TU/e Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science
Data Mining Research Group

Benchmark Studies of Various Deep Learning Architecture

Data Mining Seminar

Irfan Nur Afif

Supervisors:
Joaquin Vanschoren

version 1.0

Eindhoven, December 2017

Contents

Chapter 1

Introduction

1.1 Context & Motivation

Machine learning has become an integral part of today's technology. It has a lot of applications in our daily lifes, for example a recommender system or a prediction models. One of the machine learning technique that we often see nowadays is deep learning. It is the latest topics in the machine learning research which is proven to do well in solving complex classification task such as image and speech recognition [5]. The ability to discover intricate structure makes it strong tools for high dimensional data processing such as image data. In this research, we are primarily interested to explore deep learning on image datasets.

The challenging part for doing a deep learning is deciding the architecture to use for a given image datasets. There is no exact guidelines on designing a deep learning architecture. Several architecture has been proposed for a specific datasets, however we rarely see the performance comparison of the proposed design for the same datasets. Such comparison will tells us in what cases does an architecture performs better compared to the other. A comparison also tells us whether there exists an architecture that generally works well for a general image dataset or what kind of datasets criteria that works well in an architecture.

To solve this problem, we propose a benchmarking studies of multiple deep learning architecture on many image datasets. The goal of the research is to have a benchmark analysis of various deep learning architecture performance on multiple image datasets.

1.2 Research Question

The works tries to answer the following research question: "How does the performance comparison of deep learning architecture model looks like for a given image classification dataset?" In answering this research question, the approach that we try to use is to implement the state-of-the art and widely-used deep learning architecture in various datasets and analyze its performance. Some results from previous related research paper will also be used for benchmarking. There are also some sub-questions to solve the main research questions, which are:

1. Which architecture that works bests for a given datasets?
2. What kind of datasets characteristics that makes a deep learning architecture works well?
3. Is there any architecture that generally works well for image classification?

Chapter 2

Literature Study

In this chapter, we describe the related state-of-the-art and widely used deep learning architecture for image processing. The datasets that will be used for experiment will also explained in this section.

2.1 Deep Learning

Deep learning is a special form of neural networks that uses complex model to solve a problem. Deep learning technique that heavily used for dealing with image classification problem is convolutional neural network (CNN). In CNN, usually the model goes into three kinds of layers (other than the input and output layers).

The first layer type is convolutional layer. In this layer an element wise multiplication operation is implemented using a moving kernel/filter. A convolutional layer produces some feature maps for the next process. The output of the convolutional layer is usually connected to an activation function such as ReLU, tanh or sigmoid. The second type of layer is subsampling/pooling layer. This layer's function is to reduce the number of tuned parameters. The common operators for subsampling layer are: max-pooling and average pooling. The last type is fully-connected layers. In this layer, the networks try to determine the probability of the input falls into each class by learning the high level abstraction from convolutional and maxpooling layer output. There is also an optional layer that serves as regularization layer such as dropout layer. There are a lot of deep learning architecture that was proposed. Below, we highlight some of the most interesting architecture to be tested in the experiment.

2.1.1 LeNet-5

LeNet-5 was proposed by Yann LeCun, et al. [6] consists of seven non-input layers The first layer is a 5x5 convolutional layer with six feature maps. The second layer is a 2x2 non-overlapping subsampling layer. The third layer is a 5x5 convolutional layer with sixteen feeature maps. The fourth layer consists of 2x2 non-overlapping subsampling layer. The fifth layer is a 5x5 convolutional layer with 120 feeature maps. The sixth layer is an 84-units fully-connected layer. The output layer is composed of Euclidean RBF units, one for each class, with 84 inputs each. The architecture of LeNet-5 can be seen in figure 2.1. LeNet-5 is one of the first initial architectures of CNN that was tested to classify hand-written number using MNIST dataset.

