Employing Multi-Source Transfer Learning and Web Scraping to Enhance Model Accuracy Where Dataset is Limited

Mustapha Tidoo Yussif[1], Gbetondji Jean-Sebastien Dovonon[1]

Samuel Atule[1]

Ashesi University

[3]Department of Computer Science, Ashesi University, Berekuso, Ghana

Abstract

Machine learning and, more specifically, deep learning have recently driven many innovations. The availability of massive datasets and computation resources has made it possible to create deeper neural networks that are able to learn more meaningful representations of the data. Those new possibilities are not always accessible to the average African company trying to leverage on deep learning to increase profit. In that case, scarcity of data, especially, could be a limitation since neural networks are known to be data-hungry. When faced with the issue of unavailability of public data, a company can either increase the size of the dataset by collecting data themselves or increase the size and complexity of the model. The option studied here is to use web scraping to manage and clean a bigger dataset. In trying to increase the size and complexity of the model, to avoid overfitting, the transfer learning approach was used. This technique involves the transfer of weights from several datasets using model ensembling. All these methods were tested on a rice meal classification problem. The problem consists of classifying images of four rice-based dishes: jollof rice, fried rice, plain rice, and waakye. The dataset contains 60 train images and 20 test images for each group making up a total of 240 training images and 80 testing images. The baseline of 75% was achieved using a dense net Convolutional Neural Network (CNN). The web scraping method used to increase the dataset size attained an accuracy of 87%. A multi-source transfer learning approach was also used where models were pre-trained on the Food-101 dataset and the Food-256 dataset. The multi-source transfer learning method achieved an accuracy of 90%. Using these two methods, we implement two ways to significantly increase the efficiency of a model when the original dataset is small.

1. Introduction

The size of training data plays a vital role in Machine Learning (ML), in learning useful representations that accurately predict the task at hand. Generally, it is common knowledge that a small dataset will result in a sparse approximation of the underlying regularity, resulting in a model with abysmal performance. Techniques have been proposed in ML literature to facilitate selecting the best model in the face of a small training dataset. Methods include employing a short model with fewer parameters, cross-validation [1], transfer learning [3], among others. However, too little data size does not often yield much improvement with these approaches. Most importantly, in the African context where data is often highly unstructured and available in minimal quantities, building models with high accuracy calls for a different approach.

This paper, demonstrates how to achieve models with high predictive accuracy, when faced with data that is highly unstructured or unavailable, using Multi-Source Transfer learning (Model Ensembling) or Web Scraping to assemble a structured dataset set for most ML tasks. Multi-Source Transfer is a common practice employed by participants in ML competitions to build winning models, while Web Scraping can be applied to create datasets in a context where no data infrastructure exists to facilitate dataset creation.

2. Materials and Methods

2.1 Materials

The following are some of the technologies we used in the project.

PyTorch: Pytorch, is a deep learning package, which is known for its high-level tensor computations and building neural networks with less effort. Pytorch is Pythonic, and more importantly, highly optimized for computationally expensive operations, such as convolutional neural networks, recurrence neural networks, and complex tensor operations. Pytorch has many pre-trained models that enhance speedy model training with appreciable high accuracy. This package is still a young player compared to its competitors like TensorFlow. However, it gains momentum very fast due to its features above.

Floyd Hub: Training deep learning models is computationally expensive; it requires machines with high processing power. The cheapest way to train the models is to purchase cloud computing services if buying a GPU for your local device is expensive. Floydhub is a cloud computing option employed to train the model under study, because they already pre-installed TensorFlow, PyTorch, Keras, and many more dependencies. Quite apart from having an extensive collection of pre-installed dependencies, Floydhub is simple to use.

Google-images-download: This technology is a python package utility for conveniently scraping images from google. However, other powerful web scraping tools exist.

2.2 Multi-source Transfer Learning

Transfer learning consists of transferring knowledge from a more extensive database (source) to a smaller database (target) using weights learned from the bigger database as a starting point when training a model for the second database. The domain of the source and how closely it is related to the domain of the target is also relevant since it can increase the efficiency of the transfer. In this experiment, weights are transfered from several sources, which is known as multi-source transfer learning. One limitation of multi-source transfer learning is that multiple sets of weights cannot be used as a starting point. Hence the use of a model ensembling method to transfer knowledge from all the datasets [2].

Multi-Source Transfer or Model Ensembling consists of pooling together the predictions of a set of different models to produce better forecasts – where the final model is trained for each of the sources. To fuse the knowledge from these various models, we used an ensemble that concatenates features extracted using the models trained on the sources, and passed the concatenated features to a multilayer perceptron (MLP). The ensemble was then fine-tuned on the target. The models were pre-trained on four datasets: the original dataset, ImageNet, Food-101 [4], and Food-256 [5].

