Exploring Impact of Diversity in Multilayer, Multitask Neural Networks

Dan Goodwin
(original idea formulation with Davie Rolnick)

Motivation

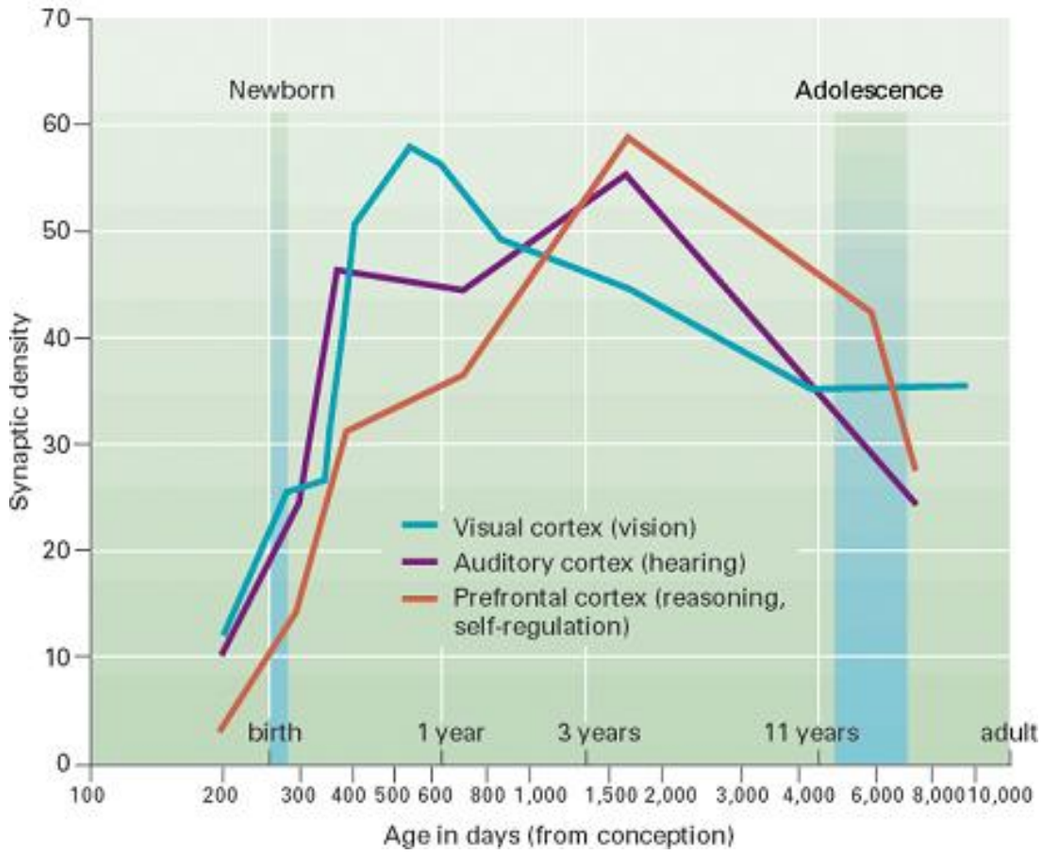
Human learning benefits from aggregating lessons across a multitude of domains, scales, tasks. How might an artificial brain also develop?

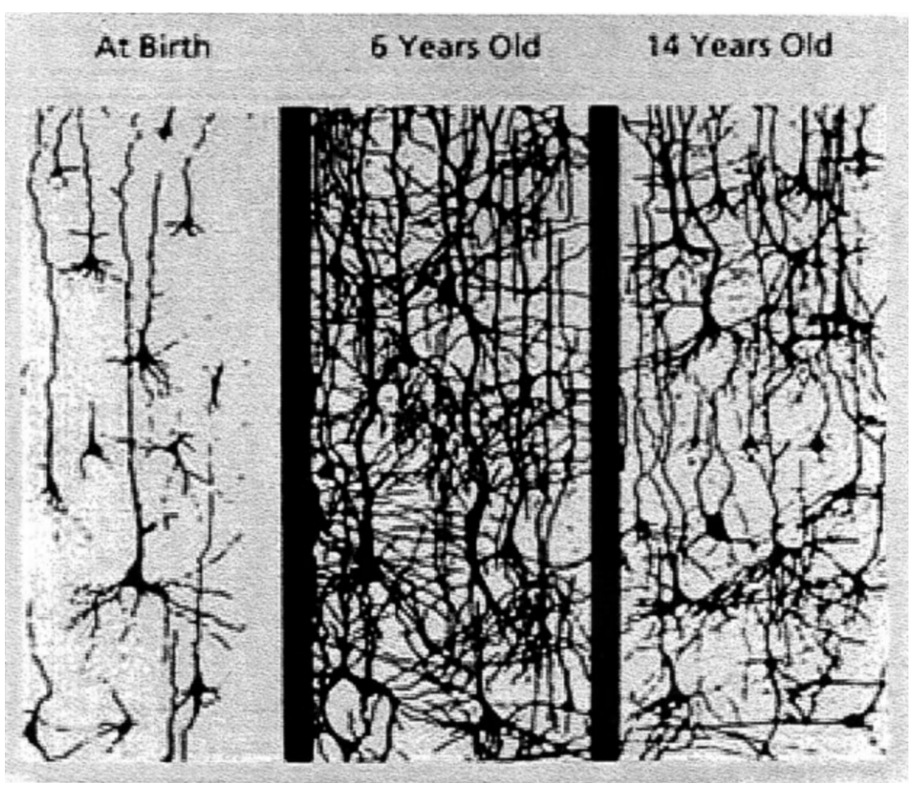
“Diversity”

- Variation of training and testing data
- Multiple tasks
- Variance in internal state

Neuroscience foundations

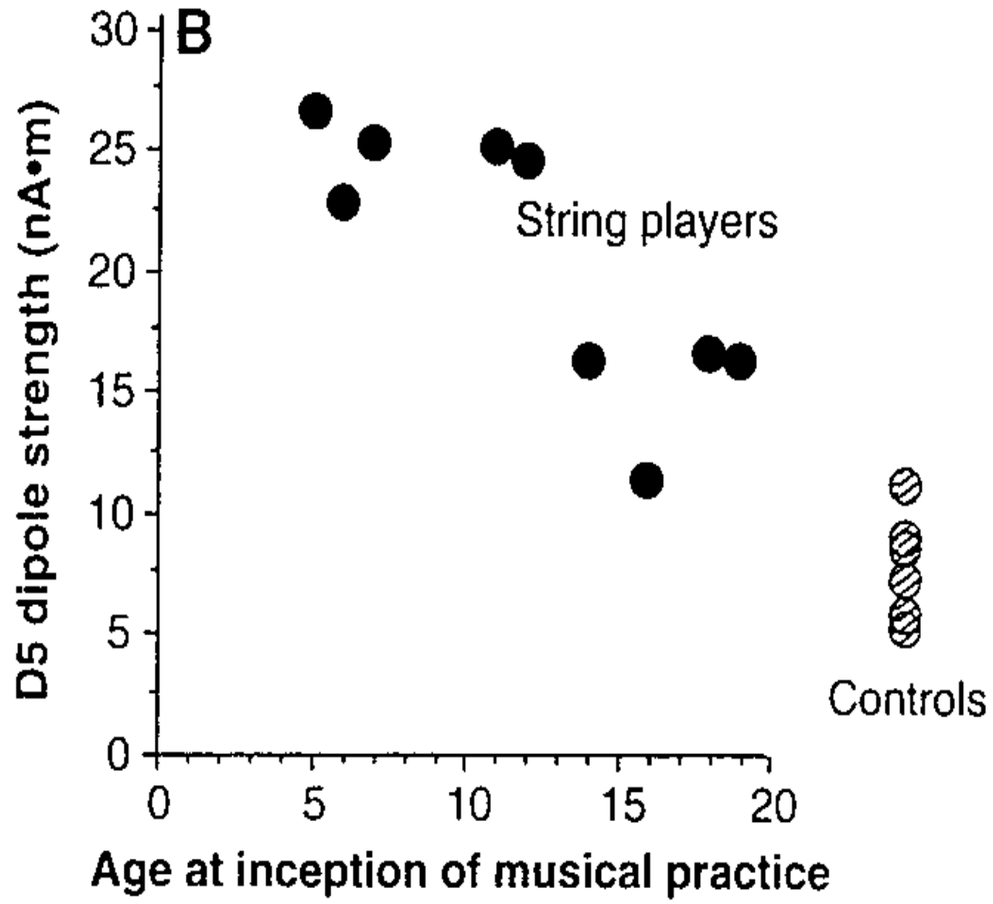
- Critical periods of neurodevelopment are marked by mass-pruning in neuronal networks





Neuroscience foundations

- Tasks and languages learned at younger ages recruit larger cortical areas.



Neuroscience foundations

- Noise is ubiquitous in biological systems. In the brain, it can be a mechanism for additional represetnation

Noise From Voltage-Gated Ion Channels May Influence Neuronal Dynamics in the Entorhinal Cortex

JOHN A. WHITE,[1] RUBY KLINK,[2] ANGEL ALONSO,[2] AND ALAN R. KAY[3]
[1]*Department of Biomedical Engineering, Center for BioDynamics, Boston University, Boston, Massachusetts 02215;*
[2]*Department of Neurology and Neurosurgery, Montreal Neurological Institute, Montreal, Quebec H3A 2B4, Canada;*
and [3]*Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242*

White, John A., Ruby Klink, Angel Alonso, and Alan R. Kay. Noise from voltage-gated ion channels may influence neuronal dynamics in the entorhinal cortex. *J. Neurophysiol.* 80: 262–269, 1998. Neurons of the superficial medial entorhinal cortex (MEC), which deliver neocortical input to the hippocampus, exhibit intrinsic, subthreshold oscillations with slow dynamics. These intrinsic oscillations, driven by a persistent Na^+ current and a slow outward current, may help to generate the theta rhythm, a slow rhythm that plays an important role in spatial and declarative learning. Here we show that the number of persistent Na^+ channels underlying subthreshold oscillations is relatively small ($<10^4$) and use a physi-

Contributions of the inherently stochastic nature of voltage-gated ion channels to neuronal noise levels are widely assumed to be minimal because of the large number of channels involved. However, evidence from experimental (Johansson and Århem 1994; Sigworth 1980; Verveen 1961), theoretical (Chow and White 1996; Lecar and Nossal 1971a,b) and computational (Rubinstein 1995; Schneidman et al. 1998; Skaugen and Walløe 1979; Strassberg and DeFelice 1993) studies indicates that noise from voltage-gated channels can have important effects at the cellular level.

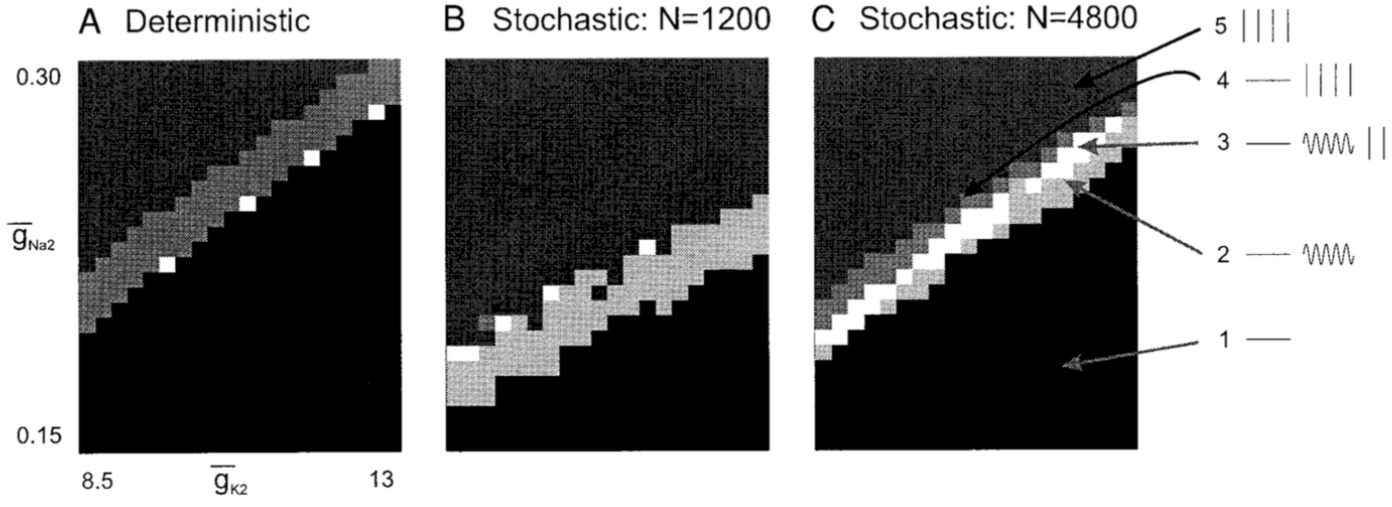


FIG. 2. Channel noise affects qualitative behavior. Shown are maps of qualitative behavior as a function of the maximal values of the persistent Na^+ conductance ($\bar{g}_{Na2}$) and slow K^+ conductance ($\bar{g}_{K2}$) for 3 model types. Changes in $\bar{g}_{Na2}$ or N were effected by changing the size of the membrane area that was simulated, thus leaving all other aspects of the model unchanged. Behavior was mapped by applying currents from 0–3 $\mu A/cm^2$ and measuring the ratio of AC to total power and the probability of spike generation per subthreshold cycle. The AC power ratio threshold for subthreshold oscillations was 1.4×10^{-4}, equivalent to a 1-mV sinusoid superimposed on a membrane potential of −60 mV. The threshold for reliable spiking was a probability of 0.1/cycle. Shade-coded regions, numerically identified to the right of *panel C*, represent different combinations of quiescent, subthreshold oscillatory and spiking behavior over the current range tested and are explained in the text.

Neuroscience foundations

- Shared architecture: many elements in a processing pipeline (ie, the ventral stream) are shared across cognitive tasks

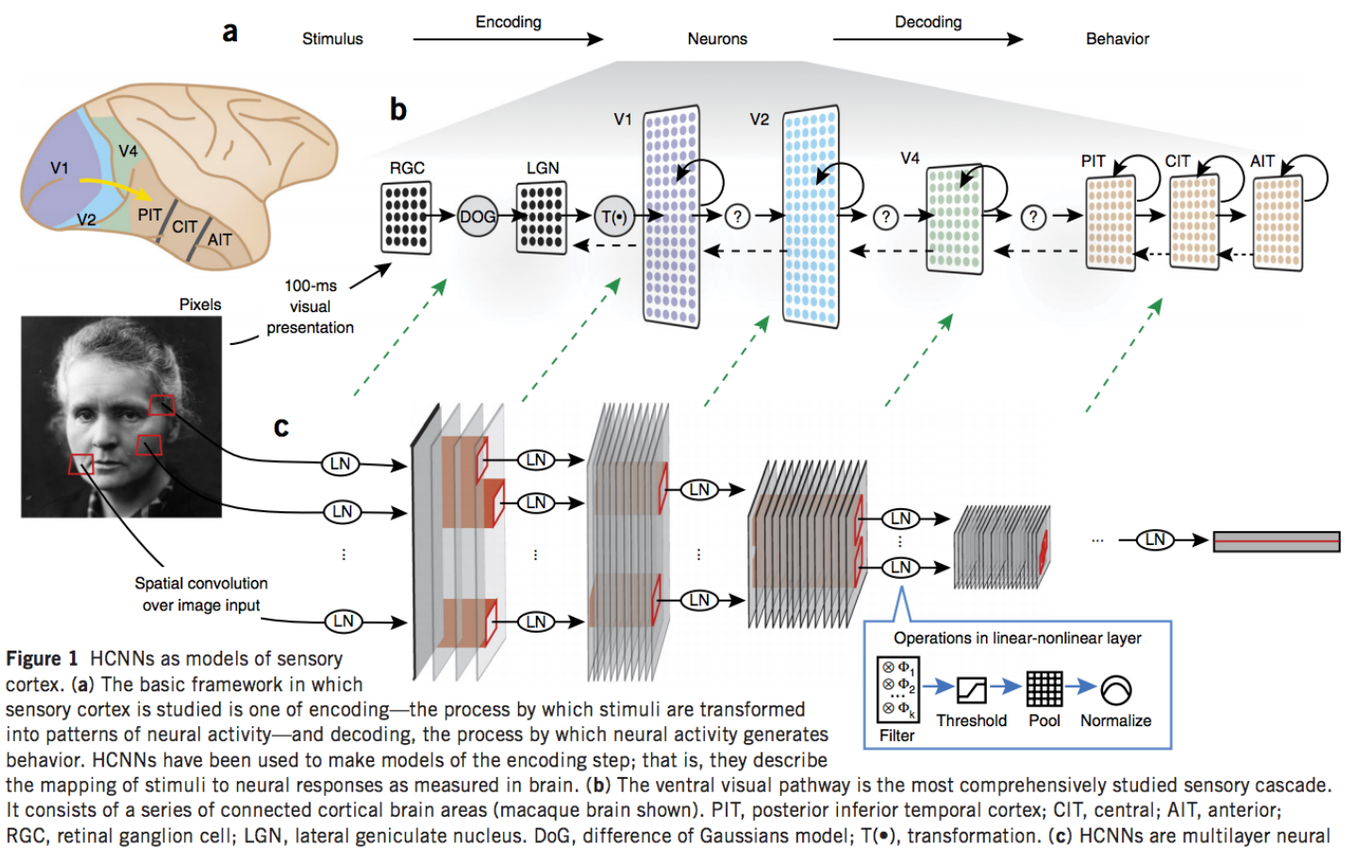


Figure 1 HCNNs as models of sensory cortex. (**a**) The basic framework in which sensory cortex is studied is one of encoding—the process by which stimuli are transformed into patterns of neural activity—and decoding, the process by which neural activity generates behavior. HCNNs have been used to make models of the encoding step; that is, they describe the mapping of stimuli to neural responses as measured in brain. (**b**) The ventral visual pathway is the most comprehensively studied sensory cascade. It consists of a series of connected cortical brain areas (macaque brain shown). PIT, posterior inferior temporal cortex; CIT, central; AIT, anterior; RGC, retinal ganglion cell; LGN, lateral geniculate nucleus. DoG, difference of Gaussians model; T(•), transformation. (**c**) HCNNs are multilayer neural

