CSC14003 - ARTIFICIAL INTELLIGENCE

PROJECT 3

Teacher: Ngô Đình Hy

Student:

- Chung Kim Khánh
- Trần Đại Hoàng Trung

Table Of Contents

I. GROUP INFORMATION

Group name: Homeless

Fullname	ID	Assignment	Completion level
Chung Kim Khánh	19127644	Code	100%
		Synthesis, report writing and checking	100%
Trần Đại Hoàng Trung	19127081	Writing report	100%

II. THREE TYPE OF SET AND OVERFITTING, UNDERFITTING PROBLEM

1. Distinguish 3 types of training set, validation set, test set.

In machine learning, dataset is something we must come across many times. Dataset is the set of data that we work with, also the set of data on which we apply AI algorithms, machine learning models to test and evaluate. And usually, this dataset is very large in size, and it is divided into smaller sets. When we train a machine learning model or a neural network, we split the dataset into three categories: training dataset, validation dataset, and test dataset.

- **Training set:** A set of examples used for learning, that is to fit the parameters of the classifier. This is usually a large dataset. Training set consists of 2 parts: Input (the input data) and Output (the results corresponding to the input set). In every iteration of training, we use the training set as the examples for the model. The model "learns" by extracting patterns from the training data and generalizing the training examples.
- **Validation set:** A set of examples used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The model occasionally sees this data, but never does it "Learn" from this. After training, we use the validation data set to check the performance of the model during training. If we are not satisfied with the result, we modify the hyperparameters and continue training using the training set again.
- **Test set:** A set of examples used to provide an unbiased evaluation of a final model fit on the training set. The Test set provides the gold standard used to evaluate the model. When the training is completed, we use the test set to check the final performance of the

model. The test set must be used only once a model is completely trained. For example: On many Kaggle competitions, the validation set is released initially along with the training set and the actual test set is only released when the competition is about to close, and it is the result of the model on the Test set that decides the winner.

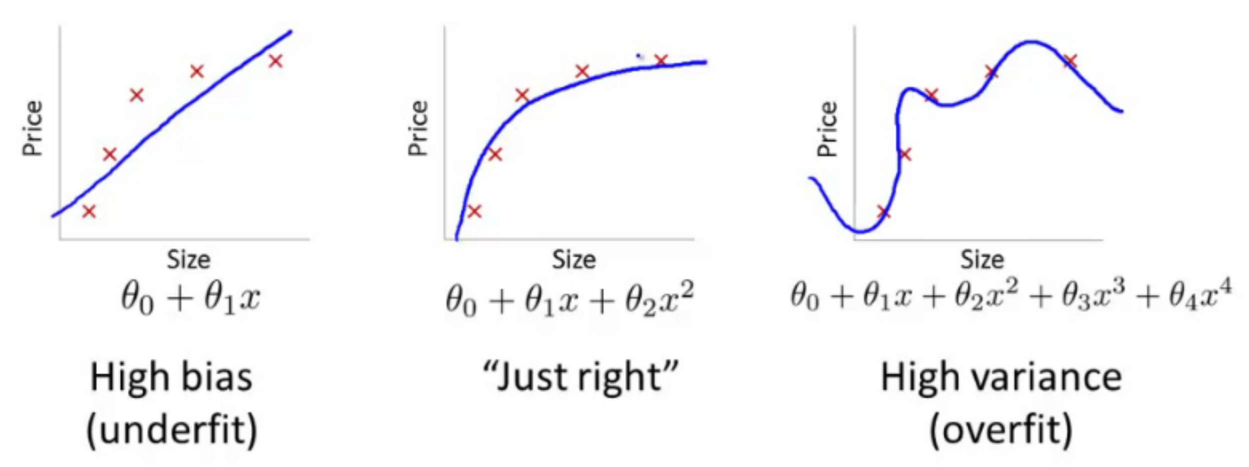
2. How to detect and prevent overfitting and underfitting problem.

A key challenge of detecting overfitting and underfitting, is that we can't know how well our model will perform on new data until we test it. To address this, we can split our initial dataset into separated different subsets to make it easy for training and testing. The data is split into two parts: training set and test set. This method can approximate how well our model will perform on new data.

