

# 1 Machine Learning Theory

Throughout this section the theory of machine learning that will be used in this thesis is described and explained. In Section 1.1 we describe the general motivation behind using machine learning. Then, in Section 1.2 we explain the three most common paradigms in machine learning and how we will use some of them. In Section 1.3, we give a brief overview of how to evaluate a developed model. Lastly, in Section 1.5, we describe the mathematics behind feedforward- and convolutional neural networks .

## 1.1 Motivation

It can be difficult for humans to recognize certain patterns and trends in data. This becomes more difficult the greater the quantity of the data is, which is becoming more and more common with the rapidly growing topic of *Big Data*. For this reason, computers are often used instead of humans to recognize patterns and trends in the data by analyzing the data, which is what is called *Machine Learning*. In this thesis, we will use machine learning in section ?? to develop a model to estimate the 2D pose of a single human in an image. Later, in section ??, we will use machine learning to improve our understanding of the model.

## 1.2 Machine Learning Paradigms

Machine learning is usually split into the at least two paradigms:

1. *Supervised learning* where the data consists of features and labels. By analyzing the data the algorithm learns to predict the labels given the features [2]. Supervised learning is further split into *classification* and *regression*.
2. *Unsupervised learning* where the data only consists of features. The algorithm then learns properties of the data, without any provided labels [2].

In this thesis we will make use of supervised learning when developing our model for pose estimation, where we will make use of regression. Later, unsupervised learning is used when we explore our developed model.

## 1.3 Evaluation of Machine Learning Models

When developing a machine learning model it is important to know how trustworthy the developed model is. This is usually done by testing how good the model is at generalizing to unseen data, which is done by making use of several dataset splits and *evaluation metrics*.

### 1.3.1 Splitting the dataset

When developing a machine learning model, the data needs to both create the model, but also to evaluate the model. For the evaluation of the model, one of the two following techniques is usually used

1. *Cross validation* where the data is split into  $K$  non-overlapping chunks. The model is then trained for  $K$  rounds on  $K - 1$  chunks, where the last chunk is used for evaluating the model [10].
2. *Train-validation-test* where the data is split into 3 random non-overlapping chunks. The training dataset is then used for training the model and the validation dataset is used for evaluating the model as it is being developed - this often means, that the *hyperparameters*, the parameters that are not possible to fit from the data, are being tweaked to yield the

---

**Algorithm 1** PCK [6][13]

---

**Require:** Ground truth heatmaps  $heatmaps_{gt}$  of keypoints

**Require:** Predicted heatmaps  $heatmaps_{pred}$  of keypoints

**Require:** Threshold radius  $r$

**Require:** Normalization constant  $c$

- 1: Let  $n = 0$  be the running total of correctly predicted keypoints
  - 2: Let  $N$  be the amount of annotated heatmaps
  - 3: **for each** annotated ground truth heatmap,  $heatmap_{gt}$ , in  $heatmaps_{gt}$  **do**
  - 4:   Let  $(x_{gt}, y_{gt})$  be the 2D index of the maximum activation of  $heatmap_{gt}$
  - 5:   Let  $(x_{pred}, y_{pred})$  be the 2D index of the maximum activation of the predicted heatmap corresponding to  $heatmap_{gt}$
  - 6:   Let  $dist$  be the Euclidean distance between  $(x_{gt}, y_{gt})$  and  $(x_{pred}, y_{pred})$ .
  - 7:   Normalize  $dist$ :  $dist = \frac{dist}{c}$
  - 8:   **if**  $dist < r$  **then**
  - 9:      $n = n + 1$
  - 10: Let  $ratio = \frac{n}{N}$  be the ratio of correctly annotated heatmaps
  - 11: **return**  $ratio$
- 

best validation loss. Lastly, the testing dataset is used as a final evaluation of the model to yield an unbiased evaluation of the ability of the model to generalize to unseen data. Once the testing dataset has been used it can no longer be used for evaluating the data, as this potentially introduces bias in future evaluations [3].

Throughout this thesis the train-validation-test technique will be used over cross validation for evaluating the developed models. This is done, since cross validation is better suited for smaller datasets, as the runtime is much greater than the runtime of the train-validation-test technique.

