Exploring diversity of task and data in artificial neural networks

Daniel Goodwin
MIT Media Lab
dgoodwin@mit.edu

David Rolnick
MIT Math Department
drolnick@mit.edu

Abstract

In this project, we used a known convolutional network as a substrate for exploring the impact of diversity in artificial neural network development. Our testbed was 16,000 annotated images on the two separate tasks of age and gender classification. Based on neurobiology and cognitive science, we define diversity as variance in data, task and shared internal representation. From this, we ran experiments testing hypotheses of sparsity, noise and shared knowledge between neural systems. In effect, we explored a microcosm of a cognitive architecture.

1. Introduction

One avenue to exploring the future of AI is to think how the successful sub-symbolic systems such as convolutional networks may used a substrate to explore pieces of a cognitive architecture. For this class project, we have explored the important of diversity in learning from the human perspective and explored if such principles could be expressed in multilayer convolutional networks.

1.1. Motivation and Prior Work

As children we are exposed to many different lessons from many different domains. We aggregate these multiple learnings into a form of general intelligence with which we improve our efficiency in future learning. There are two pieces of prior work that are of particular relevance. In one of his papers utilizing the LEABRA cognitive architecture, Randall O'Reilly demonstrates a modeled multipart neural architecture that includes four neurons that are dedicated to identifying the task. If O'Reilly et al represents a more theoretical end of the artificial neural network spectrum, then the multitask learning research is on the more applied side.

The idea of training a multilayer convolutional network on both a primary task and a secondary task was presented in 2008 by Collobert et al and in 2013 by Seltzer et al, achieving state of the art performances in natural language processing and speech recognition, respectively. The goal of this project is to chart a course down the center of the spectrum between theoretical and applied: can we utilize existing convnet infrastructure to show that a neural network can benefit from diversity just as the human brain?

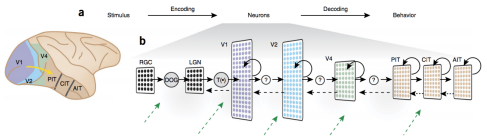

Figure 1. The human ventral stream has a many-layered hierarchy before the various cortical subregions work on specific subtasks such as facial recognition or physical expectation

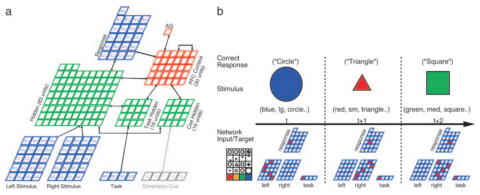

Figure 2. An application of O'Reilly's LEABRA architecture [3] that (a) includes an extra four neurons to encode and recognize the variety of tasks

1.2. Defining Diversity

For the context of this research, diversity is defined as:

1. *Variation of training and testing data* In the context of a single machine learning task, how does the variance and novelty contribute to the overall performance of the algorithm?

2. *Multiple tasks* Can a single architecture simultaneously learn and perform multiple tasks? Specifically, can the nature of learning multiple tasks improve the individual performances of each task?

3. *Variance in internal state* Can internal diversity to a learned representation or inside computational unit contribute to improved performance of the network?

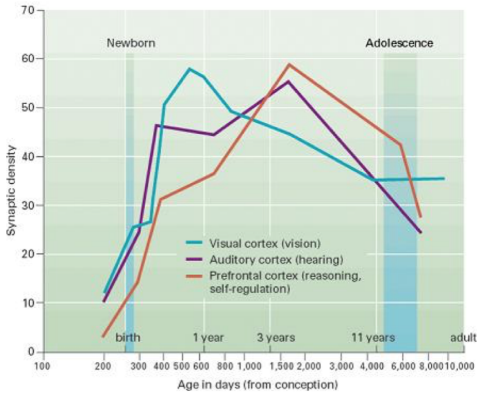

Figure 3. Synaptic pruning is a lifelong process starting after a peak synaptic density early in development. Taken from [2]

This variance could be either in the neurons recruited to a specific task or a modulation in the output.

1.3. Neuroscience Background

There are several concepts in neurobiology that inspire the three aspects of diversity enumerated above. Novelty is essential for developing critical neural circuitry, perhaps best illustrated by the classic 1970 work of Blakermore and Cooper showing that without experiencing horizontal lines early in development, the feline visual system will be unable to perceive horizontal lines later in life [6].

