Deep Learning Crash Course

- Single Layer Perceptrons
- Multiple Layer Perceptrons
- Convolutional Neural Nets
Binary Classification

Two classes shown as different colors:

- The label \( y \in \{-1,1\} \) or \( y \in \{0,1\} \).
- The samples with label 1 are called positive samples.
- The samples with label -1 or 0 are called negative samples.
Parameterizing Lines

Equation of a line

\[(u, v) \in \mathbb{R}^2, au + bv + c = 0\]

Normal vector

\[
n = \frac{1}{\sqrt{a^2 + b^2}} [a, b]
\]

\([a, b, c]\) and \(\frac{1}{\sqrt{a^2 + b^2}}[a, b, c]\) define the same line.
Normalized Parameterization

Equation of a line

\[(u, v) \in \mathbb{R}^2, au + bv + c = 0\]
with \(a^2 + b^2 = 1\)

Normal vector

\[\mathbf{l} = [a, b, c]\]

\[\mathbf{n} = [a, b]\]
Signed Distance

\[ n = [w_1, w_2] \]
\[ h = w_0 + w_1 x_1 + w_2 x_2 \]
\[ = \tilde{w} \cdot \tilde{x} \]

\( \tilde{x} = [1, x_1, x_2] \)

\( \tilde{w} = [w_0, w_1, w_2] \) with \( w_1^2 + w_2^2 = 1 \)

Notation:
\( x = [x_1, x_2] \)
\( \tilde{x} = [1, x_1, x_2] \)

h=0: Point is on the line.
h>0: Point in the normal’s direction.
h<0: Point in the other direction.
Signed Distance in 3D

\[ x = [x, y, z] \quad \tilde{x} = [1, x, y, z] \]

\[ x \in \mathbb{R}^3, \quad z = ax + by + c \]

\[ \tilde{w} = [w_0, w_1, w_2, w_3] \]

\[ \tilde{w} \cdot \tilde{x} = 0 \]

Signed distance \( h = \tilde{w} \cdot \tilde{x} \) if \( w_1^2 + w_2^2 + w_3^2 = 1 \).
Signed Distance in N Dimensions

\[ \mathbf{\tilde{w}} = [w_0, w_1, \ldots, w_n] \text{ with } \sum_{i=1}^{N} w_i^2 = 1 \]

\[ \mathbf{\tilde{x}} = [1, x_1, \ldots, x_n] \]

Notation:

\[ \mathbf{x} = [x_1, \ldots, x_n] \]
\[ \mathbf{\tilde{x}} = [1, x_1, \ldots, x_n] \]

Hyperplane:

\[ \mathbf{x} \in \mathbb{R}^n, \quad 0 = \mathbf{\tilde{w}} \cdot \mathbf{\tilde{x}} \]
\[ = w_0 + w_1 x_1 + \ldots w_n x_n \]

Signed distance:

\[ h = \mathbf{\tilde{w}} \cdot \mathbf{\tilde{x}} \]

h=0: Point is on the decision boundary.
h>0: Point on one side.
h<0: Point on the other side.
Binary Classification in N Dimensions

Hyperplane: \( x \in R^N, \tilde{w} \cdot \tilde{x} = 0 \), with \( \tilde{x} = [1 \mid x] \).

Signed distance: \( \tilde{w} \cdot \tilde{x} \), with \( \tilde{w} = [w_0 \mid w] \) and \( ||w|| = 1 \).

Problem statement: Find \( \tilde{w} \) such that

- for all or most positive samples \( \tilde{w} \cdot \tilde{x} > 0 \),
- for all or most negative samples \( \tilde{w} \cdot \tilde{x} < 0 \).
Logistic Regression

\[ y(x; \tilde{w}) = \sigma(\tilde{w} \cdot \tilde{x}) \]

\[ = \frac{1}{1 + \exp(-\tilde{w} \cdot \tilde{x})} \]

- When the noise is Gaussian, this is the maximum likelihood solution.
- \( y(x; \tilde{w}) \) can be interpreted at the probability that \( x \) belongs to positive class.

Given a **training** set \( \{(x_n, t_n)_{1 \leq n \leq N}\} \) minimize

\[ -\sum_n (t_n \ln y(x_n) + (1 - t_n)\ln(1 - y(x_n))) \]

with respect to \( \tilde{w} \).