- If our model does much better on the training set than on the test set, then we are likely overfitting. For example, our model performed with a 95% accuracy on the training set but only 48% accuracy on the test set. It is Overfitting the model and did not perform well on unseen dataset.
- If our model does much better on the test set than on the training set, then we are likely underfitting. For example, our model performed with a 96% accuracy on the test set but only 50% accuracy on the training set.

Using cross-validation is also one simple method to detect overfitting and underfitting. This attempts to examine the trained model with a new data set to check its predictive accuracy. Given a dataset, some portion of this is held back (say 30%) while the rest is used in training the model. Once the model has been trained the reserved data is then used to check the accuracy of the model compared to the accuracy of derived from the data used in training. A significant variance in these two flags overfitting.

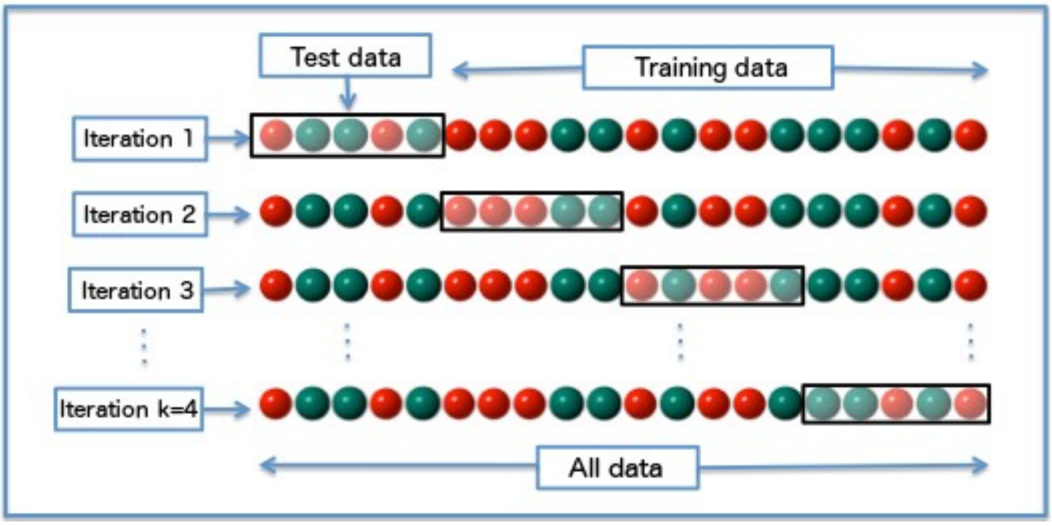
Other method to detect overfitting and underfitting is to use the bias-variance tradeoff, which can be represented like this:



- Our model is over fitted when we have a high variance.
- Our model is under fitted when we have a high bias.

We have some most popular solutions for overfitting:

- **Cross-Validation:** Cross-validation is a powerful preventative measure against overfitting. A standard way to find out-of-sample prediction error is to use k-fold cross-validation. In standard k-fold cross-validation, we partition the data into k subsets, called folds. We let one of the folds to be the test set and the others as the training set. Then repeat this process until each individual group has been used as the test set.

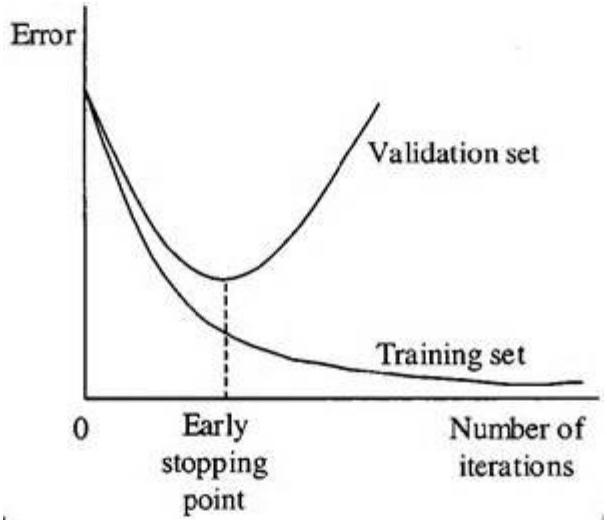


- **Regularization:** Regularization is a technique to reduce the complexity of the model. The most common techniques are known as L1 and L2 regularization. In L1 or L2 regularization, we can add a penalty term on the cost function to push the estimated coefficients for many variables to zero (and not take more extreme values). L2 regularization allows weights to decay towards zero but not to zero, while L1 regularization allows weights to decay to zero.