There are a multitude of shared architectures in the brain, perhaps best illustrated by the hierarchical perceptual systems. The ventral stream, for example, processes raw information through many layers (RGC $\rightarrow$ LGB$\rightarrow$V1$\rightarrow$V2$\rightarrow$V4) before the cortical subareas do the task-specific cognition such as facial recognition or object detection. It is an inevitability that designing shared architectures across multiple tasks and dominas will be an essential step in artificial neural networks.

Finally, we know that noise contributes to the information storage and processing ability of the brain, both on the scales of networks and individual cells. For example, Alan Kay and John White showed in a 1998 piece that varying modeled noise in the ion channels could induce qualitatively diffenent behavior modes in a neuron, indicating an improvement richness of representation that a network could contain. [1]

2. Methods

Using the Adience Benchmark dataset, we had access to 16,000 cell phone images of faces that have been annotated by age, gender and head-tilt and aligned to a consistent face position. The distribution of data used in this paper 4. We used the 6-layer convNet architecture used in [4], implemented the network in Caffe and trained on a machine with 256GB RAM and a NVIDIA GPU. With this setup, training 8000 iterations for two networks on 5GB of data took about 70 minutes, and we completed over 20 training experiments throughout this project.

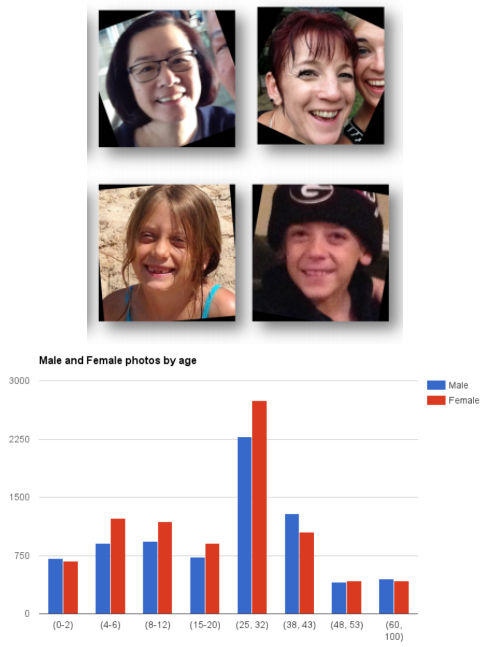

Figure 4. Distribution of face data in the Adience [5] dataset

Using this well-annotated dataset, we explored how the two seperate cognitive tasks of age classification and gender identification could benefit from differing amounts of data variance, filter sparsity, filter noise and shared number of lower-level layers. Below is an explanation of each modification based on the diversity principles enumerated above.

2.1. Shared Network layers

Extending ideas from [8] and [9], we used the same neural network architecture for the two separate tasks of age and gender classification, sharing some subset of the first convolutional layers. This is diagramed in Figure [**?**] and implemented in Caffe.

2.2. Progressive Sparsity

Extending ideas from Dropout [7] and Optimal Brain Damage [11], we explored a method of sparsity in the convolutional layers. Part of this project was discovering how many variants of this technique have been done in the field: our options were either to reduce redundant filters through a low-rank approximation such as [10] or to selective zero-out individual weights per filter. For rapid prototyping in the context of a class project, the fastest way to explore this

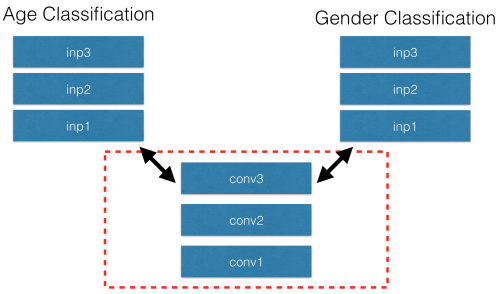

Figure 5. Showing the system

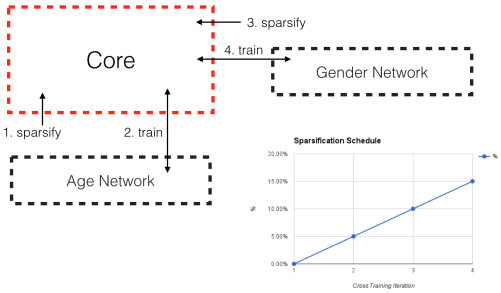

Figure 6. Sparsification loop

effect was to identify the smallest weighed elements per filter and clamp them to zero. That is, as the network trained through gradient descent of the backpropogation error, the pre-identified smalled weights that had been set to zero were kept to zero.

