The Effect of Data Content and Human Oracles on Convolutional Neural Networks and Incremental Learning.

Stanley D Rosenbaum
University of Kentucky
Davis Marksbury Building
329 Rose Street
Lexington, KY 40506-0633
srose3@uky.edu

Abstract

In building Convolutional Neural Network (CNN) models that can aid in accessibly technology for navigation for the visual impaired, this paper examines two aspect of model development regarding the training data. The first aspect is if the content of the training data can influence accuracy of the training model. In this case, we measure the differences of training on images of staircases verses 3D point-cloud data of the same object. The second aspect is if incremental training with human reinforcement can improve performance and training time for a CNN models. The results show, that although data content can have a measured effect on the training of a neural network, incremental training (retraining) when the number of identification classes stay the same does not.

Intro

As Machine Learning (ML) has progressed the consensus for training of ML models has been to train them on larger and larger data sets. Unfortunately, training sets are starting to become so large that even with advances in CPU/GPU horsepower the time and computing resources needed to train has greatly increased as well. This research is aimed to examine two aspects of training sets to determine if there is a path that would allow a more agile or leaner approach to training.

Convolutional Neural Network (CNN) models are examined in this paper, since there is an established history of their use in object identification in images. Specifically, CNN models created in this research to identify staircases for use in the vision impairment navigation assistance applications. CNN models examined in this domain offer an opportunity to determine if the underline data content has a measured effect on the training a CNN for a specific task. Additionally, the domain offered and opportunity to examine if human reinforcement can allow for more optimal approaches to structuring training sets for CNN models.

However, in the end, this paper will demonstrate two discoveries. First, the underlying data content CNN models can have a noticeable effect on training of the CNN model. Second, incremental learning on convolutional neural network with established classes training has a negligible effect, if any.

Related Work

The origin of the research for this paper stems from three domains. The first is the area of accessibility technology for the visually impaired. The research for this project is intended to expand the functionally of the SEAR-RL system. SEAR-RL is an iOS based application that uses the 3D Sensor Occipital Structure to translate

real-world environments into audio for obstacle avoidance by the visually impaired. Prior work was done during the Fall 2016 semester as part of a Masters Project [1]. The project focused on a general-purpose system for identification and representation of 3D environments via audio using an Apple iPhone and Occipital Structure 3D Scanner (similar to a Microsoft Kinect). However, there was plenty of room for SEAR-RL to examine the area of specific object identification.

Similar work in assistive technology has been done by those such as Liang Nu and colleagues' 'A Wearable Assistive Technology for the Visually Impaired with Door Knob Detection and Real-Time Feedback for Hand-to-Handle Manipulation' [2], Bor-Shing Lin and colleagues', 'Simple Smartphone-Based Guiding System for Visually Impaired People' [3], and Mekhalfi and colleagues', 'Fast indoor scene description for blind people with multiresolution random projections' [4]. However, most research for navigation using 3D sensors has focused on automaton navigation for robots [5], or drones [6]. Still, there has been some quality work in specialized algorithms development for identification of staircases [7] for said accessibly assistant systems. However, this research was curious if a generalized CNN model could be trained for object detection of stair cases.

Finally, since there was no pre-existing data set with a focus on staircases research was also interested in exploring the use of incremental learning for continually training and improvement of CNN models. The hypothesis being that when faced with a smaller or non-existent dataset (such as data for staircases) maybe use of a measure approached to training could improve overall results.

There is a body of research that does exist in this area. Most prominent, is probably 'Error-Driven Incremental Learning in Deep Convolutional Neural Network for Large-Scale Image Classification' [8] by Xiao and colleagues. There is also 'How transferable are features in deep neural networks' [9] by Yosinski and colleagues. However, both papers focused on expanding the number of identified classes from and existing trained model. This research work aimed to focus on the effect of incremental human reinforcement training for improving the performance of a model on a fixed number of classes.

Methodology

This project explored two different areas for this research. First, is there an effect of data content in the training of neural nets? The research focused on comparing different data sets of the same object on an identical CNN structure. Specifically, the research aimed to compare traditional image data from a camera verses point cloud data from a 3D sensor. This was a fairly simple comparison exercise and will be covered in detail in the experiments section.

Second, the approach for studying incremental training needs a deeper introduction. In the past convolutional neural nets (CNN) has proven themselves to be very good at identifying objects in images via supervised learning. However, most machine learning models always result in a model that produces a certain amount of error. This research project was curious if assistance from a human teacher could improve overall training of a CNN via reinforcement learning. Specifically, a human teacher (or human-oracle) whose goal was to examines a CNN's performance and identify areas of weakness. Once identified, the human can offer up additional examples in which the CNN can be trained.