2.1.2 Alex-Net

Alex Net was proposed by Alex Krizhevsky, et al. [5] based on their winning on ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) 2012. The architecture consists of eight layers: five convolutional and three fully-connected layers. The first convolutional layer is using a 11x11 filter size with a stride of 4. The other convolutional layer use 3x3 filter size. The subsampling layer

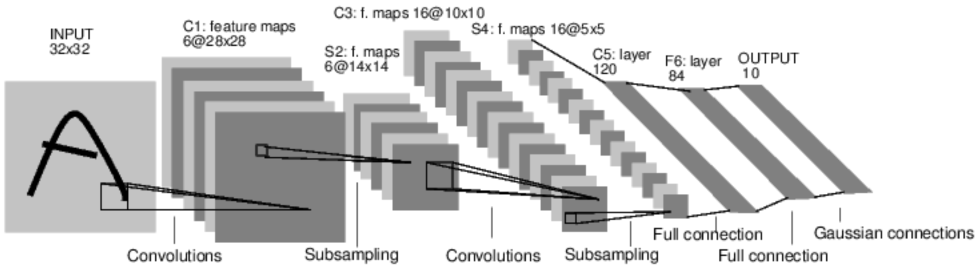


Figure 2.1: LeNet-5 Architecture [6]

that was used is max-pooling. An architecture of Alex-Net can be seen in figure 2.2. It was tested on ILSVRC 2012 dataset and achieves top 5 test error rate of 15.4%.

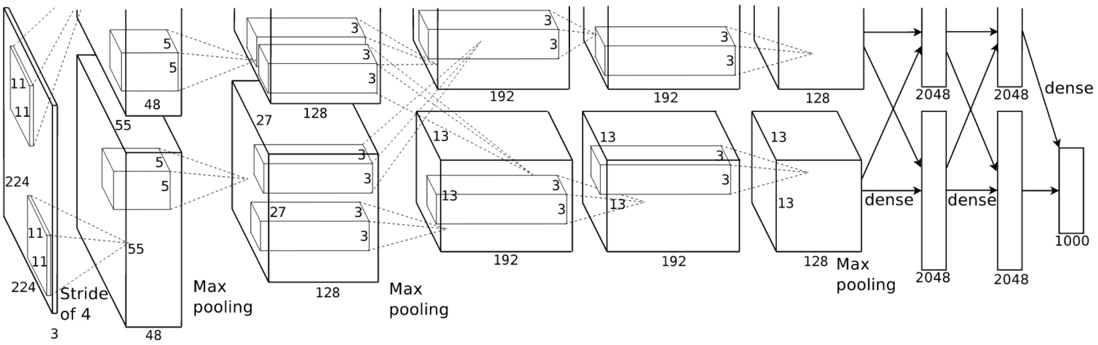


Figure 2.2: Alex Net Architecture [5]

2.1.3 ZF-Net

ZF-Net was proposed by Zeiler and Fergus [10]. It was the winner of ILSVRC 2013. ZF-Net is a modification from Alex-Net. One of the differences between ZF-Net and Alex-Net is the filter size in the first layer, instead of using 11x11 with stride 4 like Alex-Net did, ZF-Net using 7x7 filter size with stride 2. The full architecture of ZF-Net can bee seen in figure 2.3. It was tested on the ILSVRC 2012 dataset and achieved 11.2% error rate.

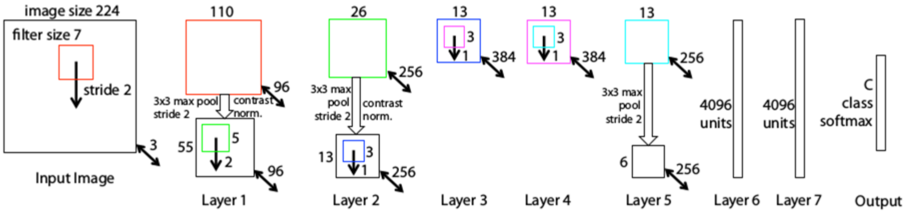
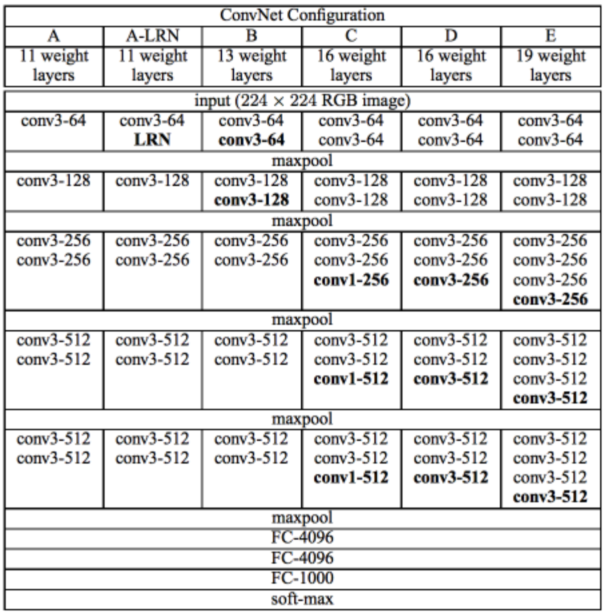


