SMALL EXACT NEURAL NETWORKS

N. P. STRICKLAND

Here we discuss some neural networks where one can specify the weights explicitly.

1. Binary boolean functions

We can represent truth values by 0 and 1. There are 16 possible boolean functions of two arguments, and they can all be represented by 2-layer networks. Explicitly, if $x, y \in \{0,1\}$ then $xy = (x+y-1)_+$. Now suppose we have $f\colon \{0,1\}^2 \to \mathbb{R}^m$ with

$$f(0,0) = u_0 \qquad f(0,1) = u_1 \qquad f(1,0) = u_2 \qquad f(1,1) = u_3.$$

We then find that

$$f(x,y) = u_0 + (u_2 - u_0)x + (u_1 - u_0)y + (u_3 - u_2 - u_1 + u_0)xy.$$

This leads to an evident encoding of f as a RELU layer (encoding $(x,y) \mapsto (x,y,xy)$) followed by a linear layer. In particular, we can apply this to the function $f\colon \{0,1\}^2 \to \mathbb{R}^{16}$ which computes all 16 boolean functions simultaneously. This is implemented in the file `bool.py`. On the other hand, we could consider an individual boolean function $f\colon \{0,1\}^2 \to \{0,1\}$. Of the 16 possibilities for f, two are constant. A further 12 can be put in the form $f(x,y) = a \pm (bx+cy+d)_+$. The remaining two are $f(x,y) = x \oplus y = (x-y)_+ + (y-x)_+$ and $f(x,y) = \not(x \oplus y) = 1 - (x-y)_+ - (y-x)_+$. One can check that these cannot be represented in the form $a \pm (bx+cy+d)_+$.

A similar construction computes any function of n boolean variables as a RELU layer with $2^n - 1$ outputs followed by a linear layer. For particular boolean functions it will be more efficient to use a deeper network. When $n = 3$, one example is

$$f(x,y,z) = x \oplus y \oplus z = (x-y-z)_+ + (-x+y-z)_+ + (-x-y+z)_+ + (x+y+z-2)_+.$$

We do not know whether this can be represented with fewer than 4 terms. We can find the above form by training but it seems we need a learning rate larger than the default for this to work.

In fact, experiment suggests that most functions $\{0,1\}^3 \to \{0,1\}$ can be expressed as a combination of 3 RELU terms with all coefficients in $\{0, \pm 1, \pm 2\}$. For example

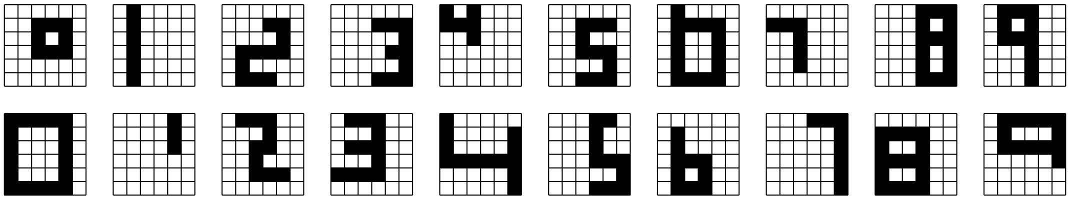
$$(x \wedge y) \oplus z = (x-y)_+ + (2y-x-z)_+ - 2(y-x-z)_+.$$

However, we have not found an expression of this form for $x \oplus y \oplus z$.

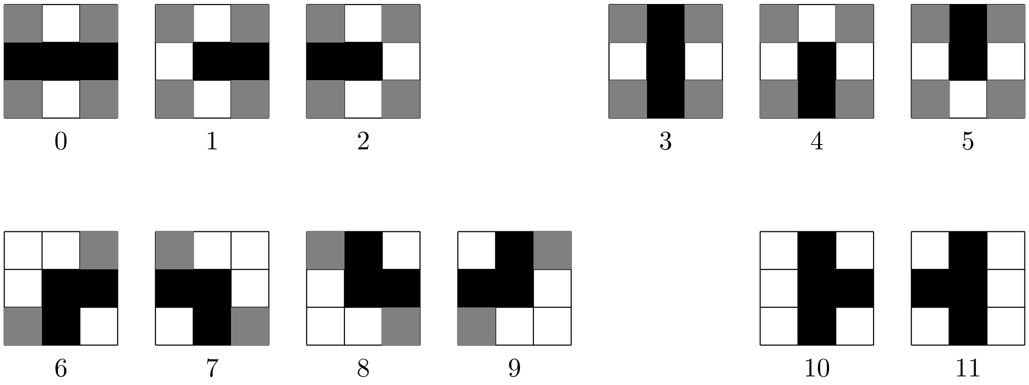
2. Digit recognition

A standard initial exercise in Machine Learning is to recognise images of handwritten digits from the mnist dataset. Here we do an even simpler version of that.

