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 recent years connectionist based approaches have gained popularity to solve computer vision tasks (Dinsmore, 2014). This has resulted in the emergence of a significant number of novel convolutional neural networks architectures, 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 research and development of such solutions. These factors include an increased amount of available training data (Simonyan & Zisserman, 2014b), 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. This poses a problem to solve computer vision tasks where the amount of data is not appropriate to utilise these emergent technologies. In addition, the amount of information extracted from images (features) also impacts the performance of a model. It follows that a given model would perform poorly on the same task using very different data.

There exists a number of data manipulation methods that aim to reduce the impact of small datasets. Some alternatives simply perform data augmentation (Krizhevsky et al., 2012), whereas others implement more advanced techniques such as data synthesise (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).

This report aims to explore how the size of datasets and the information from images impact the performance of a given convolutional neural network. The work described therein 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.

This report presents and analyses the results of reducing the training set size using two different and comparable datasets. The methodology includes performing an observation of the accuracy with regards to the proposed architecture as the training size is decreased (4); this procedure is applied to both datasets. The dataset selection criteria is based on the visual perception of differences between classes. In the former, differences are easily perceived, whereas in the latter, classes have subtle differences.

Within the remainder of this paper, a set of research questions and associated hypotheses (2) are presented. 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 research questions that are to be later explored.

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?

With regards to the first question, humans can easily identify different objects based on their visual characteristics. It is assumed that this holds for both instances where features are quite different from one another (pieces of clothing), in addition to those where differences are subtle (facial expressions).

This paper is motivated by how visually perceived similarity of different classes can affect the performance of a neural network. Although a similarity metric to evaluate instances

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

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

of data is not provided, a comment about this topic is made in future work (7).

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

Two datasets based on the visually perceived differences between their classes are selected:

- 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." It follows 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 to as the expressions dataset) contains 55767 colourful labelled images with dimensionality 256x256. The images depict six individually stylised 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 case 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

Both datasets are preprocessed in order to have comparable characteristics, as seen in Table (1).

The clothes dataset is truncated to have seven classes, as there are only seven available classes within the expressions dataset. After converting every image to grayscale, both datasets are split into training and testing sets. This results in the clothes dataset having 42000 training examples and 7000 test samples, evenly distributed among seven classes; and the expressions dataset having 42000 training examples and 6300 testing samples.

Thereafter, every image contained within both datasets were resized to 64x64. This size allows for the maintenance of relevant features for facial expressions without increasing the computational resources in the training stage too much.

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

The proposed task involves classification 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. The results of the experiments are compared 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. It follows that the rest of variables in the model such as any hyper-parameters, the activation functions, optimization strategies and others remain constant.

There are two categories of experiments to perform. In the first, one dataset is selected to train the model. The

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

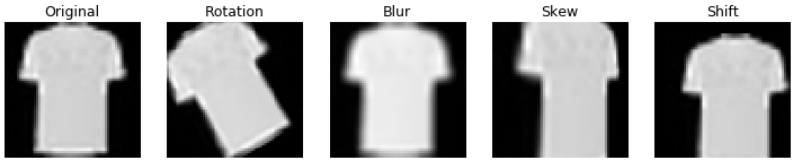

Figure 3. Transformations applied to samples in clothes test set

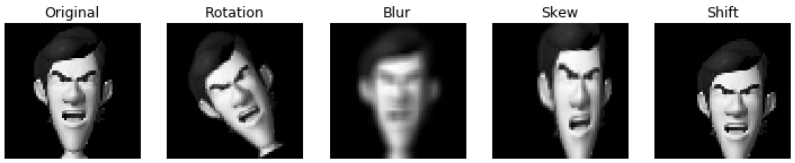

Figure 4. Transformations applied to samples in expressions test set

second category of experiment correspond to the size of the selected dataset. When performing these experiments, the same samples for the validation stage are used in order to make empirical comparisons of the accuracy. To do so, the test sets that were preprocessed earlier are utilised.

4.1. Proposed Neural Network Architecture

The proposed architecture is simple, containing one convolutional layer 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. It is important to take into consideration that the visually perceived differences in the clothes dataset are higher than in the expressions dataset. It follows that a deeper architecture might be able to extract more features from both datasets, but that information would be more relevant for expressions than for clothes.

Finally, relu and softmax activations along with max pooling layers are integrated into the structure of the network. 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 within the network is 31637.

5. Baseline Experiments

The task outlined in section 3.2 has been evaluated using the given datasets (3) in combination with the aforementioned methodology (4). The evaluation of this task provides a set of experimental results, as shown in table and . Both tables display the incremental reduction in dataset size and the classification accuracy of both employed datasets.

6. Interim Conclusions

The experimental results show that the first hypothesis stated in section 2.3 is proven to be true. To elaborate, a shallow network architecture may not be sufficient when classifying examples with subtle differences between

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

classes. This is demonstrated within table and , where the classification accuracy for the clothes dataset is much greater than the accuracy for faces dataset 3.

The classification accuracy presented within the results for both given datasets appears to remain fairly consistent regardless of the training set size. This suggests that the second hypothesis (H2) has been disproved. However, the hypothesis assumes that 1% of the dataset is not enough data to sufficiently train the network.

7. Future Work

To further investigate hypothesis two (H2), the dataset size could be reduced below 1%. This future work would be motivated by (Cho et al., 2015), where it is shown that the classification accuracy is severely reduced when the size of the dataset falls below 100 training examples.

In the second phase of the project, two different transfer learning methods will be studied. These methods will serve as an investigation into the performance benefits of transfer learning given small datasets, which is a frequent occurrence in real world scenarios.

Firstly, a large pre-trained network (VGG16) (Simonyan & Zisserman, 2014a) is transferred on our aforementioned baseline system (REF). The baseline system has pre-trained weights from a small dataset. The generality of this proposed model may be beneficial to train on images contained within the clothes dataset, since VGG16 trains on 200 types of common objects. In addition to transferring model parameters to domain-specific datasets, an investigation into transferring model parameters to datasets with unrelated and subtle differences between classes is presented. This is proposed in order to test the effectiveness of transfer learning on tasks that share little similarity with the tasks that the pre-trained (transferred) model was trained on.

Secondly, an investigation into the effectiveness of one-shot learning on small datasets is intended to be performed. To demonstrate a basic version of one-shot learning, an implementation of a Siamese network (Bromley et al., 1994) on either one of the two datasets will be executed. This will be accomplished using existing models and additional modifications to these models, due to time constrains and the consideration of potential difficulty when implementing a one-shot learning architecture from scratch.

Finally, the proposed investigations would also aim to increase the accuracy of datasets with subtle difference between classes (such as the faces dataset). This would be conducted in an attempt to bring the classification accuracy of both proposed datasets into alignment.

7.1. Backup Plans

In instances where it is no longer possible to explore all proposed research avenues, a viable backup plan we intend to fall back to would be to solely concentrate on transfer learning.

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.

Bromley, Jane, Guyon, Isabelle, LeCun, Yann, Säckinger, Eduard, and Shah, Roopak. Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pp. 737–744, 1994.

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.

Cho, Junghwan, Lee, Kyewook, Shin, Ellie, Choy, Garry, and Do, Synho. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.

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. *CoRR*, abs/1409.1556, 2014a. URL http://arxiv.org/abs/1409.1556.

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

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.