Yamins and DiCarlo, Science 2016

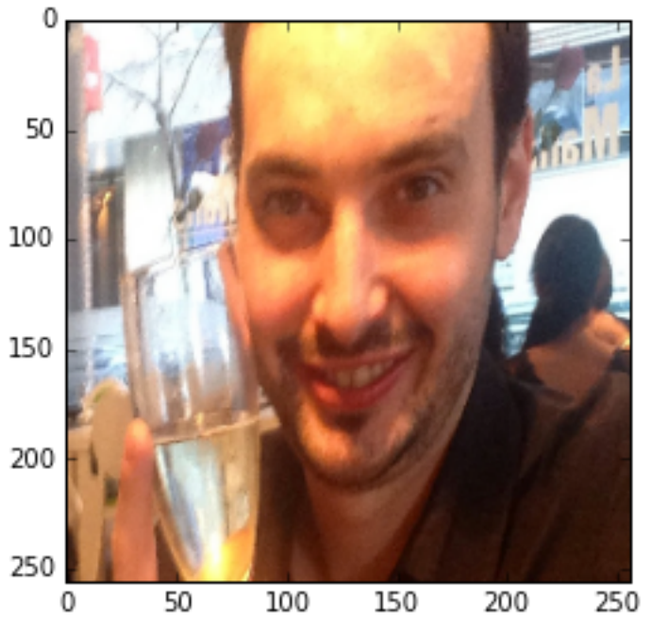
Neuroscience foundations

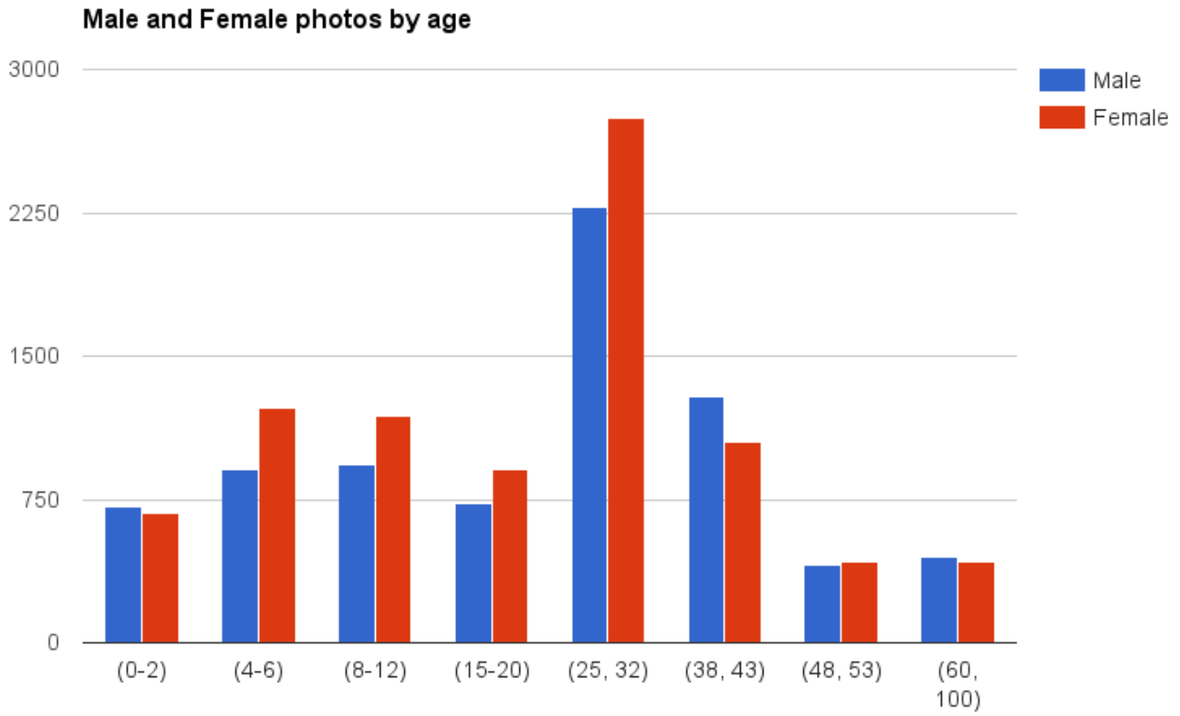
- Open question: A mature representation is a sparse representation.

Testing Diversity

- Variation of training and testing data
- Multiple tasks
- Variance in internal state

Experimental Design





Adience Benchmark

7703 male faces

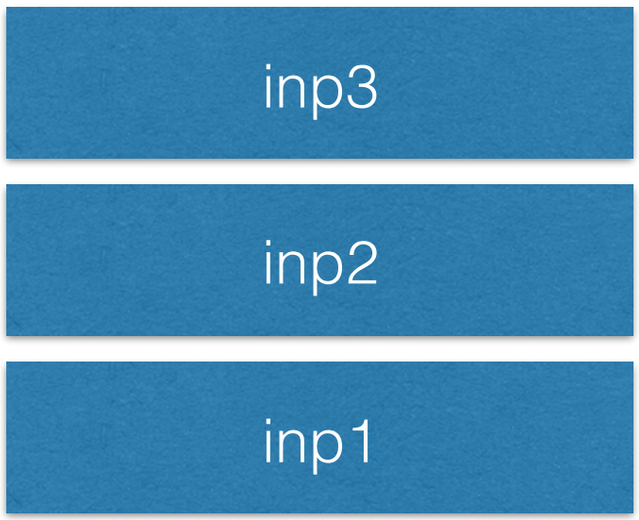
8648 female faces

Annotated by age, gender, head tilt, etc.

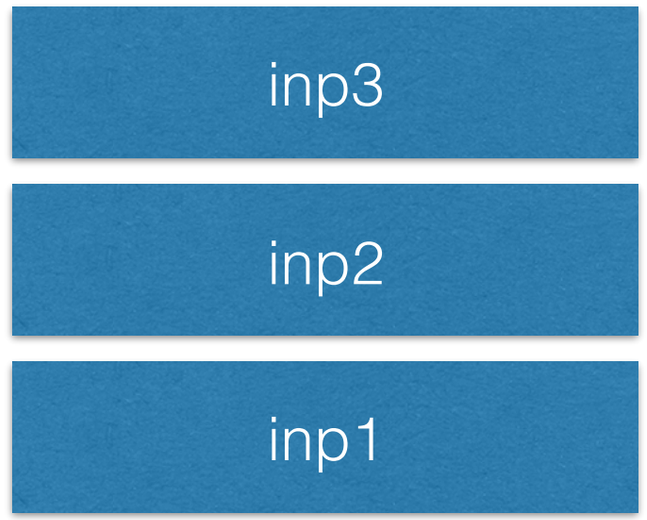
database link: http://www.openu.ac.il/home/hassner/Adience/data.html#agegender

Experimental Design

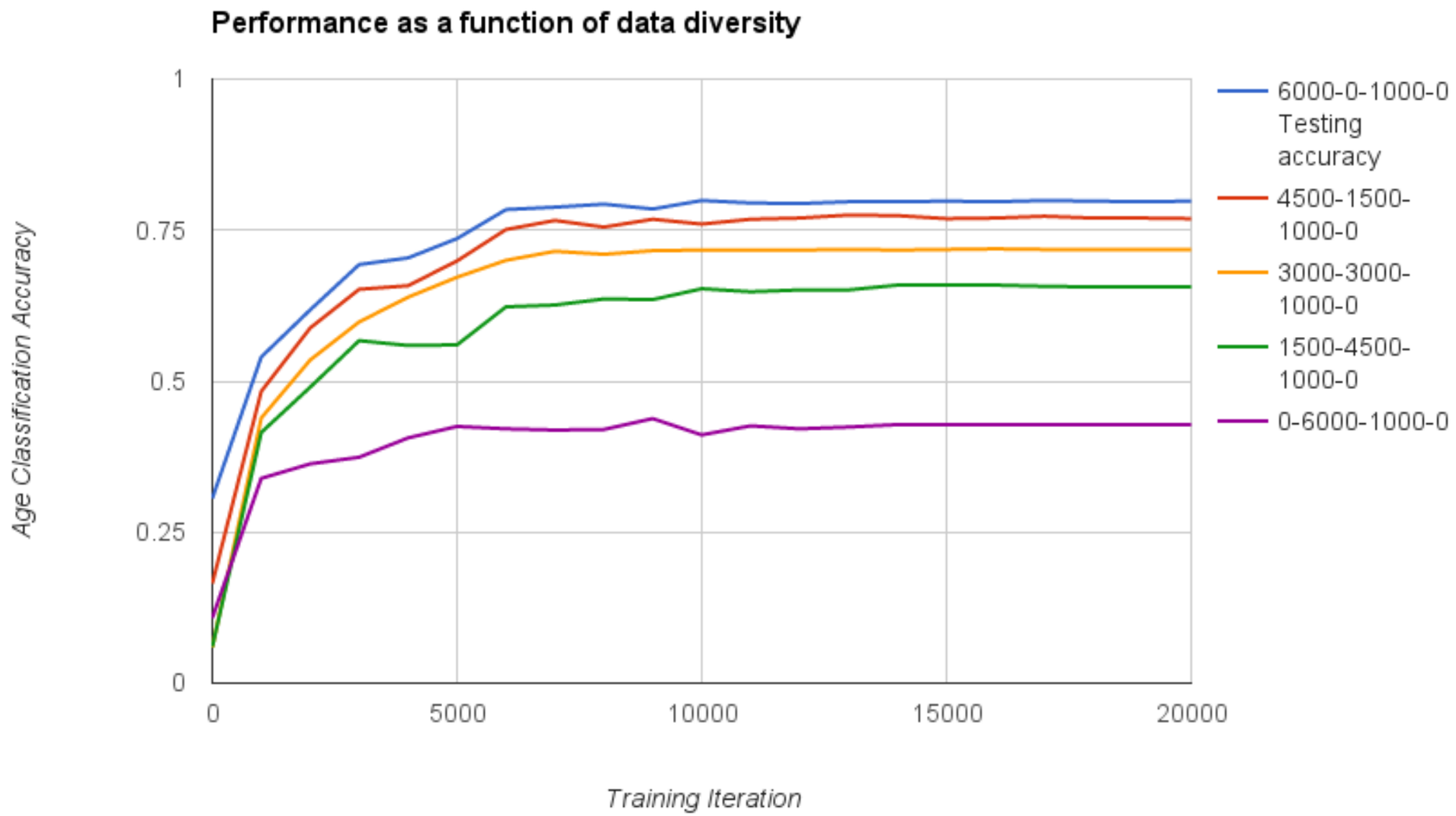
Age Classification



Gender Classification



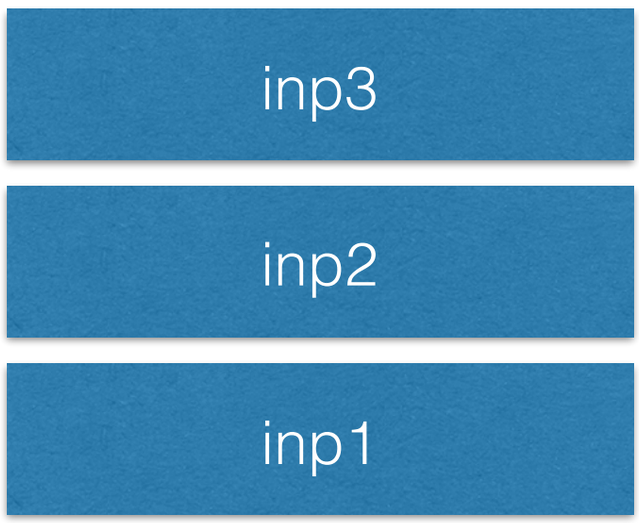
Question 1: Diversity of Data



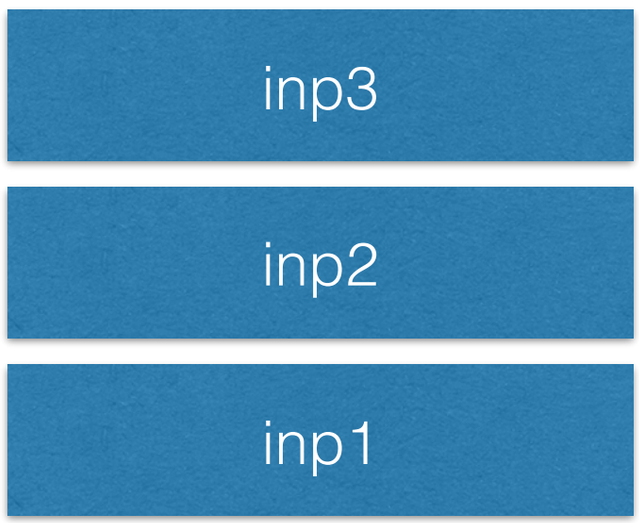
Using 6000 training images to classify age of 1000 male faces, how does the male/female split of the training data impact performance?

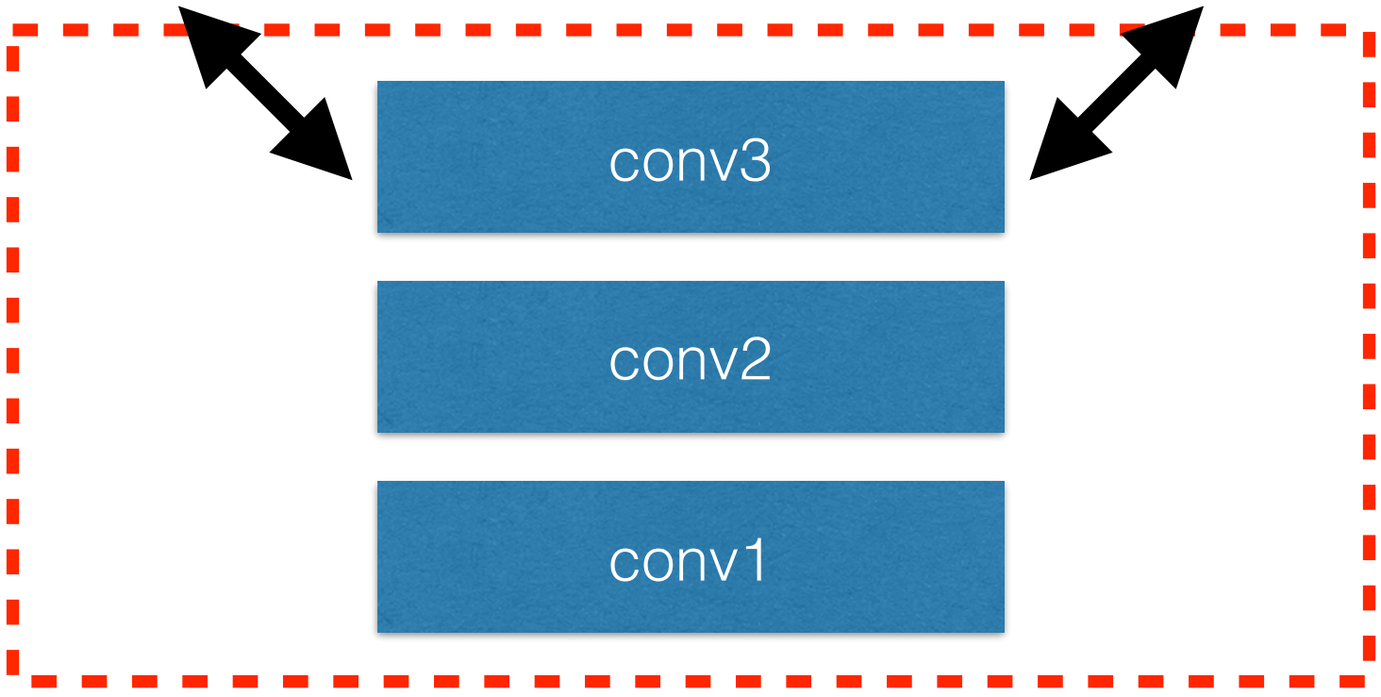
Question 2: Can learning two tasks improve each other?