This is idea is akin to a music teacher providing finger exercises for a student. If a music student is having difficulty performing a small section in a piece of music, a teacher can provide the student a short song, often referred to as an etude or similar finger exercise. By practicing and mastering the etude the student's overall skill and performance of the larger piece of music can drastically improve.

The hypothesis for a method of improvement is to apply a similar approach to a trained CNN. Starting with a CNN trained on a small base set

of samples, the initial model is examined for which classes is has the most difficulty identifying. The human trainer (or oracle in this case), then would add additional samples to the initial base training set and the CNN is trained again on the expanded set. This process is repeated until training improves to the desired accuracy.

Experiments

The project research required two sets of experiments. A first preliminary experiment to discover a common CNN that would work for various data. Then a second main experiment to test the hypothesis of incremental learning.

Collecting data.

Since there is not a proper archive of samples of point cloud depth data, original data had to be collected for this project. 1802 samples were collected divided into 4 classes consisting of 556 for UP (for data collected from the bottom of a set of stairs looking upwards), 402 for DOWN (data from the top of a staircase looking down), 408 for NA (representing flat or sloped ground without stairs), and 436 for HOLE (representing an elevation change downwards greater than the standard step rise (7 ¾ inches or less) [10]. The final data set was about 8 gigabytes in size.

Preliminary experiment – finding the right CNN

The project needed to determine an adequate general CNN structure that could be used for two primary types of data encountered in the research. This was viewed as an opportunity to examine different forms a data representation on one neural network. Working in the research's favor was the fact, that each sample of the image and point cloud data, effectively had the same resolutions. Both data sets had dimensional resolutions of 640x480 pixels which eliminated any possible resolution size problems. The bulk of the difference arose from the depth of data at each pixel.

Since there was only one channel of point cloud data, it was decided to use greyscale images (which only have one channel of data) to make a fair comparison and helped simplifying the CNNs. The data for each image came from the PNG image format used to store data on disk, which the default data storage is an unsigned 8-bit integer. For the depth data, the default storage format is arrays if 32-bit floats. Although it would seem that depth data could potentially have greater depth resolution compared to the images, this is mitigated in that the point cloud data only spreads across a range of 0 to 10,000 with a maximum of 3 decimal places of precision.

Another difference in the data is the field of view for the onboard iPhone Color Camera effectively extends into the infinite on the z-axis away from the camera. Although, all objects may not be in focus, a normal camera will capture objects as far as the resolution will allow. This is not true for the depth-camera of the Occipital Structure. The depth camera only has an effective range of 4 meters or about 15 feet. Therefore, while images contain a significant larger amount of varied data from and infinite range the depth image and point cloud data only contains about 125 cubic feet of depth data. Thus, the results, is that the one pixel of image data, with a smaller resolution, can be representative of a greater distance. While the point cloud data has a greater resolution over a smaller distance.

As an additional bonus the Occipital Structure does offer the ability to automatically convert any point clouds it records into a grayscale PNG file. This added data can act as union and difference set in bridging the PNG image data taken with the color camera with the raw point cloud data. It offers a domain to help insure our main data sets of inquiry remain disjoint.

Considering the staircases are relatively, simple geometric structures, a general CNN was

picked for use. This is because as CNN structures add more layers the deeper layer tends to detect finer features. This research only focused on detecting the general shape of ascending or descending staircases. Attempted to detect too fine of layer, could lead to overfitting.

All these factors were taken into consideration using a basic grid search on preliminary collected data. The goal was to find a CNNs structure that appeared to have good results for the image and depth data. The combinations searched were 4-512 nodes and 1-4 layers. The resulting structure that was picked consisted of: 2 standard convolution layers, the first with a 5x5 filter with a stride of 3 and the second a 3x3 filter with a stride of 2; then a final a flattened and fully connected layer of only 48 nodes.

Experimental Techniques for Incremental learning

To measure the effect on incremental learning the experiment was divided into three evaluations. First, a base case to determine how well a CNN can learn features using batch training. The incremental learning began by training the CNN with a traditional batch method. Each data flavor (image from color camera, image of point cloud, raw point cloud data) was ran separately. The data was split via a 60/20/20 and accuracy for each class and measured.

