Experiments on connection sparseness without retraining

实验一：prune 掉绝对值小于阈值的 connection，观察 WER 与阈值的关系

实验数据：

1	variable: abs threshold				
	layer pruned: all				
	pruning method: abs				
	pruning threshold: ?				
group	pruned layer	pruning method	pruning threshold	sparse rate	best_wer_tgpr
10	all	abs	0.05	26.6452	8.33
30	all	abs	0.06	31.3210	8.54
16	all	abs	0.07	35.7423	8.85
9	all	abs	0.08	39.9287	10.46
15	all	abs	0.09	43.9048	25.10
8	all	abs	0.10	47.6445	73.91
7	all	abs	0.15	63.4790	99.96

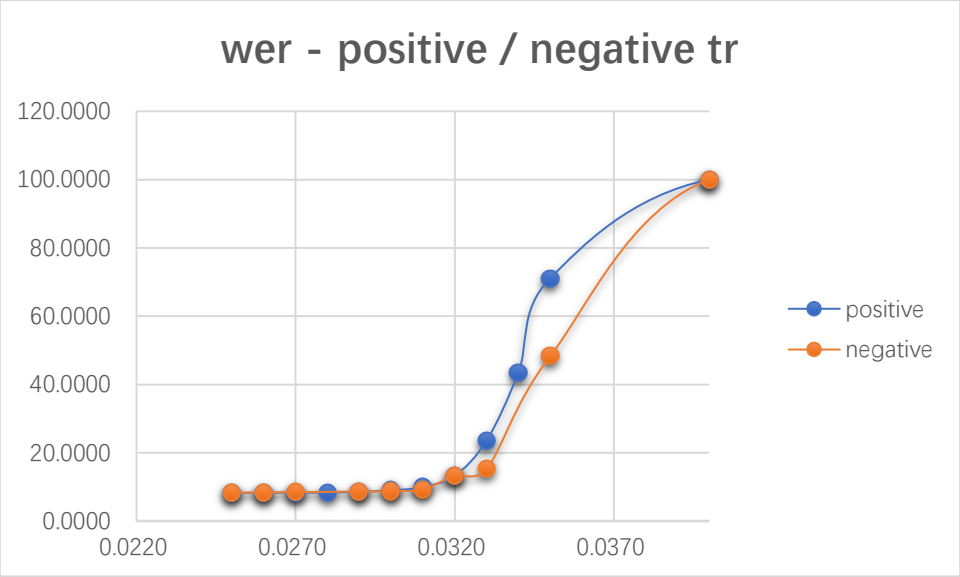
实验二、实验三：prune 绝对值小于给定阈值的大于/小于 0 的 connection，观察 WER 与阈值的关系

实验数据：

2	variable: positive threshold				
	layer pruned: all				
	pruning method: +				
	pruning threshold: ?				
group	pruned layer	pruning method	pruning threshold	sparse rate	best_wer_tgpr
49	all	+	0.0250	6.8878	8.2200
51	all	+	0.0260	7.1543	8.2900
33	all	+	0.0270	7.4196	8.3400
48	all	+	0.0280	7.6838	8.3500
46	all	+	0.0290	7.9459	8.5900
21	all	+	0.0300	8.2089	8.9800
44	all	+	0.0310	8.4679	9.9600
42	all	+	0.0320	8.7283	13.2300
28	all	+	0.0330	8.9857	23.5500
40	all	+	0.0340	9.2408	43.4700
27	all	+	0.0350	9.4953	71.0900
25	all	+	0.0400	10.7511	100.0000
23	all	+	0.0500	13.1516	100.0000
20	all	+	0.0700	17.5480	100.0000
19	all	+	0.0800	19.5568	100.0000
18	all	+	0.0900	21.4476	100.0000

3	variable: negative threshold				
	layer pruned: all				
	pruning method: -				
	pruning threshold: ?				
group	pruned layer	pruning method	pruning threshold	sparse rate	best_wer_tgpr
50	all	-	0.0250	6.9676	8.3000
52	all	-	0.0260	7.2389	8.3600
34	all	-	0.0270	7.5109	8.4700
47	all	-	0.0290	8.0517	8.6400
22	all	-	0.0300	8.3229	8.7300
45	all	-	0.0310	8.5927	9.2900
43	all	-	0.0320	8.8645	13.2300
29	all	-	0.0330	9.1317	15.2000
39	all	-	0.0350	9.6597	48.2800