Age Classification



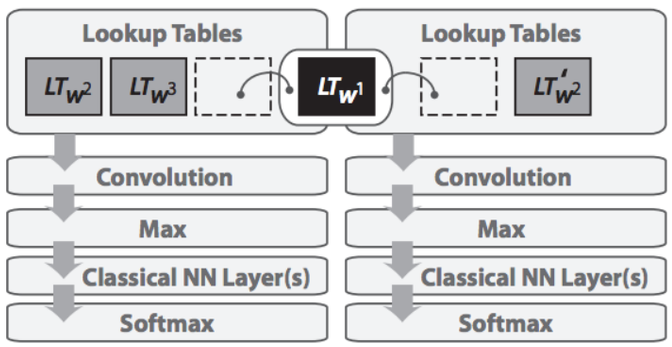
Gender Classification



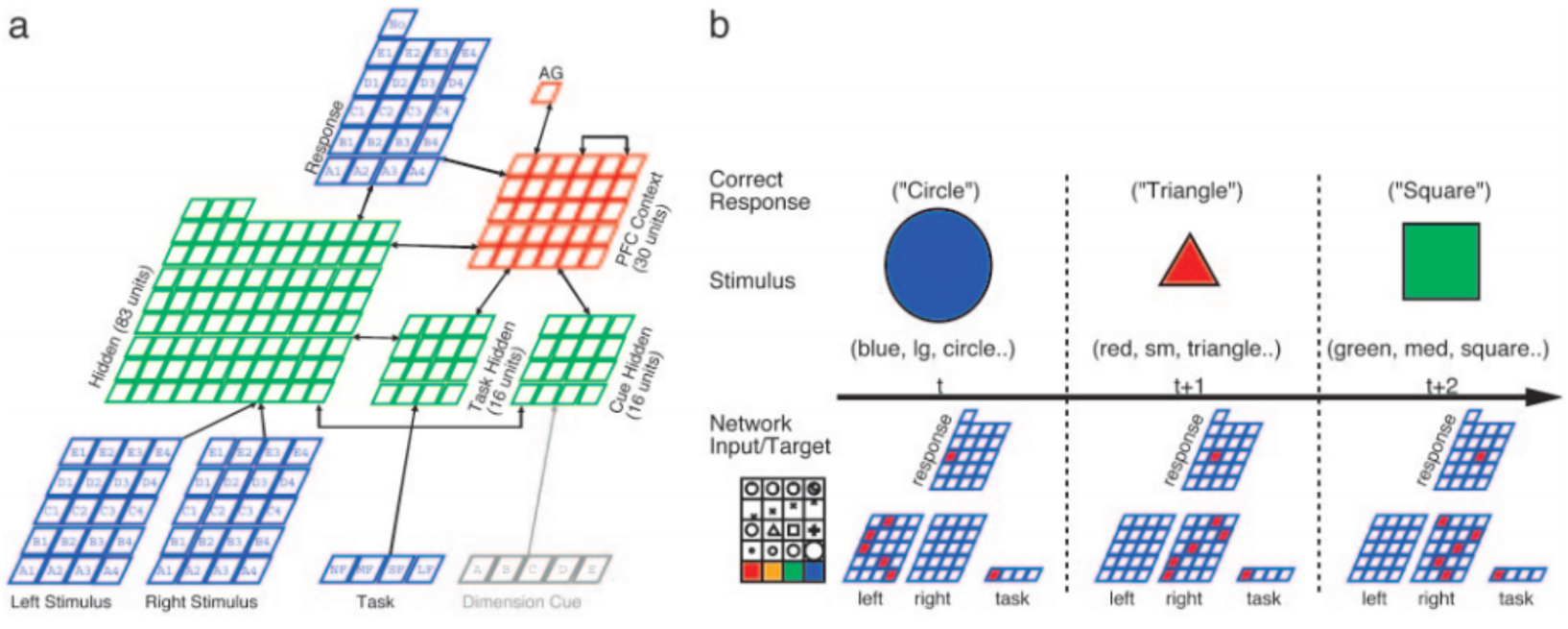


Question 2: Can learning two tasks improve each other?

- Multitask Learning (Seltzer 2013, Collobert 2008)



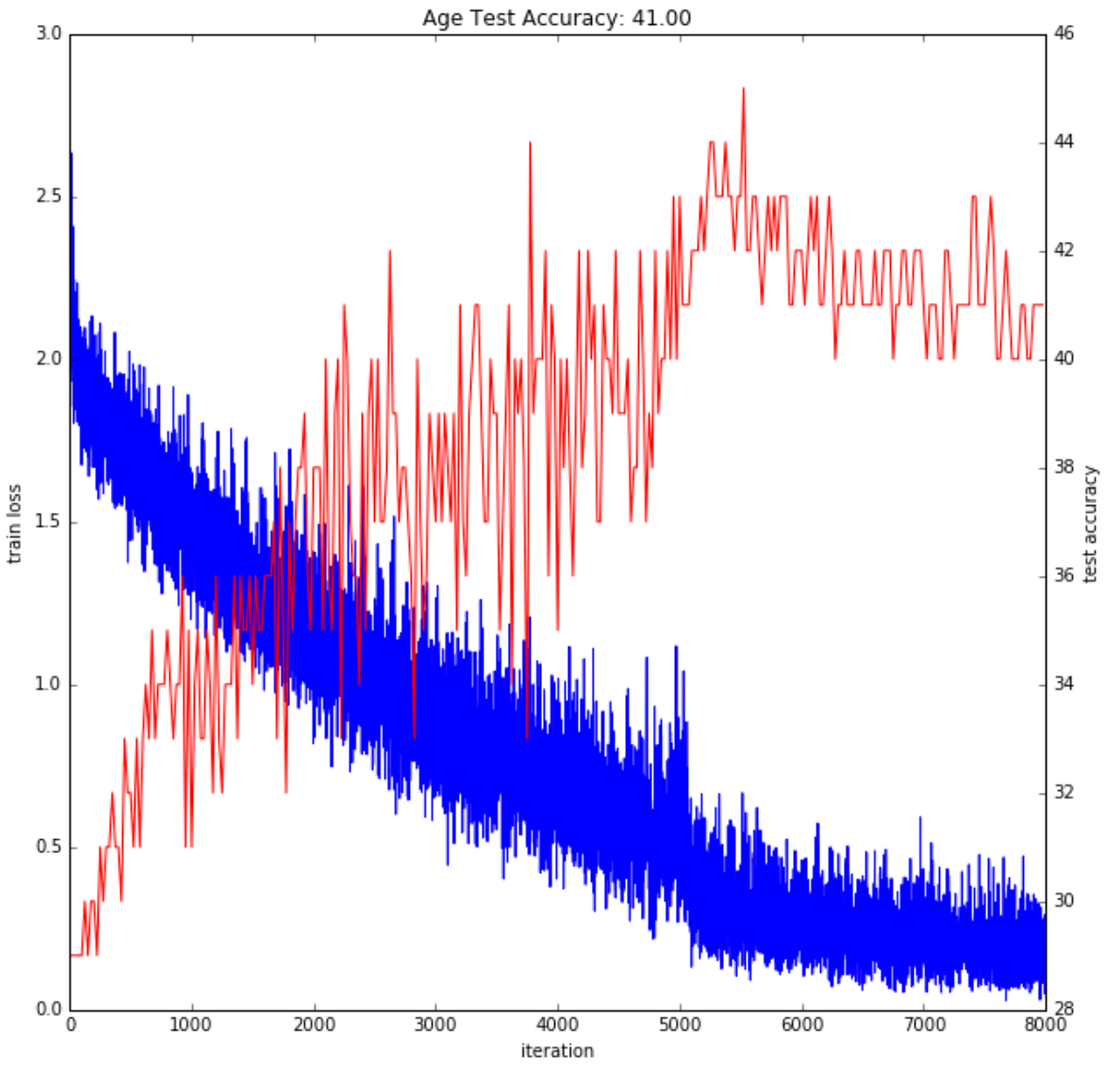
- LEABRA



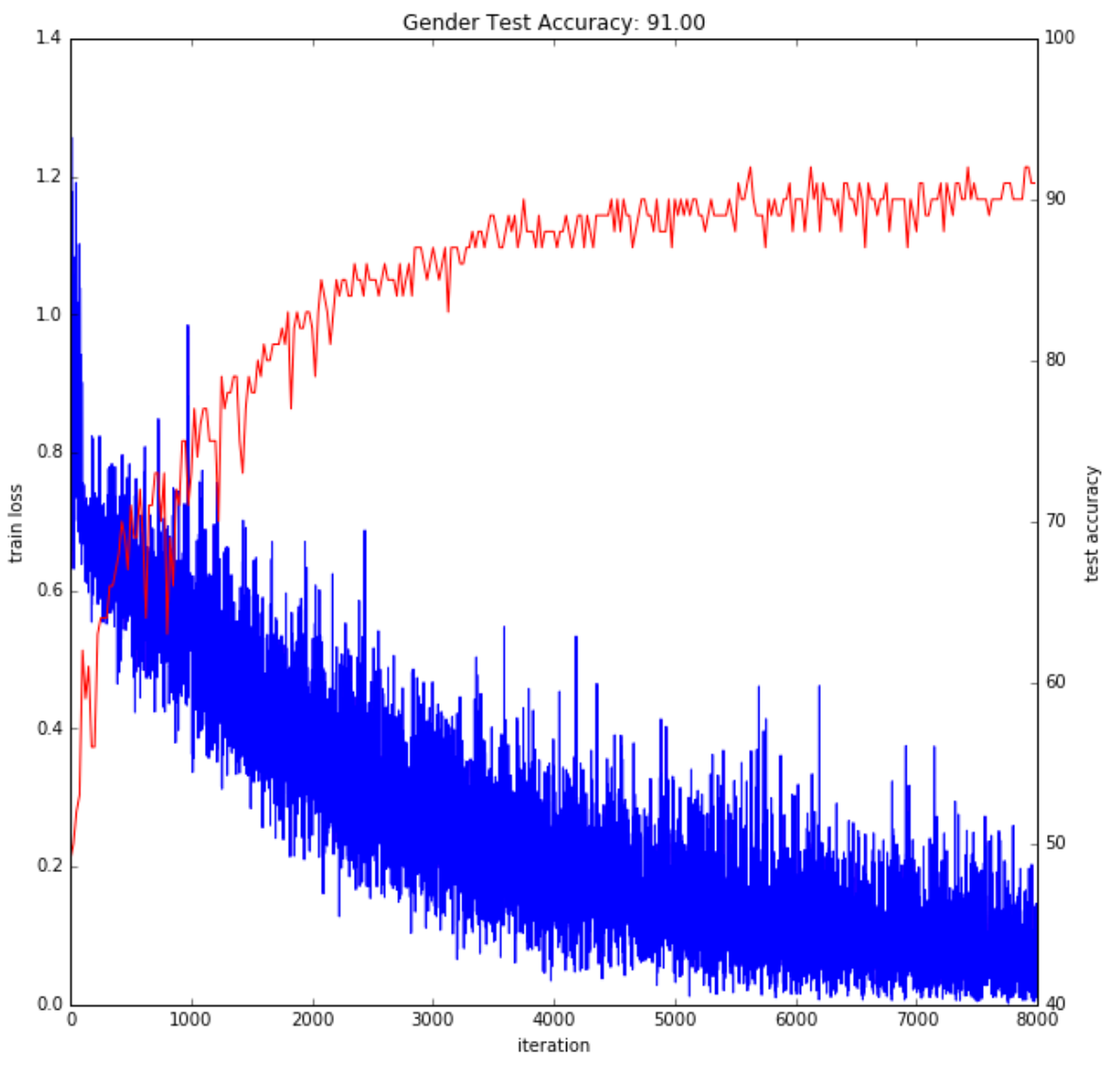
Question 2: Can learning two tasks improve each other?

(benchmark: training the two independently)

6000 females train
1000 males test

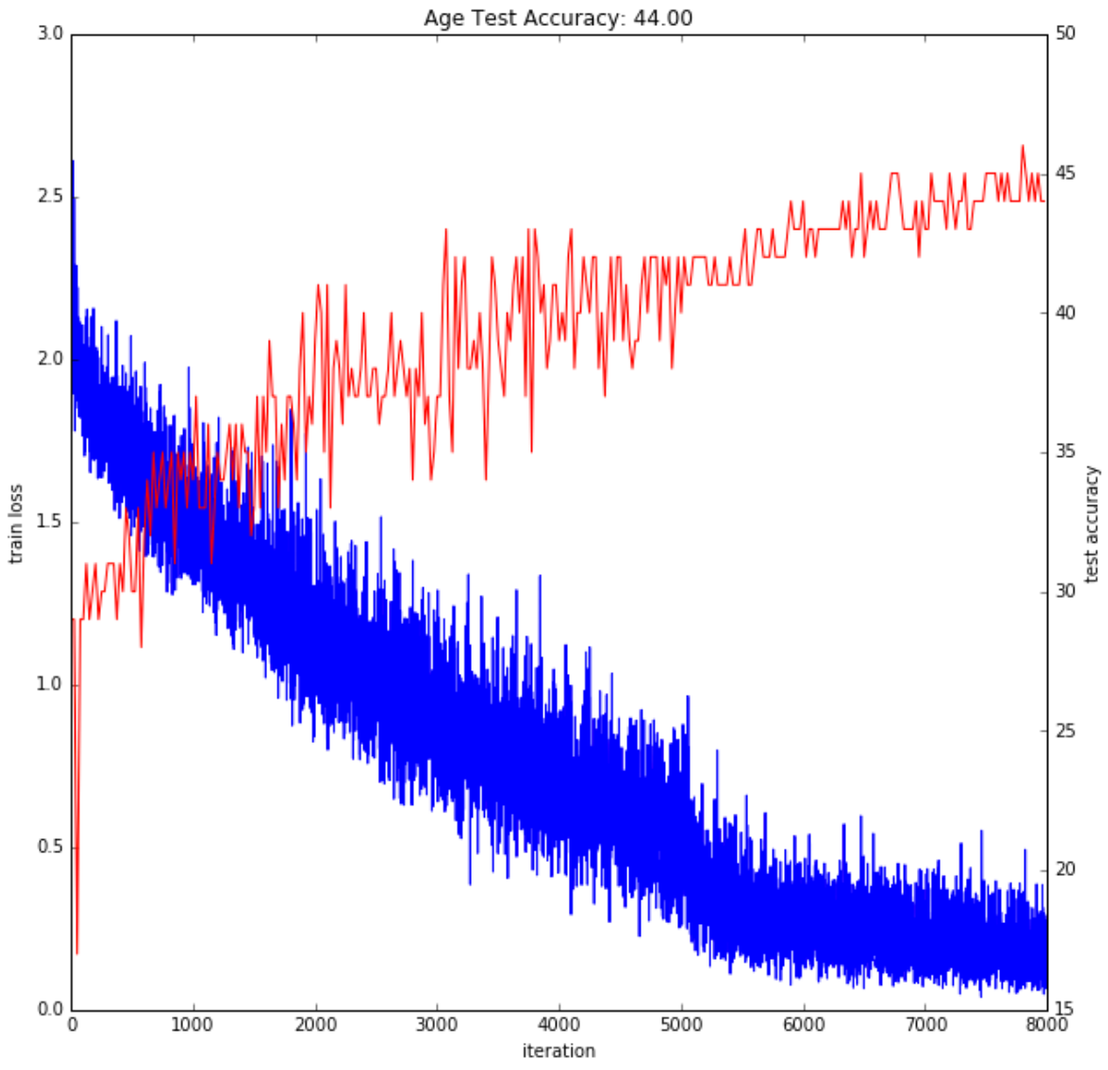


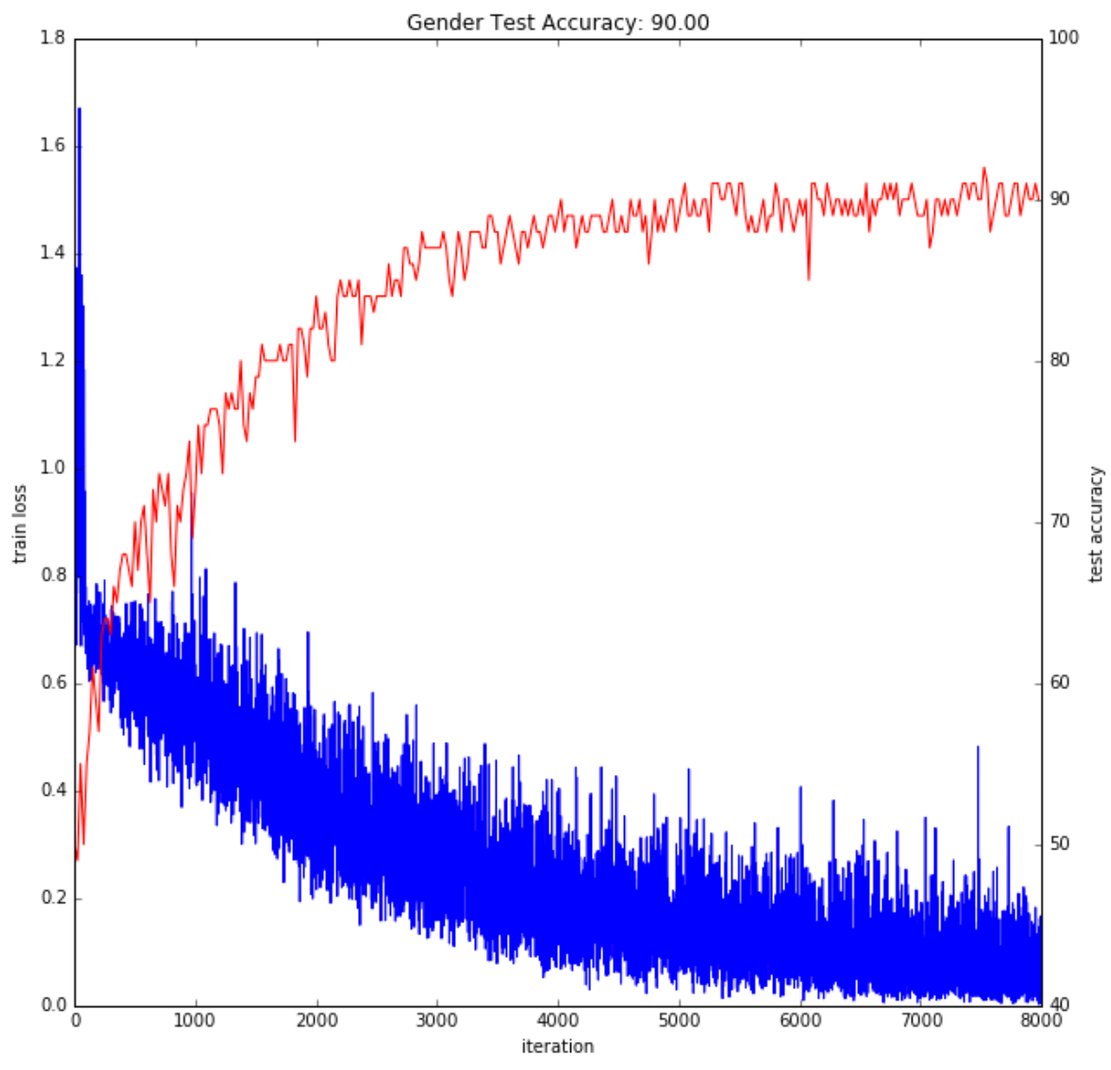
3000/3000 train
750/750 test



Question 2: Can learning two tasks improve each other?

