MLP Coursework 3: Interim Report
Dataset size and performance of a convolutional neural network
Group G74

s1700260, s1457374, s1784849

Abstract

In this report, two initial research questions are proposed to evaluate the effects on the performance of a convolutional neural network on two datasets. First, the impact of the reduction of the sizes of the datasets is investigated. Then, the performance of the model is evaluated based on the visually perceived differences among the classes in the datasets. In both cases, the conditions of the network like the hyperparameters, optimization methods, and initialization strategies are maintained. The evaluation is done by comparing the validation accuracy and validation error with both datasets. The experimental results demonstrate a correlation between the performance of the model and the size of one dataset. Moreover, the performance of the model is also affected by the visually perceived differences in the classes of both datasets. Finally, two further research questions have been proposed to investigate the potential effect of transfer learning and feature extraction methods on small datasets which serve as the main objective for next phase of the project.

1. Introduction and Motivation

During recent years connectionist based approaches have gained popularity to solve computer vision tasks (Dinsmore, 2014). Such popularity increase has resulted in the emergence of a significant number of novel deep neural network architectures, especially since they demonstrated their power in the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2012). Several factors have contributed to extending the research and development of such approaches. 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 are available. This poses a problem to solve computer vision tasks where the amount of data might not be appropriate to utilize these emergent technologies. In addition, the relevant information (features) extracted from raw data impacts the performance of a machine learning model as well (Blum & Langley, 1997). 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 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).

This report aims to explore how the size of datasets and the features extracted from images impact the performance of a proposed convolutional neural network architecture (hereafter also referred as the architecture or the model.) 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 the architecture 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 validation accuracy and validation error of the model as the training size is decreased (4); this procedure is applied to both datasets. The dataset selection criteria are based on the visual perception of differences between classes. In one of them the differences are easily perceived, whereas, in the other, 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).

[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 after preprocessing

2. Research Questions

There are two aspects of data that can impact the performance of neural networks: size and features extracted from samples. These variables are the base for research questions that are to be later explored.

Here, we present objectives to pursue (2.1), research questions addressed within this report (2.2), future research questions for the next stage of the project(2.3), and hypotheses(2.4).

2.1. Objectives

The main objective is to observe the experimental results to establish the impact of the reduction of the size of the datasets and the performance of the model when the rest of conditions like hyperparameters, initialization strategies, and optimization methods are kept the same. The fulfillment of this objective will be served as the foundation to explore machine learning techniques that can reduce the affectation of the performance of the model due to the size of the datasets.

Additionally, this reports aims to explore changes in the performance of the model with two different and comparable datasets. The main distinction between the datasets is the visually perceived differences of the samples among their classes. The observation of the results of these experiments will be served to make conclusions about the datasets.

2.2. Interim Research Questions

1. How does reducing the size of the training datasets affect the validation accuracy and validation error of the proposed convolutional neural network architecture?
2. How do visually perceived differences among classes affect the validation accuracy and validation error of the proposed convolutional neural network architecture (4.1)?

It is commonly thought that small datasets lead to poor generalization (lot). The first research question links into our future research questions (2.3) associated with using machine learning techniques to improve the performance of neural network architectures under small data conditions.

In regards to the second 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).

The experiments in this report are motivated by how visually perceived similarity of different classes can affect the performance of the model. Although a similarity metric to evaluate instances of data is not provided, a comment about this topic is made in future work (7).

2.3. 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 methods related to features extraction (e.g. our proposed Similarity network for one-shot learning) affect the performance of the proposed neural network architecture under different sizes of the datasets?

2.4. Hypotheses

H.1 The proposed convolutional neural neural network (4.1) reduces its performance when the size of the training set are reduced.

H.2 The proposed convolutional neural neural network (4.1) has a better performance with the dataset where the visually perceived differences among classes are high.

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 evenly 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 colorful labeled 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 require 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, and the expressions dataset having 42000 training examples and 6300 testing samples. In both datasets, the number of samples is evenly distributed among the seven classes.

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 were 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 a classification based on the labels provided by the datasets. Each experiment is evaluated based on the validation accuracy and validation error metrics obtained in the last iteration of the training process. The results of the experiments are compared in two ways: comparison of the metrics through different sizes of each dataset, and comparison of the metrics between the same sizes of both datasets.

4. Methodology

We examine the interim research questions (2.2) 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-parameter, 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 with and without data augmentation. The second category of experiment corresponds 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 metrics. To do so, the test sets that were preprocessed earlier are utilized.

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 convolutions are performed with five kernels of size 5x5. 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 initialized with uniform distributions, whereas the parameters of the dense layer are initialized using the glorot uniform strategy. The Adam learning rule is used to reduce the error of the system. The error is calculated by using the cross-entropy loss function. 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 2 and 3. Both tables display the incremental reduction in dataset size and the validation accuracy and validation error of both employed datasets.