So the system, shown in Figure 6 is explained in steps below. Note that the sparisification schedule, also shown in Figure 6 was chosen to be $[0.0\%, 5\%, 10\%, 15\%,]$ across the four training iteractions

1. Initialize networks
   (a) Train Age network (training data: 6000 female faces, testing data 1000 male faces)
   (b) Initialize Gender network with the first few convolutional layers of the Age network
   (c) Train Gender network (training data: 3000 male faces, 3000 female faces. Testing data 1000 faces of both gender)
2. For training age network:
   (a) Copy the trained core layers from Gender network back into the age network
   (b) Identify the smallest magnitude $x\%$ weights per filter
   (c) In each step of gradient descent in training, clamp those weights to zero

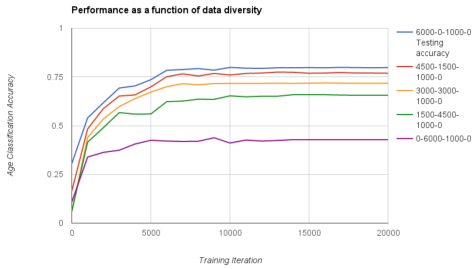

Figure 7. Testing results on an all-male image set based on varying the training data from all-male (blue line) to all-female (purple line). The legend is in the format $N_{male,training}N_{female,training}N_{male,testing}N_{female,testing}$

3. For training gender network:
   (a) Copy the trained core layers from Age network back into the gender network
   (b) Identify the smallest magnitude $x\%$ weights per filter
   (c) In each step of gradient descent in training, clamp those weights to zero
4. Repeat steps (2) and (3) for 4 iterations.

2.3. Noise

In a very similar implementation to the sparsification process, progressive amounts of noise are added. In this case, the progressive noise parameter was a function of statistics of the weights per filter. So, a 20% noise parameter would mean a gaussian distribution centered at the mean of the filter with a variance of 20% of the mean weight. The progressive noise addition follows the steps sparsification except with the addition of noise at every step in the gradient descent.

3. Results

The summary table of relevant experiments can be seen in Table 1. Below is an explanation of the results.

3.1. Diversity of Data

The Age classification network was trained with a varying amount of male/female faces and testing on all-male corpus and the results are plotted in Figure 7. The dependency on male faces showed to be very sensitive to including a few male faces: the marginal improvement on adding the first 1500 male faces was bigger than the effect of adding the next 4500 combined.

3.2. Sharing layers of the Core between tasks

To highlight the maximal effect of learning from multiple tasks and data, we only trained the Age network on

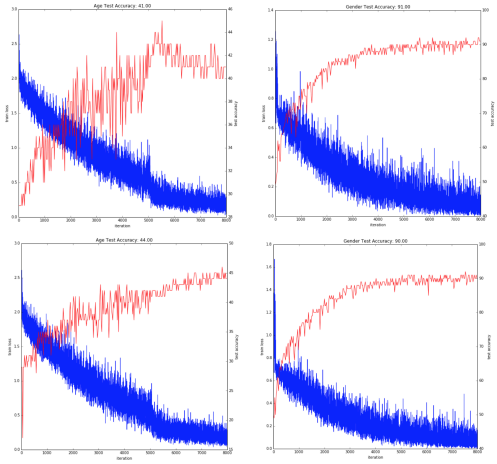

Figure 8. The left column of plots is for the Age network, the right column is the Gender network. The blue training loss and the red testing accuracy across different shared core architectures. The top row is independent: ie, the Age network and Gender network are completely separate. The bottom row shows the benefit of sharing two of the three convolutional layers.

women and tested on men, whereas the Gender network was trained on an even split of men and women. The hypothesis here is that we should see an improvement on the Age network as it benefits from the representations learned by the gender network. As can be seen in Figure 8, the testing performance of the shared network architecture shows a better classification for the Age network, and the slope of the testing accuracy indicates that it would have continued to improve with more iterations.

3.3. Sparsification

To confirm that Caffe was indeed performing as expected, we set the sparsification to 100%, zeroing out every weight in the first two layers of the network. As expected, this set the network to constantly predict the same classification label. Then we explored the effect of differing layers of sparsity. Figure 9 shows the impact of the sparsification event, and also the rapid rate of recovery for the network. This is of particular interest given that the training parameters of the network, consistent across all experiments, is such that the learning rate is decreasing by an order of magnitude every 2000 iterations as well.

