Deep Preference Neural Network for Move Prediction in Board Games

Canada-France-Iceland Workshop, Reykjavik, July 11th, 2018

Thomas Philip Runarsson

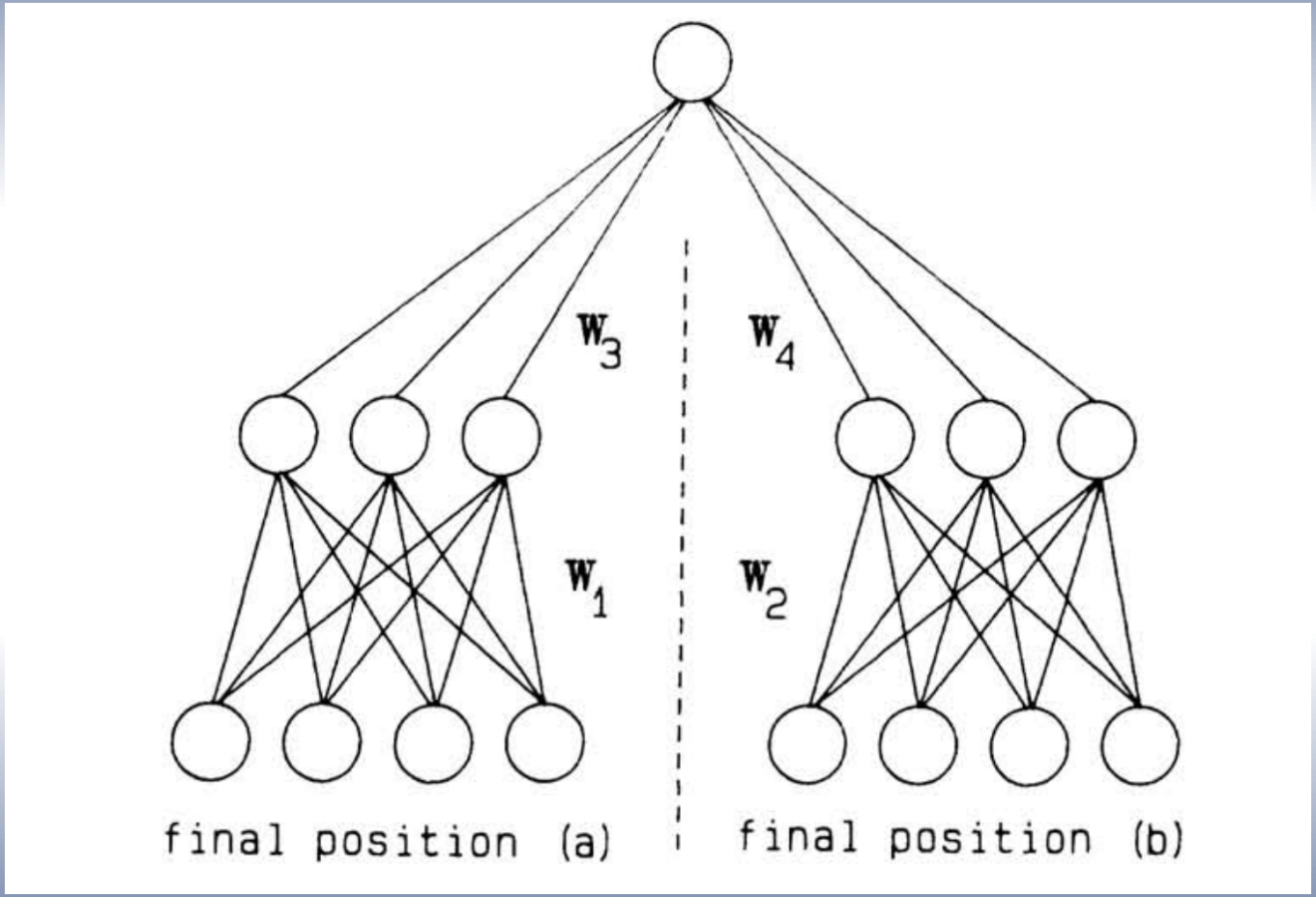
School of Engineering and Natural Science

University of Iceland

Introduction

- Training deep preference neural network using Othello expert games.
- Beat our n-tuple linear SVM using a deep preference networks?
- Based on a twenty year old model by Tesauro (comparison training).