### 1.3.2 Evaluation Metrics for Supervised Machine Learning

When we have trained a model, we need to somehow evaluate how well the model performs on unseen data. This is usually done by making use of evaluation metrics. In articulated human pose estimation, one commonly used evaluation metric is the *Percentage of Correct Keypoints* (PCK), which states the percentage of predictions that are within a normalized distance of the ground truth.

The pseudocode of PCK is visualized in Algorithm 1. The algorithm works by iterating over each annotated ground truth heatmap and the corresponding predicted heatmap. It then finds the Euclidean distance between the maximum activation of a ground truth heatmap and the corresponding predicted heatmap. The distance is then normalized by a constant  $c$  and compared to a threshold radius  $r$ . The ratio of normalized distances that are less than the threshold  $r$  are then computed and returned, yielding the PCK accuracy between the ground truth heatmaps and the corresponding predicted heatmaps [6] [13]. The aim is thus to maximize the PCK accuracy.

## 1.4 General Machine Learning Terminology

### Loss Functions

The training of a machine learning model is done by minimizing a given *loss function*, which measures the error of the model. There are many different loss functions, each with their own

advantages and disadvantages. One of the most common loss functions for regression is the *Mean Squared Error (MSE)*, defined as

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where  $y_i$  is the true value of the  $i$ th observation and  $\hat{y}_i$  is the estimated value of the  $i$ th observation. Thus,  $MSE$  measures the average squared difference between the true observation and the estimated observation. The aim of a model is thus to make the  $MSE$  as small as possible [4].

### Overfitting and Regularization

The main goal of a machine learning model is to generalize well on unseen data. This can often be difficult, as the model simply "remembers" the training data instead of learning the patterns in the training data. In other words, the gap between the training error and the test error is too large, which is a concept called *overfitting*. Certain techniques are designed to reduce the test error - these techniques are collectively called *regularization* [1]. One commonly used Regularization technique is *batch normalization*, which we will be using when we in Section ?? will be developing our model. Batch normalization will be described further in Section 1.5.

---

#### Algorithm 2 Gradient Descent [15]

---

**Require:** Learning rate  $\eta$

**Require:** Initial parameter  $\theta$

- 1: **while** stopping criterion not met **do**
  - 2:     Sample an observation from the training set  $x$  with corresponding target  $y$
  - 3:     Apply update:  $\theta = \theta - \eta \nabla_{\theta} L(f(x; \theta), y)$
- 

### Gradient Descent

The goal of a machine learning model when training is to minimize its loss. There are different methods to do so, however, the most common algorithms are variants of *gradient descent*, whose algorithm is described in Algorithm 2. The algorithm works by taking a learning rate  $\eta$  and the initial parameter  $\theta$  as input. It then computes  $\nabla_{\theta} L(f(x; \theta), y)$  and subtracts the gradient times  $\eta$  from  $\theta$ . This is done until a stopping condition is met, such as when the magnitude of the gradient  $|\nabla_{\theta} L(f(x; \theta), y)|$  is small or until a maximum amount of iterations has been reached [15].

### Online, Mini-batch and Batch methods

When gradient descent is used in machine learning, computing  $\nabla_{\theta} L(f(x; \theta), y)$  can be done by averaging the gradient of each of the  $n$  observations of the training set, which is called a *batch gradient method* and is computational inefficient, as the cost is  $\mathcal{O}(n)$ . It is therefore common to use variants of gradient descents, that reduces the cost of computing the gradient. In *online gradient methods* a single observation from the dataset is used to compute the gradient, which brings the cost down to  $\mathcal{O}(1)$ . In *mini-batch gradient methods* a subset of the dataset is used to compute the gradient, making the cost  $\mathcal{O}(|\mathcal{B}|)$ , where  $|\mathcal{B}|$  is the mini-batch size. Thus, batch gradient descent uses the fewest iterations, however, each iteration takes the longest to compute, whereas in online gradient descent each iteration is computed quickly, however, it also uses more iterations [15].

Choosing the right batch size can be difficult, however, there are a few guidelines which one can follow [7] [15]

1. A batch size that is of power of 2 can offer in better runtime for some hardware. A batch size that is often used for larger models is 16, however, they typically range between 32 and 256.