Consider digits drawn as 6×6 pixel images using straight horizontal and vertical lines, as illustrated below.



We can build a neural network to recognise such digits as described below. The first step is a convolutional layer with twelve kernels of shape 3×3, whose effect is illustrated by the pictures below.

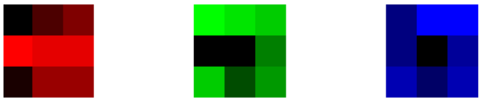


In more detail, we say that a pixel has type i if the surrounding 3×3 block matches the above picture marked i, in the sense that black pixels must be black, white pixels must be white, and grey pixels can be either colour. To interpret this for pixels on the edge, we assume that the image is surrounded by white pixels. For a valid digit image, each black pixel must have type i for some i. If we start with an image represented as an element of $M_6(\{0,1\})$, then the convolutional layer will replace it by an element of $M_6(\{0,1\}^{12})$. If a pixel has type i for some i, then the corresponding entry in the new matrix will be e_i. If a pixel does not have type i for any i then the corresponding entry will be 0.

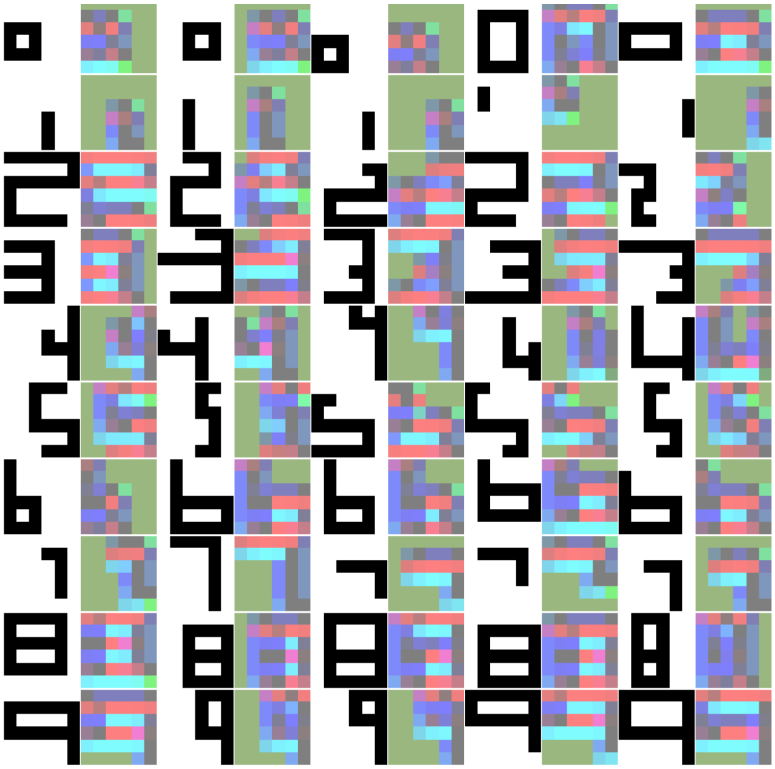
We can now take the sum over all matrix entries to count the total number of pixels of each type. The numbers for types 0 and 3 will depend on the precise geometry of the represented digit, but the numbers for the other types will not. These numbers are sufficient to distinguish all digits except for 2 and 5. To distinguish these, we need to check whether the pixels of types 6 and 8 appear above or below those of types 7 and 9, which is also easily done by a neural network layer. All this is implemented in `digits.py`. The method used there does not always reject non-digits; ideally it should be enhanced to do that.

We have also trained a neural network in the conventional way to perform the same task. We can achieve perfect performance with a smaller model than the handcrafted model described above. Specifically, we have a model consisting of a 3-channel convolutional layer with kernels of size 3×3, followed by a dense layer with 3 outputs, followed by a dense layer with 10 outputs. We train this model to predict the one-hot encoding of the relevant digit, and then apply argmax to predict the digit itself.

