00084	0.0009	0.002	0.0013	0.0047	0.0056	0.61	0.0046	0.0084	0.005	0.35
10	0.013	0.0035	0.038	0.01	0.012	0.012	0.031	0.026	0.12	0.11	0.29	0.032	0.03	0.27
11	0.022	0.026	0.073	0.022	0.072	0.032	0.097	0.079	0.13	0.066	0.26	0.053	0.028	0.044
12	0.045	0.027	0.096	0.027	0.053	0.032	0.057	0.045	0.13	0.059	0.33	0.035	0.022	0.049
13	0.023	0.009	0.0093	0.0096	0.0099	0.021	0.0097	0.015	0.011	0.027	0.026	0.018	0.03	0.78

Step 3000 attention weights at layer 0 head 1, with accuracy 0.596

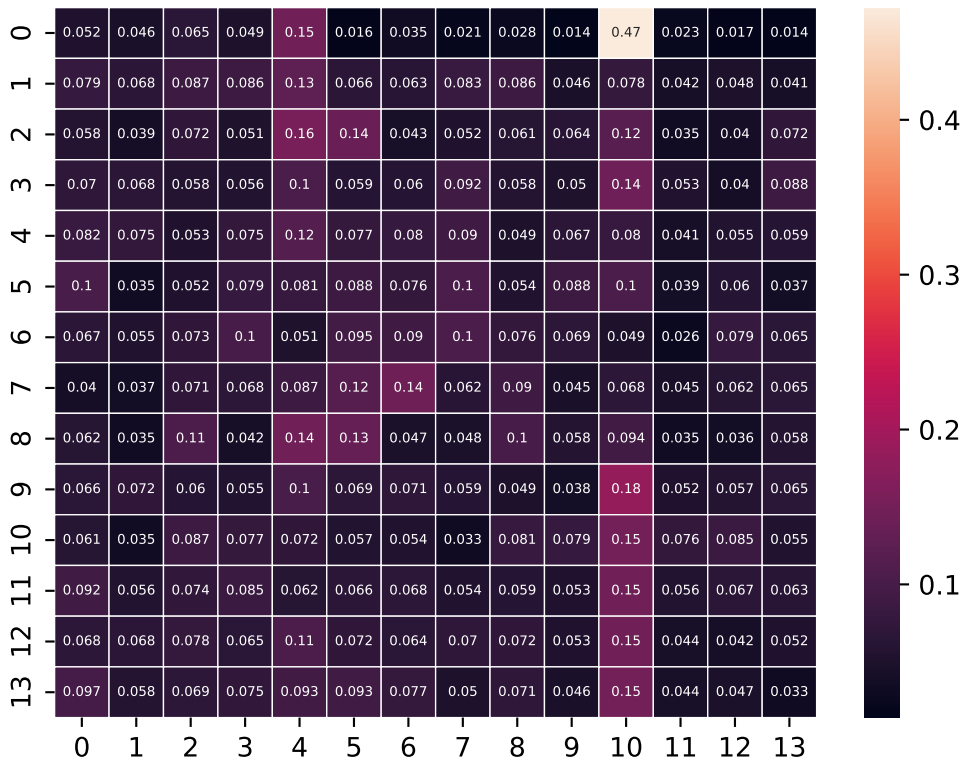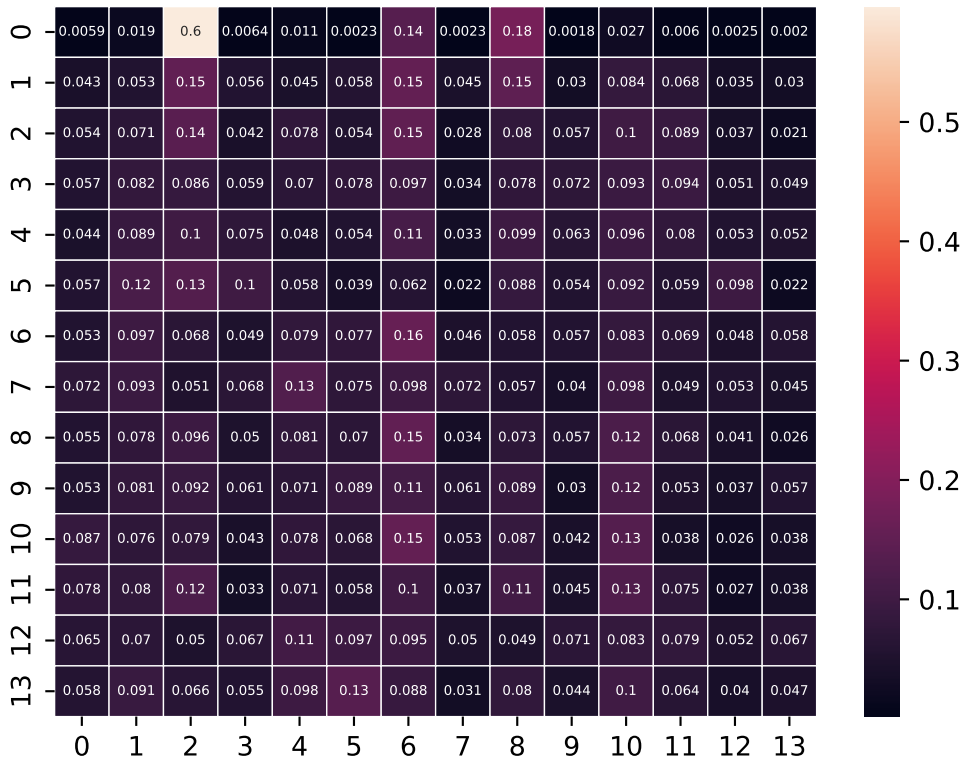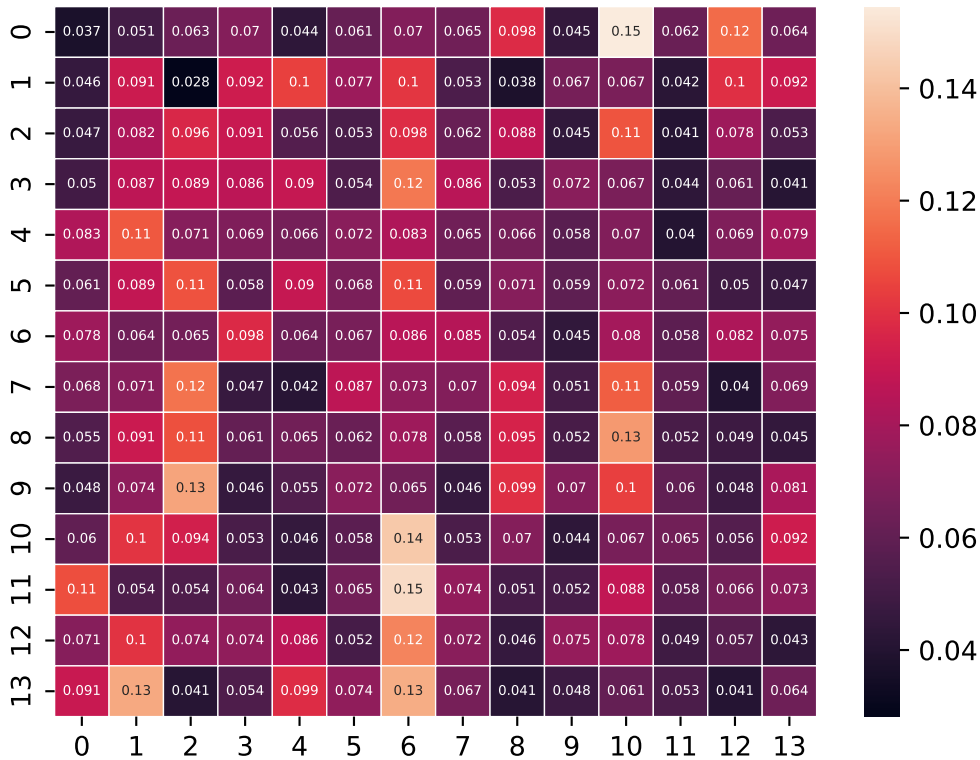|	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.0032	0.015	0.39	7.8e-05	0.023	0.0002	0.043	0.0017	0.48	0.00085	0.02	0.022	0.00028	0.00011
1	0.0056	0.75	0.019	0.0026	0.013	0.0047	0.029	0.01	0.042	0.011	0.02	0.078	0.0076	0.0042
2	0.011	0.023	0.058	0.011	0.042	0.023	0.11	0.04	0.37	0.052	0.075	0.16	0.016	0.016
3	0.026	0.041	0.062	0.23	0.033	0.033	0.048	0.05	0.28	0.037	0.03	0.092	0.025	0.012
4	0.021	0.027	0.057	0.0073	0.065	0.034	0.15	0.13	0.21	0.031	0.07	0.16	0.018	0.019
5	0.052	0.031	0.074	0.012	0.046	0.089	0.071	0.073	0.23	0.061	0.062	0.059	0.062	0.08
6	0.029	0.02	0.033	0.017	0.044	0.055	0.09	0.15	0.073	0.038	0.069	0.069	0.026	0.29
7	0.041	0.012	0.027	0.018	0.024	0.028	0.031	0.14	0.039	0.053	0.034	0.029	0.015	0.51
8	0.013	0.0075	0.0065	0.012	0.0054	0.027	0.01	0.026	0.096	0.024	0.012	0.0087	0.019	0.73
9	0.039	0.02	0.03	0.022	0.015	0.041	0.021	0.061	0.1	0.088	0.047	0.13	0.028	0.36
10	0.027	0.037	0.058	0.03	0.047	0.03	0.08	0.072	0.13	0.037	0.088	0.036	0.016	0.31
11	0.035	0.1	0.093	0.021	0.036	0.053	0.069	0.048	0.18	0.051	0.056	0.05	0.034	0.17
12	0.047	0.043	0.086	0.011	0.051	0.098	0.075	0.066	0.2	0.058	0.076	0.047	0.069	0.077
13	0.034	0.035	0.019	0.023	0.014	0.04	0.024	0.04	0.063	0.034	0.03	0.02	0.021	0.6