2. Smaller batch sizes can offer a regularizing effect, as it is difficult for the model to "remember" the complete dataset from batches that does not represent the whole dataset.

---

**Algorithm 3** Stochastic Gradient Descent with Momentum [1]

---

**Require:** Learning rate  $\eta$

**Require:** Momentum parameter  $\alpha$

**Require:** Initial parameter  $\theta$

**Require:** Initial velocity  $v$

- 1: **while** Stopping criterion not met **do**
  - 2:   Sample a minibatch of  $m$  random observations from the training set  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$  with corresponding targets  $\mathbf{y}^{(i)}$
  - 3:   Compute gradient estimate:  $\mathbf{g} = \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}^{(i)}; \theta), \mathbf{y}^{(i)})$
  - 4:   Compute velocity update:  $\mathbf{v} = \alpha \mathbf{v} - \eta \mathbf{g}$
  - 5:    $\theta = \theta + \mathbf{v}$
- 

---

**Algorithm 4** RMSProp [1]

---

**Require:** Learning rate  $\eta$

**Require:** Decay rate  $\rho$

**Require:** Starting position  $\theta$

**Require:** Small constant  $\delta$ , usually  $10^{-6}$

- 1: Initialize accumulation variables  $\mathbf{r} = \mathbf{0}$
  - 2: **while** stopping criterion not met **do**
  - 3:   Sample a minibatch of  $m$  random observations from the training set  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$  with corresponding targets  $\mathbf{y}^{(i)}$
  - 4:   Compute gradient:  $\mathbf{g} = \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}^{(i)}; \theta), \mathbf{y}^{(i)})$
  - 5:   Accumulate squared gradient:  $\mathbf{r} = \rho \mathbf{r} + (1 - \rho) \mathbf{g} \odot \mathbf{g}$
  - 6:   Compute parameter update:  $\Delta \theta = -\frac{\eta}{\sqrt{\delta + \mathbf{r}}} \odot \mathbf{g}$
  - 7:   Apply update:  $\theta = \theta + \Delta \theta$
- 

## Optimization Algorithms

An example of a mini-batch based optimization algorithm is *Stochastic Gradient Descent* (SGD). SGD is very closely related to the algorithm behind gradient descent, however, instead of updating the parameters for each sample, SGD instead uses the mean gradient of  $n$  samples to update its parameters.

One problem of the gradient descent algorithm is, that the learning rate can be difficult to choose. Therefore, there have been developed a range of various optimization algorithms that uses a separate learning rate for each parameter and automatically adapt these learning rates. One of which is *RMSProp*, which is visualized in Algorithm 4. The algorithm works by using an decaying average that discards knowledge from the past, so that it can converge after finding a convex bowl of the loss function. The algorithm uses a hyperparameter  $\rho$ , that controls the length scale of the moving average [1].

## Momentum

One problem with SGD is how slow it often can be. For this reason *momentum* is often used to accelerate learning. Momentum works by accumulating a decaying moving average of the past gradients and continuing to move in their direction. This is done by introducing two new variable;  $\mathbf{v}$ , which is the direction and speed of which the parameters move through the parameter space, and  $\alpha \in [0, 1)$ , which describes how quickly the contribution of previous

gradients decay. Common values of  $\alpha$  are 0.5, 0.9 and 0.99 [1].

### Epoch

An *epoch* is an iteration through the complete dataset during fitting of the network. Multiple epochs are often needed to reach the minimum of the loss function [15].

## 1.5 Neural Networks

In recent years *deep learning* and *neural networks* have revolutionized the use of machine learning. In this thesis a neural network will be used for performing the human pose estimation. Throughout subsection 1.5 the theory and mathematics behind neural networks is described and explained.