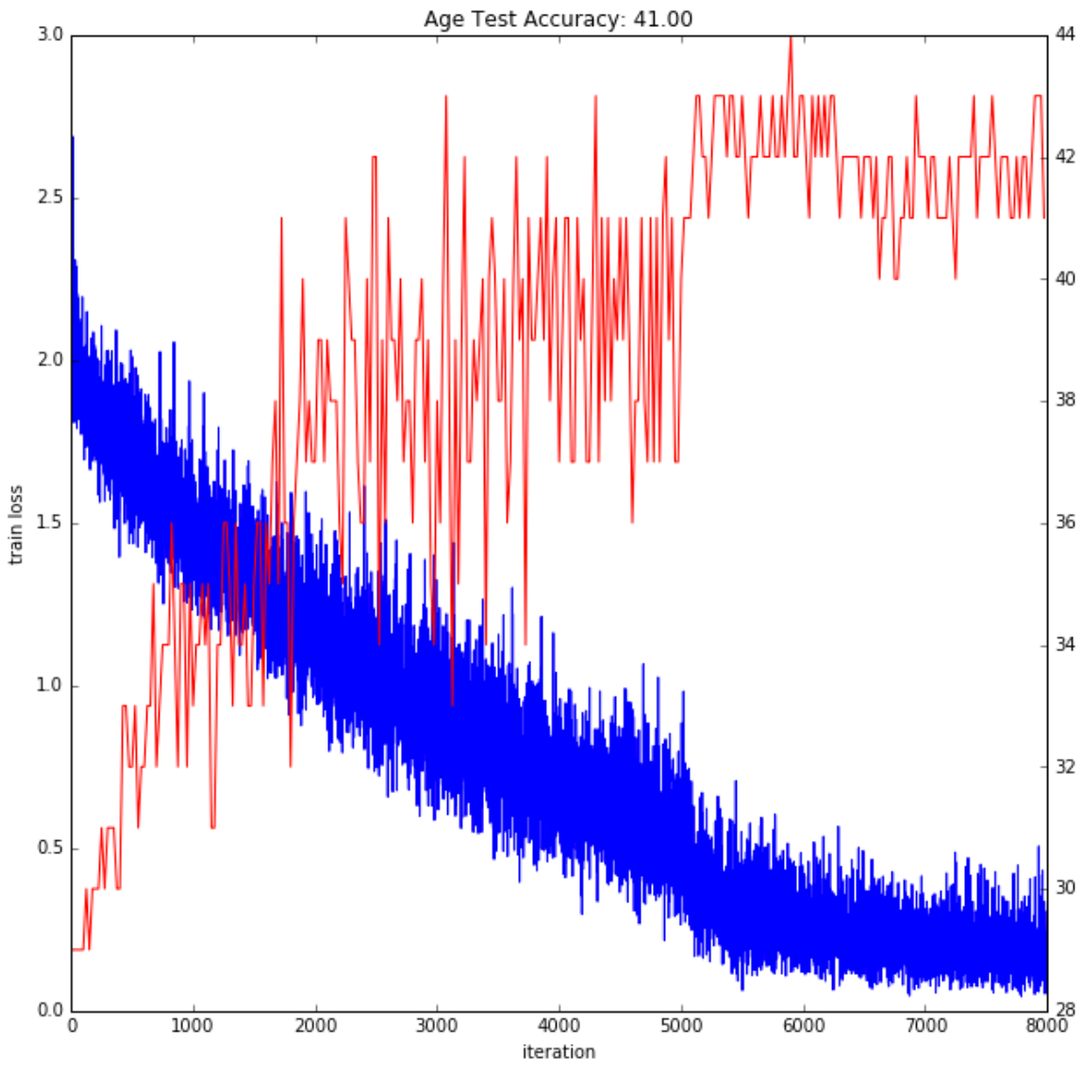
(Sharing two layers)

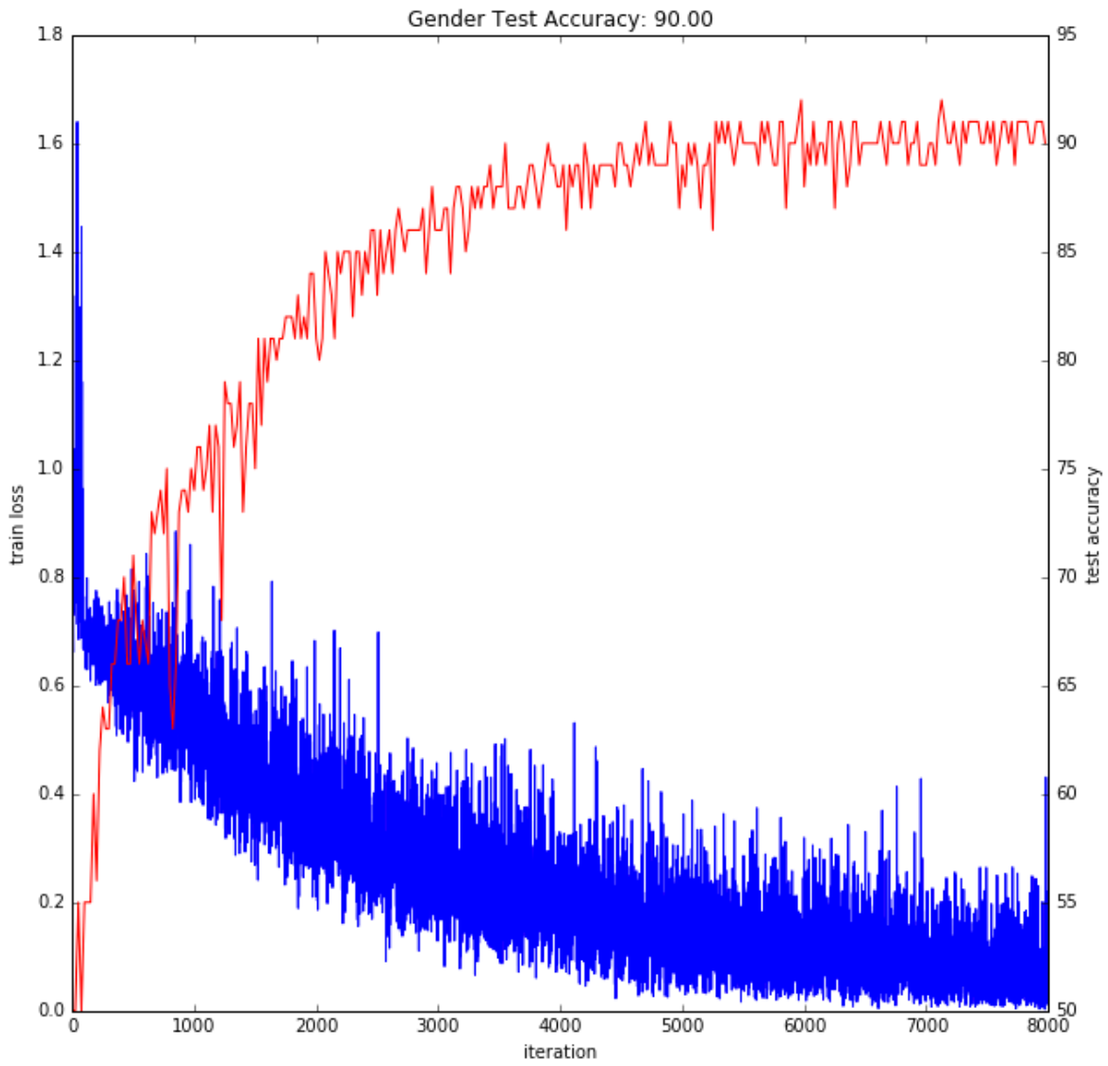




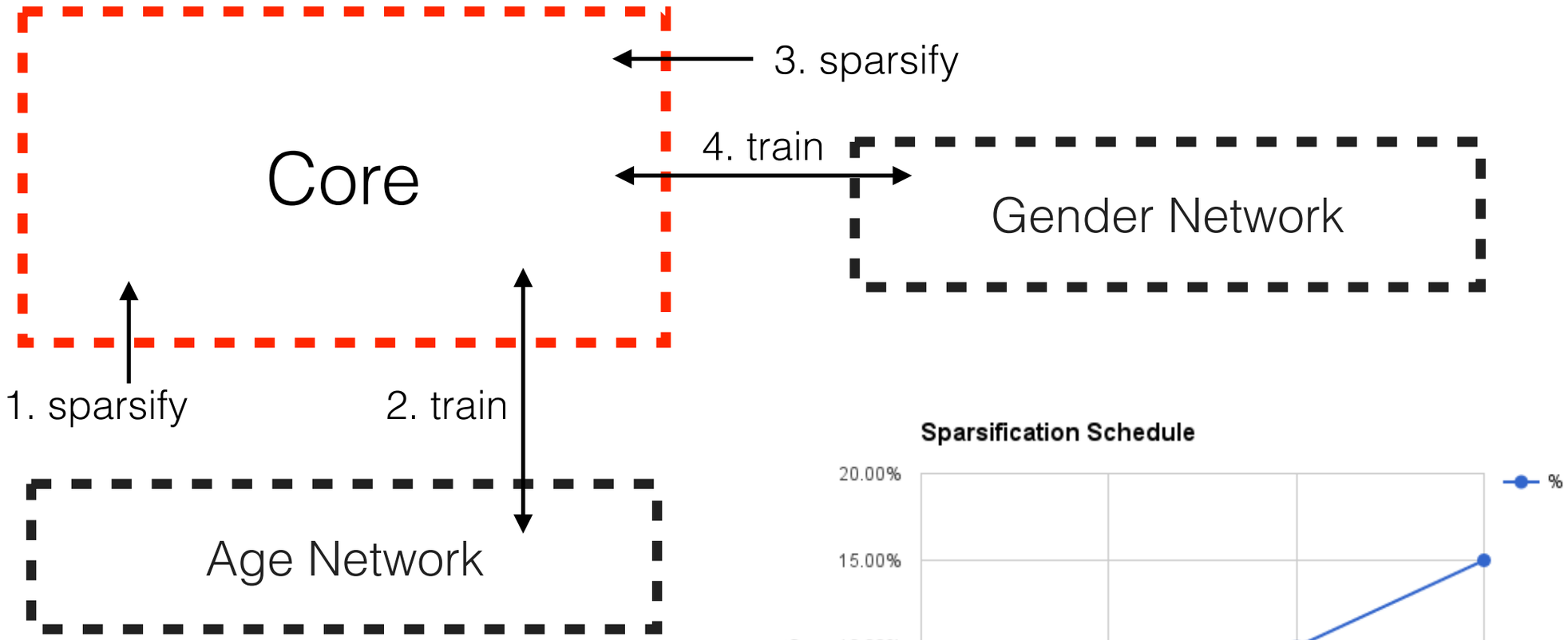
Question 2: Can learning two tasks improve each other?

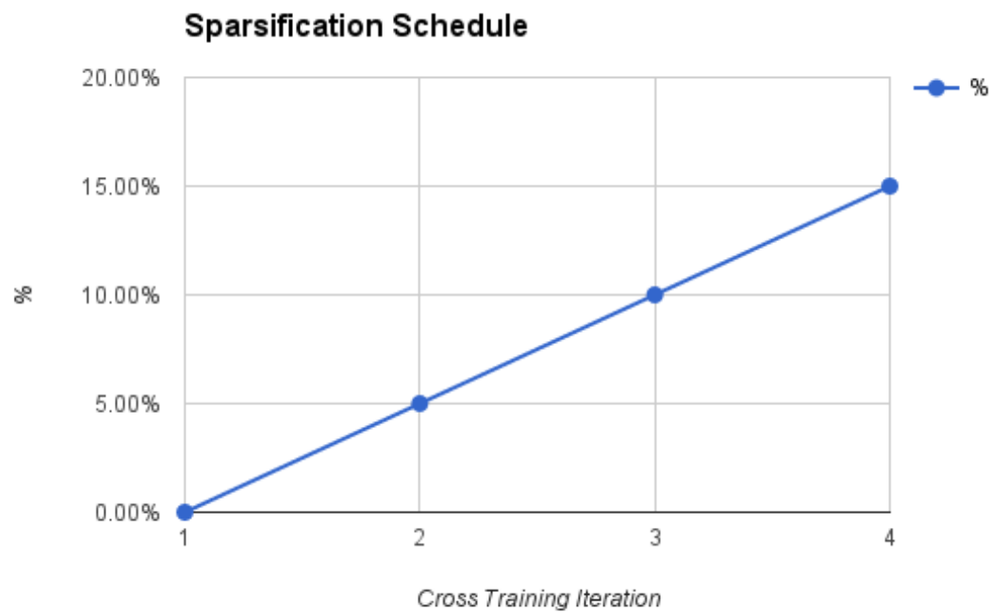
(Sharing three layers)





Question 3: Can progressive sparsification improve performance?