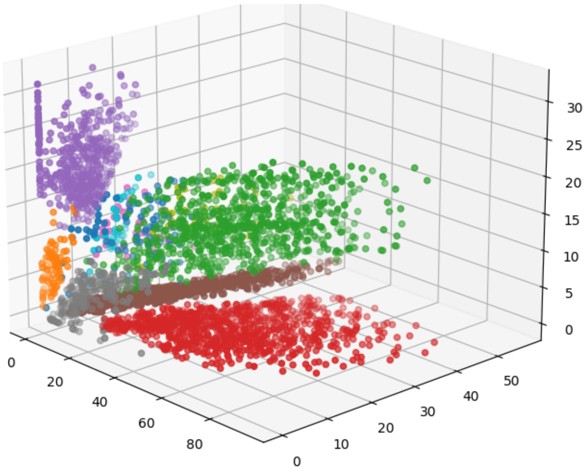
For the convolutional layer, we can interpret the three output channels as red, green and blue, so the layer has the effect of colouring the image. The kernels can be displayed as follows:



The effect on a random selection of examples is shown below.



The output of the second layer gives a point in $\mathbb{R}^3$ for each digit image. These points are best displayed using a 3D viewer which allows the diagram to be rotated interactively. The figure below shows the view from one angle (with different digits marked in different colours).



3. Noughts and crosses

We would like to train a neural network to play noughts and crosses (also known as Tick Tack Toe). Code for this is in the file `ox.py` (and also the Maple worksheet `ox.mw`).

The game is played on a 3×3 board; we number the board positions as $0, \ldots, 8$ in the obvious way. We have two players: player 0 places Os on the board and player 1 places Xs on the board. We represent the game state as an array of shape 9×2, with a 1 at position (i, j) if player j has placed a counter at position i. There are 8 horizontal, vertical or diagonal lines on the board, corresponding to the following sets of positions:

$$W = \{\{0,3,6\}, \{1,4,7\}, \{2,5,8\}, \{0,1,2\}, \{3,4,5\}, \{6,7,8\}, \{0,4,8\}, \{2,4,6\}\}.$$

A player wins if they have covered at least one of these lines.

We are mostly interested in game states such that

(a) The state is empty or could be reached by a legal sequence of moves with the last move played by player 1.

(b) The game has not finished, so neither player has won, and there is at least one blank spot on the board.

Let X be the set of states of this type. We find that $|X| = 4519$. We want a function F that accepts an element of X and generates a probability distribution on the board positions. Player 0 can then sample randomly from that distribution to decide where to play. This may result in an illegal move, where the selected position is already occupied; if so we declare that player 0 has lost.

There is a well-known function F_1 which is optimal: if a player uses that algorithm then they are guaranteed to win or draw.

The most obvious approach is just to train a model to replicate the behaviour of F_1. We have successfully done this, with two hidden layers of size 20 between the input layer (of shape 9×2) and the output layer (of size 9).

We have also attempted to do two different flavours of reinforcement learning: a policy gradient approach and a deep Q-learning approach. Neither has so far been successful. It is not clear whether our implementation is flawed or whether our training hyperparameters are inappropriate.

4. Conway's Game of Life

Conway's Game of Life can be described as follows. Fix N and let X be the set of $N \times N$ matrices with entries in $\{0,1\}$. For $A \in X$ we have $A_{ij} \in \{0,1\}$ for $0 \leq i,j < N$, and we take A_{ij} to be 0 for all other pairs (i,j). In this extended matrix, every entry A_{ij} has 8 adjacent entries; we define $(NA)_{ij}$ to be the sum of these. We then define $(PA)_{ij}$ to be 1 if $0 \leq i,j < N$ and $((A_{ij} = 0$ and $(NA)_{ij} = 3)$ or $(A_{ij} = 1$ and $(NA)_{ij} \in \{2,3\}))$. This defines $P\colon X \to X$, and the idea is to study the sequence $(P^n A)_{n \in \mathbb{N}}$. Here, however, we just want a convolutional neural network to calculate the function P. The first step is to define $QA = A + 2NA$; we then have $(PA)_{ij} = 1$ iff $(QA)_{ij} \in \{5,6,7\}$. Also, if we put

$$f(x) = (x-4)_+ - (x-5)_+ - (x-7)_+ + (x-8)_+,$$

we find that for $x \in \mathbb{Z}$ we have $f(x) = 1$ if $x \in \{5,6,7\}$ and $f(x) = 0$ otherwise; so $(PA)_{ij} = f((QA)_{ij})$. This description of P can be translated in a straightforward way into a convolutional neural network, as is done in the file `life.py`.

5. Resettable running total

Example 5.1. Given an infinite sequence of natural numbers $(x_i)_{i \geq 0}$, we can define $y_0 = x_0$ and

$$y_{i+1} = \begin{cases} y_i + x_{i+1} & \text{if } x_{i+1} > 0 \\ 0 & \text{if } x_{i+1} = 0. \end{cases}$$

In other words, our output is the running total of the inputs, except that the total resets to zero whenever the input is zero. This can be computed by a Gated Recurrent Unit network (GRU), as is done in `rnn.py`.