Comparison training (Tesauro)



Deep Preferences Learning

Let DeepNN be some deep neural network of your choice and

$$\boldsymbol{z}_i = \text{DeepNN}(\boldsymbol{x}_i)$$

be the output of the network at its very last layer and the network output (must be anti-symmetric)

$$\hat{y} = \tanh\left(\boldsymbol{w}^\top(\boldsymbol{z}_i - \boldsymbol{z}_j)\right)$$

The training data are preference pair as $\{(\boldsymbol{x}_i, \boldsymbol{x}_j), y_{ij}\}$ with

$$y_{ij} = \begin{cases} 1 & \boldsymbol{x}_i \succ \boldsymbol{x}_j \\ -1 & \boldsymbol{x}_i \prec \boldsymbol{x}_j \end{cases}$$

$\boldsymbol{x}_i \succ \boldsymbol{x}_j \Rightarrow f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j)$, where

$$f(\boldsymbol{x}) = \tanh\left(\boldsymbol{w}^\top \boldsymbol{z}\right) = \tanh\left(\boldsymbol{w}^\top \text{DeepNN}(\boldsymbol{x})\right)$$

Training preference networks

The training of preference networks is challenging since

- we minimize the mean square error, but
- the goal is to select a single best move out of many.

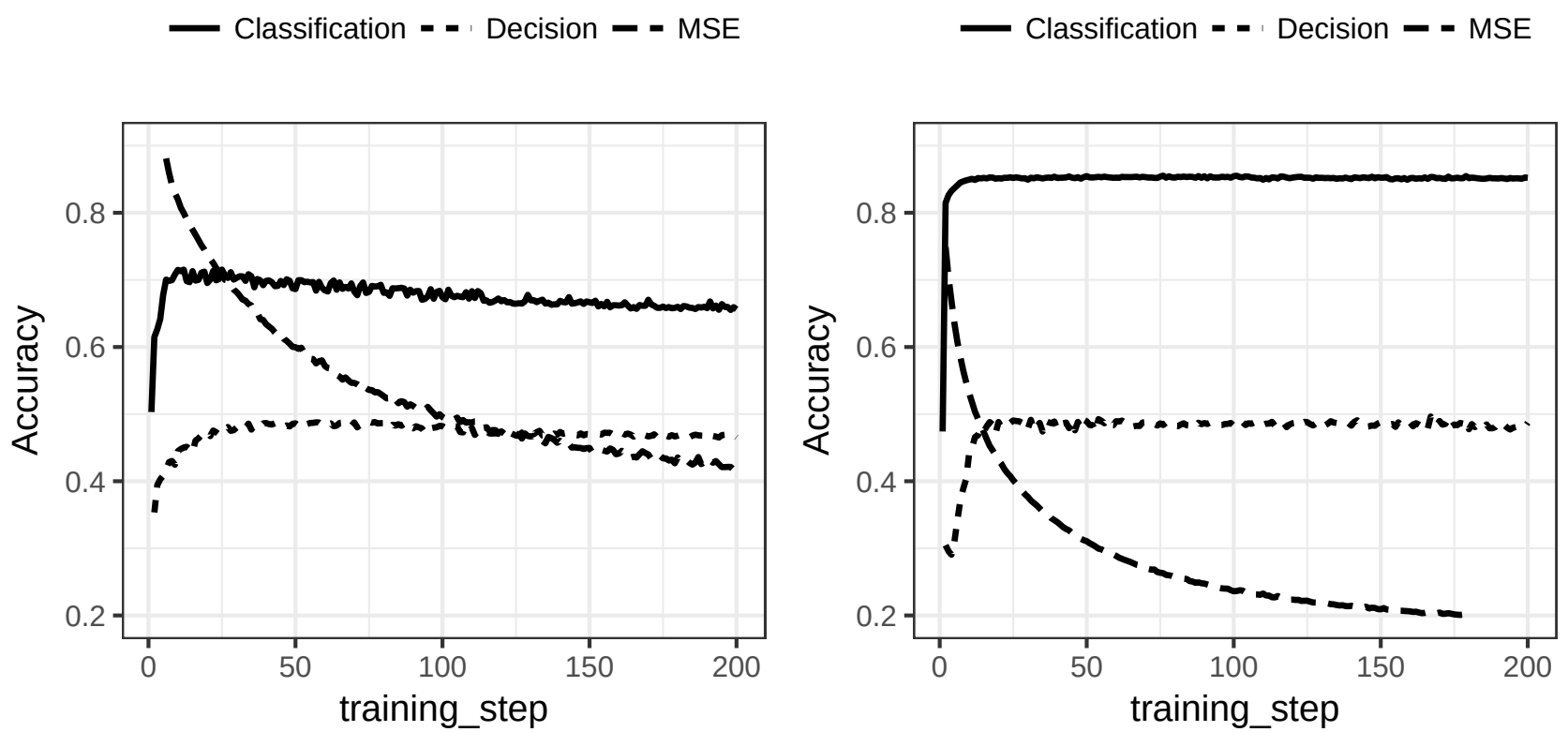
Single-label classification

- Moves selected by the expert players are added to the training set with a positive label.
- The rest are also added to the training set with a negative label.
- Care must be take to have balanced data set.