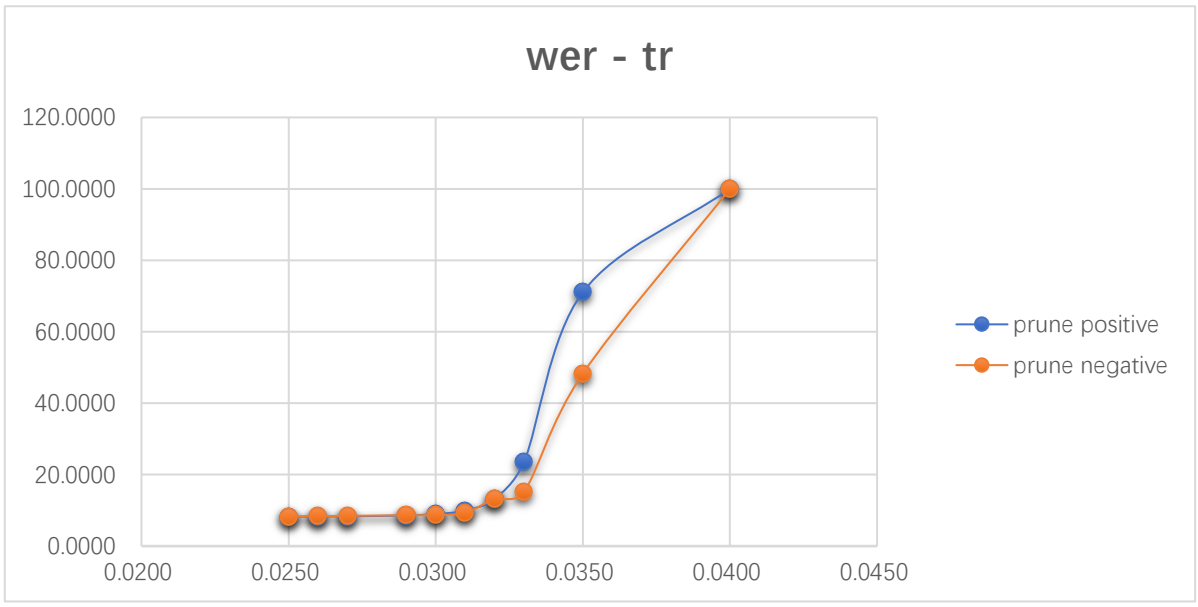
26	all	-	0.0400	10.9585	100.0000
24	all	-	0.0500	13.4937	100.0000
32	all	-	0.0700	18.1943	100.0000
31	all	-	0.0900	22.4572	100.0000