6. Dropping zeros

Example 6.1. Suppose we have a sequence $x = (x_0, \dots, x_{p-1})$ in $[0,q]$. Let $\delta(x)$ be the sequence obtained by moving all zeros to the end. Hector's solution is essentially as follows:

- For $x \in [0,q]^p$ we put $f_0(x)_i = I(x_i > 0)$.
- For $x \in \{0,1\}^p$ we put $f_1(x)_i = \sum_{j \leq i} x_j$.
- For $x \in [0,p]^p$ we define $f_2(x) \in (\{0,1\}^{p+1})^p$ so that the i'th row of $f_2(x)$ is e_{x_i}.
- For $x \in [0,q]^p$ we put $g_0(x)_i = (f_0(x)_i, f_2(f_1(f_0(x)))_i) \in \{0,1\} \times \{0,1\}^{p+1}$
- For $(a,x) \in \{0,1\} \times \{0,1\}^{p+1}$ we define $f_3(x)_i = (x_i + a - 1)_+$. This has the effect that $f_3(0,e_i) = 0$ and $f_3(1,e_i) = e_i$.
- For $x \in [0,q]^p$ we put $g_1(x)_i = (x_i, f_3(g_0(x))_i) \in [0,q] \times \{0,1\}^p$.

- For $u \in ([0,q] \times \{0,1\}^p)^p$ with $u_i = (a_i, x_i)$
$$f_4(u)_{ij} = \begin{cases} a_0 & \text{if } i = j = 0 \\ 0 & \text{if } i = 0 \text{ and } j > 0 \\ f_4(u)_{i-1,j} & \text{if } i > 0 \text{ and } x_{ij} = 0 \\ a_i & \text{if } i > 0 \text{ and } x_{ij} = 1. \end{cases}$$
- Finally, we define $g_2(x) = f_4(g_1(x))_{p-1}$. One can check that this is $\delta(x)$.

These functions can be computed as follows:

- The function $f_0(x)_i = 1 - (1 - x_i)_+ = x_i - (x_i - 1)_+$ is evidently the composite of a RELU layer followed by a linear layer.
- The function $f_1(x)$ can be computed by a linear layer or a GRU. For the latter, we define $z_i = 1 - x_i$ and $\widehat{h}_i = h_i + 1$ and $h_i = z_i h_{i-1} + (1 - z_i)\widehat{h}_i$; this gives $f_1(x) = h$.
- The formula $f_2(x)_{ij} = (j + 1 - x_i)_+ - 2(j - x_i)_+ + (j - 1 - x_i)_+$ expresses f_2 as a composite of a RELU layer followed by a linear layer.
- The function $f_3(x)$ is evidently a RELU layer.
- The function $f_4(u)$ can be translated quite directly into a GRU layer.
- We now just need plumbing for g_0, g_1 and g_2.

I think that this is a simpler way:

Example 6.2. We make a GRU with the history vector h of the form (m, l) with $m \in \{0,1\}^p$ and $l \in [0,q]^p$. Initially we have $h_0 = (e_0, 0)$. We update with a pair $(x, a) \in [0,q] \times \{0,1\}$, where it is assumed that $a = I(x > 0)$. Put $e = (1, \dots, 1) \in \{0,1\}^p$. Put $r = x \in \mathbb{Z}$. Let $R \in M_p(\mathbb{Z})$ be the right shift matrix and put $\widehat{h} = (Rm, l + rm)$. Put $z = (ae, (m + a - 1)_+) \in \{0,1\}^{2p}$. Put $f((x,a),h) = z\widehat{h} + (1 - z)h$. If we update h_0 successively with $(x_i, I(x_i > 0))$ for $i = 0, \dots, p-1$, we end up with $(0, \delta(x))$. I have implemented this in Maple but not in tensorflow.

7. Miscellaneous other examples

Example 7.1. For integer x we find that
$$\begin{aligned} (n - x)_+ - (n - 1 - x)_+ &= I(x \geq n) \\ 1 - (n - x)_+ + (n - 1 - x)_+ &= I(x < n) \\ (n + 1 - x)_+ - 2(n - x)_+ + (n - 1 - x)_+ &= I(x = n). \end{aligned}$$
It follows that these functions can be computed by 2-layer networks with RELU activation in the first layer. The same therefore holds for composites of these functions with additive functions $\mathbb{Z}^n \to \mathbb{Z}$.

It seems that standard training algorithms struggle to find these weights.