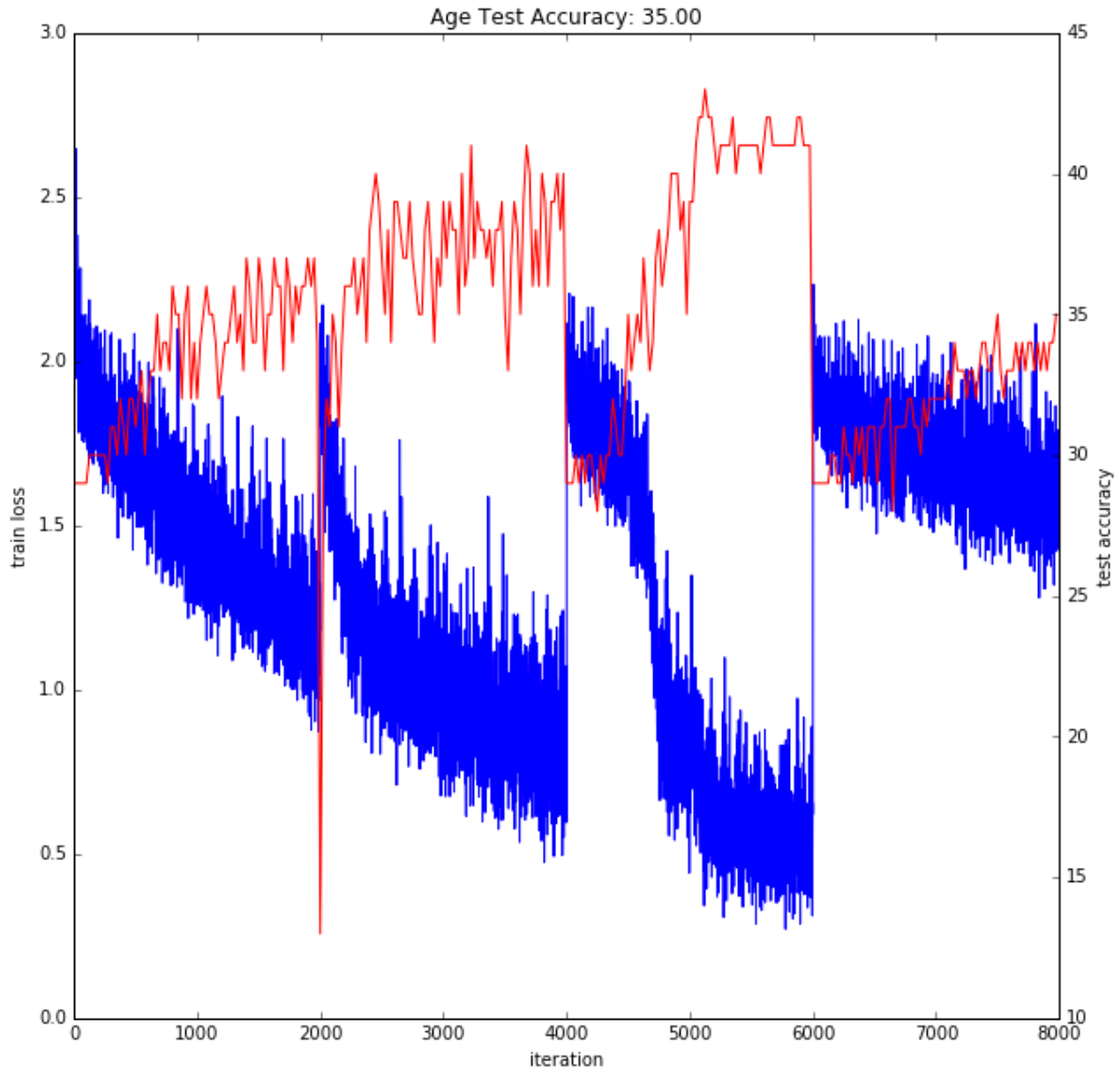


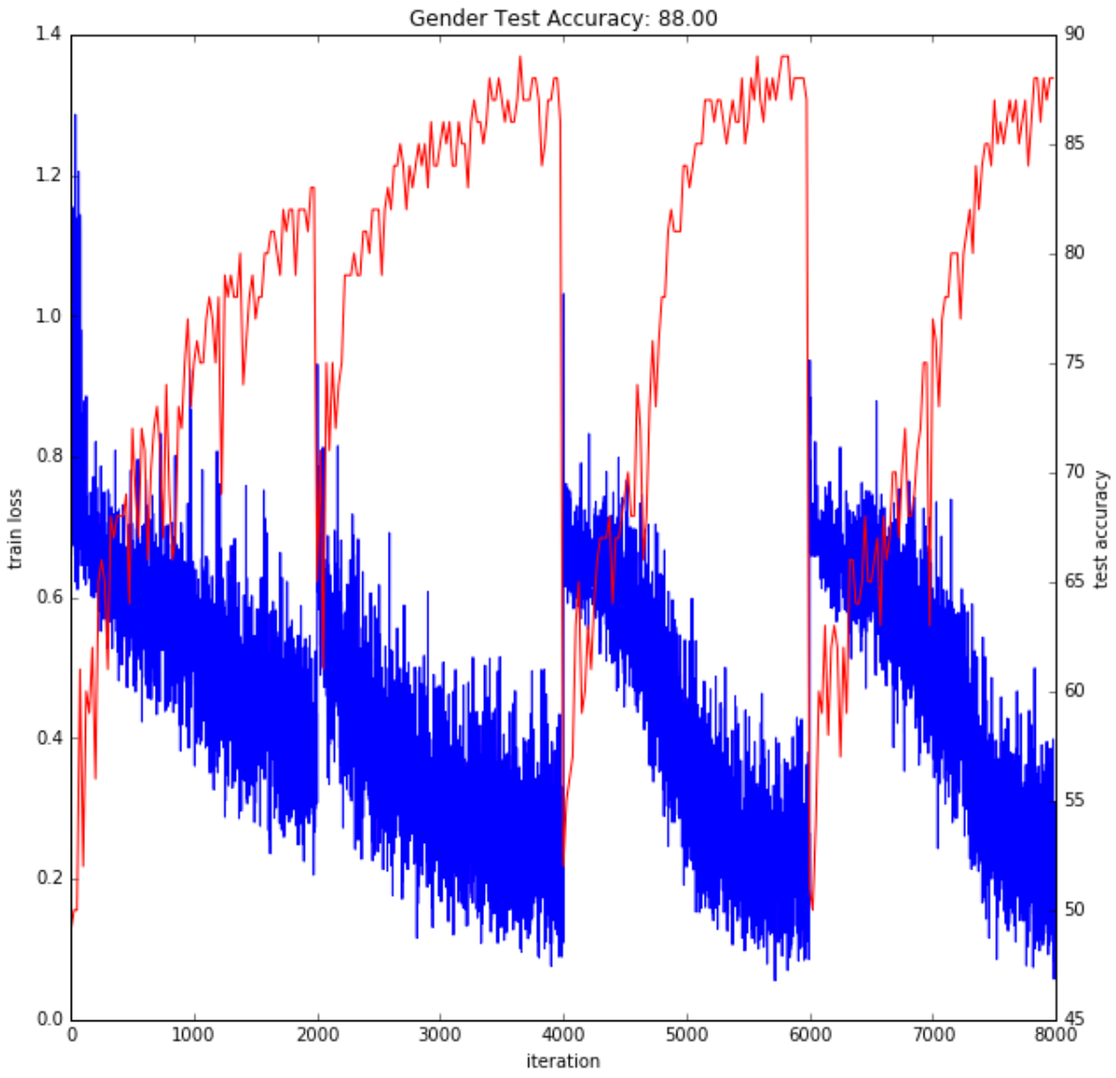
Question 3: Can progressive sparsification improve performance?

- Dropout (Srivastava 2014)
- Maxout (Goodfellow 2013)
- Low-rank expansions of filters (Jaderberg et al 2014)
- Optimal Brain Damage (Lecun et al, 1990)

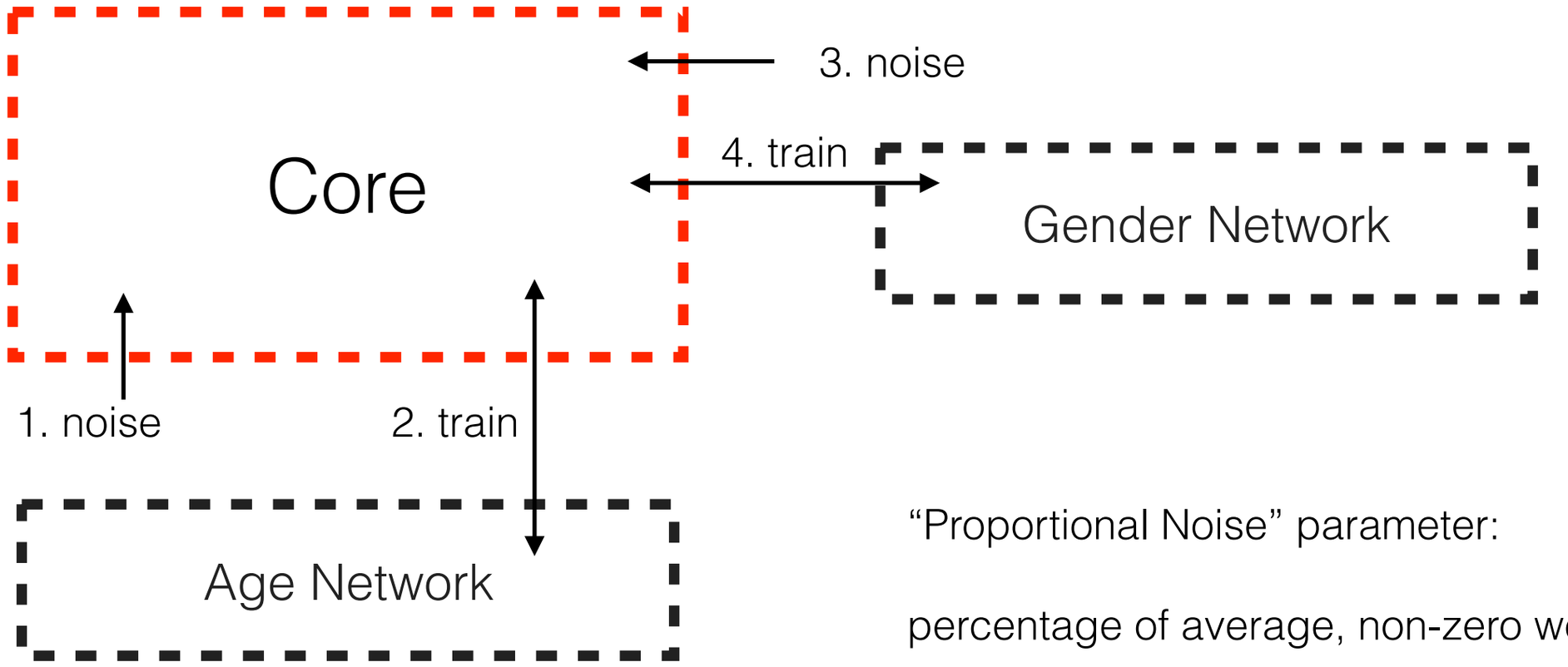
Question 3: Can progressive sparsification improve performance?

(Sharing first two layers)





Question 4: Does noise improve neural networks?

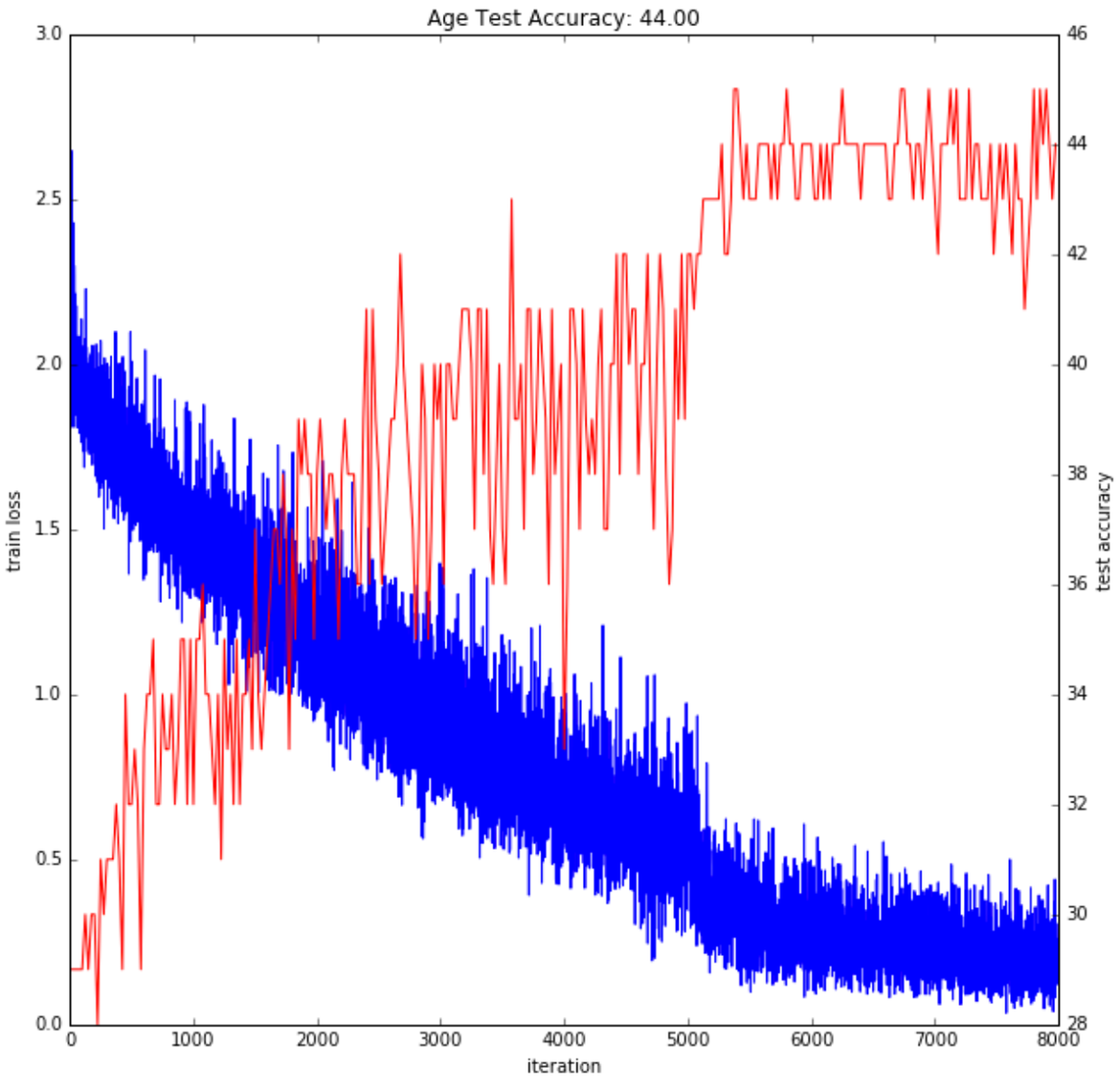


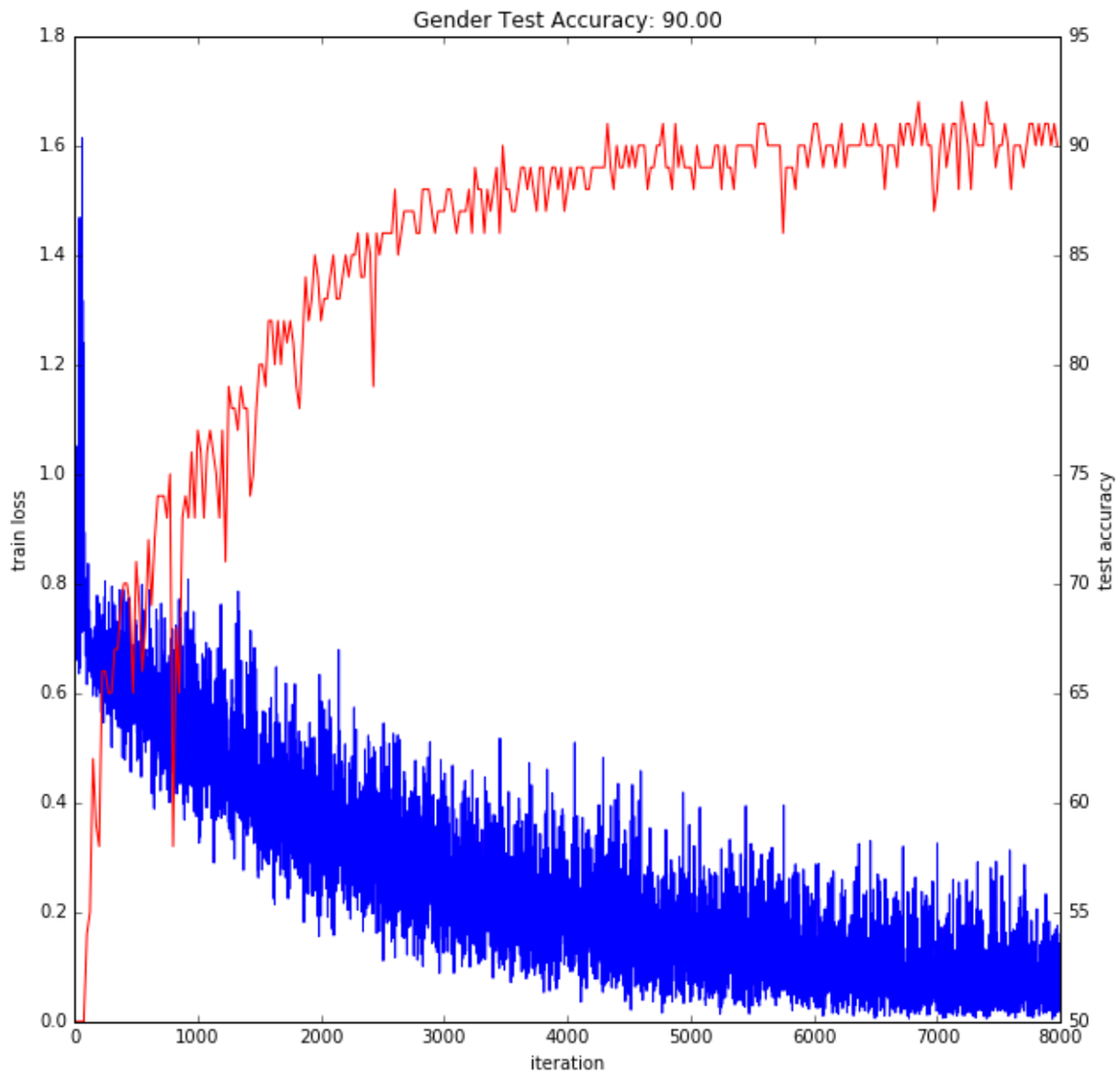
“Proportional Noise” parameter:

percentage of average, non-zero weight per filter

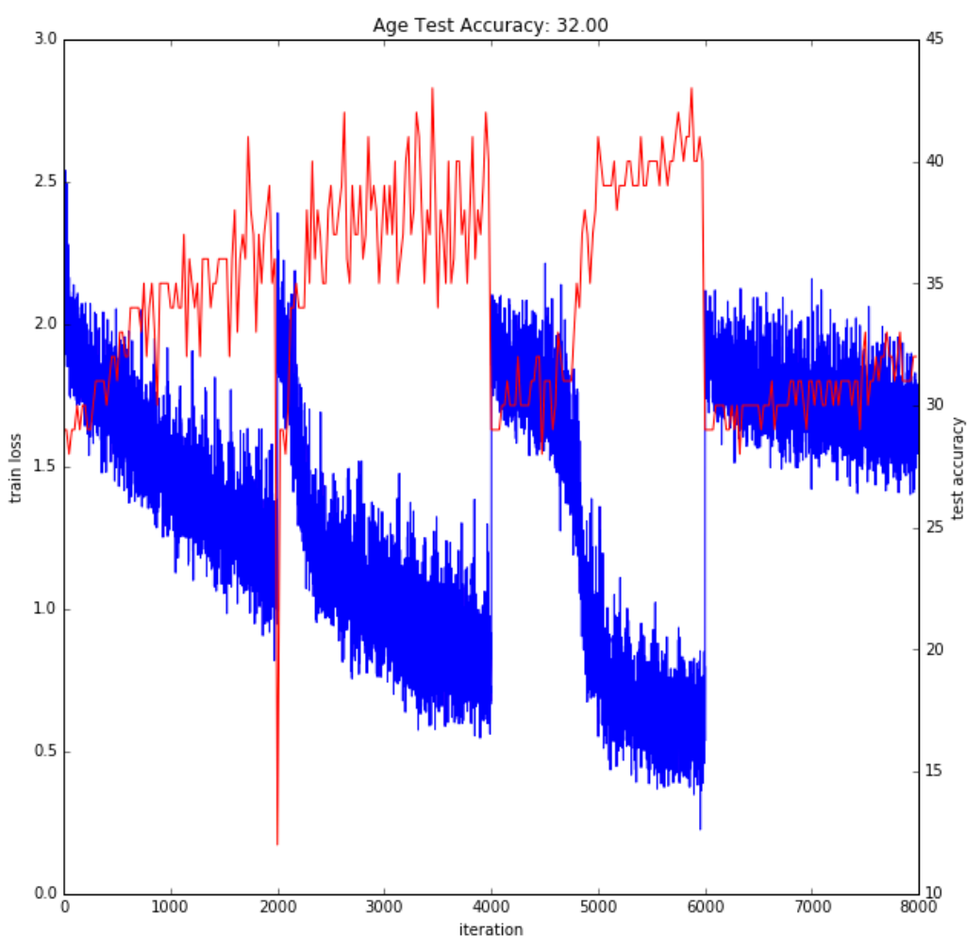
Question 4: Does noise improve shared networks?