L1 Regularization	L2 Regularization
1. L1 penalizes sum of absolute values of weights.	1. L2 penalizes sum of square values of weights.
2. L1 generates model that is simple and interpretable.	2. L2 regularization is able to learn complex data patterns.
3. L1 is robust to outliers.	3. L2 is not robust to outliers.

- **Train with more data:** It won’t work every time, but training with more data can help algorithms detect the signal better. As the user feeds more training data into the model, it will be unable to overfit all the samples and will be forced to generalize to obtain results. However, if we just add more noisy data, this technique won’t help. You should always ensure your data is relevant and clean.

- **Data augmentation:** This is an alternative method to training with more data. If you can not continuously collect more data, you can make available datasets to be diverse. Data augmentation means increasing the size of the data, increasing the number of images present in the dataset. Because we add more data, the model is unable to overfit all the samples, and is forced to generalize.
- **Early Stopping:** When you're training a learning algorithm iteratively, you can measure how well each iteration of the model performs. Up until a certain number of iterations, new iterations improve the model. After that point, however, the model's ability to generalize can weaken as it begins to overfit the training data. Early stopping refers stopping the training process before the learner passes that point.



- **Ensembling:** Ensembles are machine learning methods for combining predictions from multiple separate models. There are a few different methods for ensembling, but the two most common are: Bagging and Boosting.
    - Bagging attempts to reduce the chance of overfitting complex models.
    - Boosting attempts to improve the predictive flexibility of simple models.

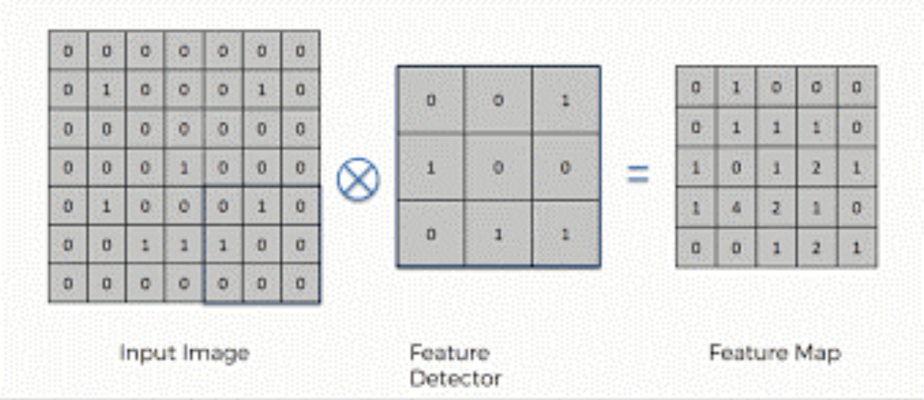
Solutions to prevent underfitting: We should move on, try alternate machine learning algorithms, increase more features and parameters. Increase more features and parameters expands the hypothesis space.