- When the noise is Gaussian, this is the maximum likelihood solution.
- \( y(x; \tilde{w}) \) can be interpreted at the probability that \( x \) belongs to positive class.
Non Separable Distribution

Positive: $100(x_2 - x_1^2)^2 + (1 - x_1)^2 < 0.5$

Negative: Otherwise

• Logistic regression can handle a few outliers but not a complex non-linear boundary.

• How can we learn a function $y$ such that $y(x; \tilde{w})$ is close to 1 for positive samples and close to 0 or -1 for negative ones?

$\Rightarrow$ Use LOTS of hyperplanes.
Reformulating Logistic Regression

\[ y(x) = \sigma(w \cdot x + b) \]

\[ x = [x_1, x_2, \ldots, x_n]^T \]

\[ w = [w_1, w_2, \ldots, w_n]^T \]
Repeating the Process

\[ h_1 = \sigma(w_1 \cdot x + b_1) \]
\[ w_1 = [w_{11}, w_{12}, w_{13}, w_{14}]^T \]

\[ h_2 = \sigma(w_2 \cdot x + b_2) \]
\[ w_2 = [w_{21}, w_{22}, w_{23}, w_{24}]^T \]

\[ \vdots \]

\[ h_H = \sigma(w_H \cdot x + b_H) \]
\[ w_H = [w_{H1}, w_{H2}, w_{H3}, w_{H4}]^T \]
Repeating the Process

\[ h = \sigma(Wx + b), \]

with \( W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_H \end{bmatrix} \)

and \( b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_H \end{bmatrix} \).
Multi-Layer Perceptron

The process can be repeated several times to create a vector \( h \).

\[
\begin{align*}
  h &= \sigma_1(W_1x + b_1) \\
  y &= \sigma_2(W_2h + b_2)
\end{align*}
\]
The process can be repeated several times to create a vector $h$.

It can then be done again to produce an output $y$.

$\begin{align*}
h &= \sigma_1(W_1x + b_1) \\
y &= \sigma_2(W_2h + b_2)
\end{align*}$

$\longrightarrow$ This output is a **differentiable** function of the weights.
ReLU

- Each node defines a hyperplane.
- The resulting function is piecewise linear affine and continuous.

\[
\begin{align*}
  h &= \sigma_1(W_1 x + b_1) \\
  y &= \sigma_2(W_2 h + b_2)
\end{align*}
\]

\[
\sigma(x) = \max(0, x)
\]
ReLu Behavior

\[ h = ReLu(W_1 x) \]

\[ y = w_2^T h + b_2 \]
One Single Hyperplane

\[ y = \max(\mathbf{w}^\top \mathbf{x} + b, 0) \]

\[ y = 0 \]

\[ \mathbf{w}^\top \mathbf{x} + b = 0 \]

\[ y = \mathbf{w}^\top \mathbf{x} + b \]
Two Hyperplanes

\[ h = \max(Wx + b, 0) \] with \( W = \begin{bmatrix} w_1^T \\ w_2^T \end{bmatrix} \) and \( b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \)

\[ y = w'^T h + b' \]
Three Hyperplanes

\begin{align*}
\begin{cases}
    h &= \max(\mathbf{W}x + b, 0) \\
y &= \mathbf{w}^T h
\end{cases}
\end{align*}

with \( \dim(h) = 3 \)
Sigmoid and Tanh

\[
h = \sigma(W_1x + b_1)
\]
\[
y = \sigma(W_2h + b_2)
\]

- Each node defines a hyperplane.
- The resulting function is continuously differentiable.
Non-Linear Binary Classification

Positive: $100(x_2 - x_1^2)^2 + (1 - x_1)^2 < 0.5$
Negative: Otherwise

Problem statement: Find $\mathbf{w}$ such that

- for all or most positive samples $y(\mathbf{x}; \mathbf{w}) > 0.5$,
- for all or most negative samples $y(\mathbf{x}; \mathbf{w}) < 0.5$.

$y(\mathbf{x}; \mathbf{w})$ is now a non-linear function implemented by the network.
Binary Case

- Let the training set be \( \{(x_n, t_n)_{1 \leq n \leq N}\} \) where \( t_n \in \{0,1\} \) is the class label and let us consider a neural net with a 1D output.