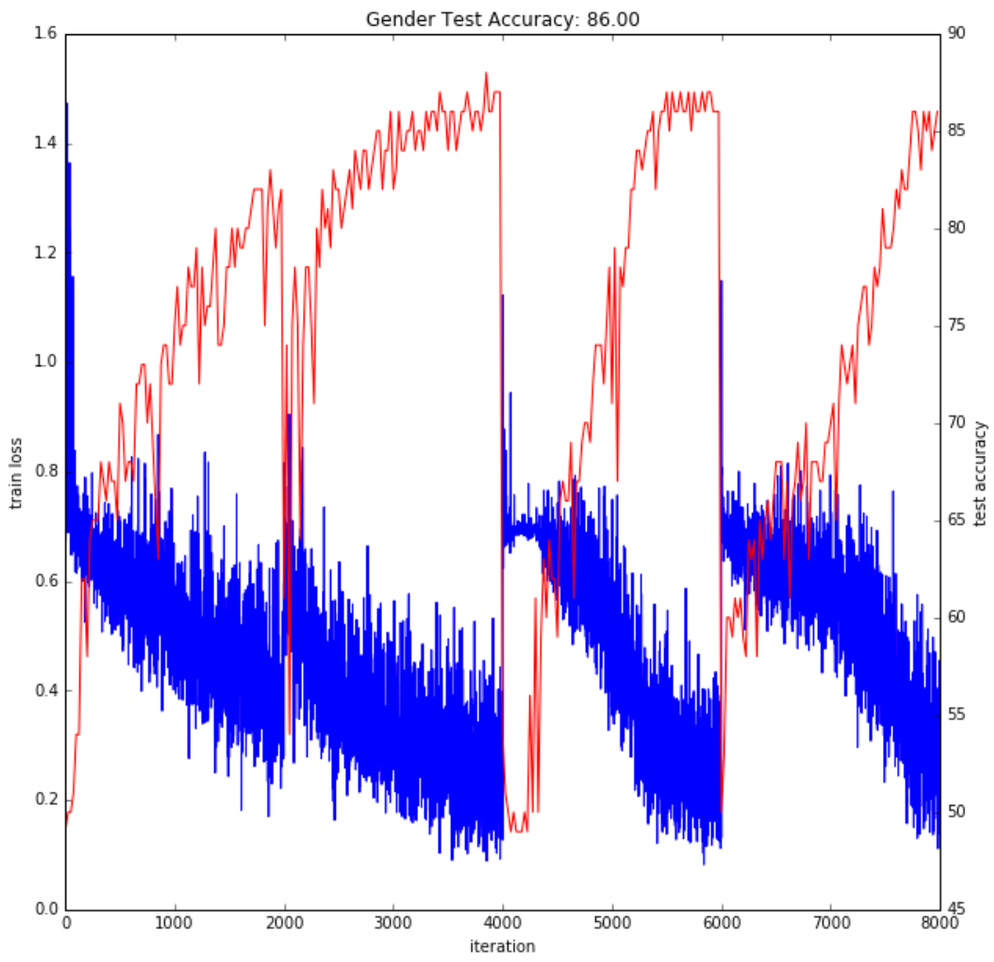
(Sharing first two layers, noise in first three)





Question 5: What happens when we put it all together?





[0,.05,.1,.15 sparsity] to conv1 conv2.
Adding noise (.2) to conv1, conv2, conv3.

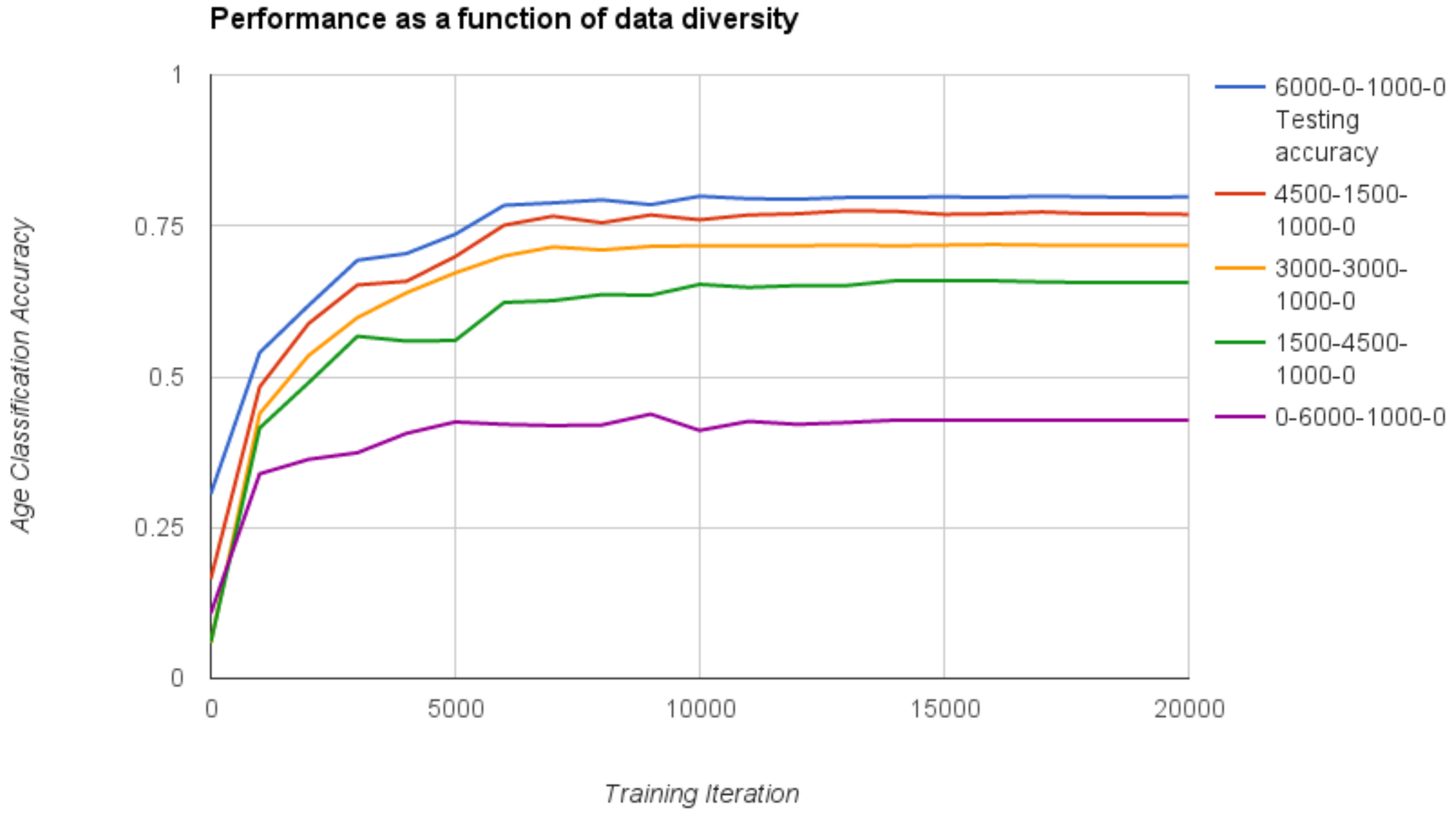
Summary

- Sparsity plus a small amount of noise is better than just sparsity
- The speed of the network performance recovery after sparsification events is notable.
- Sharing content between layers in a convNet feels exciting but requires deeper (…) thinking. Designing a dynamic architecture to do this more elegantly would be interesting.
- Sparsity might have made more sense in the fully connected layers, or by removing the filters which added least unique information
- Future idea: Learning Using Privelidged Information + Neural Nets

Raw Data

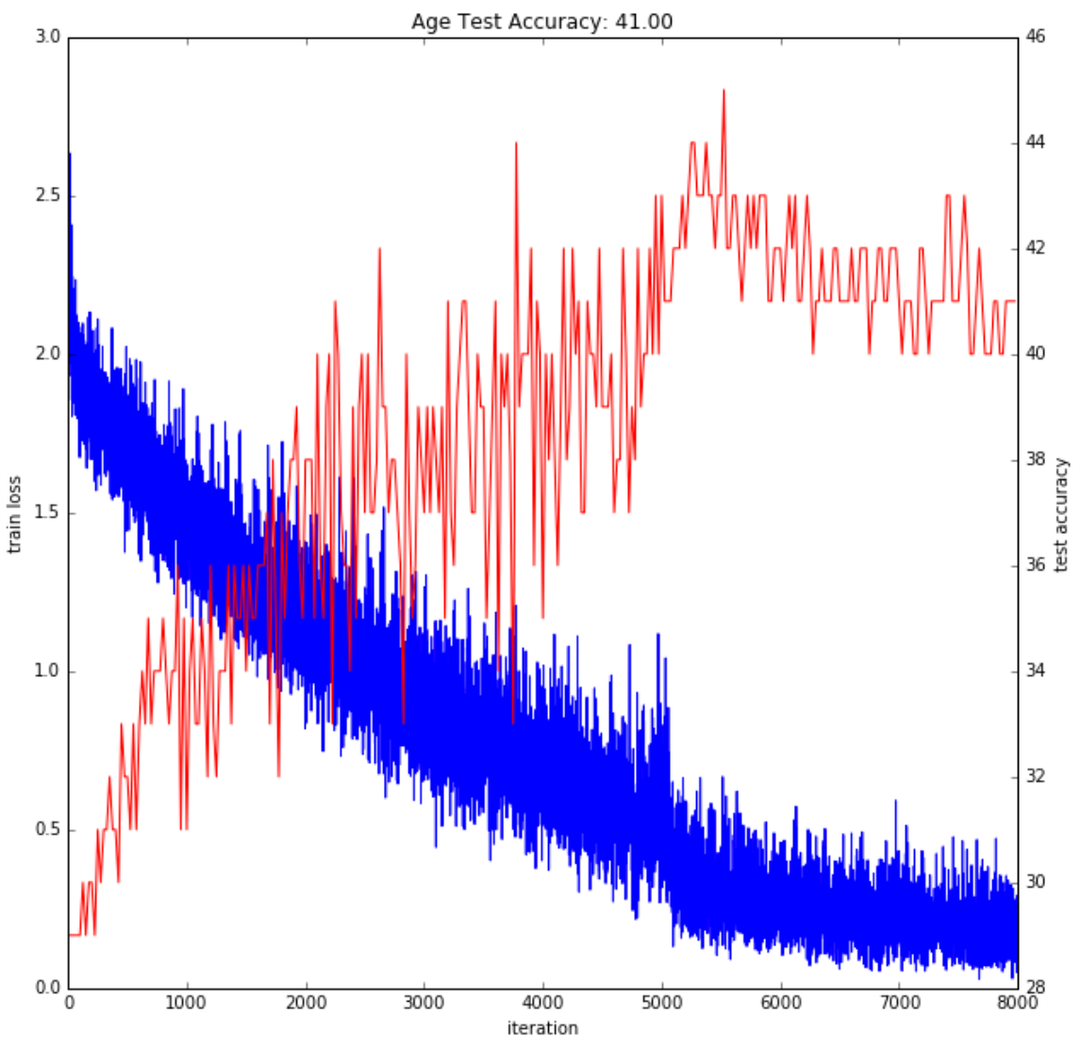
```
GLOG_logtostderr=1 ../caffe/build/tools/convert_imageset --
resize_height=256 --resize_width=256 --shuffle ./ models_gender/trained/
gender_3000-3000-750-750_train_meta.txt gender_train_lmdb
Making the training db with 6000 examples via the convert_imageset caffe
tool
Done!
GLOG_logtostderr=1 ../caffe/build/tools/convert_imageset --
resize_height=256 --resize_width=256 --shuffle ./ models_gender/trained/
gender_3000-3000-750-750_test_meta.txt gender_test_lmdb
Making the testing db with 1500 examples
Done!
```

```
GLOG_logtostderr=1 ../caffe/build/tools/convert_imageset --resize_height=256 --
resize_width=256 --shuffle ./ models/trained/age_0-6000-1000-0_train_meta.txt
age_train_lmdb
Making the training db with 6000 examples via the convert_imageset caffe tool
Done!
GLOG_logtostderr=1 ../caffe/build/tools/convert_imageset --resize_height=256 --
resize_width=256 --shuffle ./ models/trained/age_0-6000-1000-0_test_meta.txt
age_test_lmdb
Making the testing db with 1000 examples
Done!
```