Step 3000 attention weights at layer 0 head 2, with accuracy 0.596

Step 3000 attention weights at layer 0 head 3, with accuracy 0.596

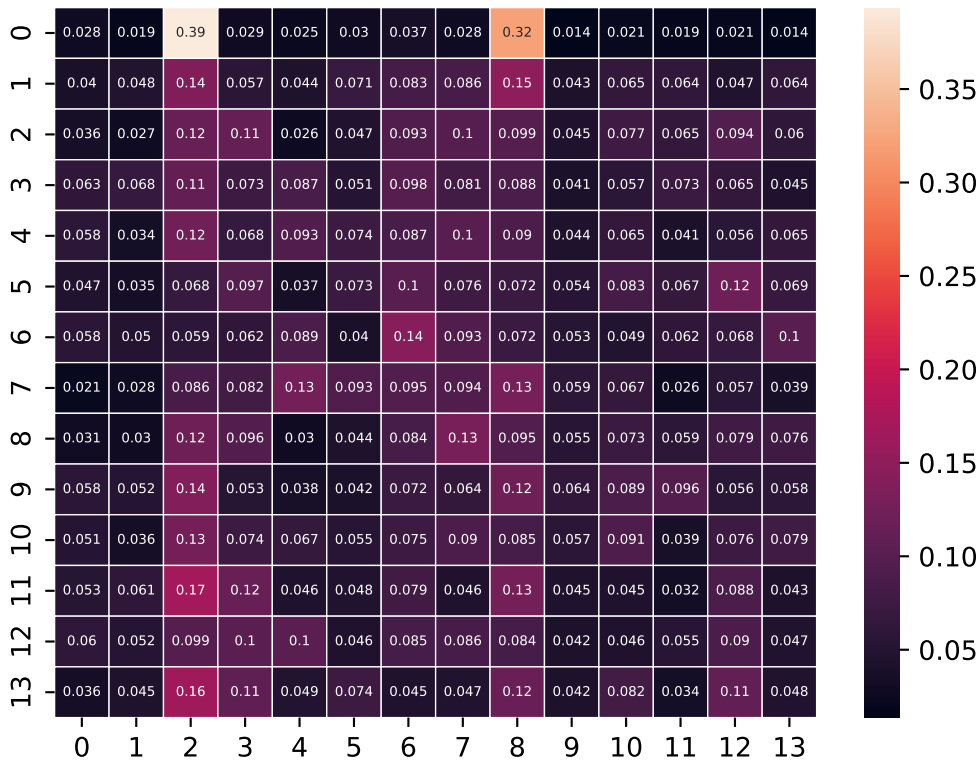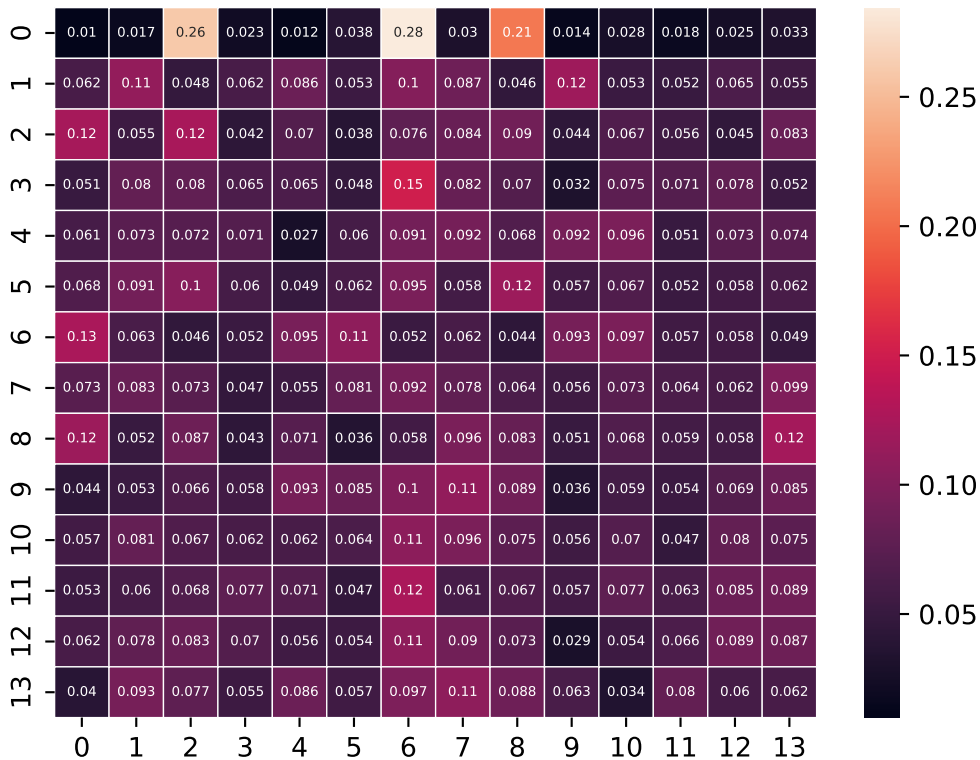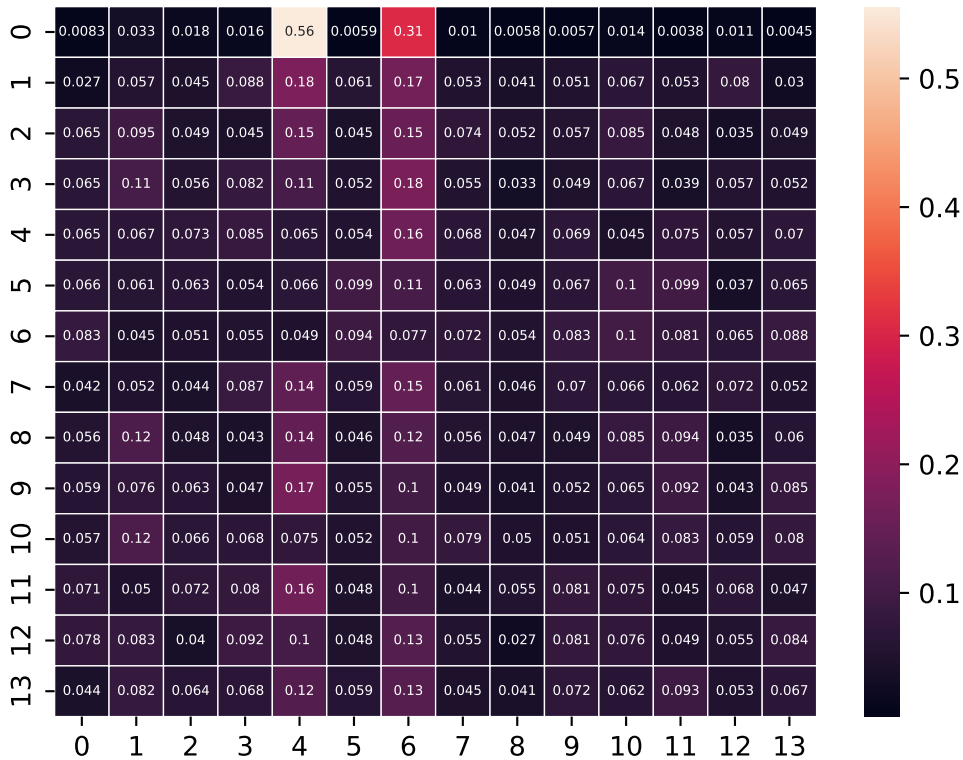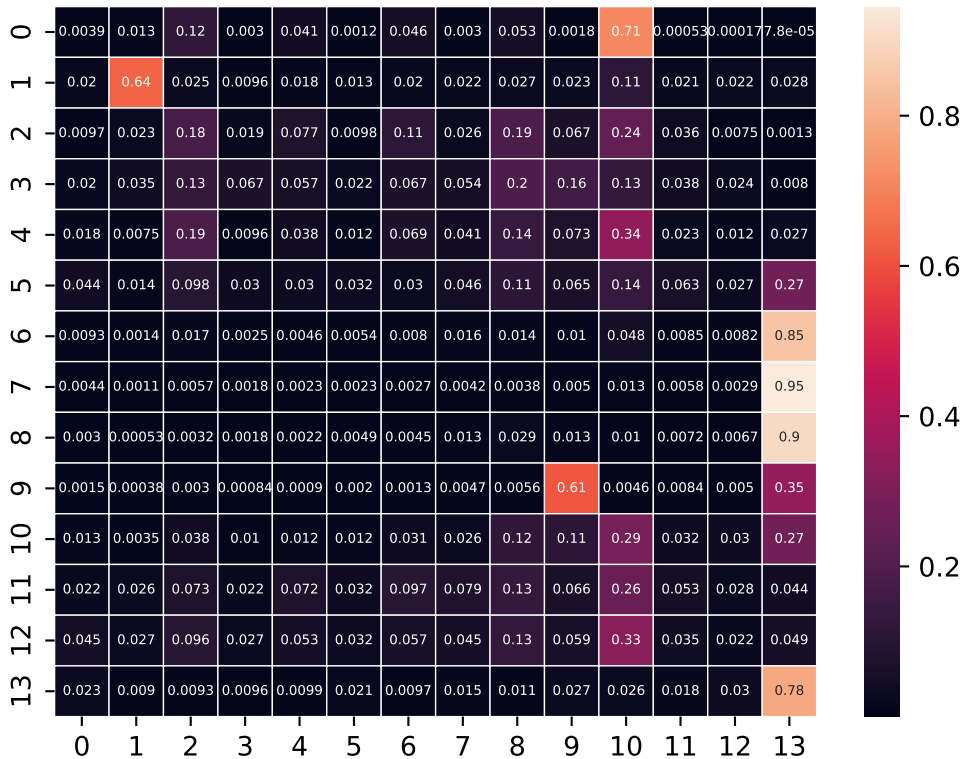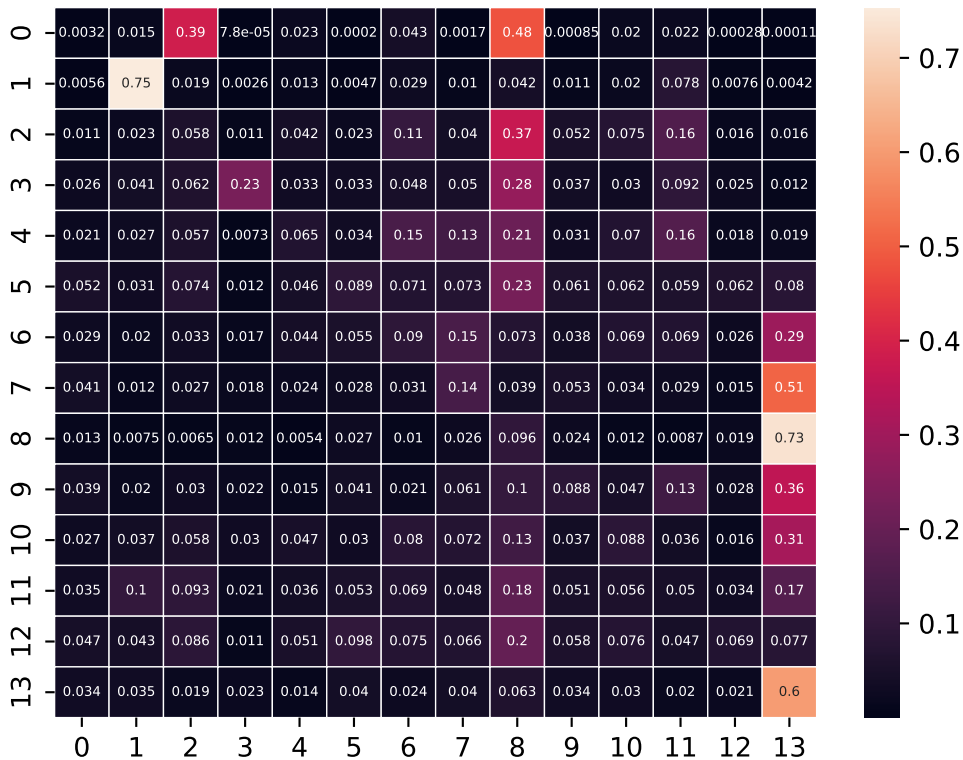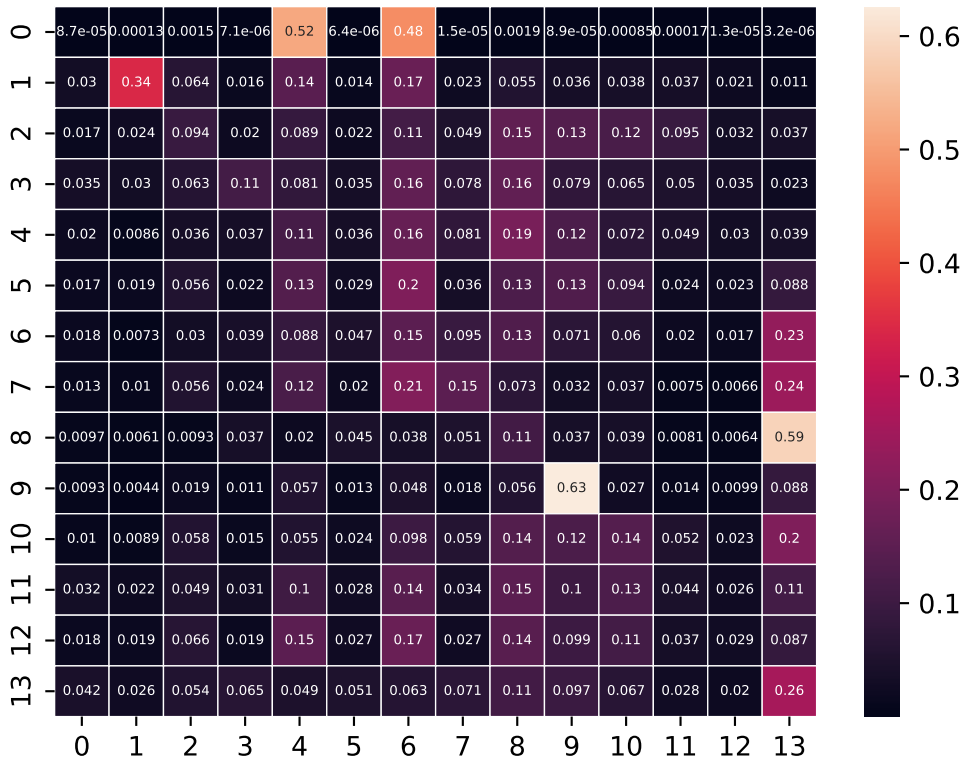|	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.0064	0.036	0.63	0.00012	0.028	0.00013	0.032	0.00088	0.23	0.00094	0.026	0.003	9.9e-05	0.00022
1	0.05	0.097	0.086	0.017	0.14	0.017	0.17	0.034	0.13	0.042	0.082	0.092	0.017	0.022
2	0.027	0.042	0.076	0.021	0.092	0.038	0.13	0.03	0.24	0.051	0.11	0.087	0.028	0.027
3	0.048	0.076	0.15	0.027	0.046	0.039	0.052	0.04	0.21	0.073	0.075	0.095	0.044	0.021