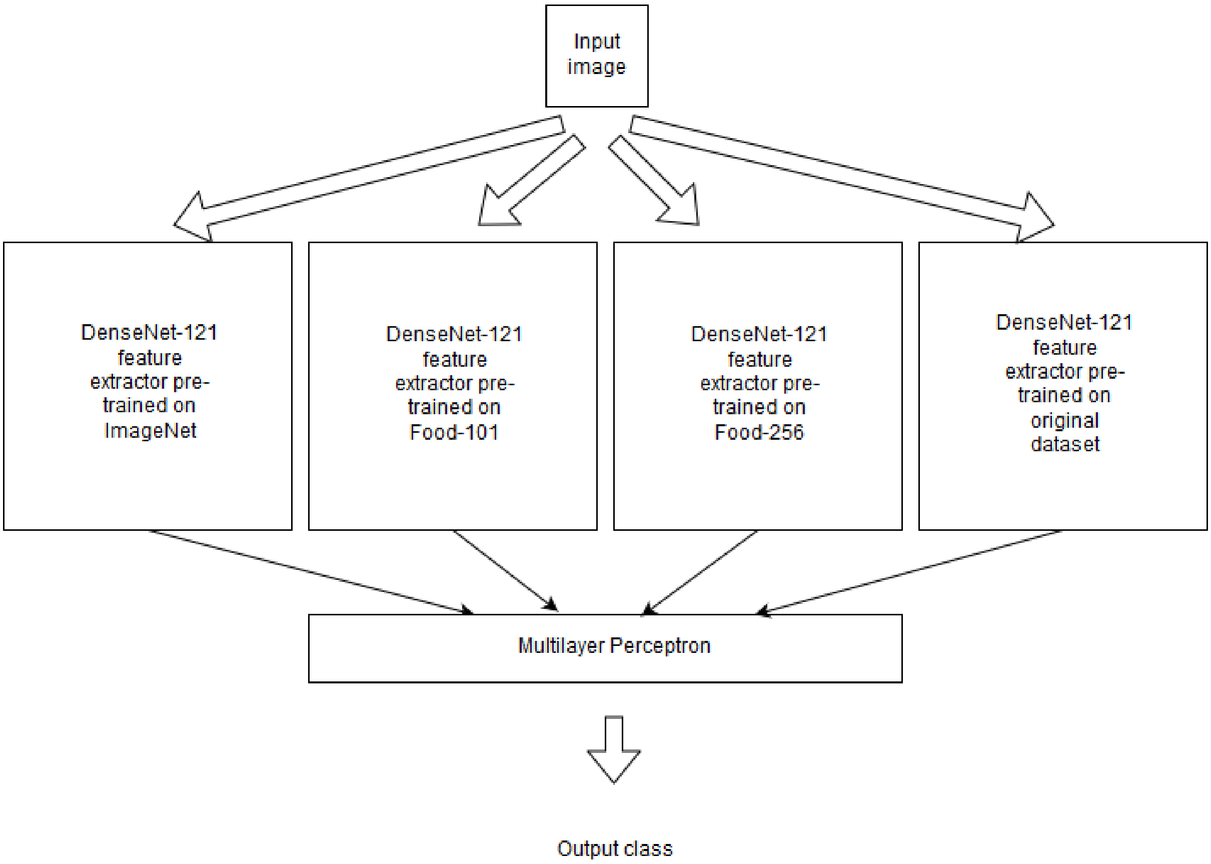
Input image
DenseNet-121 feature extractor pre-trained on ImageNet
DenseNet-121 feature extractor pre-trained on Food-101
DenseNet-121 feature extractor pre-trained on Food-256
DenseNet-121 feature extractor pre-trained on original dataset
Multilayer Perceptron
Output class

Figure 1: Diagram of the model ensemble used for multi-source transfer learning

2.3 Web Scraping

Web Scraping or web data extraction is used for extracting human-readable data from websites. It is a handy technique used in scenarios where data for a specific ML task is complicated to come by. There are not a lot of data infrastructure that exists in most organizations in the African context. Furthermore, due to bureaucracy and the lack of trust of organizations exposing their data, this technique is powerful for collecting data for various ML tasks in the Africa context.

Using this method, two thousand training images of local rice dishes were scrapped in less than thirty (30) minutes. Since a lot of junk images are often downloaded during web scraping, some manual work was done to remove unwanted files and clean up the data into a suitable format.

2.4 Implementation details

The methods tested were implemented using PyTorch and run on a virtual machine with a Tesla V100 GPU (16 GB ram) and 8 CPUs. All models were trained with a learning rate of 1e-3[1][m2]. The pre-trained models for Food-101 and Food-256 were trained for five epochs. The models (baselines and ensemble) were trained for 30 epochs.

3. Results

As a baseline method, a randomly initialized densenet-121 model on the initial dataset was trained. Using this method, an accuracy of 87% was achieved. Table 1 provides a summary of the results obtained.