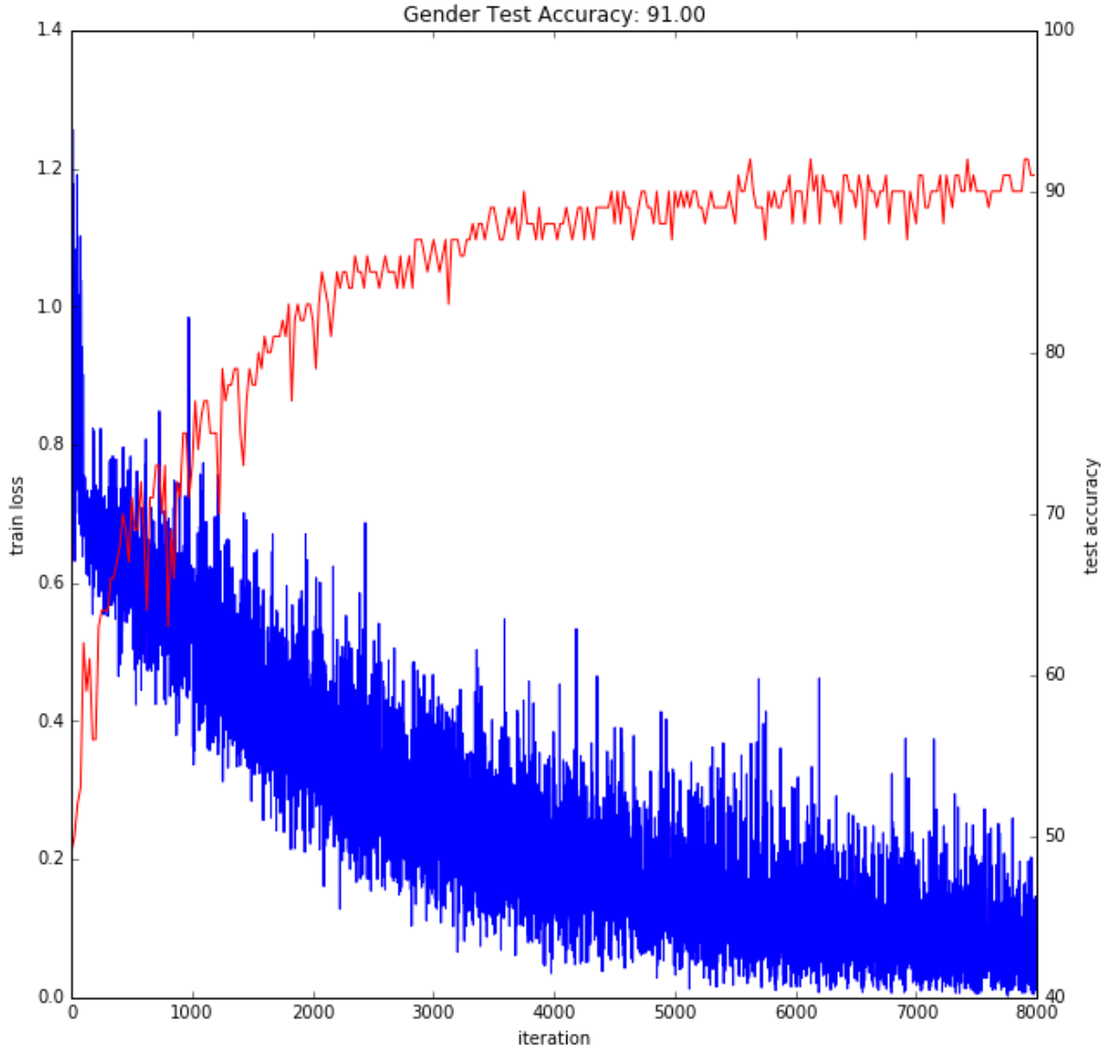


Not entirely relevant because this was tested on men only

6000 females train
1000 males test

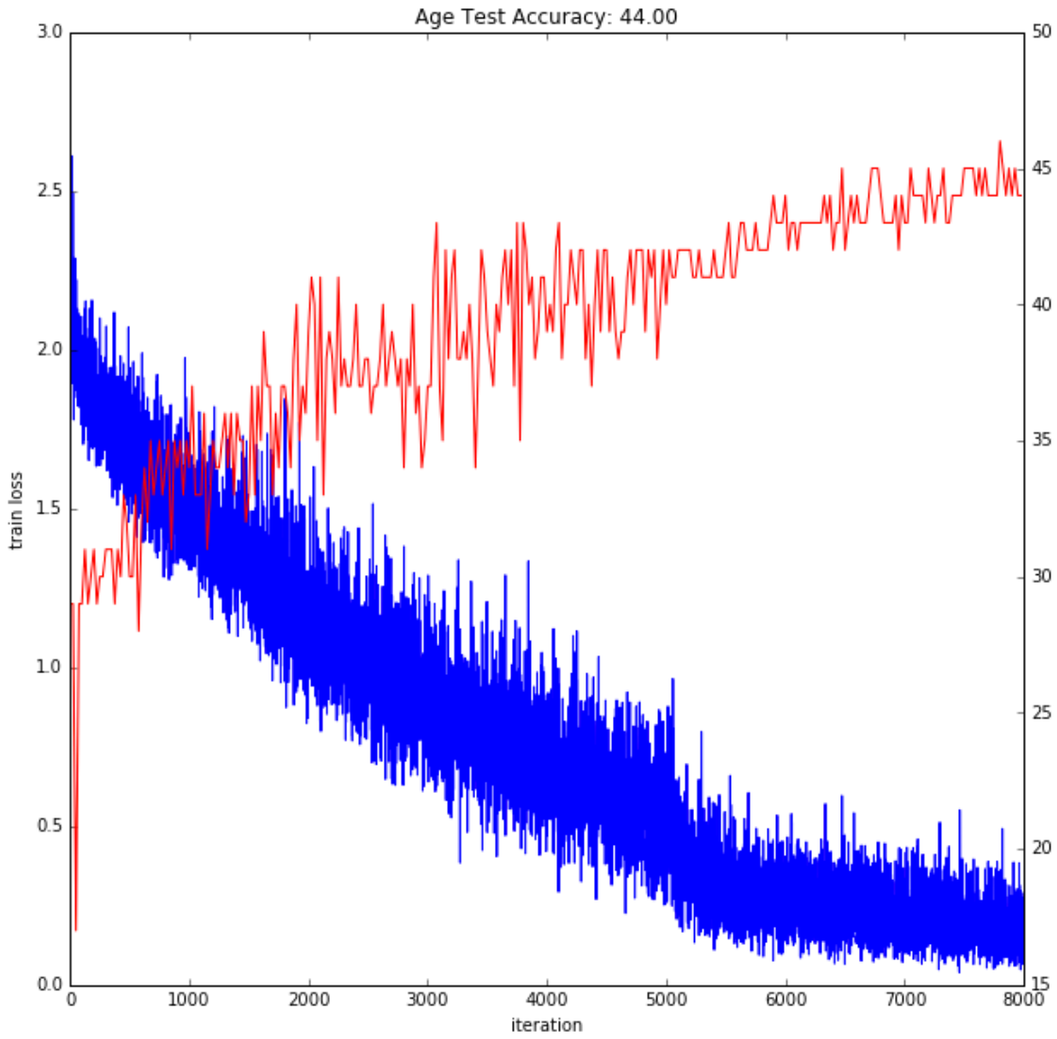


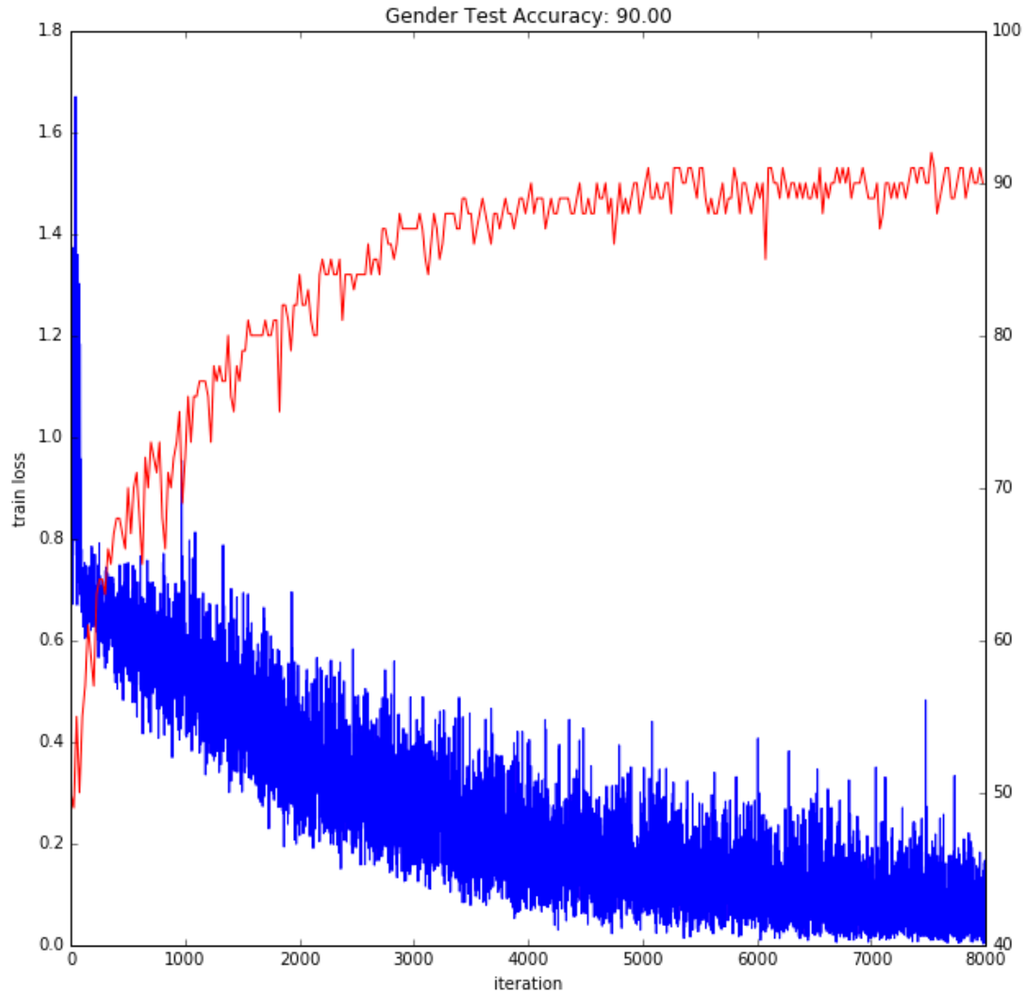
3000/3000 train
750/750 test



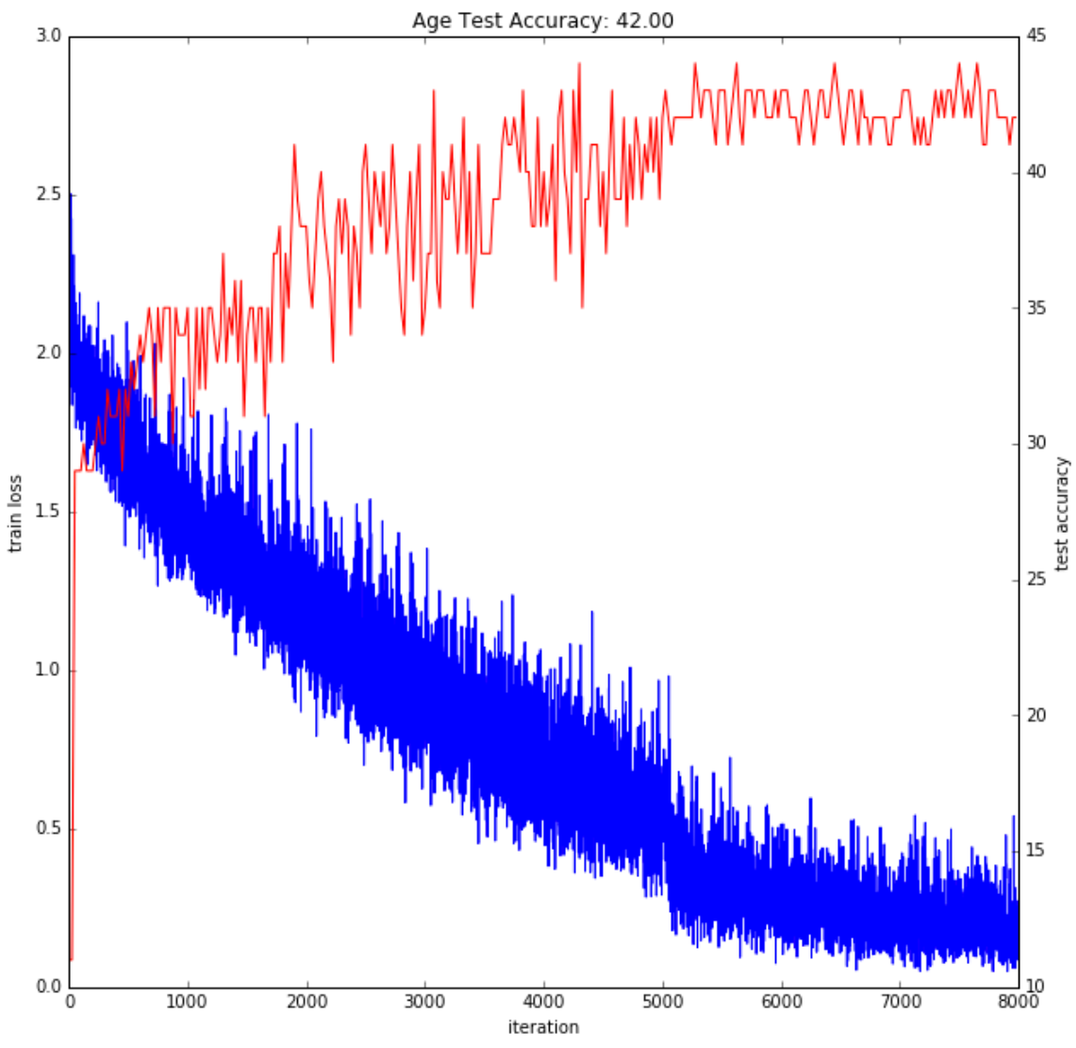
Experiment 1: cores independent (sanity check)

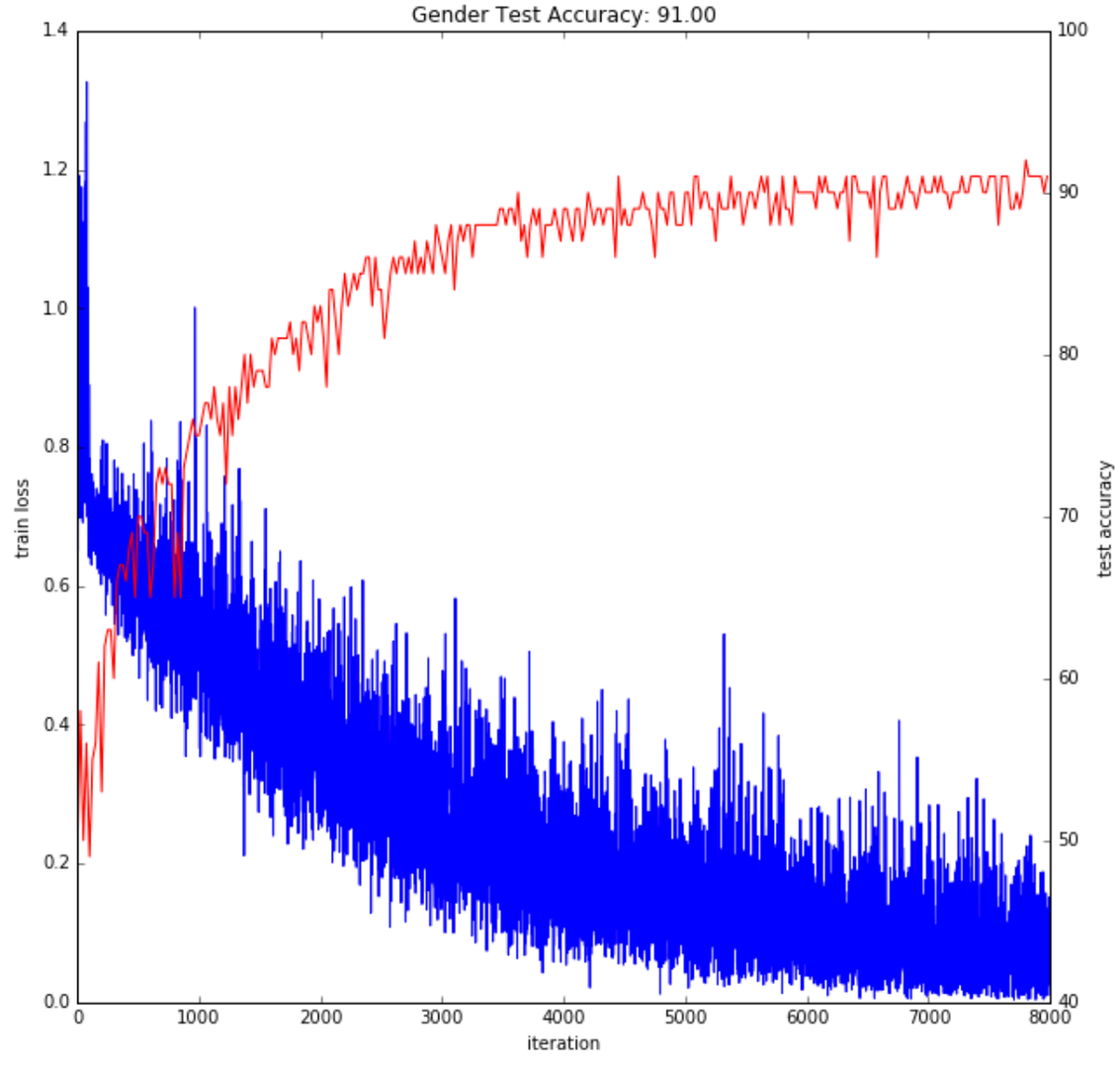
Note that the age vector seems to plateau fairly quickly



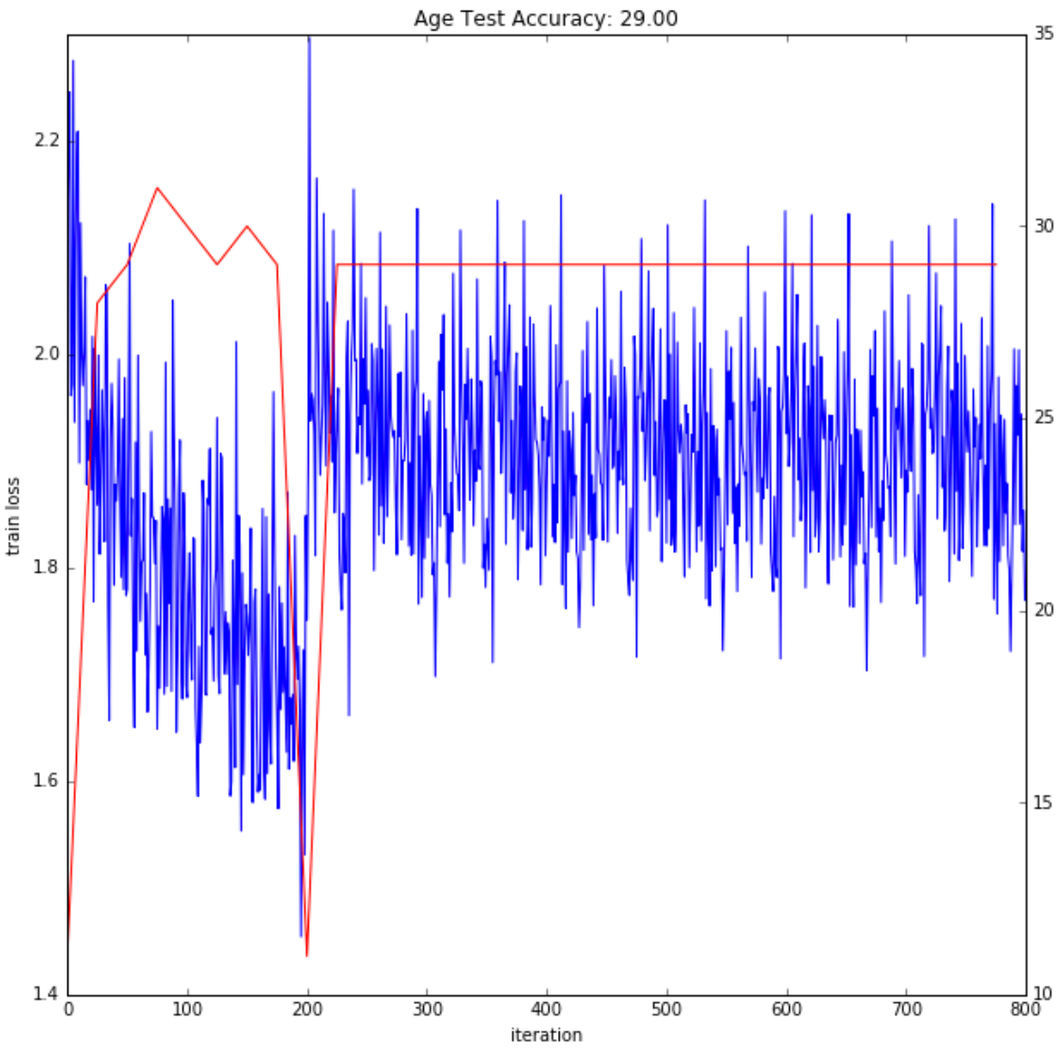


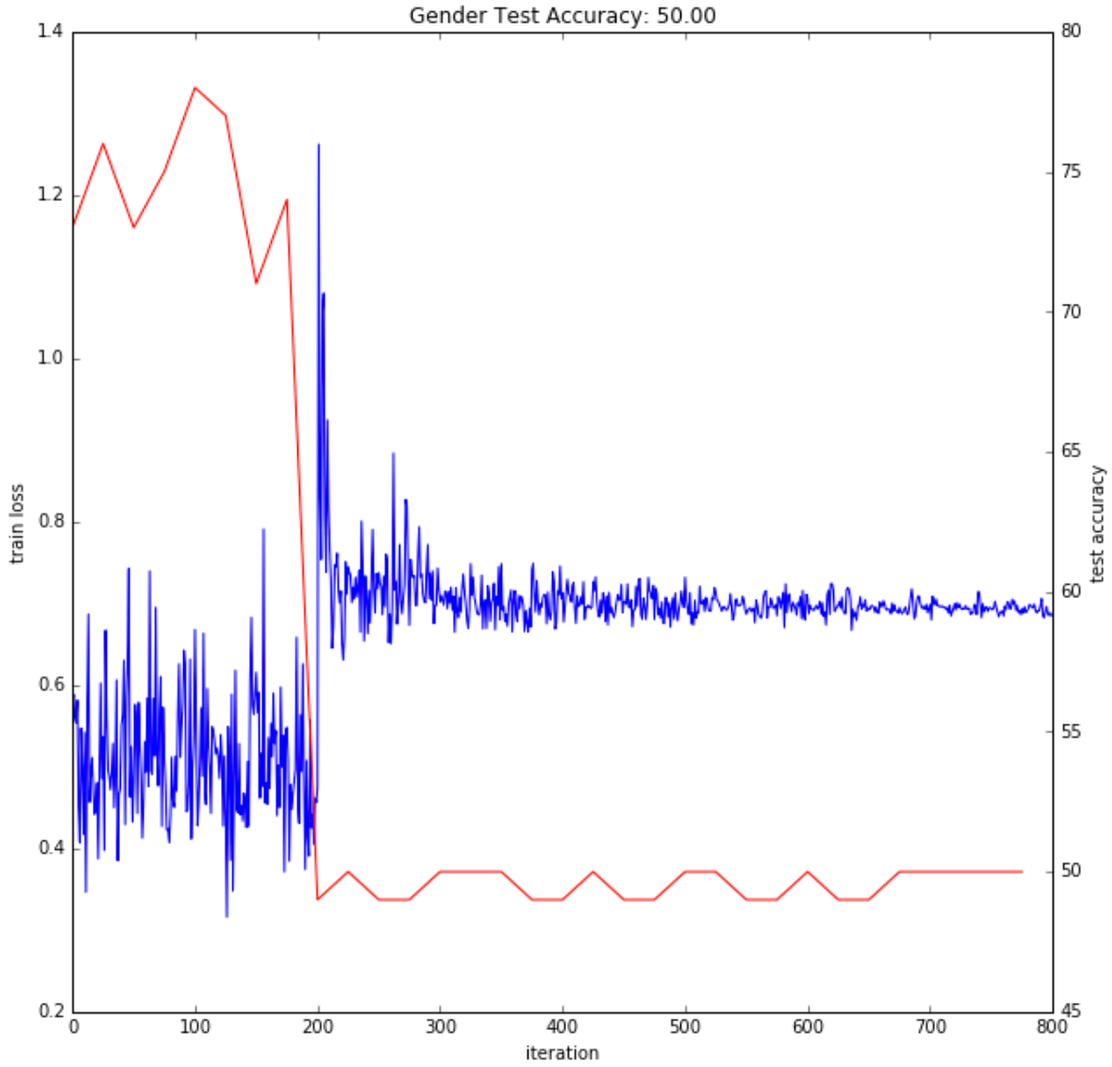
Experiment 0: passing the core two layers back and forth.