- We write

\[
y_n = \sigma(w_2(\sigma(W_1 x_n + b_1)) + b_2) \in [0.1]
\]

- We want to minimize the binary cross entropy

\[
E(W_1, w_2, b_1, b_2) = \frac{1}{N} \sum_{n=1}^{N} E_n(W_1, w_2, b_1, b_2)
\]

\[
E_n(W_1, w_2, b_1, b_2) = - (t_n \ln(y_n) + (1 - t_n)\ln(1 - y_n))
\]

with respect to the coefficients of \( W_1, w_2, b_1, \) and \( b_2 \).

- \( E \) is a different function and this can be done using a gradient-based technique.
Multi-Class Case

Let the training set be \( \{(x_n, [t_n^1, \ldots, t_n^K])_{1 \leq n \leq N}\} \) where \( t_n^k \in \{0,1\} \) is the probability that sample \( x_n \) belongs to class \( k \).

- We write

\[
y_n = \sigma(W_2(\sigma(W_1x_n + b_1)) + b_2) \in R^K
\]

\[
p_n^k = \frac{\exp(y_n[k])}{\sum_j \exp(y_n[j])}
\]

- We want to minimize the cross entropy

\[
E(W_1, W_2, b_1, b_2) = \frac{1}{N} \sum_{n=1}^{N} E_n(W_1, W_2, b_1, b_2),
\]

\[
E_n(W_1, W_2, b_1, b_2) = - \sum t_n^k \ln(p_n^k),
\]

with respect to the coefficients of \( W_1, W_2, b_1, \) and \( b_2 \).
Problem statement: Find $\tilde{w}$ such that

$$y(x; \tilde{w}) \approx 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

Positive: $100(x_2 - x_1^2)^2 + (1 - x_1)^2 < 0.5$

Negative: Otherwise

$y(x; \tilde{w})$ is now a non-linear function implemented by the network.
From Classification to Regression

Minimize \( \sum_i [t_i \log(\text{sigm}(f(x_i, y_i))) + (1 - t_i) \log(1 - \text{sigm}(f(x_i, y_i))) \right]

with respect to \( W_1, w_2, b_x, b_y, b_z \).

Minimize \( \sum_i (z_i - f(x_i, y_i))^2 \),

with respect to \( W_1, w_2, b_x, b_y, b_z \).
Approximating a Surface

\[ z = 100 \times (y - x^2)^2 + (1 - x)^2 \]

Minimize \( \sum_i (z_i - f(x_i, y_i))^2 \), with respect to \( W_1, W_2, b_x, b_y, b_z \).
Interpolating a Surface

\[ z = 100 \times (y - x^2)^2 + (1 - x)^2 \]

3-node hidden layer

loss: 2.192508e+01
Interpolating a Surface

\[ z = 100 \times (y - x^2)^2 + (1 - x)^2 \]

4-node hidden layer

loss: 1.089789e+00
Adding more Nodes

\[ z = 100 \times (y - x^2)^2 + (1 - x)^2 \]
Adding more Nodes

\[ z = \sin(x) \sin(y) \]

- 2 nodes -> loss 2.61e-01
- 3 nodes -> loss 2.51e-04
- 4 nodes -> loss 3.07e-07
More Complex Surface

\[ I = f(x, y) \]
More Complex Surface

\[ I = f(x, y) \]

- 50 nodes -> loss 3.65e-01
- 100 nodes -> loss 2.50e-01
- 125 nodes -> loss 2.40e-01
- 300 nodes -> loss 1.92e-01
Universal Approximation Theorem

A feedforward network with a linear output layer and at least one hidden layer with any 'squashing' activation function (e.g. logistic sigmoid) can approximate any Borel measurable function (from one finite-dimensional space to another) with any desired nonzero error.

Any continuous function on a closed and bounded set of $\mathbb{R}^n$ is Borel-measurable.

$\rightarrow$ In theory, any reasonable function can be approximated by a one-hidden layer network as long as it is continuous.

[Hornik et al, 1989; Cybenko, 1989]
In Practice

- It may take an exponentially large number of parameters for a good approximation.
- The optimization problem becomes increasingly difficult.

→ The one hidden layer perceptron may not converge to the best solution!
• The network takes as input 28x28 images represented as 784D vectors.
• The output is a 10D vector giving the probability of the image representing any of the 10 digits.
• There are 50’000 training pairs of images and the corresponding label, 10’000 validation pairs, and 5’000 testing pairs.
MNIST Results

- nIn = 784
- nOut = 10
- 20 < hidden layer size < 120

- MLPs have many parameters.
- This has long been a major problem.
  --> Was eventually solved by using GPUs.