4	0.029	0.082	0.079	0.04	0.075	0.044	0.099	0.087	0.099	0.045	0.13	0.11	0.031	0.054
5	0.042	0.086	0.16	0.05	0.067	0.077	0.072	0.064	0.1	0.051	0.095	0.057	0.043	0.032
6	0.041	0.088	0.045	0.063	0.075	0.062	0.099	0.14	0.053	0.034	0.094	0.087	0.035	0.079
7	0.068	0.15	0.082	0.035	0.041	0.067	0.058	0.21	0.044	0.039	0.044	0.046	0.048	0.058
8	0.044	0.096	0.065	0.083	0.056	0.13	0.069	0.076	0.11	0.028	0.027	0.11	0.066	0.043
9	0.055	0.062	0.063	0.083	0.092	0.098	0.072	0.059	0.039	0.086	0.041	0.11	0.047	0.09
10	0.064	0.11	0.084	0.13	0.075	0.056	0.092	0.099	0.042	0.048	0.041	0.054	0.037	0.068
11	0.077	0.084	0.11	0.034	0.089	0.047	0.11	0.055	0.12	0.029	0.072	0.081	0.042	0.045
12	0.048	0.068	0.17	0.046	0.076	0.055	0.073	0.069	0.11	0.049	0.1	0.06	0.033	0.035
13	0.058	0.078	0.078	0.16	0.062	0.07	0.067	0.069	0.055	0.047	0.05	0.078	0.048	0.085

