MLP Coursework 3: Project Interim Report

s1, s2, s3

Abstract

The abstract should be 100–200 words long, providing a concise summary of the contents of your report. **still needs to be written.**

1. Introduction and Motivation

During the last years, connectionist based approaches have gained popularity to solve computer vision tasks (Dinsmore, 2014). A significant number of convolutional neural networks architectures have emerged since they demonstrated their power in the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2012). Several factors have contributed to extend the development of this kind of solutions including an increase in the amount of available training data (Simonyan & Zisserman, 2014), more compute power (Szegedy et al., 2015), and better software abstraction layers.

Although deep neural networks work well with large amounts of training data (lot), the performance of these models typically decreases in situations where only small amounts of training data is available [INSERT REFERENCE TO LEARNING CURVE HERE]. This poses a problem to solve computer vision tasks where the amount of data is not appropiate to utilise these emergent technologies. In addtion, the amount of information extracted from images (features) also impacts the performance of a model. This situation means that a given system would not be suitable to solve the same task with other data.

There exists a number of data manipulation methods that try to reduce the impact of small datasets. Some alternatives simply perform data augmentation (Krizhevsky et al., 2012), whereas others implement more advanced techniques like data synthesize (Hu et al., 2018). Recently, some approaches have tried to address this problem using novel techniques within the domain of deep neural networks including transfer learning (Ng et al., 2015), (Oquab et al., 2014). and deep features extraction (Chen et al., 2016).

In this report, we aim to explore how the size of datasets and the information from images impact in the performance of a given convolutional neural network. The work described here will be the foundation for further investigations of machine learning techniques e.g., transfer learning and deep features extraction [1], to boost the performance of deep neural network architectures under conditions of small data.

[1] These techniques have other use cases that are not necessarily explored within this paper.

We present and analyse the results of reducing the training set size using two different and comparable datasets. The methodology includes performing an observation of the accuracy of the proposed architecture as the training size is decreased (4); this procedure is applied to both datasets. The datasets were selected based on the visual perception of differences between classes; in the first one the differences are easily perceived, whereas in the second one, classes have subtle differences.

Within the remainder of this paper, we present a set of research questions and associated hypotheses (2). Later, an overview of the selected datasets and the task is documented (3). Subsequently, the methodology employed to address the aforementioned research questions and hypotheses are outlined (4) and experimental results are documented (5) which are then draw upon to derive a set of initial conclusions (6). Finally, details of any associated risks, backup plans and further work are provided (7).

2. Research Questions

As described in (4.1), there are two aspects of data that can impact the performance of neural networks: size and features. These variables are the base for the research questions that we explore later on.

Here, we present research questions addressed within this report (2.1), future research questions for the next stage of the project(2.2), and hypotheses(2.3).

2.1. Interim Research Questions

1. How do visually perceived differences among classes affect the accuracy of the proposed convolutional neural network architecture (4.1)?

2. How does reducing the size of a training dataset affect the performance of the proposed convolutional neural network architecture?

In regards to the first question, humans can easily identify different objects based on their visual characteristics. This situation is true for elements which attributes are quite different from one to another (pieces of clothing) as well as for those which differences are subtle (facial expressions).

We are interested in explore how the visually perceived similarity of different classes can affect the performance of a neural network. Although we do not provide a similarity metric among instances of data, we comment about this topic as a potential future work (7).

Dataset	Number of classes	Training size	Test size	Sample per class	Samples size	Format
Clothes	7	42000	7000	7000	64x64	Grayscale
Faces	7	42000	6300	6900	64x64	Grayscale

Table 1. Characteristics of datasets

It is commonly thought that small datasets lead to poor generalisation (lot). The second research question links into our future research questions (2.2) associated with using machine learning techniques to improve the performance of neural network architectures under small data conditions.

2.2. Future Research Questions

1. How does the application of transfer learning affect the performance of the proposed neural network architecture under different sizes of the datasets?

2. How does the application of deep features extraction affect the performance of the proposed neural network architecture under different sizes of the datasets?

2.3. Hypotheses

H.1 The proposed convolutional neural neural network (4.1) has a better performance when the visually perceived differences among classes are high.

H.2 The proposed convolutional neural neural network (4.1) reduces its performance when the size of the training set is reduced.

3. Datasets and Task

We selected two datasets based in the visually perceived differences among their classes:

- Fashion-MNIST (Xiao et al., 2017)
- Facial Expression Research Group Database (FERG-DB)(Aneja et al., 2016)

Fashion-MNIST (hereafter referred as clothes dataset) as stated by their authors "it is intended to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms." This situation means that the dataset shares the same characteristics with its predecessor. It contains 70000 grayscale images of size 28x28 distributed among 10 classes. The training and test sets have 60000 and 10000 samples, respectively.

FERG-DB (hereafter referred as expressions dataset) contains 55767 colorful labeled images of size 256x256. The images depict six individually stylized characters with one of out seven facial expressions: anger, disgust, fear, joy, neutral, sadness and surprise.

As mentioned, the visually perceived differences among the classes in the datasets are different. In the cases of the clothes dataset, it is clearly evident that the main difference between the samples of different classes is the contour as seen in Figure (1). Unlike clothes, facial expressions requires subtle changes in the geometry of a face. The overall structure of a face is similar in every expression, however, the shapes of its components (mouth, eyes, eyebrows, etc) change as seen in Figure (2).