III. CONVOLUTIONAL NEURAL NETWORK (CNN)

1. Components and Reason

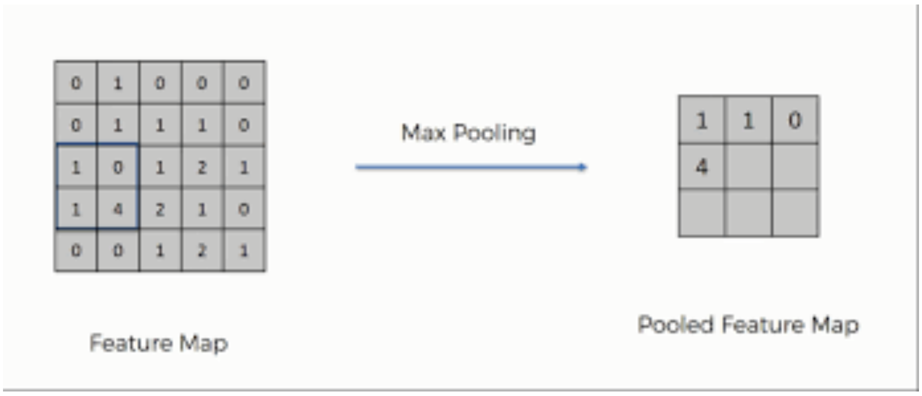
- **Input layer:** This layer contains the actual image – a matrix/tensor/whatever you wish to call it of shape (width, height, depth)
- **Convolution layer:** This is the first layer that comes after the input layer. Convolution layer is used to extract the various features from the input images. In this layer, there is a linear mathematical operation involving the multiplication of weights with the input.

The multiplication of convolution is performed between an array of input data and a filter (a 2D array of weights). The filter is always smaller than input data and the dot product is performed between the filter and the parts of the input data with respect to the size of the filter. The output is the Feature map which gives us information about the corners and edges of input image. Then the Feature map is fed to other layers to learn several other features of the input image. Note that the output of Convolution layer will pass through the activation function before becoming the input of the next layer.



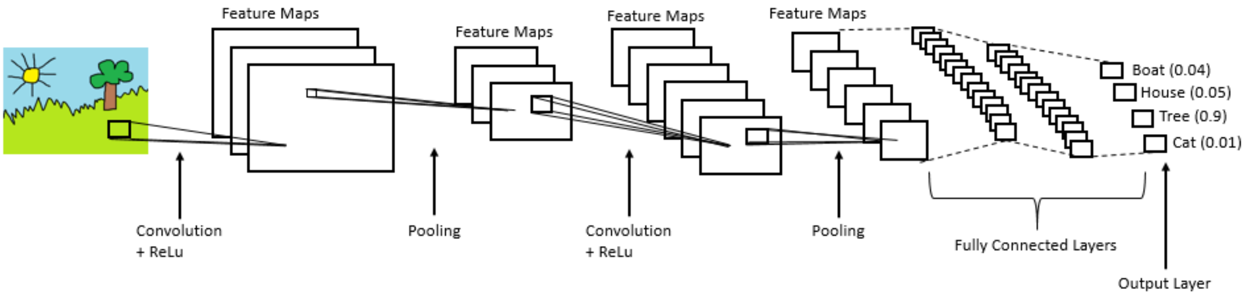
➔ The activation function is added to help CNN learn complex patterns in the data. The main purpose of activation function is to add non-linearity into the neural network.

- **Pooling layer:** The main purpose of this layer is to decrease the size of the convolved feature map to reduce the number of parameters and the computational costs. It progressively reduces the spatial size of the network and thus controls overfitting. This is performed by decreasing the connections between layers and independently operates on each feature map. There are two types of Pooling operations: Average pooling and Maximum pooling. In this project, we use Maximum pooling. In Max Pooling, the largest element is taken from Feature map. This is possible with the help of filters sliding through the Feature map and at each stride, the maximum parameter will be taken out and the rest will be dropped. The Pooling layer usually is often seen as a bridge between the Convolutional Layer and Fully connected layer. Unlike in the Convolutional layer, the Pooling layer does not modify the depth of the network. The Pooling layer usually is often seen as a bridge between the Convolutional Layer and Fully connected layer.



- **Fully connected layer:** The Fully connected layer consists of the weights and biases along with the neurons. This layer is used to connect the neurons between two different layers. This layer is usually placed before the output layer. The output from the final Pooling layer are flattened and fed to the Fully connected layer. The neurons present in the fully connected layer detect a certain feature and preserves its value then communicates the value to both the dog and cat classes. Then the dog and cat classes check out the feature and decide if the feature is relevant to them. Like convolutional layers, this layer will also have non-linear activation functions.
- **Output layer:** The Output layer will probably use the so-called softmax activation function because it allows us to direct output into a class.