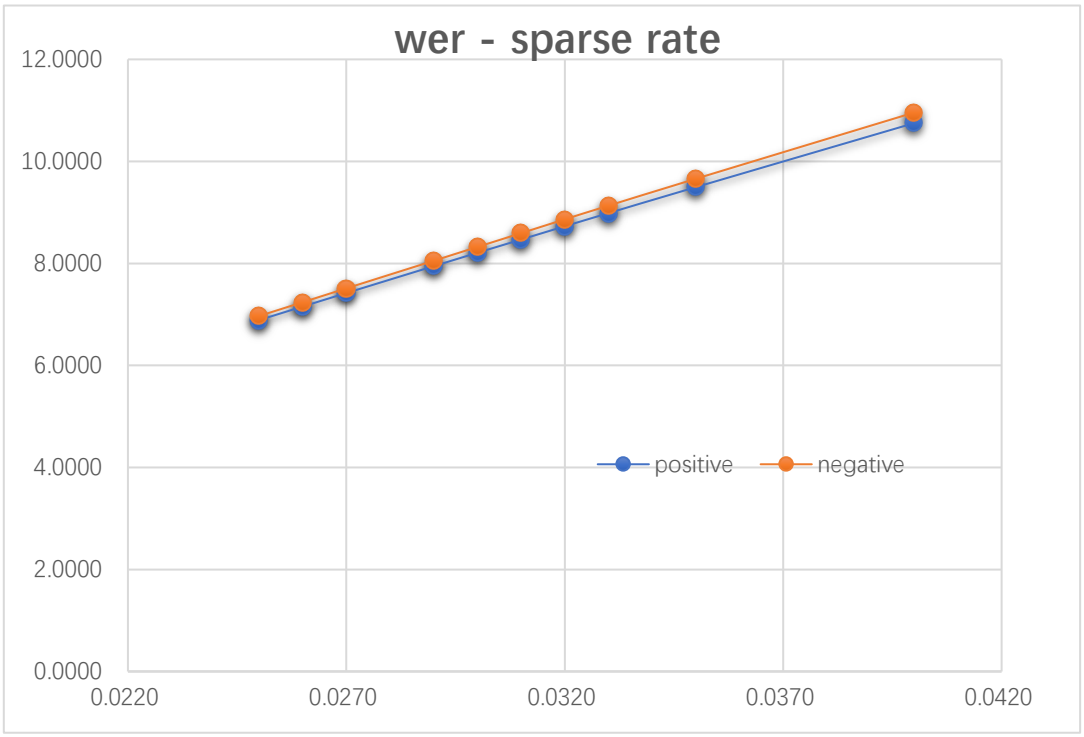


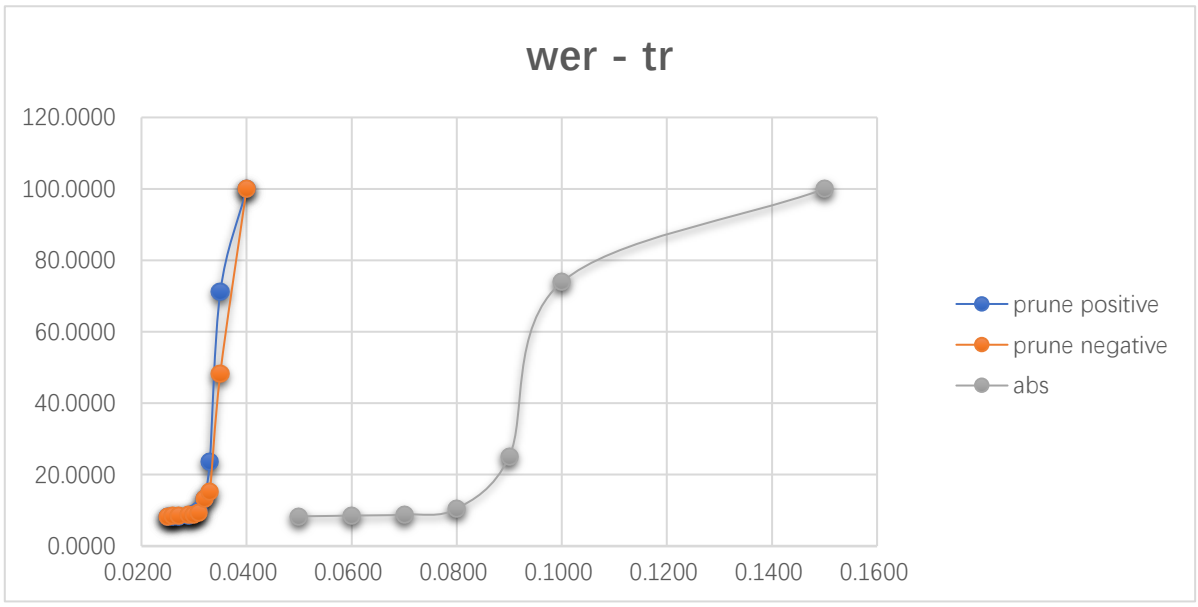
实验四：对比 prune 大于 0/小于 0 的值的效果

实验数据：

4	variable: pruning method				
	layer pruned: all				
	pruning method: ?				
	pruning threshold: ?				
group	pruned layer	pruning method	pruning threshold	sparse rate	best_wer_tgpr
49	all	+	0.0250	6.8878	8.2200
51	all	+	0.0260	7.1543	8.2900
33	all	+	0.0270	7.4196	8.3400
46	all	+	0.0290	7.9459	8.5900
21	all	+	0.0300	8.2089	8.9800
44	all	+	0.0310	8.4679	9.9600
42	all	+	0.0320	8.7283	13.2300
28	all	+	0.0330	8.9857	23.5500
27	all	+	0.0350	9.4953	71.0900
25	all	+	0.0400	10.7511	100.0000
group	pruned layer	pruning method	pruning threshold	sparse rate	best_wer_tgpr
50	all	-	0.0250	6.9676	8.3000
52	all	-	0.0260	7.2389	8.3600
34	all	-	0.0270	7.5109	8.4700
47	all	-	0.0290	8.0517	8.6400
22	all	-	0.0300	8.3229	8.7300
45	all	-	0.0310	8.5927	9.2900
43	all	-	0.0320	8.8645	13.2300
29	all	-	0.0330	9.1317	15.2000
39	all	-	0.0350	9.6597	48.2800
26	all	-	0.0400	10.9585	100.0000


wer - tr
120.0000
100.0000
80.0000
60.0000
40.0000
20.0000
0.0000
0.0200
0.0250
0.0300
0.0350
0.0400
0.0450
prune positive
prune negative


wer - sparse rate
12.0000
10.0000
8.0000
6.0000
4.0000
2.0000
0.0000
0.0220
0.0270
0.0320
0.0370
0.0420
positive
negative



