On Finding the Best Strategy for Limited Dataset Deep Learning

Neelabh Pareek[#1], Nicholas Christman[*2]

[#]*Department of Computer Science, Columbia University in the City of New York, New York, NY 10027*

[1]np2647@columbia.edu

[2]nc2677@columbia.edu

***Abstract*— Many companies and machine learning practitioners do not have access to extremely large datasets. This constraint makes it very difficult to achieve model effectiveness. The research covered herein provides a perspective on three strategies that have been known to excel on jobs that are constrained by limited data -- namely, data augmentation, transfer learning, and one-shot learning. Each method is compared to the performance of a baseline ResNet-18 model that is trained using the entire dataset. In the end, we show that there is no "one size fits all" approach to choosing a model or process for limited dataset deep learning.**

***Keywords*— deep learning, limited dataset, data augmentation, transfer learning, one-shot learning.**

I. Introduction

It would seem that in the past couple of decades, industries have generated more data than they know how to use, consuming a superfluous amount or resources to acquire data that *might* be important. At the turn of the century as data storage technologies continued to evolve, companies now had the ability to store all this seemingly valuable data but the technology to process these large amounts of data had not yet matured [10]. As we are all aware, however, the re-invention of artificial intelligence (AI), machine learning (ML), and deep learning (DL), have significantly improved a company's ability to leverage *big data* and make more informed business decisions -- a data-driven approach to solve complex problems using the large amounts of data that had been collected since the turn of the century [10].

The objective of this work, on the contrary, is to provide a perspective into the complexities of dealing with small datasets while leveraging the resources developed to process big data. The impact of different novel learning strategies utilized for limited dataset applications will be surveyed. Specifically, we plan to explore the following hypotheses for how stakeholders might overcome challenges with limited data:

(i) using data augmentation (pre-processing) to extend a smaller dataset,
(ii) applying transfer learning using a base model, and
(iii) exploring one-shot learning for difference detection (limited use-cases).

For all hypotheses, a larger dataset will be sampled as evenly as possible to create a limited set with the same training/test data split. This makes it possible to have a base model trained as a control variable during the experiments, providing ground-truth results to compare performance improvements and degradations.

A. Problem Motivation

As alluded to above, many companies and ML practitioners may not have access to extremely large datasets -- for example, a startup company in a niche or new industry may only have a very limited and likely biased dataset from a beta launch. Where state-of-the-art neural networks generally form deep, complex networks, we know that to effectively train these networks generally requires a large amount of data [2]. It is thus easy to recognize that businesses wishing to leverage ML technology with limited data will have difficulty achieving model effectiveness. The question that motivates this research is simply, what are some ways for a company to overcome this constraint?

Conducting the experiments outlined above will provide insight into the different techniques and