### 1.5.1 Feedforward Neural Networks

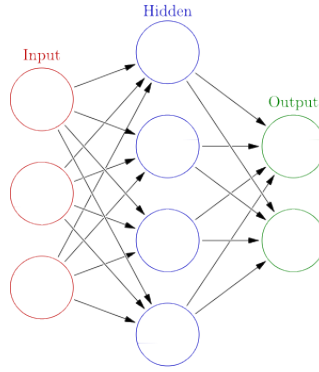


Figure 1: Visualization of a feedforward neural network with a single hidden layer [14]

#### The Architecture and Forwardpropagation

One of the most common types of neural networks are *feedforward neural networks*, where the data flows unidirectionally through the network. Such a network is visualized in Figure 1. The network is a directed acyclic graph and is built up of three types of components: the *input layer*, the *hidden layers* and the *output layer*. Each layer is built up of *units*, also called *neurons* (which are visualized as circles in Figure 1), where each neuron has a *bias* assigned to it, and is connected to one or two other layers through *edges* (which are visualized as arrows in Figure 1), where each edge has a *weight* assigned to it. Hidden layers are connected to two other layers - one before the hidden layer and one after the hidden layer - where the input layer is only connected to the next layer in the network and the output layer is only connected to the previous layer in the network.

We can define the network mathematically by letting  $a_n^{(i)}$  denote the value of the  $n$ th node in the  $i$ th layer,  $w_{m,n}$  denote the value of the weight of the edge connecting the  $n$ th node in the  $i$ th layer to the  $m$ th node in layer  $i + 1$  and  $b_n^{(i)}$  denote the bias corresponding to the  $n$ th node in the  $i$ th layer.

When data flows through the model it follows the following formula

$$\mathbf{a}^{(i+1)} = g^{(i+1)} \left( \mathbf{z}^{(i+1)} \right)$$

where

$$\mathbf{z}^{(i+1)} = \mathbf{W}^{(i+1)} \mathbf{a}^{(i)} + \mathbf{b}^{(i+1)},$$

$\mathbf{W}^{(i+1)}$  is the weights between layer  $i$  and layer  $i + 1$  defined by

$$\mathbf{W}^{(i+1)} = \begin{pmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \cdots & w_{m,n} \end{pmatrix},$$

$\mathbf{a}^{(i)}$  is the values of the nodes in the  $i$ th layer defined by

$$\mathbf{a}^{(i)} = \begin{pmatrix} a_0^{(i)} \\ a_1^{(i)} \\ \vdots \\ a_n^{(i)} \end{pmatrix},$$

$\mathbf{b}^{(i+1)}$  is the values of the biases of layer  $i + 1$  defined by

$$\mathbf{b}^{(i+1)} = \begin{pmatrix} b_0^{(i+1)} \\ b_1^{(i+1)} \\ \vdots \\ b_m^{(i+1)} \end{pmatrix}$$

and  $g$  is a non-linear *activation function*, that is typically applied element-wise [1] [12]. The purpose behind using non-linear activation functions is to make it possible for the output of the network to be non-linear. One often used activation function is the *rectified linear activation function* (or ReLU for short) defined by

$$g(x) = \max\{0, x\}.$$

The ReLU-function is very close to being linear, making the function keep many of the properties of linear functions that make them easy to optimize and generalizing, which are two great advantages of using the ReLU-function. Another great advantage of using the ReLU-function is stated by the *universal approximation theorem* which states, that a feedforward network with a linear output layer and at least one hidden layer with the ReLU-function (or another non-linear activation function from a wide class of activation functions) can approximate any continuous function on a closed and bounded subset of  $\mathbb{R}^n$  (and actually some functions outside of this class), as long as the network has enough hidden neurons [1].

## Backpropagation

Backpropagation is an algorithm used to compute the gradient of the network. It is used together with an optimization algorithm, such as RMSProp, to train the model by minimizing the training loss of the model. Backpropagation happens after data has flowed through the model from the input to the output, and works by computing the gradient of each parameter sequentially from the output to the input of the model. The procedure makes heavily use of the *chain rule* from calculus, which states, that if we let  $\mathbf{x} \in \mathbb{R}^m$ ,  $\mathbf{y} \in \mathbb{R}^n$ ,  $g$  be a function that maps from  $\mathbb{R}^m$  to  $\mathbb{R}^n$  and  $f$  be a function that maps from  $\mathbb{R}^n$  to  $\mathbb{R}$ , then, if we let  $\mathbf{y} = g(\mathbf{x})$  and  $z = f(\mathbf{y})$ , we can then compute  $\frac{\partial z}{\partial x_i}$  by