Deep network architecture

- We have 3 inputs per board square, $(1, 0, 0)$ denotes that the square is occupied by black, $(0, 1, 0)$ as empty and $(0, 0, 1)$ as white.
- The maximum depth of the network is 5 and so the full net, described in terms of the number of units, is denoted by $192 - 4096 - 256 - 128 - 64 - 32 - 1$.
- Same evaluation function for both players, but the color of the board is reversed whenever it is White's turn to move.

Single-label versus preference learning

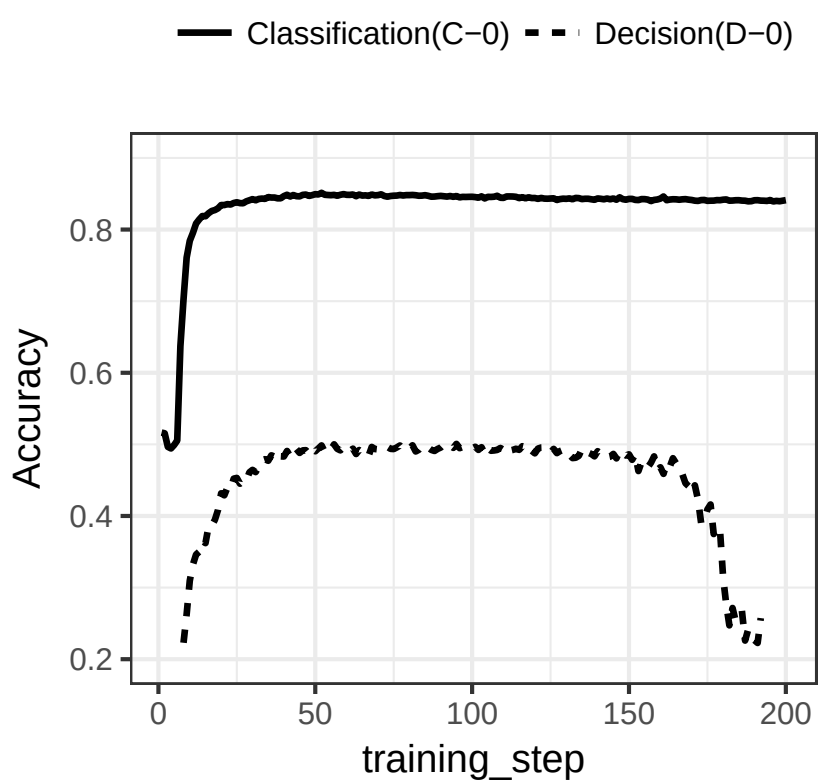


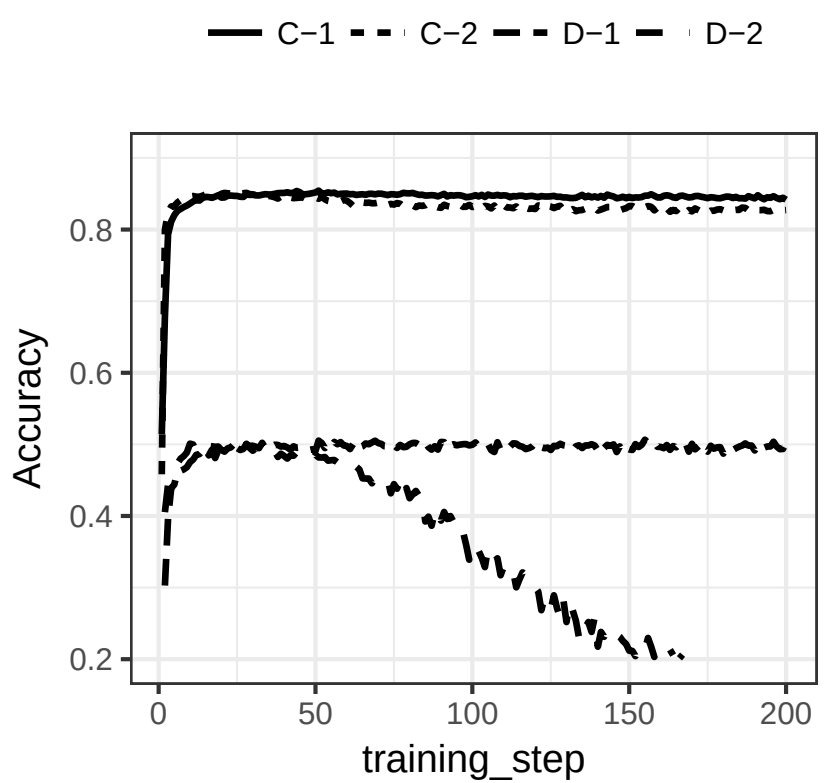
The training mean square error (MSE) and the validation accuracy for classification and decision (move selection) are also depicted for the two training methods.