Figure 2.3: ZF Net Architecture [10]

2.1.4 VGG Net

VGG Net was proposed by Karen Simonyan and Andrew Zisserman [8] based on their winning on ILSVRC 2014. The architecture has 6 variations as shown in figure 2.4. The key idea of this architecture is using a 3x3 filter size with a stride of 1 for every convolutional layer. In subsampling layer, VGG-Net uses 2x2 max pooling layer with stride size of 2. It was tested on ILSVRC-2012 dataset and can achieve a top-5 test error of 6.8%.



ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 2.4: VGG Net Configuration [8]

2.1.5 ResNet

In 2015, Microsoft proposed a very deep architecture called ResNet [3]. It can go up to 1202 layers, but the best results for CIFAR-10 dataset is 6.43% error rate by using 110 layers. The idea behind ResNet is stacking convolutional layer with 3x3 filter size until it reaches a very deep network.

2.2 Dataset

Benchmarking the proposed deep learning architecture can be done by collecting some datasets to evaluate the architecture that was presented before. Here we presents some interesting image datasets.

2.2.1 MNIST

The MNIST Datasets (http://yann.lecun.com/exdb/mnist/) is a dataset of handwritten digits with 784 features (28x28 grayscale images). There are 10 classes, 60,000 training examples and 10,000 testing examples. An example of MNIST dataset can be seen in figure 2.5.

We choose this dataset because it is heavily used as validation datasets for image recognition learning algorithm. Also it doesn't need preprocessing and formatting steps since the data is quite clean, thus we can focus on the learning implementation.

2.2.2 Fashion-MNIST

Fashion-MNIST [9] is a dataset of Zalando's article images consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated

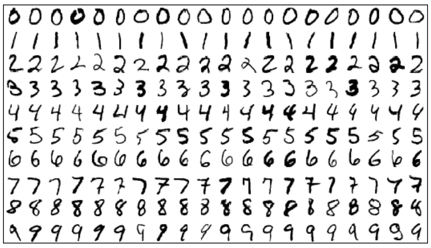


Figure 2.5: An example of MNIST dataset

with a label from 10 classes. Zalando intends Fashion-MNIST to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits.

Compared to MNIST, this datasets are quite new. Thus, unlike MNIST, these datasets are not heavily studied. We choose this datasets because of the similarities with MNIST datasets in terms of image size and representation.

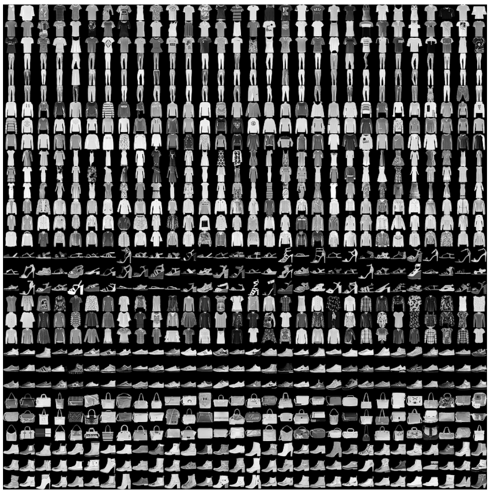


Figure 2.6: Fashion-MNIST sprite. Each three rows in the sprite corresponds to a single class example. [9]

2.2.3 CIFAR-10

CIFAR-10 is a labeled subset of the 80 million tiny images dataset that were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton [4]. It consists of 32x32 color images representing 10 classes of objects: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. An example of CIFAR-10 dataset can be seen in figure 2.7.