The first thing to notice from the results is the increase of the accuracy and decrease of the error when training the model with data augmentation. Although this occurs for every size in both datasets, the increase of accuracy is higher for the clothes one (about 50% higher) than for the expressions one (about 40% higher.) Something similar occurs with the error, in average, it is reduced to about one third and one fourth of its original value for the clothes and expressions datasets respectively.

For the different sizes of both datasets, the accuracy of the model does not significantly vary. However, a closer inspection of the error shows a constant increment every time the size of the clothes dataset is reduced and no data augmentation is performed. It is also noticeable that the error is higher for the size of 1% than for the other sizes of the clothes dataset when data augmentation is performed.

Finally, for every dataset, the validation accuracy is fairly similar for each size. This occurs when the model is trained with and without data augmentation.

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

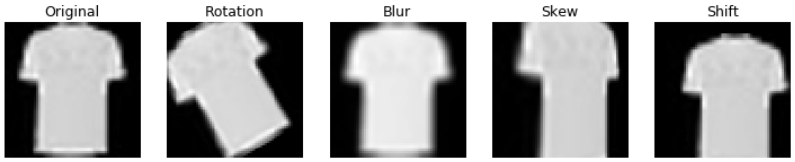

Figure 3. Transformations applied to samples in clothes test set

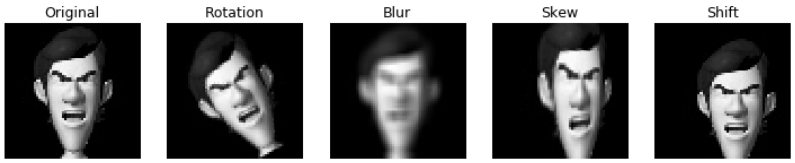

Figure 4. Transformations applied to samples in expressions test set

Size	Accuracy for clothes	Accuracy for expressions	Error for clothes	Error for expressions
100%	0.51	0.34	2.60	7.26
75%	0.53	0.34	2.64	6.78
50%	0.51	0.35	3.13	6.40
25%	0.52	0.34	3.42	6.81
10%	0.55	0.34	3.97	6.77
1%	0.50	0.32	4.69	6.78

Table 2. Validation accuracies for different sizes of the datasets

Size	Accuracy for clothes	Accuracy for expressions	Error for clothes	Error for expressions
100%	0.76	0.49	0.69	2.19
75%	0.78	0.49	0.70	2.08
50%	0.76	0.48	0.67	2.12
25%	0.74	0.47	0.67	1.92
10%	0.78	0.50	0.71	2.11
1%	0.76	0.44	1.18	2.29

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

6. Interim Conclusions

The validation error presented within the results for the expressions dataset when no data augmentation is executed demonstrates a correlation between the performance of the model and the size of that dataset. This suggests that the first hypothesis stated in section 2.4 has been proved to be true for the clothes dataset.

The experimental results show that the second hypothesis is proven to be true. To elaborate, the features extracted by a shallow network have more relevance when the visually perceived differences between the classes of a dataset are high. A deeper convolutional neural network might extract more relevant features from a dataset with subtle differences between its classes.

The results also show that data augmentation helped to maintain the performance fairly steady when the size of the clothes dataset was reduced. However, it seems to be a threshold where the benefits of this technique start to diminish.

The fairly steady value of the validation accuracy throughout all the sizes of the datasets gives a clue about the characteristics of them. It seems that even 1% of the total samples of both datasets is representative enough, which means that the elements among the classes are quite homogeneous. It follows that the predictions of the model could be drastically dropped with test samples from other sources.

7. Future Work

To further investigate hypothesis one (H1), 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 validation 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 model 4.1. The baseline architecture 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 high 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 feature extraction methods, implementing Siamese network for 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. The principle of Siamese network is first to train a neural network to learn similarity measures by discriminating between a set of same/different pairs of image (Koch & Salakhutdinov, 2015).

The trained neural network is then implemented on one test image and all possible classes of images in a pairwise manner, in an unseen dataset, and classification is done by finding the class with highest similarity score (probability). A version of the siamese network will be accomplished using existing models and additional modifications to these models, due to time constraints 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 differences between classes (such as the expressions 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 intended to fall back to would be to solely concentrate on transfer learning (transferring pre-trained model or other feature extraction methods) instead of the Siamese network for one-shot learning.

References

Andrew ng: Why "deep learning" 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.

Blum, Avrim L and Langley, Pat. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997.

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.

Koch, Zemel and Salakhutdinov. Siamese neural networks for one-shot image recognition. In *PhD thesis*, 2015.

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.