Evaluations two and three consisted of a human-oracle examining the performance of training and adding the appropriate data to the training set. Stage two (human-oracle) focused on adding samples specifically complimenting the weakest performing classes of the previous training. Stage three (random-oracle) focused on adding random samples without regard to performance serving as a control to stage two.

One term needs to be defined. To understand the set of all training set for the reinforcement stages the concept of an ‘Age’ is introduced. Machine Learning traditionally describes each pass of training set as an ‘Epoch’. The training of a machine learning model, can usually consist of several epochs until the training loss and accuracy reaches desired levels. For this research, the analysis of several sets of a series of epochs had to be measured. To differentiate between each series the meta-term ‘Age’ will be use to describe one series of epochs. There ended up being two types of Ages which were run. The first was the initial Age which ran for 24 epochs and sub-sequential ages which ran for 6 epochs. The shorten series for the sub-sequential Ages was because of observations that test loss and accuracy score would flatten after only a few epochs, and therefore those Ages were cut short.

The human-oracle, would compare the training of the previous age to a set of validation samples which also contains an equal number of samples from each class. The class in which the trained model scored the lowest was noted. The human-oracle would then select from an unlabeled set that matched the weakest performing class and add 20 samples of that class to the training set. At that point another Age of training would take place using the existing model and weights from the previous age. To provide a control for the random-oracle would select 20 samples from all 4 classes indiscriminately. Each evaluation ran for 20 ages (the maximum that data set would allow).

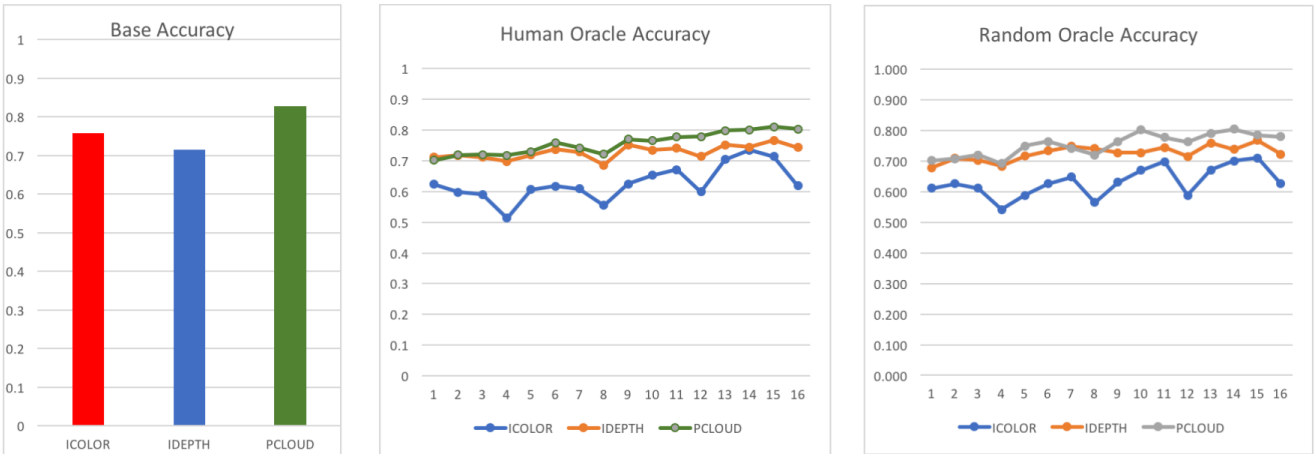

Figure 1 - Training Comparisons of the Base, Human-Oracle, Random-Oracle

We finally note the content of each initial training set. Each training set for each evaluation initially contained an equal number of samples of each class to be identified. Also where the base evaluation consisted of 60% of the total data samples, the human oracle stage (evaluation 2) and random-oracle (evaluation 3) each began with only 20% of the total data samples as a start set size. The human and random oracle were ran with 5 differently initial training sets, and compared to 5 different validation sets for a total 25 different test which results were collected and averaged together.

Results & Discussions

The results showed that although data content, can have an effect on the training of a CNN model incremental learning does not.