Example 7.2. Given a list $(x_0, \dots, x_{n-1}) \in \{0,1\}^n$, we can define $y_i = p\left(x_i - \sum_{j<i} x_j\right)$. If $x = 0$ then $y = 0$. Otherwise, if i is least such that $x_i = 1$ then $y = e_i$.

Example 7.3. Suppose that $u \in \{0,1\}^n$, and define $f\colon \{0,1\}^n \to \{0,1\}$ to be the delta function concentrated at u. Put $U = \{i \mid u_i = 1\}$ and $e = (1, \dots, 1) \in \{0,1\}^n$. We then find that
$$\begin{aligned} f(x) &= \left(1 + \sum_{i \in U}(x_i - 1) - \sum_{i \notin U} x_i\right)_+ \\ &= \left(1 + (x - e).u - (e - u).x\right)_+ = \left(1 - e.u + (2u - e).x\right)_+. \end{aligned}$$
Thus, f can be calculated by a one-layer RELU network. This is used as an ingredient in many of the examples described later.

Suppose we wanted to train a network to calculate f. If we work in terms of the variables $2x_i - 1 \in \{1, -1\}$ then a change in u just corresponds to multiplying some weights by -1. Thus, we can restrict attention to the case $u = e$, where $f(x) = (e.x - (n-1))_+$.

Put $X = \{x \in \{0,1\}^{n+1} \mid x_0 = 1\}$ and $e' = (1,\dots,1) \in X$. We want to find $a \in \mathbb{R}^{n+1}$ such that $(a.x)_+ = \delta_{x,e'}$ for all $x \in X$, so we use the loss function

$$L(a) = \sum_{x \in X} ((a.x)_+ - \delta_{x,e'})^2/2.$$

Note that there are actually many solutions a with $L(a) = 0$: they are given by $a_0 = 1 - \sum_{i=1}^n a_i$ with $a_i \geq 1$ for $i = 1,\dots,n$. Any reasonable regularisation penalty should pick out the solution where $a_0 = 1 - n$ and $a_i = 1$ for $i > 0$.

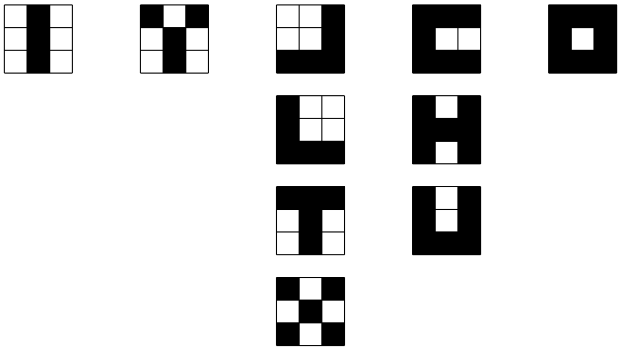
Put $X_i = \{x \in X \mid x_i = 1\}$. We find that the partial derivatives are

$$L_{,i}(a) = \sum_{x \in X_i} (a.x)_+ - H(a.e'),$$

where H is the Heaviside function. When a is close to a solution then the terms $(a.x)_+$ will be small for $x \in X_i \setminus \{e'\}$, and $H(a.e')$ will be 1, and $(a.e')_+ - 1$ will also be small, so $L_{,i}(a)$ will be small. There is code for this analysis in `maple/train_delta.mw`.

Example 7.4. It should be possible to write an explicit network to translate English to German, where the only allowed English words are "zero" and "one" (to be translated as "null" and "eins"). Text processing networks typically start with a word tokenizer, which would make the problem trivial. Instead we should assume that the input is given as a sequence of characters.

Example 7.5. The letters C,H,I,J,L,O,T,U,X,Y can be represented using as 3×3 pixel images. These are shown below, grouped according to the number of black pixels.



It should be possible to write an explicit network to recognise them, and then make a convolutional version to find them in a larger grid. Then we could perhaps try to count the number of appearances of each letter. The file `words_3x3.txt` contains 181 words that use only the letters listed here; we could also try to detect and count these words.

Example 7.6. Consider a piecewise-linear function $f\colon \mathbb{R} \to \mathbb{R}$ with finitely many corners, so there are real numbers $a_1 < \cdots < a_r$ and slopes $m_0,\dots,m_r$ and a constant c such that

(a) $f(x) = m_0 x + c$ on $(-\infty, a_1]$
(b) $f(x)$ has slope m_i on $[a_i, a_{i+1}]$ for $i = 1,\dots,r-1$
(c) $f(x)$ has slope m_r on $[a_r, \infty)$.

Then

$$f(x) = c + m_0 x + \sum_{i=1}^r (m_i - m_{i-1})(x - a_i)_+.$$

However, the situation with piecewise-linear functions in more than one variable seems unclear.