Method	**Accuracy**
Baseline (randomly initialized densenet-121)	75%
Increased dataset	87%
Multi-source transfer learning	90%

Table 1: Summary of results

The pre-trained ensemble model achieved the highest accuracy and had a consistently high accuracy across epochs.

N0	Pre-trained models ensemble
N1	Pre-trained on Food-101 only
N2	Pre-trained on Food-256 only
N3	Pre-trained on ImageNet only
N4	Baseline (no pre-training)

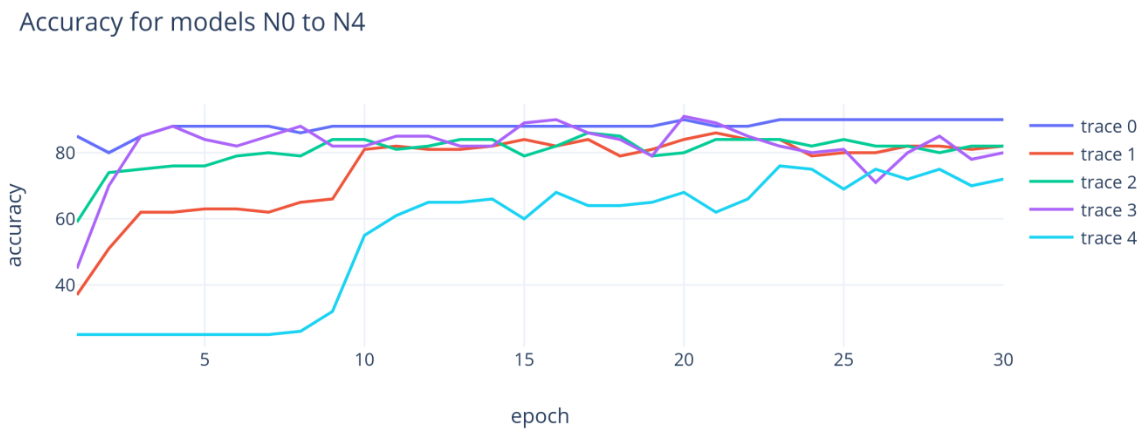

Figure 2: Accuracies for models N0 to N4

4. Discussion

After increasing the size of the dataset, with the same method, there was a substantial increase in the level of accuracy. This improvement shows the importance of large data sets, and the limitations of deep learning techniques when dealing with small data. Therefore, for deep learning to be used to its full potential in Africa, it is important that quality local datasets are made public, in order to foster research and make it easier for companies to profit from the deep learning advancement. It also showcases how useful web scraping can be as a data science tool for dataset creation.

Multi-source transfer learning resulted in a model that was able to reach high accuracies quite fast. The proposed ensemble also seems not to overfit and shows a stable increase in accuracy. Sharing pre-trained models can greatly help when faced with limited data, especially with pretrained models on various datasets. Our approach could have been made easier to implement if all the pre-trained models were already available. Also, so far this approach would not be suited for all types of inferences because of the big size of the models and their high latency. This can, however, be solved using model distillation and model compression.

5. Conclusion

In summary, this study showed has how easy it is to obtain a significant improvement in accuracy (up to 15%) over the baseline results, using Multi-Source Transfer Learning. The study also showed how to achieve a significant improvement in accuracy (up to 12%) over the baseline results by employing Web Scraping technique to acquire more dataset for a specific task. There is still a great avenue to improve upon these accuracies. We estimate that we can achieve a very high overall accuracy when these two techniques are combined. In our future work, we hope to test this hypothesis by combining these two techniques.

Acknowledgement

We acknowledge Dennis A. Owusu, our ML lecturer at Ashesi University for the guidance and support during this project. The experiment of this model made possible by FloydHub and their cloud platform.

References

[1] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *Research gate*, 2001. Available: https://www.researchgate.net/profile/Ron_Kohavi/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection/links/02e7e51bcc14c5e91c000000.pdf. [Accessed 12 June 2019].

[2] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe and S. Mougiakakou, "Multisource Transfer Learning With Convolutional Neural Networks for Lung Pattern Analysis", *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 76-84, 2017. Available: 10.1109/jbhi.2016.2636929.

[3] Ng. Hong-Wei & Nguyen, D. Vonikakis, V. Winkler, Stefan. " Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning". Available: 10.1145/2818346.2830593.

[4]L. Bossard, M. Guillaumin and L. Van Gool, "Food-101 -- Mining Discriminative Components with Random Forests", *Vision.ee.ethz.ch*, 2019. [Online]. Available: https://www.vision.ee.ethz.ch/datasets_extra/food-101/. [Accessed: 13- Jun- 2019].

[5]L. Bossard, M. Guillaumin and L. Van Gool, "Food-101 -- Mining Discriminative Components with Random Forests", Vision.ee.ethz.ch, 2019. [Online]. Available: https://www.vision.ee.ethz.ch/datasets_extra/food-101/. [Accessed: 13- Jun- 2019].