- Around 2005, SVMs were often felt to be superior to neural nets.
- This is no longer the case ....
The descriptive power of the net increases with the number of layers.

In the case of a 1D signal, it is roughly proportional to $\prod_{n} W_n$ where $w_n$ is the width of layer $n$. 

$h_1 = \sigma_1(W_1x_1 + b_1)$

$h_2 = \sigma_2(W_2h_2 + b_2)$

$\ldots$

$y = \sigma_n(W_nh_n + b_n)$
One Layer: Two Hyperplanes

\[ h = \max(Wx + b, 0) \text{ with } W = \begin{bmatrix} w_1^T \\ w_2^T \end{bmatrix} \text{ and } b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \]

\[ y = w'^T h + b' \]
Two Layers: Two Hyperplanes

\[ h = \max(Wx + b, 0) \]
\[ h' = \max(W'h + b', 0) \]
\[ y = w''^T h' + b'' \]
The function learned by an MLP using the ReLU, Sigmoid, or Tanh operators is:

- piecewise affine or smooth;
- continuous because it is a composition of continuous functions.

Each region created by a layer is split into smaller regions:

- Their boundaries are correlated in a complex way.
- Their descriptive power is larger than shallow networks for the same number of parameters.
Second Layer for Approximation

\[ I = f(x, y) \]

1 Layer: 125 nodes -> loss $2.40 \times 10^{-1}$

2 Layers: 20 nodes -> loss $8.31 \times 10^{-2}$

501 weights in both cases
Adding a Third Layer

\[ I = f(x, y) \]

2 Layers: 20 nodes -> loss 8.31e-02  3 Layers: 14 nodes -> loss 7.55e-02

501 weights  477 weights
Adding a Third Layer

\[ I = f(x, y) \]

3 Layers: 15 nodes -> loss 5.93e-02  
541 weights

3 Layers: 19 nodes -> loss 4.38e-02  
837 weights
Multi Layer Perceptrons

- Adding layers often yields better convergence properties.
- In current practice, deeper is usually better.
Further improvements in the convergence properties have been obtained by adding a bypass, which allows the final layers to only compute residuals.

\[ x + l_2(\sigma(l_1(x))) \]
Improving the Network

Original 51x51 image: 2601 gray level values.

MLP 10/20/10 Interpolation: 471 weights, loss 6.43e-02.

MLP 10/20/10/10 Interpolation: 581 weights, loss 5.30e-2.
Improving the Network

- Relatively small improvement in this case.
- We will see a different behavior for large networks.
- The problem is probably too small.

--> Networks can behave very differently for small and large problems!
Tanh vs ReLu

Original 51x51 image: 2601 gray level values.

MLP 10/20/10 Interpolation:
- Tanh, loss 6.43e-02.
- ReLU, loss 3.07e-1.

Tanh, 1 layers
Tanh, 2 layers
Tanh, 3 layers
ReLU, 3 layers
Tanh, 4 layers
Tanh vs ReLu

- Tanh works better than ReLU in this case.
- ReLU is widely credited with eliminating the vanishing gradient problem in large networks.

--> There is no substitute for experimentation!
A MxN image can be represented as an MN vector, in which case neighborhood relationships are lost.

By contrast, treating it as a 2D array preserves neighborhood relationships.
Image Specificities

- In a typical image, the values of neighboring pixels tend to be more highly correlated than those of distant ones.
- An image filter should be translation invariant.

→ These two properties can be exploited to drastically reduce the number of weights required by CNNs using so-called convolutional layers.
The descriptive power of the net increases with the number of layers.

In the case of a 1D signal, it is roughly proportional to \( \prod W_n \) where \( W_n \) represents the width of a layer.
Convolutional Layer

\[
\sigma \left( b + \sum_{x=0}^{n_x} \sum_{y=0}^{n_y} w_{x,y} a_{i+x,j+y} \right)
\]
Feature Maps

28 × 28 input neurons

first hidden layer: 3 × 24 × 24 neurons

Filters:
Filters

Derivatives

Learned filters
Pooling Layer

- Reduces the number of inputs by replacing all activations in a neighborhood by a single one.
- Can be thought as asking if a particular feature is present in that neighborhood while ignoring the exact location.
Adding the Pooling Layers

The output size is reduced by the pooling layers.
Adding a Fully Connected Layer

- Each neuron in the final fully connected layer is connected to all neurons in the preceding one.
- Deep architecture with many parameters to learn but still far fewer than an equivalent multilayer perceptron.
LeNet (1989-1999)
Lenet Results

- SVM: 98.6
- LN5: 99.05
- Knn: 96.8
AlexNet (2012)

Task: Image classification
Training images: Large Scale Visual Recognition Challenge 2010
Training time: 2 weeks on 2 GPUs

Major Breakthrough: Training large networks has now been shown to be practical!!