3.1. Preprocessing

We preprocessed both datasets to be able to have similar characteristics as seen in Table (1). Since the expressions dataset contains seven classes, we reduced the number of classes of the clothes dataset to the same value. Later we split the expressions dataset into training and test sets. After this process, the clothes dataset ended up with 42000 and 7000 samples for the training and test sets evenly distributed among the seven classes.

A similar situation occured for the expressions dataset with 42000 and 6300 samples for the training and test sets respectively. In addition, the images in this dataset were converted from colorful to grayscale. Finally, all the images from both datasets were resized to 64x64, this size allows to maintain the relevant features for the facial expressions without increasing too much the computational resources in the training stage.

Finally, a series of transformations was performed in the test sets samples in order to increase the variability of the images. Rotation, blur, skew and shift transformations were randomly applied to each image as seen in Figures (3) and (4).

3.2. Task and Evaluation

We set a classification task based on the labels provided by the datasets. Each experiment is evaluated based on the validation accuracy obtained in the last iteration of the training process. We compare the results of the experiments in two ways: comparison of the accuracy through different sizes of each dataset, and comparison of the accuracy between the same sizes of both datasets.

4. Methodology

We examine the interim research questions (2.1) to create a neural network architecture (4.1). Based on our hypotheses, there are two main variables that we handle in every experiment: the dataset and its size. This means that the rest of elements in the system like the hyper-parameters, activation functions, optimization strategies and else are kept constant.

There are two kind of experiments to perform. In the first one we select one dataset to train the model. The other

Figure 1. Samples of classes in clothes dataset. For easy understanding, we have labeled the classes with words

Figure 2. Samples of classes in expressions dataset. Each of the six characters is displayed

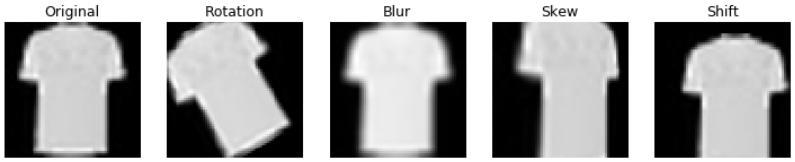

Figure 3. Transformations applied to samples in clothes test set

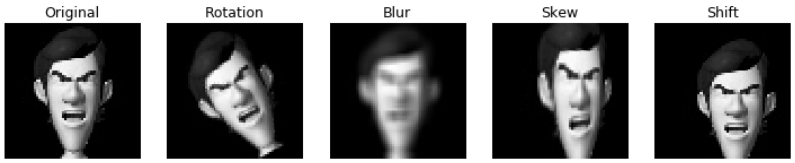

Figure 4. Transformations applied to samples in expressions test set

type of experiments correspond to the size of the selected dataset. When performing the second kind of experiments, we use the same samples for the validation stage to make proper comparisons of the accuracy. To do so, we utilise the test sets that we preprocessed earlier.

4.1. Proposed Neural Network Architecture

We use a simple architecture containing one convolutional which is flattened and connected to a dense layer. The objective of this shallow structure is to have a model that extracts features at the same level in different datasets. We have to remember that the visually perceived differences in the clothes dataset are higher than in the expressions dataset. Based on this, a deeper architecture might be able to extract more features from both datasets but that information would be more relevant for the expressions than for clothes.

Finally, relu and softmax activations along with max pooling layers are integrated into the structure of the network. The weights and biases of the convolutional layer are initialised with uniform distributions, whereas the parameters of the dense layer are initialised using the glorot strategy. The Adam learning rule is used to reduce the error of the system. The total number of training parameters of the network is 31637.

5. Baseline Experiments

5.1. Further Experiments

TODO

6. Interim Conclusions

TODO

7. Future Work

TODO

Size	Accuracy for clothes	Accuracy for expressions
100%	0.51	0.34
75%	0.53	0.34
50%	0.51	0.35
25%	0.52	0.34
10%	0.55	0.34
1%	0.50	0.32

Table 2. Validation accuracies for different sizes of the datasets

Size	Accuracy for clothes	Accuracy for expressions
100%	0.76	0.49
75%	0.78	0.49
50%	0.76	0.48
25%	0.74	0.47
10%	0.78	0.50
1%	0.76	0.44

Table 3. Validation accuracies for different sizes of the datasets using data augmentation

7.1. Backup Plans

TODO

References

Andrew ng: Why âĂŸdeep learningâĂŹ is a mandate for humans, not just machines. https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/. Accessed: 2015-05-01.

Aneja, Deepali, Colburn, Alex, Faigin, Gary, Shapiro, Linda, and Mones, Barbara. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pp. 136–153. Springer, 2016.

Chen, Yushi, Jiang, Hanlu, Li, Chunyang, Jia, Xiuping, and Ghamisi, Pedram. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6232–6251, 2016.

Dinsmore, John. *The symbolic and connectionist paradigms: closing the gap*. Psychology Press, 2014.

Hu, Guosheng, Peng, Xiaojiang, Yang, Yongxin, Hospedales, Timothy M, and Verbeek, Jakob. Frankenstein: Learning deep face representations using small data. *IEEE Transactions on Image Processing*, 27(1): 293–303, 2018.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Ng, Hong-Wei, Nguyen, Viet Dung, Vonikakis, Vassilios, and Winkler, Stefan. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443–449. ACM, 2015.

Oquab, Maxime, Bottou, Leon, Laptev, Ivan, and Sivic, Josef. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1717–1724. IEEE, 2014.

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

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Xiao, Han, Rasul, Kashif, and Vollgraf, Roland. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.