结论：

1. 在 WER 随着阈值迅速增加的阶段，按照某一阈值 prune 掉负值比正值有更高的稀疏度和更低的 WER
2. 设定的阈值相同时，与只 prune 掉正/负值相比，同时 prune 掉正值和负值可以得到更好的 WER

实验五：对比在不同的层设置相同的阈值进行 prune 的效果

实验数据：

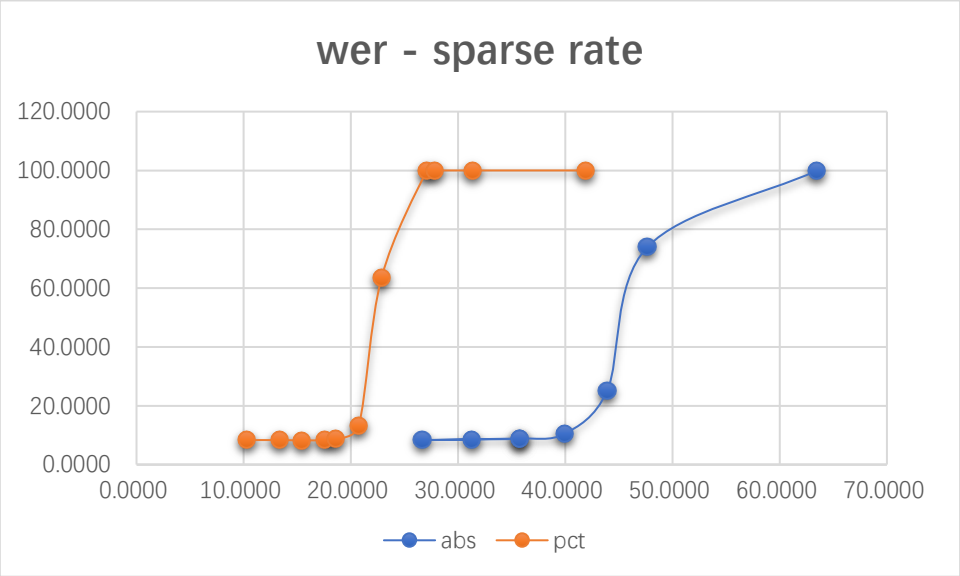
5	variable: layer				
	layer pruned: ?				
	pruning method: abs				
	pruning threshold: 0.15				
group	pruned layer	pruning method	pruning threshold	sparse rate	best_wer_tgpr
5	affine1	abs	0.1500	11.4362	8.9905
35	affine2	abs	0.1500	9.7697	8.7900
36	affine3	abs	0.1500	4.5725	10.6900
37	affine4	abs	0.1500	11.5901	11.2700
38	affine5	abs	0.1500	5.5003	8.8900
2	affine6	abs	0.1500	7.6499	31.6108
3	final-affine	abs	0.1500	9.7697	8.7637

结论：

在不同层进行 prune 操作有着不同的效果。例如，affine3（或 affine6）与 affine1 相比，设定了相同的阈值 0.15，却得到了更低的稀疏度和更高的 WER，显然在 affine1 进行 prune 操作既有利于稀疏度的增加，也有利于 WER 的控制。

实验六：两种 **prune** 方案的比较

6	variable: pct threshold				
	layer pruned: all				
	pruning method: pct				
	pruning threshold: ?				
group	**pruned layer**	**pruning method**	**pruning threshold**	**sparse rate**	**best_wer_tgpr**
55	all	pct	0.1000	10.2606	8.3400
58	all	pct	0.1300	13.3710	8.3200
54	all	pct	0.1500	15.4653	8.2400
56	all	pct	0.1700	17.5622	8.2900
57	all	pct	0.1800	18.6232	8.6400
53	all	pct	0.2000	20.7268	13.2000
59	all	pct	0.2200	22.8501	63.4600
60	all	pct	0.2400	27.0895	100.0000
61	all	pct	0.2665	27.7733	100.0000
63	all	pct	0.3000	31.3323	100.0000
62	all	pct	0.3993	41.9482	100.0000
group	**pruned layer**	**pruning method**	**pruning threshold**	**sparse rate**	**best_wer_tgpr**
17	all	abs	0.01	35.7423	8.8600
10	all	abs	0.05	26.6452	8.3260
30	all	abs	0.0600	31.3210	8.5400
16	all	abs	0.07	35.7423	8.8469
9	all	abs	0.08	39.9287	10.4570
15	all	abs	0.09	43.9048	25.1044
8	all	abs	0.1000	47.6445	73.9055
7	all	abs	0.1500	63.4790	99.9559



结论：

prune 方式一：每层 prune 相同比例的值（正值和负值都按照该比例进行 prune 操作）

prune 方式二：对于整个神经网络，按照设定的 connection 的阈值进行 prune 操作

方式二更优。