3.4. Adding noise

Adding noise to a the first three convolutional layers that is sharing the first two convolutional layers between the two tasks showed a slight increase in performance. The results can be seen in Figure 10

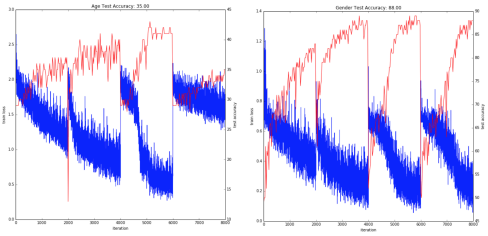

Figure 9. The effect of progressive sparsifying a network as the system goes between training the Age and Gender networks. In this case every 2000 iterations the training is paused for one network and continued for the other. When the network is continued to train, the system chooses an additional set of weights to sparsify.

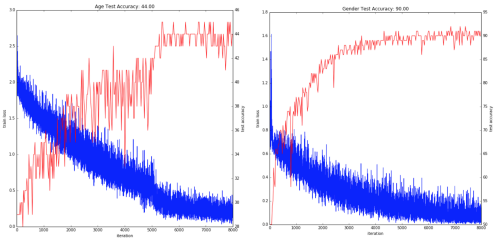

Figure 10. The effect of adding noise to the system at every step in the gradient descent t

Network	Age Test Acc.	Gender Test Acc.
Independent	41%	91%
Share2	44%	90%
Share3	41%	90%
Share2+sparsity2	35%	88%
Share2+noise3	44%	90%
Share2+noise3+sparse2	32%	86%

Table 1. Summary of results: Shared layers with progressive noise yielded the best result.

4. Discussion and Conclusion

This work is a first step towards some more focused exploration in the relevant parameters discovered through this project. On one hand, most results here are not entirely surprising: sparsification may not hurt testing performance but doesn't necessarily add to it. Adding a bit of noise improves the stochastic gradient descent and a shared representation between two tasks learning on two different sets of data can demonstrably improve results. While we see some signal pointing towards the benefit of noise and sharing representation, more work would be necessary to demonstrate statistical significance.

But this work uncovered a lot of subtlety here for future explorations. Could an optimized sparsity algorithm help generalize to novel data? Could an additional task have

nothing to do with image but still show a benefit to the age classification task, a step toward the Learning Using Privileged Information paradigm that to date has only worked with SVM? What is the most biomimetic way of training the core network while performing two differing tasks?

Finally, and personally speaking, this was our first attempt at using a deep convolutional network to prototype a cognitive architecture. These two networks could be thought of as the fusiform face area (face detection) and lateral occipital complex (shape detection); they are two networks that sit roughly at the same hierarchy of cognitive processing. An interesting path would be to explore the use of hippocampal circuitry into these multi-piece networks.

5. Acknowledgements

We would like to thank all the people involved in the MIT Media Lab's Special Seminar: Future AI for the stimulating discussions. We would also like to thank Aditya Khosla for his support with access to powerful computing machines and his deep knowledge of Caffe.

References

[1] White et al "Noise from voltage-gated ion channels may influence neuronal dynamics in the entorhinal cortex." J Neurophysiol. 1998 Jul;80(1):262-9.

[2] Blakemore "The social brain in adolescence" Nature Reviews Neuroscience 9, 267-277 (April 2008) — doi:10.1038/nrn2353

[3] Rougier, N.P., Noelle, D., Braver, T.S., Cohen, J.D. & O'Reilly, R.C. (2005). "Prefrontal Cortex and the Flexibility of Cognitive Control: Rules Without Symbols. Proceedings of the National Academy of Sciences", 102, 7338-7343.

[4] Gil Levi, Tal Hassner, "Age and Gender Classification using Convolutional Neural Networks, IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG)", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, June 2015

[5] Publication: E. Eidinger, R. Enbar, and T. Hassner, "Age and Gender Estimation of Unfiltered Faces," Database: `http://www.openu.ac.il/home/hassner/Adience/data.html#agegender`

[6] Blakemore, Colin, and Grahame F. Cooper. ?Development of the brain depends on the visual environment.? (1970): 477-478.

[7] Srivastava, Nitish and Hinton, Geoffrey and Krizhevsky, Alex and Sutskever, Ilya and Salakhutdinov, Ruslan, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting" J. Mach. Learn. Res. 2014

[8] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 6965-6969.

[9] R. Collobert and J. Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In International Conference on Machine Learning, ICML, 2008.

[10] Jaderberg, Max, Andrea Vedaldi, and Andrew Zisserman. "Speeding up convolutional neural networks with low rank expansions." arXiv preprint arXiv:1405.3866 (2014).

[11] Yann Le Cun, John S. Denker, and Sara A. Solla. 1990. Optimal brain damage. In Advances in neural information processing systems 2, David S. Touretzky (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 598-605.