Tuning parameters for prefix

Marinò Giosuè Cataldo[1]

[1]Dipartimento di Informatica, Università degli Studi di Milano,
Via Celoria 18 Milano, 20133, Italy

February 14, 2020

1 Model

In this tests we used a feed-forward Neural Network with 64 input neurons and 1 output neuron. The fixed settings are:

- The activation function of output neuros is Leaky-ReLU ($\alpha = 0.05$)
- Optimizer = SGD with Momentum ($\mu = 0.9$)
- Initializer for weights connection and bias is Random Normal with $\mu = 0$ and $\sigma = 0.05$
- Stopping Criteria: the mean absolute error/log sum exponential don't improve with patience 10 or number of epochs is 20000
- Fixed Learning Rate

We try to find a better loss function than common Mean Square Error (MSE). So we try also the Log Sum Exponential (LSE) with many learning rate (η) and batch size (mb) on the three dataset we have.

2 Results

Table 1: File3

MSE		**LSE**	
$\eta = 1.0 \times 10^{-5}$			
`mb`	ϵ	`mb`	ϵ
64	149	64	23
128	97	128	16
256	79	256	13
512	72	512	11
$\eta = 1.0 \times 10^{-4}$			
64	12	64	8
128	9	128	8
256	8	256	8
512	8	512	8
$\eta = 1.0 \times 10^{-3}$			
64	8	64	8
128	8	128	8
256	8	256	8
512	8	512	8
$\eta = 1.0 \times 10^{-2}$			
64	8	64	8
128	8	128	8
256	8	256	8
512	8	512	8
$\eta = 1.0 \times 10^{-1}$			
64	8	64	8
128	8	128	8
256	8	256	8
512	8	512	8

Table 2: File7

MSE		**LSE**	
$\eta = 1.0 \times 10^{-5}$			
`mb`	ϵ	`mb`	ϵ
64	44	64	42
128	43	128	42
256	43	256	42
512	43	512	42
1024	43	1024	42
8192	43	8192	42
$\eta = 1.0 \times 10^{-4}$			
64	42	64	42
128	42	128	42
256	42	256	42
512	42	512	42
1024	42	1024	42
8192	42	8192	42
$\eta = 1.0 \times 10^{-3}$			
64	42	64	42
128	42	128	42
256	42	256	42
512	42	512	42
1024	42	1024	42
8192	42	8192	42
$\eta = 1.0 \times 10^{-2}$			
64	42	64	42
128	42	128	43
256	42	256	42
512	43	512	43
1024	42	1024	42
8192	42	8192	42
$\eta = 1.0 \times 10^{-1}$			
64	56	64	55
128	49	128	48
256	45	256	43
512	41	512	41
1024	41	1024	41
8192	42	8192	43

Table 3: File10

MSE		**LSE**	
$\eta = 1.0 \times 10^{-5}$			
`mb`	ϵ	`mb`	ϵ
64	788	64	766
128	788	128	766
256	787	256	766
512	787	512	765
1024	787	1024	765
1048576	787	1048576	765
$\eta = 1.0 \times 10^{-4}$			
64	624	64	623
128	626	128	625
256	628	256	627
512	629	512	628
1024	629	1024	628
1048576	629	1048576	628
$\eta = 1.0 \times 10^{-3}$			
64	626	64	627
128	619	128	619
256	616	256	615
512	620	512	620
1024	619	1024	619
1048576	622	1048576	622
$\eta = 1.0 \times 10^{-2}$			
64	635	64	636
128	636	128	594
256	660	256	659
512	617	512	615
1024	616	1024	614
1048576	620	1048576	621
$\eta = 1.0 \times 10^{-1}$			
64	984	64	979
128	717	128	696
256	661	256	667
512	603	512	**573**
1024	671	1024	716
1048576	636	1048576	633

3 After Model Selection

After comparison between MSE and LSE we found a good configuration with LSE on file10 (mb=512 and $\eta = 1.0 \times 10^{-1}$). With this configuration we try to compress this model with Pruning (table 4), Weight Sharing (table 5) and Pruning+Weight Sharing (table 6). With this three compression we re-train with the same loss of original network (in this case LSE).

Table 4: Pruning

file10, $\eta = 1.0 \times 10^{-1}$, mb=512

$\%Pruning$	ϵ
10	618
20	626
30	628
40	628
50	630
60	744
70	1608
80	8837
90	66552

Table 5: Weight Sharing

file10, $\eta = 1.0 \times 10^{-1}$, mb=512

$Cluster$	ϵ
10	1600
12	925
14	699
16	693

Table 6: Pruning + Weight Sharing

file10, $\eta = 1.0 \times 10^{-1}$, mb=512

$\%Pruning$	ϵ
$Cluster = 10$	
10	1745
20	1859
30	1996
40	2250
50	2624
$Cluster = 12$	
10	966
20	988
30	1016
40	1115
50	1179
$Cluster = 14$	
10	694
20	724
30	758
40	811
50	864
$Cluster = 16$	
10	703
20	675
30	667
40	654
50	631