Krizhevsky, NIPS’12
AlexNet Results

- At the 2012 ImageNet Large Scale Visual Recognition Challenge, AlexNet achieved a top-5 error of 15.3%, more than 10.8% lower than the runner up.
- Since 2015, networks outperform humans on this task.
Feature Maps

- Some of the convolutional masks are very similar to oriented Gaussian or Gabor filters.
- The trained neural nets compute oriented derivatives, which the brain is also believed to do.
Filter Banks

Derivatives of order 0, 1, and 2.

Learned
VGG19, 3 weeks of training.

“It was demonstrated that the representation depth is beneficial for the classification accuracy, and that state-of-the-art performance on the ImageNet challenge dataset can be achieved using a conventional ConvNet architecture.”

Simonyan & Zisserman, ICLR’15
ResNet to U-Net

ResNet block

\[ x + l_2(\sigma(l_1(x))) \]

U-Net

\[ \text{Downsampling} \quad \text{Upsampling} \]

\[ \text{Skip connection} \]

\[ \text{Add skip connection to produce an output of the same size as the input.} \]
Regression

\[
\min_{W_l, B_l} \sum_{i} \|F(x_i, W_1, \ldots, W_L, b_1, \ldots, b_L) - y_i\|^2
\]

using
- stochastic gradient descent on mini-batches,
- dropout,
- hard example mining,
- ............
Input: Depth image.

Output: 3D pose vector.
Roc Hunting

DeepFace
Taigman et al. 2014

Deep Edge Detection
Shen et al. 2015
Deeper and Deeper

He et al., CVPR’16

Resnet

\[
X + l_2(\sigma(l_1(x)))
\]
Recurrent Auto Encoder

\[ q(z_{t-1} | x_{t-1}) \]
\[ \mu_{t-1} \]
\[ \sigma_{t-1} \]
\[ \varepsilon \sim N(0, I) \]
\[ z_{t-1} \]
\[ h_{t-1} \]
\[ p(y_{t-1} | h_{t-1}) \]
\[ x_{t-1} \]
\[ y_{t-1} \]

\[ q(z_t | x_t) \]
\[ \mu_t \]
\[ \sigma_t \]
\[ \varepsilon \sim N(0, I) \]
\[ z_t \]
\[ h_t \]
\[ p(y_t | h_t) \]
\[ x_t \]
\[ y_t \]

\[ q(z_{t+1} | x_{t+1}) \]
\[ \mu_{t+1} \]
\[ \sigma_{t+1} \]
\[ \varepsilon \sim N(0, I) \]
\[ z_{t+1} \]
\[ h_{t+1} \]
\[ p(y_{t+1} | h_{t+1}) \]
\[ x_{t+1} \]
\[ y_{t+1} \]
Hand Pose Estimation from Video (2019)

- This is considerably more difficult than estimating from range images.
- It requires a large training database.
Alpha Go

- Uses Deep Nets to find the most promising locations to focus on.
- Performs Tree based search when possible.
- Relies on reinforcement learning and other ML techniques to train.

--> Beat the world champion in 2017.
Visual Cortex
Recognize And Classify: Animal /No Animal

Subjects must raise their hand if they see an animal:

- 60 images
- 1 image per second

Measure their reaction time.

“Shape stimuli are optimally reinforcing each other when separated in time by ~60 ms, suggesting an underlying recurrent circuit with a time constant (feedforward + feedback) of 60 ms.”

Drewes et al., *Journal of Neuroscience*, 2016
Adversarial Images

Szegedy et al. 2013
XKCD’s View On The Matter

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.

https://xkcd.com/
Deep Nets in Short

• Deep Neural Networks can handle huge training databases.
• When the objective function can be minimized, the results are outstanding.
• There are failure cases and performance is hard to predict.

—> Many questions are still open and there is much theoretical work left to do.