Step 3000 attention weights at layer 1 head 0, with accuracy 0.596

Step 3000 attention weights at layer 1 head 1, with accuracy 0.596

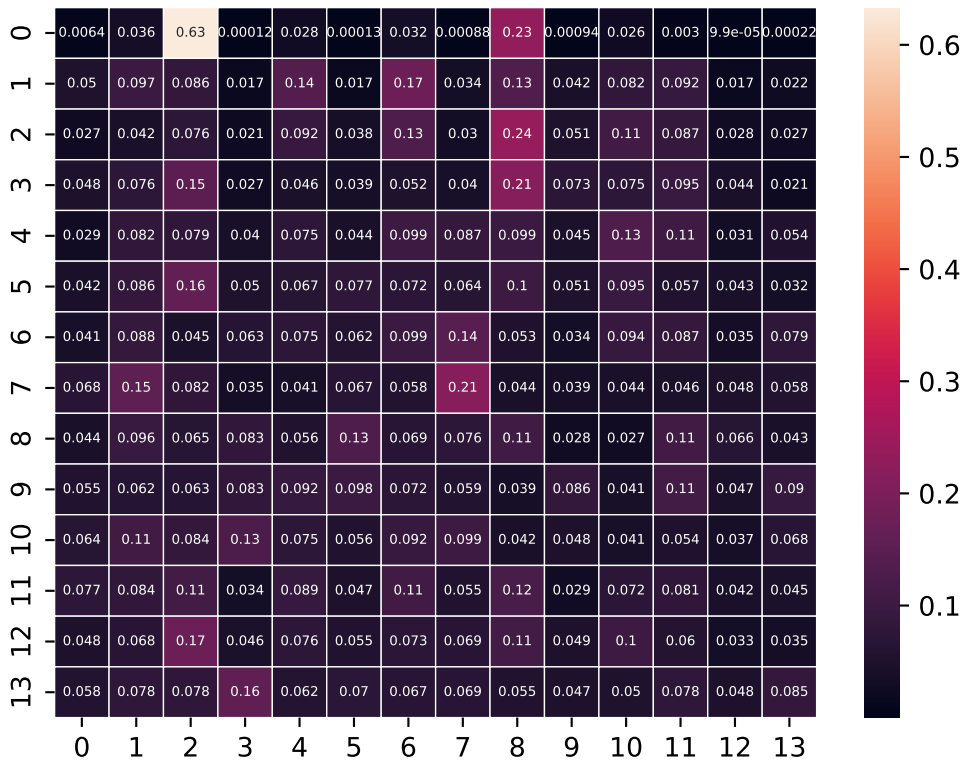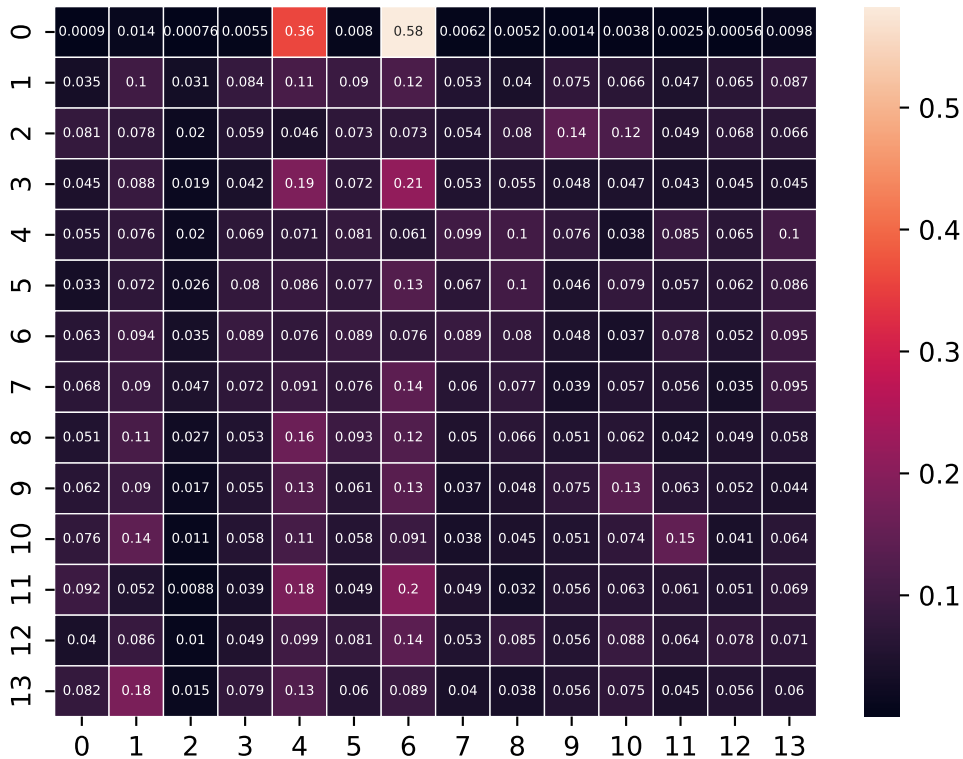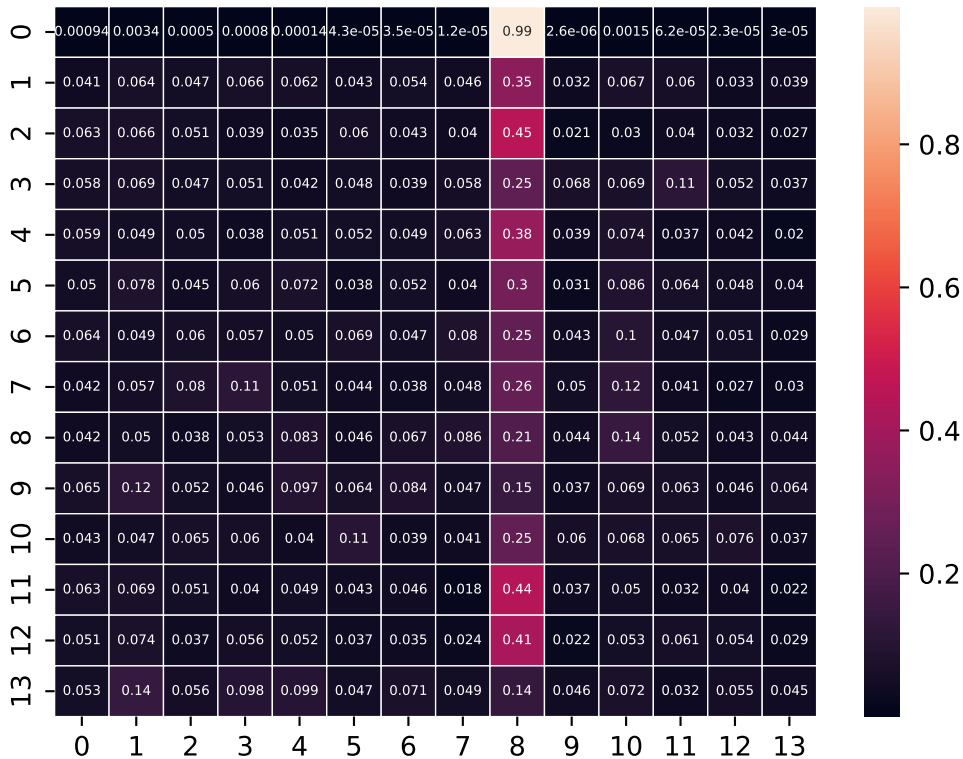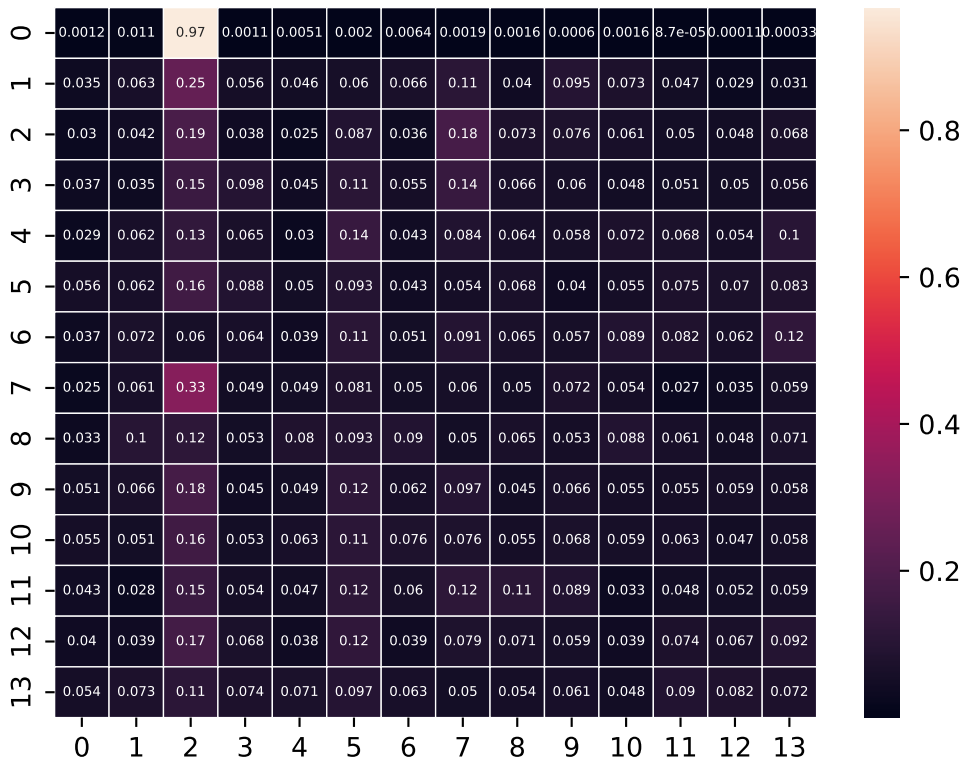|	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.00094	0.0034	0.0005	0.0008	0.00014	4.3e-05	3.5e-05	1.2e-05	0.99	2.6e-06	0.0015	6.2e-05	2.3e-05	3e-05
1	0.041	0.064	0.047	0.066	0.062	0.043	0.054	0.046	0.35	0.032	0.067	0.06	0.033	0.039
2	0.063	0.066	0.051	0.039	0.035	0.06	0.043	0.04	0.45	0.021	0.03	0.04	0.032	0.027
3	0.058	0.069	0.047	0.051	0.042	0.048	0.039	0.058	0.25	0.068	0.069	0.11	0.052	0.037
4	0.059	0.049	0.05	0.038	0.051	0.052	0.049	0.063	0.38	0.039	0.074	0.037	0.042	0.02