Data Content

There is no question there was a difference in the training accuracy between using images, verses images of the depth data, verses the straight point cloud depth data. Raw point cloud data performed the best, followed by images of depth data, followed by raw images. There are several factors which could results in this. First the sample depth at the pixel level is just greater for the point cloud compared to using images. The greater range would allow the neural net a bit more room in discriminating between different data sets for identifying classes. However, this does not completely explain why the images of the depth data performed better then straight images since the pixel data depth for these two categories are equivalent. The theory for the increased performance in the depth images, could be attributed to the data contained in each depth-image is for a fix volume of space. The images from the camera contains a snap shot of everything in its field of view stretch to infinity from the camera, the depth data contain a much smaller constrained area. Anything beyond ~15 feet is not registered by the Occcipital Structure sensor, and therefore is not recorded. Using this focused data set lends to improve accuracy as there is less data for the neural net to sift through.

Incremental Learning.

These results require a bit more finesse to explain. It is not that the results for this paper‘s research show a contradiction to the use of it incremental and transfer learning. It shows there is a qualification to it.

Earlier papers which explored incremental and transfer learning attempted to expand the recognized classes of a trained neural network model. As models were trained, it is reasonable to think overall training for previous classes would be improved. Samples which existed on a decision

boundary in a previous neural net model would be resolved into more definite classes as classes are added. However, for this research there is negligible improvement, if any, when trying to improve on existing training. This is most likely because the size of the decision boundary for determining a class remains fixed. A sample that falls in an ambiguous boundary would remain so with each training age as nothing is being introduced to radically shift the decision boundaries of the neural net. It is the equivalent of trying to slice a half and half pizza into an odd number of equal sized slices while attempting to ensure no slice contain elements from each half.

Conclusions

Incremental learning for convolutional neural networks is an area that still has room for exploration. However, in this context where the number of classes remain the same throughout training there is no significant benefit. Still a direction for future work would be to explore adding a method of classes at the same time of as adding training samples for each subsequent Age. This may provide an avenue to improve training. Until then, it seems the de facto standard of batch training for neural networks will remain so.

However, research does show that data content can have a significant different on the quality of trained models. Implying for 3D object identification, there is a motivation to use 3D data verses trying to abstract 3D data from a 2D image.

Study Recreation Information

Developed code can be found at: https://github.com/ForeverTangent/CS-660-Semester-Project-V2

Source Data can be found at: https://drive.google.com/drive/folders/0ByeMvlBTKM3vd3dQZUtMZEV1Skk?usp=sharing

References

[1] S. Rosenbaum, "Constructive Noise," Constructive Noise, 17 09 2017. [Online]. Available: http://constructive-noise.info/?cat=28. [Accessed 10 Dec 2017].

[2] L. Niu, C. Qian, J.-R. Rizzo, T. Hudson, Z. Li, S. Enright, E. Sperling, K. Conti, E. Wong and Y. Fang, "A Wearable Assistive Technology for the Visually Impaired with Door Knob Detection and Real-Time Feedback for Hand-to-Handle Manipulation," 2017.

[3] B.-S. Lin, C.-C. Lee and P.-Y. Chiang, "Simple Smartphone-Based Guiding System for Visually Impaired People," *Sensors,* vol. 17, no. 6, p. 1371, 13 Jun 2017.

[4] M. L. Mekhalfi, F. Melgani, Y. Bazbi and N. Alajlan, "Fast indoor scene description for blind people with multiresolution random projections," *Journal of Visual Communication and Image Representation,* vol. 44, pp. 95-105, 30 April 2017.

[5] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki and K. Fujimura, "The intelligent ASIMO: System overview and integratio," in *Proceedings of the 2002 IEEE/RSJ*, Lausanne, 2002.

[6] D. K. Kim and T. Chen, "Deep Neural Network for Real-Time Autonomous Indoor Navigation," *arXiv preprint arXiv,* 15 Nov 2015.

[7] T. Tang, W. L. D. Lui and W. H. Li, "Plane-based detection of staircases using inverse depth," *Australasian conference on robotics and automation,* 2012.

[8] T. Xiao, J. Zhang, K. Yang, Y. Peng and Z. Zhang, "Error-Driven Incremental Learning in Deep Convolutional Neural Network for Large-Scale Image Classification," *Proceedings of the 22nd ACM international conference on Multimedia,* pp. 177-186, 3 Nov 2014.

[9] J. Yosinski, J. Clune, Y. Bengio and H. Lipson, "How transferable are features in deep neural networks?," *Advances in neural information processing systems,* pp. 3320-3328, 2014.

[10] Stairbuilders and Manufacturers Association, "Visual Interpretation of the International Residential Code 2006 Stair Building Code," 2006. [Online]. Available: https://www.stairways.org/Resources/Documents/2006%20Stair%20IRC%20SCREEN%20web%20download.pdf. [Accessed 1 10 2017].