CIFAR-10 contains 6000 images per class. The original train-test split randomly divided these into 5000 train and 1000 test images per class. The classes are completely mutually exclusive. For example, there is no overlap between automobiles and trucks. "Automobile" includes sedans, SUVs, things of that sort. "Truck" includes only big trucks. Neither includes pickup trucks.

2.2.4 CIFAR-100

CIFAR-100 [4] dataset is just like the CIFAR-10, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. The 100 classes in the

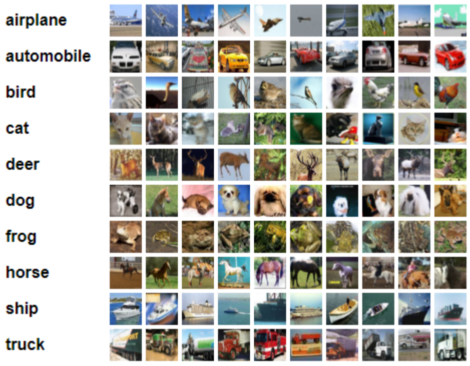


Figure 2.7: Example of CIFAR-10 dataset [4]

CIFAR-100 are grouped into 20 superclasses. Each image comes with a ”fine” label (the class to which it belongs) and a ”coarse” label (the superclass to which it belongs).

2.2.5 STL-10

STL-10 dataset [2] is an image recognition dataset that were acquired from labeled examples on ImageNet. It is inspired by the CIFAR-10 dataset but with some modifications. In particular, each class has fewer labeled training examples than in CIFAR-10, but a very large set of unlabeled examples is provided to learn image models prior to supervised training.

There are 10 classes in this dataset: airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck. Each images are colored 96x96 pixels. 500 training images (10 pre-defined folds), 800 test images per class. 100000 unlabeled images for unsupervised learning. These examples are extracted from a similar but broader distribution of images. For instance, it contains other types of animals (bears, rabbits, etc.) and vehicles (trains, buses, etc.) in addition to the ones in the labeled set.

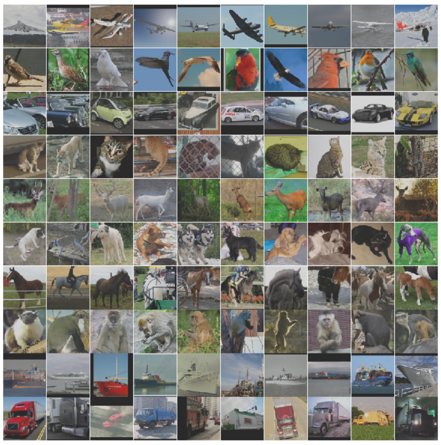


Figure 2.8: Examples of STL-10 datasets [2]

2.2.6 SVHN

The Street View House Numbers (SVHN) Dataset [7] is a dataset obtained from house numbers in Google Street View images. It is similar to MNIST (e.g., the images are of small cropped digits), but incorporates an order of magnitude more labeled data (over 600,000 digit images) and comes from a significantly harder, unsolved, real world problem (recognizing digits and numbers in natural scene images).

There are 10 classes, 73257 digits for training, 26032 digits for testing, and 531131 additional, somewhat less difficult samples, to use as extra training data. There are two formats of this datasets. The one that we are consider is the second format which is an MNIST-like dataset with 32-by-32 images centered around a single character (many of the images do contain some distractors at the sides).

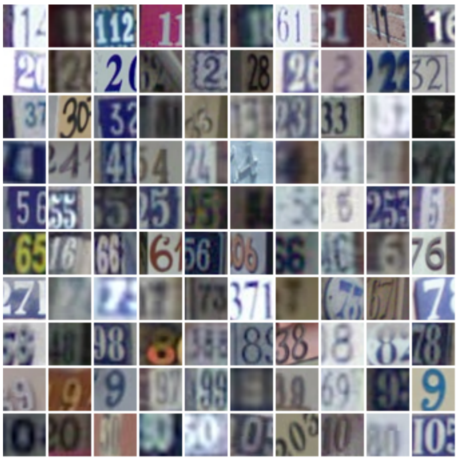


Figure 2.9: Examples of SVHN datasets [7]

Bibliography

[1] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.

[2] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011. 6

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[4] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. 5, 6

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2, 3

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2, 3

[7] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011. 7

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 4

[9] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 4, 5

[10] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3