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_i}$$

[1]. If we use this to find the gradient of each parameter, we will find, that the partial derivative of the loss  $L$  for each weight is

$$\frac{\partial L}{\partial w_{jk}^{(i)}} = \frac{\partial z_j^{(i)}}{\partial w_{jk}^{(i)}} \frac{\partial a_j^{(i)}}{\partial z_j^{(i)}} \frac{\partial L}{\partial a_j^{(i)}} = a_k^{(i-1)} g'^{(i)}(z_j^{(i)}) \frac{\partial L}{\partial a_j^{(i)}}$$

and the partial derivative of each bias is

$$\frac{\partial L}{\partial b_j^{(i)}} = \frac{\partial z_j^{(i)}}{\partial b_j^{(i)}} \frac{\partial a_j^{(i)}}{\partial z_j^{(i)}} \frac{\partial L}{\partial a_j^{(i)}} = g'^{(i)}(z_j^{(i)}) \frac{\partial L}{\partial a_j^{(i)}}$$

where for both cases

$$\frac{\partial L}{\partial a_j^{(i)}} = \sum_{k=0}^{n_i-1} w_{jk}^{(i+1)} g'^{(i+1)}(z_j^{(i+1)}) \frac{\partial L}{\partial a_j^{(i+1)}}.$$

if  $a^{(i)}$  is not the output-layer. Once the partial derivative of all weights and biases has been found, the gradient vector can be formed and an optimization method can be used to optimize the parameters of the model [11].

### Batch Normalization

*Batch normalization* is a reparametrization method, which is applied to individual layers in a neural network. If  $x \in \mathcal{B}$  is an input to the batch normalization, BN, then batch normalization is done by the following

$$\text{BN}(x) = \gamma \odot \frac{x - \hat{\mu}_{\mathcal{B}}}{\hat{\sigma}_{\mathcal{B}}} + \beta$$

where

$$\hat{\mu}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}_x}$$

and

$$\hat{\sigma}_{\mathcal{B}} = \sqrt{\frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} (x - \hat{\mu}_{\mathcal{B}})^2 + \epsilon},$$

which makes the minibatch have 0 mean and unit variance.  $\gamma$  and  $\beta$  are often trainable parameters, that are used to make the mini-batch have an arbitrary mean and standard deviation. This helps the network to converge, as the batch normalization keeps centering the mean and standard deviation of the mini-batches. The small constant  $\epsilon > 0$  is simply used to avoid division by zero. [15].

### 1.5.2 Convolutional Neural Networks

Feedforward neural networks introduced in 1.5.1 can be used for pattern recognition within images, however, they are usually not used for this task. Consider a colored input image with 3 channels of dimension  $64 \times 64$ . If we were to use a feedforward neural network on this image, each neuron in the first hidden layer would be connected to the input layer through 12.288 weights. Not only would this use a lot of computational power and time to train, however, a network of this size would also be prone to overfitting. Instead, *convolutional neural networks* (also known as CNN's) are usually used. By using a CNN the high correlation between a pixel and its neighbours is also captured, essentially leading to better results. A CNN usually consists of *convolutional layers*, *pooling layers* and *fully-connected layers*, where the fully-connected layers are analogous to the layers in a feedforward neural network [5].

#### Convolutional layers

A convolutional layer is composed of a set of *kernels* (also known as *filters*), which are matrices of weights of dimension  $c \times k \times k$ , where  $c$  is the amount of channels (or amount of filters to be applied) and  $k$  kernel-size, usually 5, 3 or 1. The convolutional layer thus consists of  $c$  filters, each of  $k \times k$  parameters, where each weight is a parameter for the model to be learned [8]. Each kernel is used on the input to produce a *feature map*. The kernels are applied to the input