Full CNN overview example:



There are several reasons for us to choose Convolutional neural network for this projoect:

- CNN provides very high accuracy in image recognition problems.
- The building blocks of CNN are filters, which are used to extract the relevant features from the input image using the convolution operation.
- CNN's ability is to develop an internal representation of a two-dimensional image. This ability is important when working with images. It will allows the model to learn position and scale in variant structures in the data.

2. The advantages and disadvantages

Advantages:

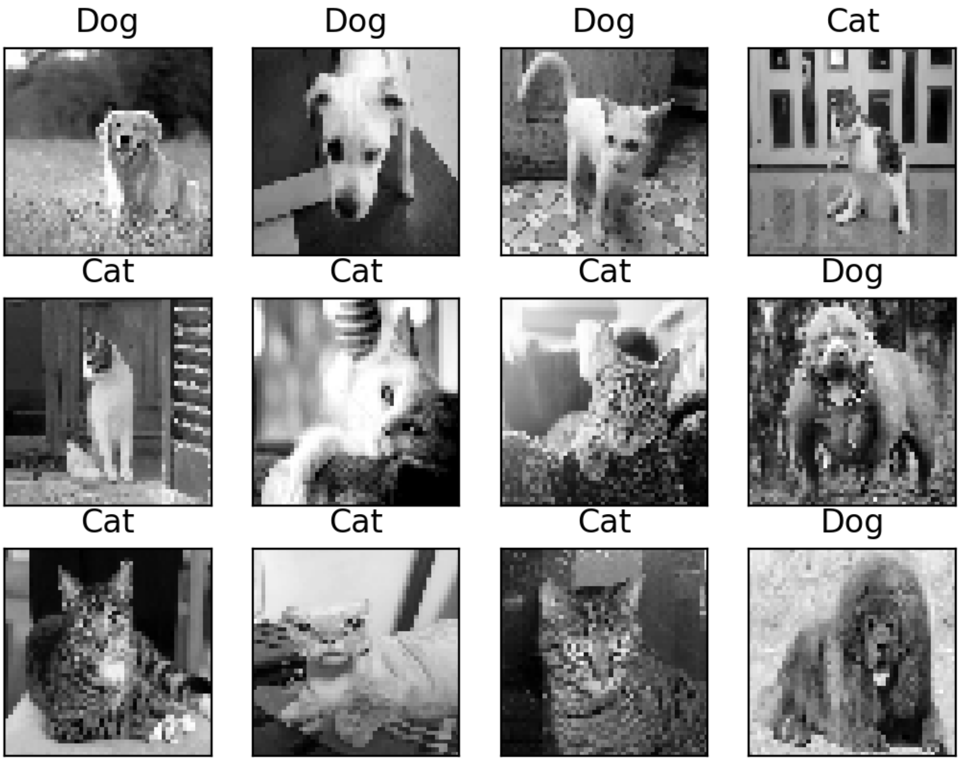
- CNN learns the filters automatically without mentioning it explicitly. These filters help in extracting the right and relevant features from the input data
- CNN is also computationally efficient. It uses special convolution and pooling operations, performs parameter sharing and dimensionality reduction, which makes models easy and quick to deploy. This enables CNN models to run on any device, even on smartphones.
- Very High accuracy in image recognition problems. CNNs have now been dominating for a very long time in most cases and tasks regarding image and video recognition and similar tasks.

Disadvantages:

- Requires a large Dataset to process and train the neural network.
- Adversarial attacks: These are cases that provide the network with ‘bad’ examples (like slightly modified images) to cause misclassification. For example, criminals can fool a CNN-based face recognition system and pass unrecognized in front of the camera
- Do not encode the position and orientation of object.

IV. RESULT

Ideas for improvement: We can train with more data to increase accuracy of model. Data augmentation is a great method. Data augmentation can increase the size of training set.



⇨ Sai 1 kết quả

V. REFERENCE SOURCE

(1) Towardsdatascience
(2) DataSciencefoundation
(3) AnalyticsVidhya
(4) Serokell