5	0.05	0.078	0.045	0.06	0.072	0.038	0.052	0.04	0.3	0.031	0.086	0.064	0.048	0.04
6	0.064	0.049	0.06	0.057	0.05	0.069	0.047	0.08	0.25	0.043	0.1	0.047	0.051	0.029
7	0.042	0.057	0.08	0.11	0.051	0.044	0.038	0.048	0.26	0.05	0.12	0.041	0.027	0.03
8	0.042	0.05	0.038	0.053	0.083	0.046	0.067	0.086	0.21	0.044	0.14	0.052	0.043	0.044
9	0.065	0.12	0.052	0.046	0.097	0.064	0.084	0.047	0.15	0.037	0.069	0.063	0.046	0.064
10	0.043	0.047	0.065	0.06	0.04	0.11	0.039	0.041	0.25	0.06	0.068	0.065	0.076	0.037
11	0.063	0.069	0.051	0.04	0.049	0.043	0.046	0.018	0.44	0.037	0.05	0.032	0.04	0.022
12	0.051	0.074	0.037	0.056	0.052	0.037	0.035	0.024	0.41	0.022	0.053	0.061	0.054	0.029
13	0.053	0.14	0.056	0.098	0.099	0.047	0.071	0.049	0.14	0.046	0.072	0.032	0.055	0.045