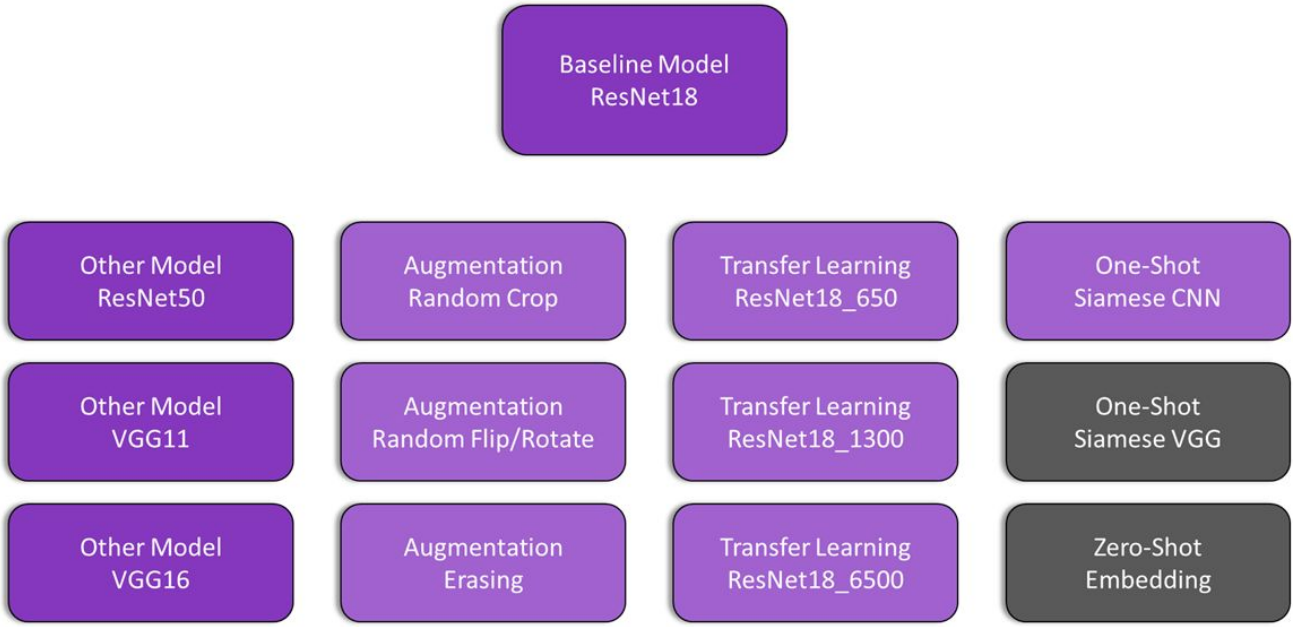

Fig. 1 Hierarchical architecture showing the flow of the solution. At the top is the baseline model (a ResNet-18 trained on the full dataset. The baseline results were used as "ground truth," indicating the absolute maximum performance achievable on the MNIST dataset as a whole. To the left (in dark purple) are a small colleton of other well known models that were also trained on the full dataset -- these additional models were included as an additional reference point, with the intention to further illustrate the complexities of limited dataset training. From left to right are each of the tasks required to achieve the project goals: data augmentation, transfer learning, and one-shot learning. Note: zero-shot learning was included for completeness, but is outside the scope if this project.

their effectiveness at addressing the concerns of using ML on limited datasets.

II. Background

Despite the notion of "big data", many companies are limited in their ability to produce reliable model results and performance due to the lack of quality training data. As such, this area of limited data and training with limited labelled data is well researched.

TBC

III. Technical Challenges

The challenges of this project were primarily encompassed with curating a dataset.

IV. Approach

A. Theory

The theory...

B. Architecture

The architecture…

V. Solution & Implementation

This section provides more details regarding the approach taken to achieve our goal. First, the solution architecture developed for this project is illustrated in Figure 1 below. Observe that we selected the ResNet-18 as our baseline model -- this choice was in part due to the positive reputation of the ResNet structure in addition to the lower complexity of the ResNet-18 structure (making it a good choice for the low number of classes in the chosen MNIST dataset).

A. Full dataset

B. Limited dataset

C. Data Augmentation

D. Transfer Learning

E. One-Shot Learning

Fig. 1 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

VI. Experimental Results

Here are our results...

A. Network training results

In the table below is a summary of the the training results for each of the approaches.

TABLE I
SHOWING THE TRAINING RESULTS FOR EACH MODEL.

Model	**Training Results**		
	Best Accuracy	**Final Loss**	**Total Time**
Resnet18			
Resnet50			
VGG-11			
VGG-16			
Augmentation			
Transfer Learning			
One-Shot			

B. Network validation results

The table below summarizes the validation results for each of the limited data training approaches.

C. Prediction results

1) *Level-1 Heading*:

VII. CONCLUSION

An amazing conclusion drawn from our results...

A. Page Layout

REFERENCES

[1] Oh, Yujin, et al. "Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets." IEEE Transactions on Medical Imaging, vol. 39, no. 8, Aug. 2020, pp. 2688–700. DOI.org (Crossref), doi:10.1109/TMI.2020.2993291.

[2] Peng, Xi, et al. "Learning Face Recognition from Limited Training Data Using Deep Neural Networks." 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 1442–47. DOI.org (Crossref), doi:10.1109/ICPR.2016.7899840.

[3] Transfer Learning for Computer Vision Tutorial — PyTorch Tutorials 1.7.1 Documentation. https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html. Accessed 04 Dec. 2020

[4] Pan, Sinno Jialin, and Qiang Yang. "A Survey on Transfer Learning." IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, Oct. 2010, pp. 1345–59. DOI.org (Crossref), doi:10.1109/TKDE.2009.191.

[5] Li Fei-Fei, et al. "One-Shot Learning of Object Categories." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, Apr. 2006, pp. 594–611. DOI.org (Crossref), doi:10.1109/TPAMI.2006.79.

[6] Lamba, Harshall. "One Shot Learning with Siamese Networks Using Keras." Medium, 17 Feb. 2019, https://towardsdatascience.com/one-shot-learning-with-siamese-networks-using-keras-17f34e75bb3d.

[7] Cheng, Ta-Ying. "Building a One-Shot Learning Network with PyTorch." Medium, 31 May 2020, https://towardsdatascience.com/building-a-one-shot-learning-network-with-pytorch-d1c3a5fafa4a.

[8] Holländer, Branislav. "Siamese Networks: Algorithm, Applications And PyTorch Implementation." Medium, 24 Sept. 2018, https://becominghuman.ai/siamese-networks-algorithm-applications-and-pytorch-implementation-4ffa3304c18.

[9] Taylor-Sakyi, Kevin. "Big Data: Understanding Big Data." ArXiv:1601.04602 [Cs], Jan. 2016. arXiv.org, http://arxiv.org/abs/1601.04602.