Depth of the network

#discs	BF	#N	n-tuple	D_1	D_2	D_3	D_4	D_5
1–16	7.1	73133	78.3	62.2	83.9	83.8	83.3	25.4
17–20	11.0	40045	52.4	71.7	68.7	68.5	68.1	16.1
21–24	11.5	42194	49.6	59.1	57.5	57.1	57.5	20.4
25–28	11.9	43796	45.1	47.8	45.6	45.2	45.6	19.3
29–32	11.7	42818	40.5	41.6	40.2	39.7	38.4	19.1
33–36	11.3	41319	40.1	37.2	34.7	34.1	33.7	20.1
37–40	10.6	38318	41.8	37.0	35.0	34.1	33.5	20.2
41–44	9.6	34308	41.5	38.1	35.3	35.6	33.8	20.2
45–48	8.4	29412	43.6	41.0	38.3	34.9	34.8	21.6
49–52	7.1	23784	44.0	44.6	42.6	39.5	38.9	23.4
53–56	5.5	17385	49.0	50.3	46.8	44.6	43.0	23.6
57–60	4.0	10960	53.9	54.8	53.7	51.2	50.6	22.5
61–64	2.5	3411	62.5	60.7	62.8	58.6	59.4	21.4
$\sum$	8.6	(437883)	50.6	49.9	51.7	50.8	50.3	21.0
Wins	against	WPC	73.9	44.9	49.6	48.6	45.2	26.6

Overfitting





The failed test performance (C-0,D-0) of a 5-layered network using a dropout rate of 0.5 on the left and on the right with a successful dropout rate of 0.7 (C-1,D-1). Additionally on the right we show a failed run (C-2,D-2) with a dropout rate of 0.5 applied only to the first layer units.

More games

#discs	BF	#N	n-tuple	D_5^{800}	D_5^{1800}	D_5^{8800}
1–16	7.1	73133	78.3	83.4	86.6	88.6
17–20	11.0	40045	52.4	69.2	76.2	79.1
21–24	11.5	42194	49.6	56.3	62.3	70.0
25–28	11.9	43796	45.1	45.1	53.9	63.2
29–32	11.7	42818	40.5	39.8	44.1	52.6
33–36	11.3	41319	40.1	33.8	37.9	45.6
37–40	10.6	38318	41.8	33.6	38.0	43.8
41–44	9.6	34308	41.5	34.0	38.4	44.2
45–48	8.4	29412	43.6	35.5	39.6	46.0
49–52	7.1	23784	44.0	40.0	41.8	49.2
53–56	5.5	17385	49.0	45.0	47.5	54.2
57–60	4.0	10960	53.9	50.9	52.5	60.4
61–64	2.5	3411	62.5	60.4	61.7	64.6
$\sum$	8.6	(437883)	50.6	50.6	55.3	61.4
Wins	against	WPC	73.9	49.1	57.7	58.5

Summary

- Preference training clearly outperforms single-label classification.
- Some success achieved with the careful use of dropout.
- Better test accuracy obtained when training using more game data.
- Evaluation function learned does not necessarily create a stronger game player when compared to our n-tuple evaluation function.
- We are overfitting?!

Summary

- Preference training clearly outperforms single-label classification.
- Some success achieved with the careful use of dropout.
- Better test accuracy obtained when training using more game data.
- Evaluation function learned does not necessarily create a stronger game player when compared to our n-tuple evaluation function.
- We are overfitting?!