Step 3000 attention weights at layer 1 head 2, with accuracy 0.596

Step 3000 attention weights at layer 1 head 3, with accuracy 0.596

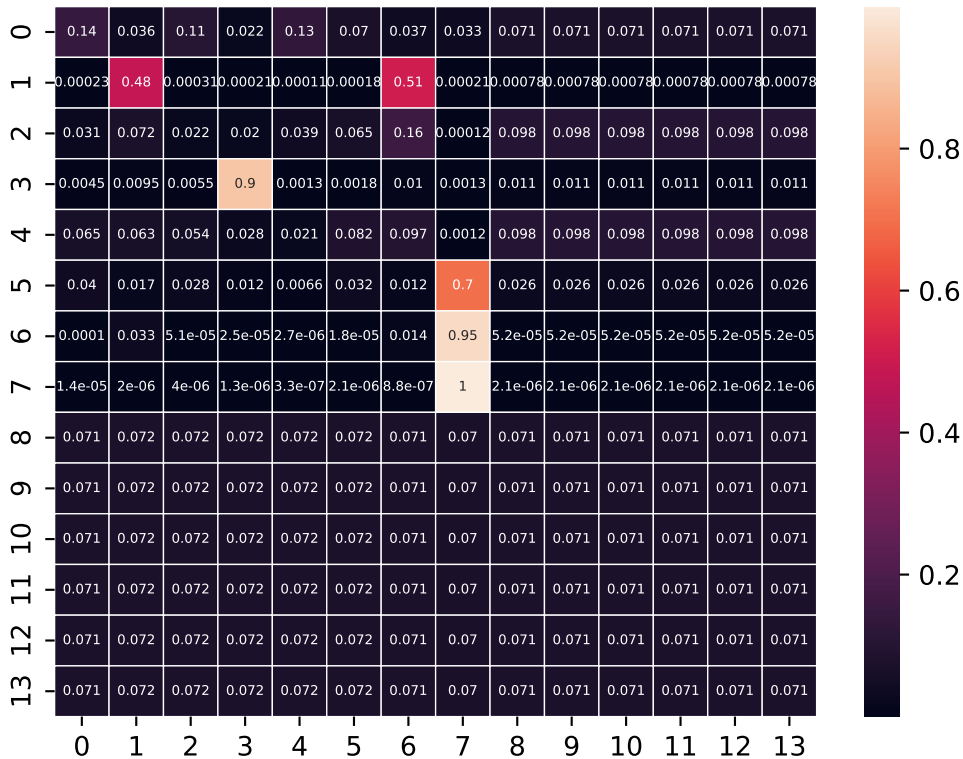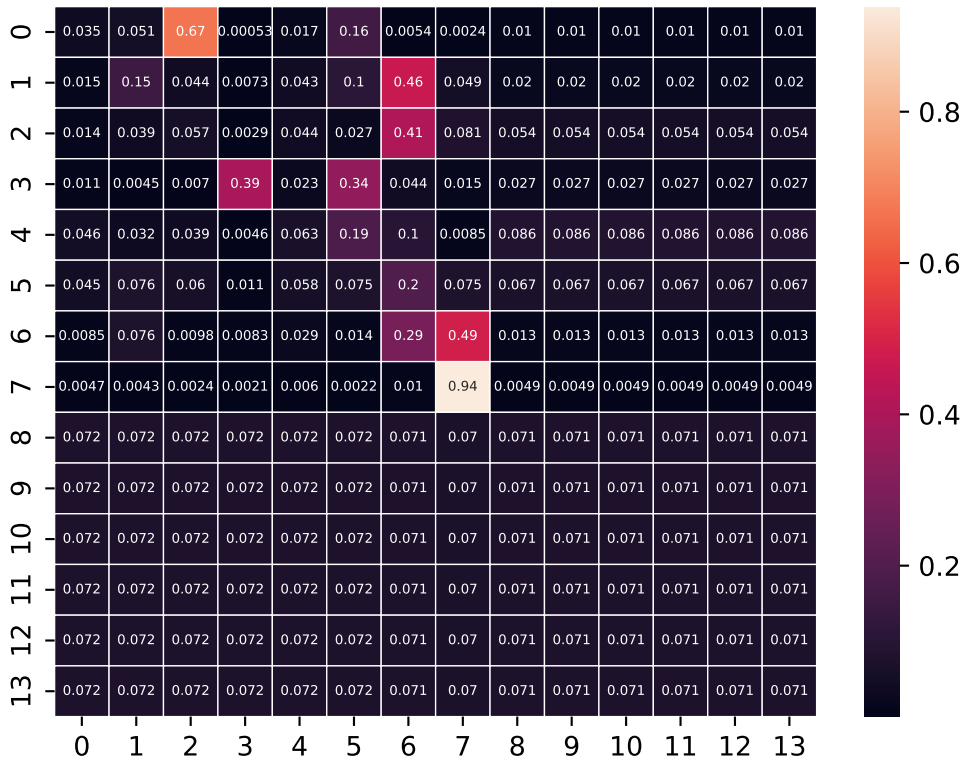|	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.0012	0.011	0.97	0.0011	0.0051	0.002	0.0064	0.0019	0.0016	0.0006	0.0016	8.7e-05	0.00011	0.00033
1	0.035	0.063	0.25	0.056	0.046	0.06	0.066	0.11	0.04	0.095	0.073	0.047	0.029	0.031
2	0.03	0.042	0.19	0.038	0.025	0.087	0.036	0.18	0.073	0.076	0.061	0.05	0.048	0.068
3	0.037	0.035	0.15	0.098	0.045	0.11	0.055	0.14	0.066	0.06	0.048	0.051	0.05	0.056
4	0.029	0.062	0.13	0.065	0.03	0.14	0.043	0.084	0.064	0.058	0.072	0.068	0.054	0.1
5	0.056	0.062	0.16	0.088	0.05	0.093	0.043	0.054	0.068	0.04	0.055	0.075	0.07	0.083
6	0.037	0.072	0.06	0.064	0.039	0.11	0.051	0.091	0.065	0.057	0.089	0.082	0.062	0.12
7	0.025	0.061	0.33	0.049	0.049	0.081	0.05	0.06	0.05	0.072	0.054	0.027	0.035	0.059
8	0.033	0.1	0.12	0.053	0.08	0.093	0.09	0.05	0.065	0.053	0.088	0.061	0.048	0.071
9	0.051	0.066	0.18	0.045	0.049	0.12	0.062	0.097	0.045	0.066	0.055	0.055	0.059	0.058
10	0.055	0.051	0.16	0.053	0.063	0.11	0.076	0.076	0.055	0.068	0.059	0.063	0.047	0.058
11	0.043	0.028	0.15	0.054	0.047	0.12	0.06	0.12	0.11	0.089	0.033	0.048	0.052	0.059
12	0.04	0.039	0.17	0.068	0.038	0.12	0.039	0.079	0.071	0.059	0.039	0.074	0.067	0.092
13	0.054	0.073	0.11	0.074	0.071	0.097	0.063	0.05	0.054	0.061	0.048	0.09	0.082	0.072