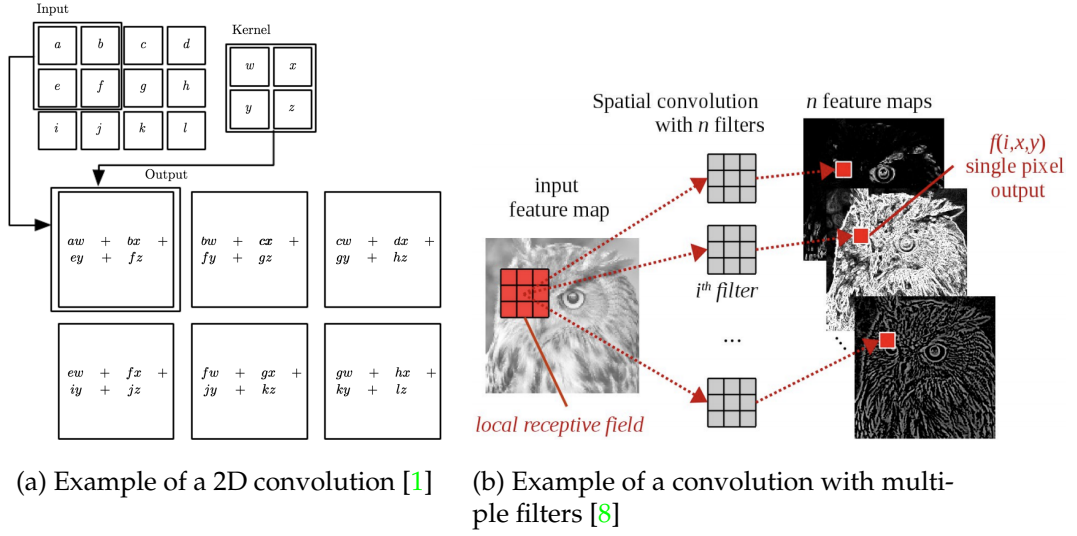


Figure 2: Convolutions visualized

by "sliding" over the input (where the step size is called *stride*). Each  $k \times k$  grid of the input (called the *local receptive field*) is then used to compute the dot-product between the grid and each kernel, which is then placed in the corresponding feature map of each kernel, as visualized in Figure 2. When all of the feature maps have been computed of a convolutional layer an activation function can be applied.

As described previously, by using a convolutional layer we can dramatically decrease the amount of weights used by the layer. If we were to use a  $3 \times 3 \times 3$  convolution on a colored image with 3 channels, we would reduce the amount of weights on each neuron in the following layer from 12.288 down to just 27, reducing both the training time and making the network less prone to overfit [5].

### Pooling layers

Pooling layers are layers used to reduce the dimension of the input. The most common pooling layer is the *maxpooling*-operation. The operation works by considering each  $k \times k$  grid, like in the case with the convolutional layer, in which the maximum entry in that grid is being inserted into the output [5].

---

#### Algorithm 5 Nearest Neighbour Upsampling [9]

---

**Require:** Input image  $X$  of size  $m \times n$

**Require:** Wanted output size  $xm \times yn$ , where  $x, y \in \mathbb{Z}^+$

- 1: Create empty image  $O$  of size  $xm \times yn$
  - 2: **for each** pixel,  $p$ , in  $X$  **do**
  - 3:    $i, j = \text{index of } p \text{ in } X$
  - 4:   Insert  $p$  at index  $(xi, yj)$  in  $O$
  - 5: **for each** empty pixel  $p \in O$  **do**
  - 6:   Let  $p$  be the value of the nearest neighbour
  - 7: **return**  $O$
- 

### Nearest Neighbour Upsampling

Sometimes we want to increase the size of an image. This can be done by making use of *upsampling* (also known as *interpolation*) techniques. One of the most common upsampling



techniques is *nearest neighbour upsampling*, whose pseudocode has been written in Algorithm 5. The algorithm starts off by taking an image,  $X$ , of size  $m \times n$ , as input, which we wish to upsample to size  $xm \times yn$ , where  $x$  and  $y$  are positive whole numbers. The algorithm then loops over each pixel,  $p$ , in  $X$ , finds the corresponding index,  $i, j$ , of  $p$  in  $X$ , and places  $p$  at index  $(xi, yj)$  in the output image  $O$  of size  $xm \times yn$ . When this is done, it assigns each of the empty pixel in  $O$  the value of their nearest neighbour, making each pixel in  $O$  have a value, and then returns  $O$  [9].