Step 3500 attention weights at layer 0 head 0, with accuracy 0.875

Step 3500 attention weights at layer 0 head 1, with accuracy 0.875

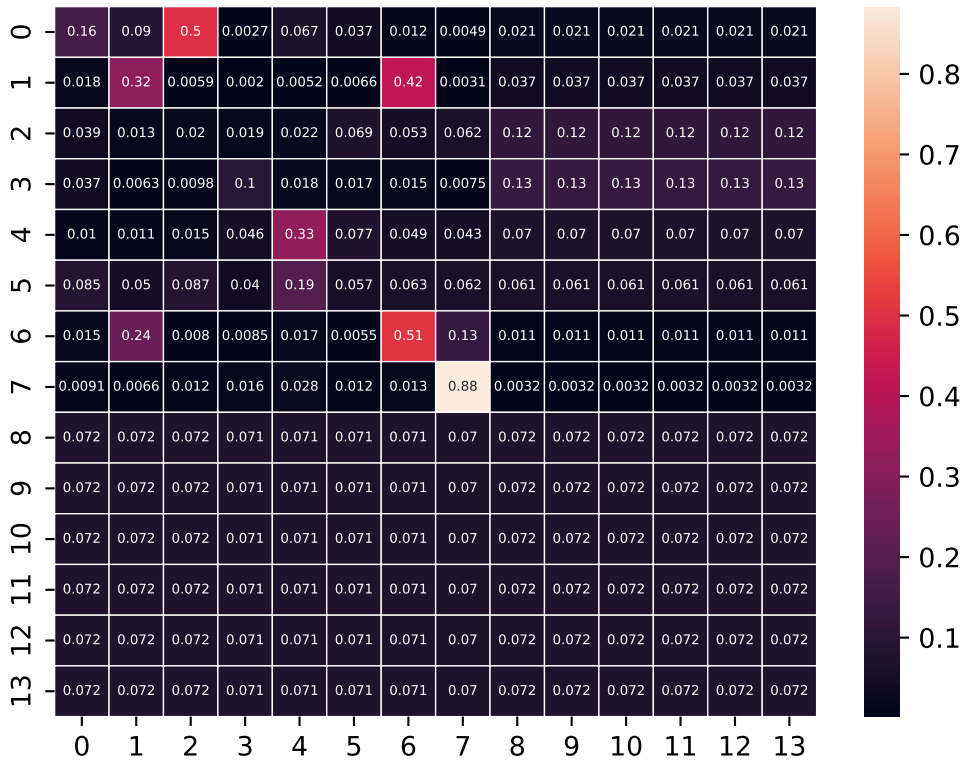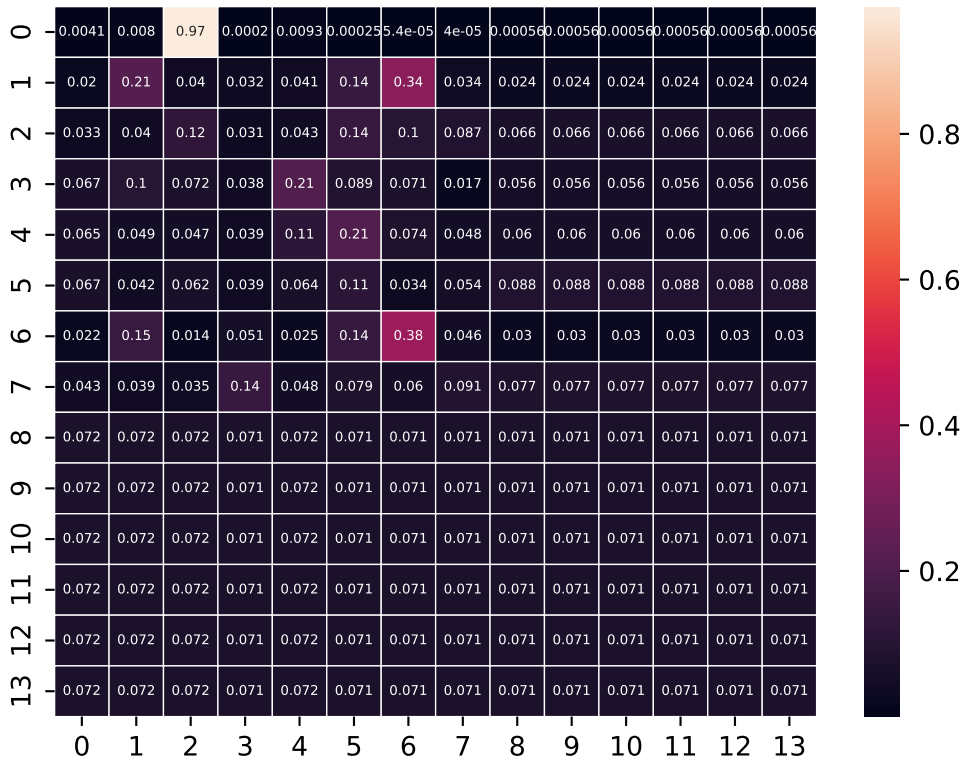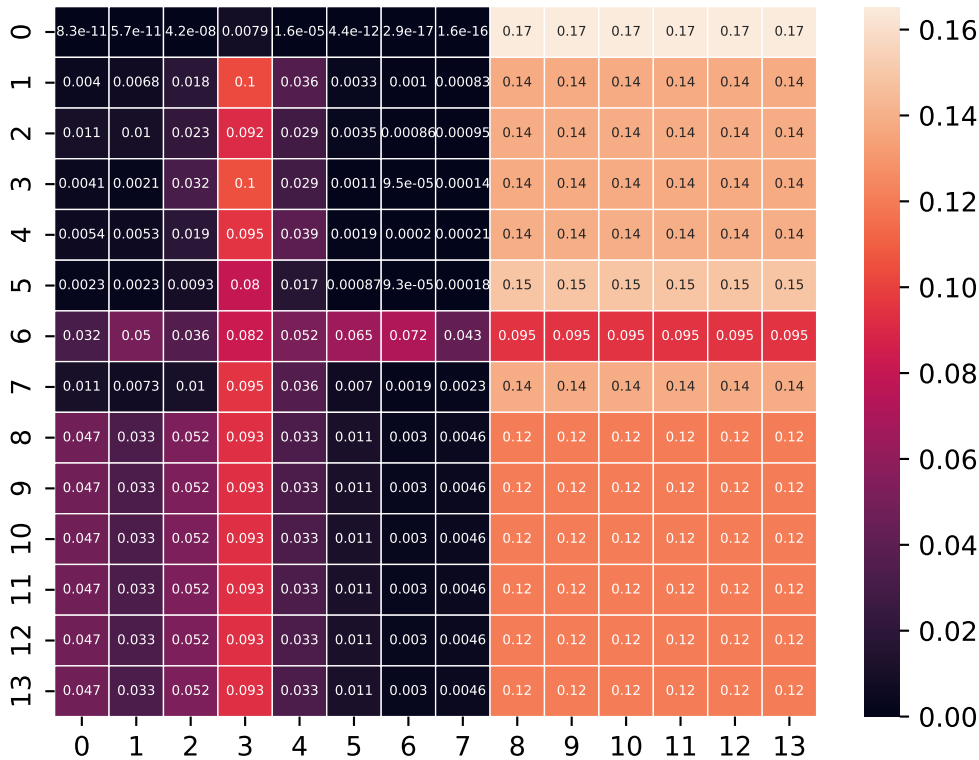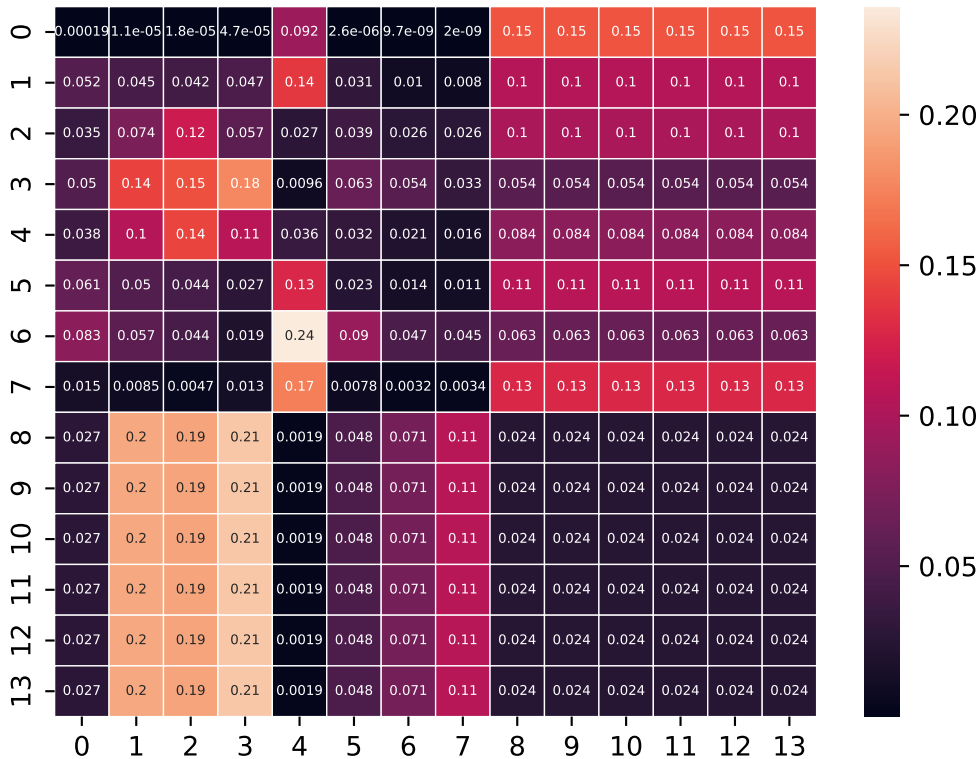|	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.035	0.051	0.67	0.00053	0.017	0.16	0.0054	0.0024	0.01	0.01	0.01	0.01	0.01	0.01
1	0.015	0.15	0.044	0.0073	0.043	0.1	0.46	0.049	0.02	0.02	0.02	0.02	0.02	0.02
2	0.014	0.039	0.057	0.0029	0.044	0.027	0.41	0.081	0.054	0.054	0.054	0.054	0.054	0.054
3	0.011	0.0045	0.007	0.39	0.023	0.34	0.044	0.015	0.027	0.027	0.027	0.027	0.027	0.027
4	0.046	0.032	0.039	0.0046	0.063	0.19	0.1	0.0085	0.086	0.086	0.086	0.086	0.086	0.086
5	0.045	0.076	0.06	0.011	0.058	0.075	0.2	0.075	0.067	0.067	0.067	0.067	0.067	0.067
6	0.0085	0.076	0.0098	0.0083	0.029	0.014	0.29	0.49	0.013	0.013	0.013	0.013	0.013	0.013
7	0.0047	0.0043	0.0024	0.0021	0.006	0.0022	0.01	0.94	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
8	0.072	0.072	0.072	0.072	0.072	0.072	0.071	0.07	0.071	0.071	0.071	0.071	0.071	0.071
9	0.072	0.072	0.072	0.072	0.072	0.072	0.071	0.07	0.071	0.071	0.071	0.071	0.071	0.071
10	0.072	0.072	0.072	0.072	0.072	0.072	0.071	0.07	0.071	0.071	0.071	0.071	0.071	0.071
11	0.072	0.072	0.072	0.072	0.072	0.072	0.071	0.07	0.071	0.071	0.071	0.071	0.071	0.071
12	0.072	0.072	0.072	0.072	0.072	0.072	0.071	0.07	0.071	0.071	0.071	0.071	0.071	0.071
13	0.072	0.072	0.072	0.072	0.072	0.072	0.071	0.07	0.071	0.071	0.071	0.071	0.071	0.071

Step 3500 attention weights at layer 0 head 2, with accuracy 0.875

Step 3500 attention weights at layer 0 head 3, with accuracy 0.875

Step 3500 attention weights at layer 1 head 0, with accuracy 0.875

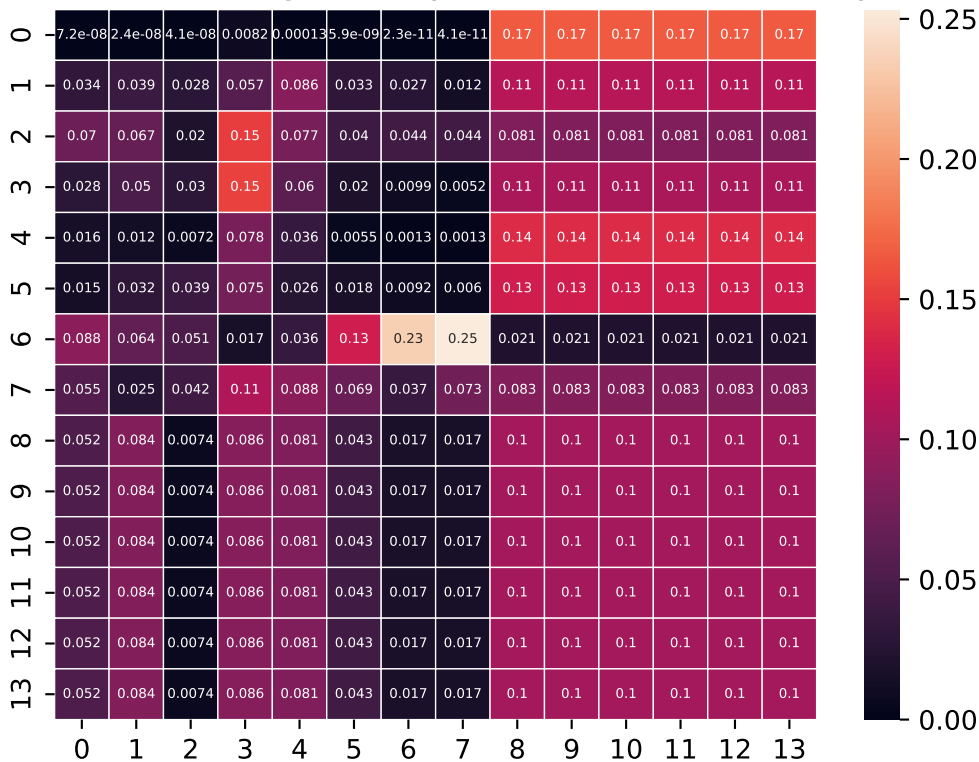|	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	8.3e-11	5.7e-11	4.2e-08	0.0079	1.6e-05	4.4e-12	2.9e-17	1.6e-16	0.17	0.17	0.17	0.17	0.17	0.17
1	0.004	0.0068	0.018	0.1	0.036	0.0033	0.001	0.00083	0.14	0.14	0.14	0.14	0.14	0.14
2	0.011	0.01	0.023	0.092	0.029	0.0035	0.00086	0.00095	0.14	0.14	0.14	0.14	0.14	0.14
3	0.0041	0.0021	0.032	0.1	0.029	0.0011	9.5e-05	0.00014	0.14	0.14	0.14	0.14	0.14	0.14
4	0.0054	0.0053	0.019	0.095	0.039	0.0019	0.0002	0.00021	0.14	0.14	0.14	0.14	0.14	0.14
5	0.0023	0.0023	0.0093	0.08	0.017	0.00087	9.3e-05	0.00018	0.15	0.15	0.15	0.15	0.15	0.15
6	0.032	0.05	0.036	0.082	0.052	0.065	0.072	0.043	0.095	0.095	0.095	0.095	0.095	0.095
7	0.011	0.0073	0.01	0.095	0.036	0.007	0.0019	0.0023	0.14	0.14	0.14	0.14	0.14	0.14
8	0.047	0.033	0.052	0.093	0.033	0.011	0.003	0.0046	0.12	0.12	0.12	0.12	0.12	0.12
9	0.047	0.033	0.052	0.093	0.033	0.011	0.003	0.0046	0.12	0.12	0.12	0.12	0.12	0.12
10	0.047	0.033	0.052	0.093	0.033	0.011	0.003	0.0046	0.12	0.12	0.12	0.12	0.12	0.12
11	0.047	0.033	0.052	0.093	0.033	0.011	0.003	0.0046	0.12	0.12	0.12	0.12	0.12	0.12
12	0.047	0.033	0.052	0.093	0.033	0.011	0.003	0.0046	0.12	0.12	0.12	0.12	0.12	0.12
13	0.047	0.033	0.052	0.093	0.033	0.011	0.003	0.0046	0.12	0.12	0.12	0.12	0.12	0.12

Step 3500 attention weights at layer 1 head 1, with accuracy 0.875

Step 3500 attention weights at layer 1 head 2, with accuracy 0.875

Step 3500 attention weights at layer 1 head 3, with accuracy 0.875