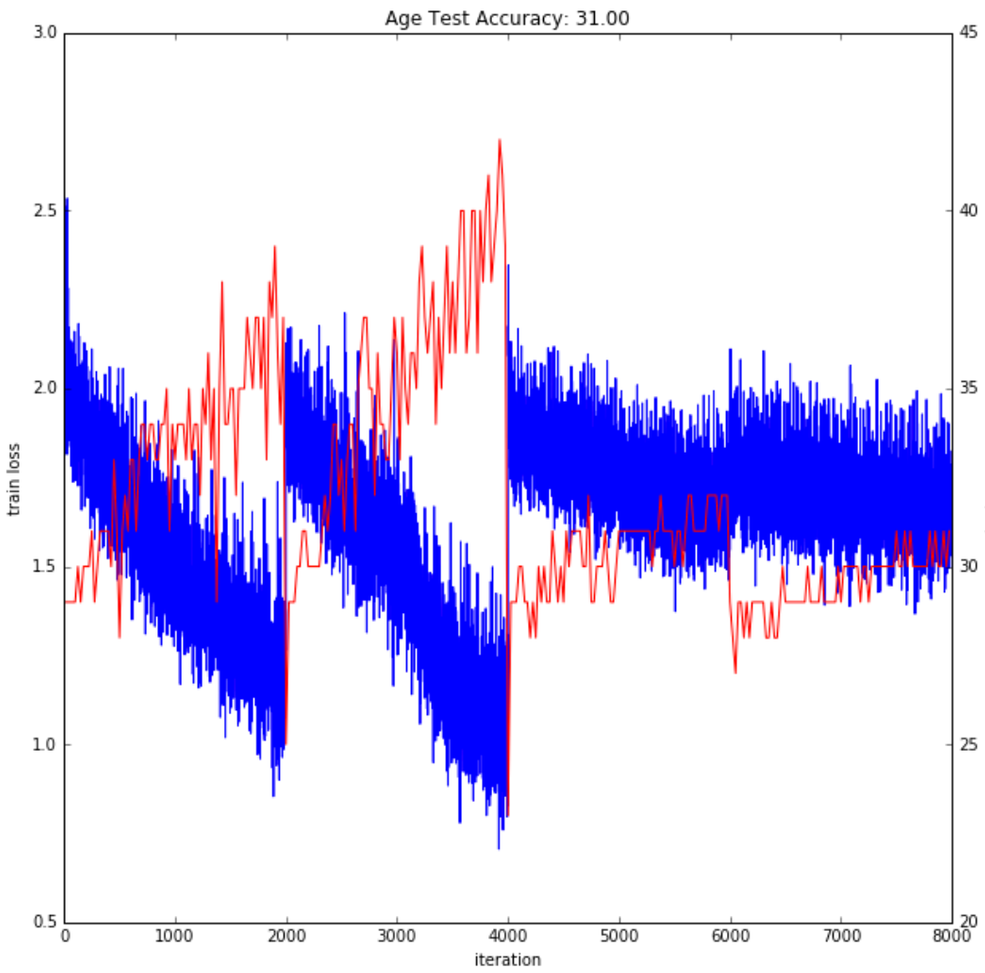


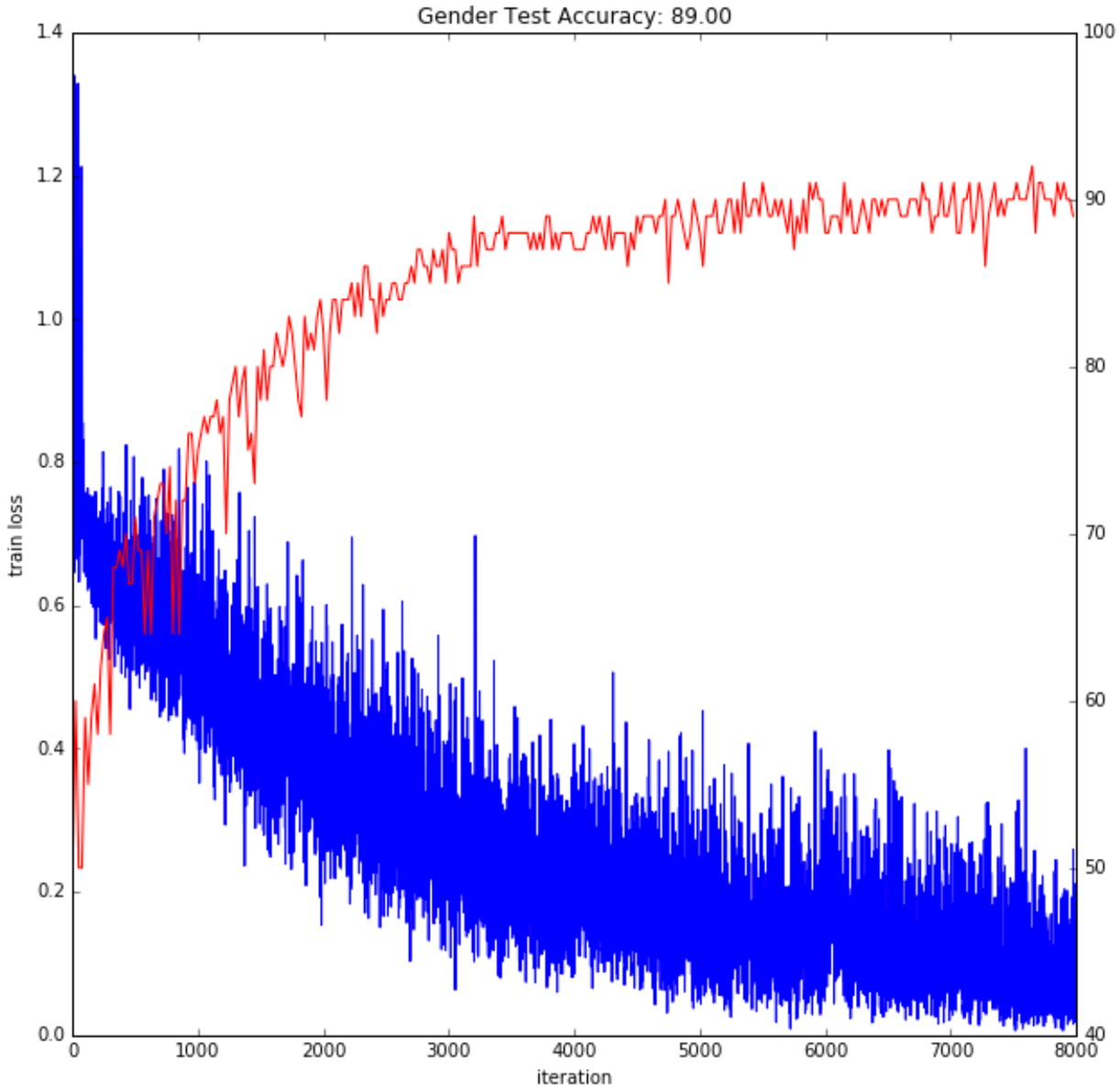
Experiment 2: Acciedntally running the same code twic



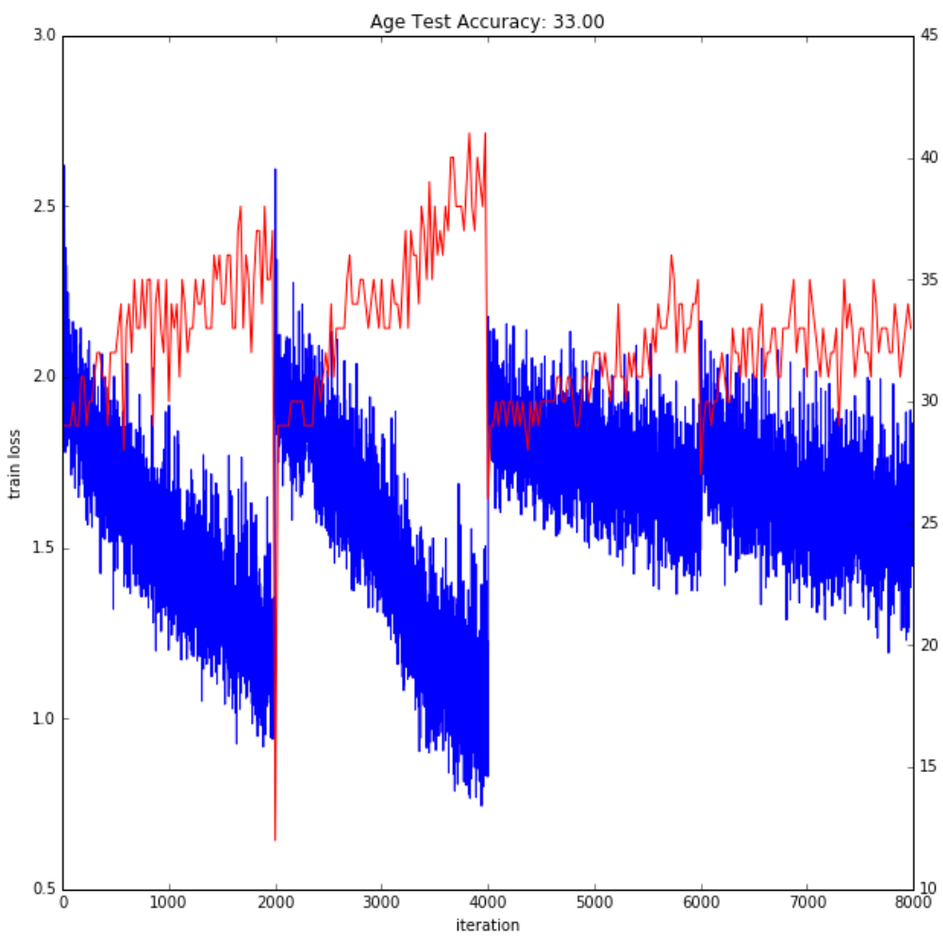


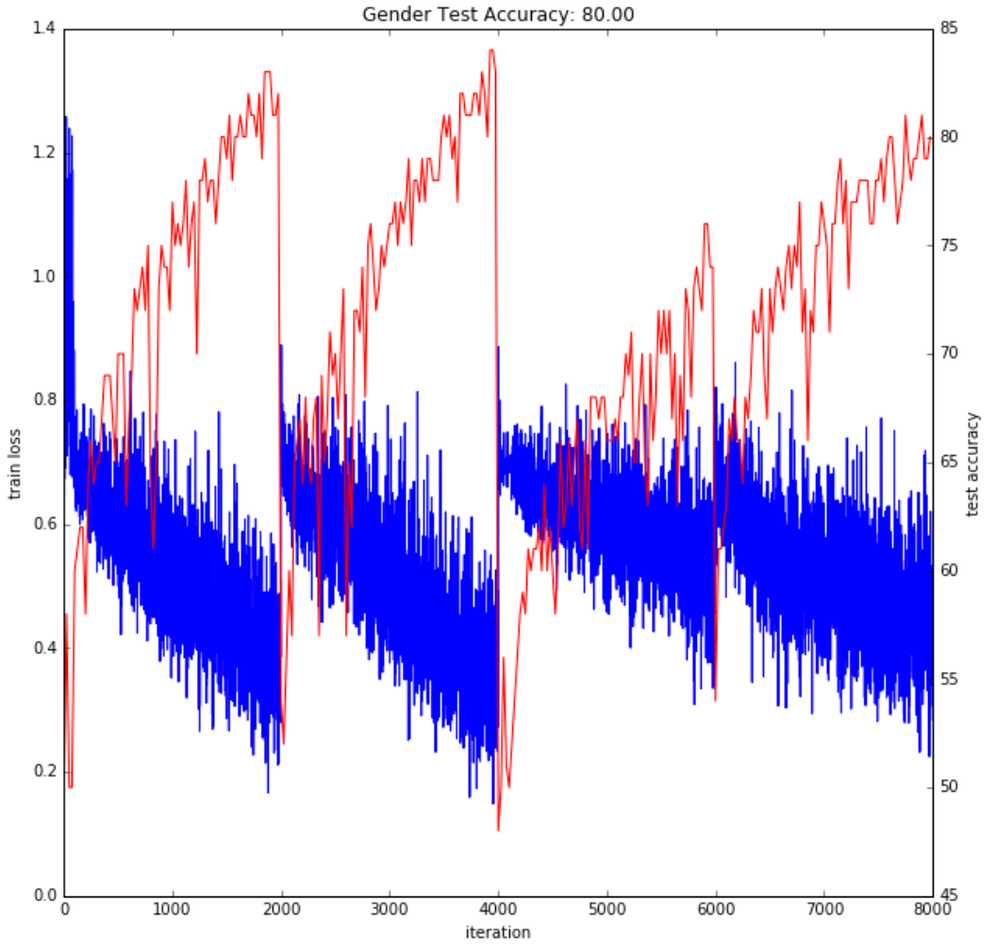
Experiment 3: Proof that 100% sparsification ruins everything



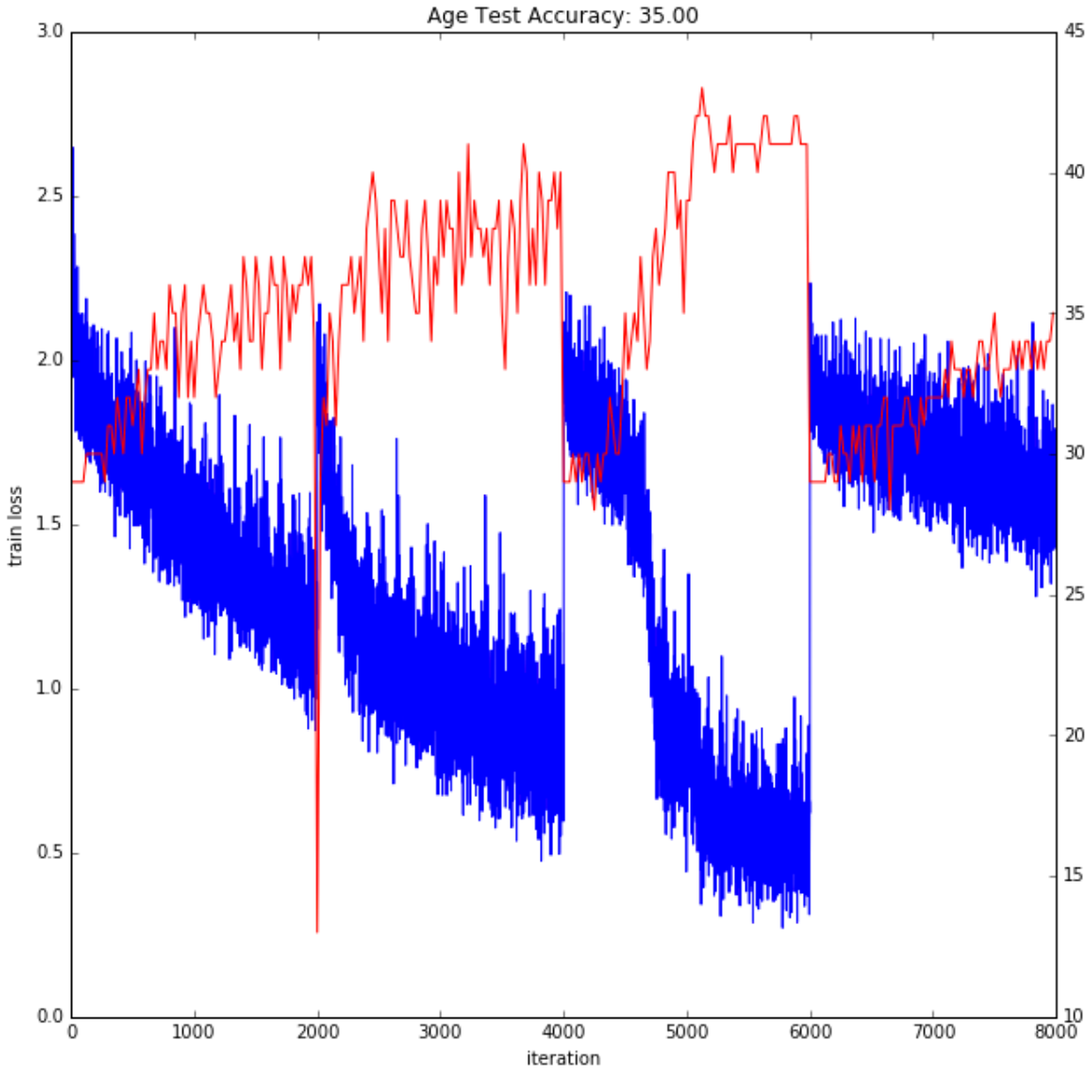


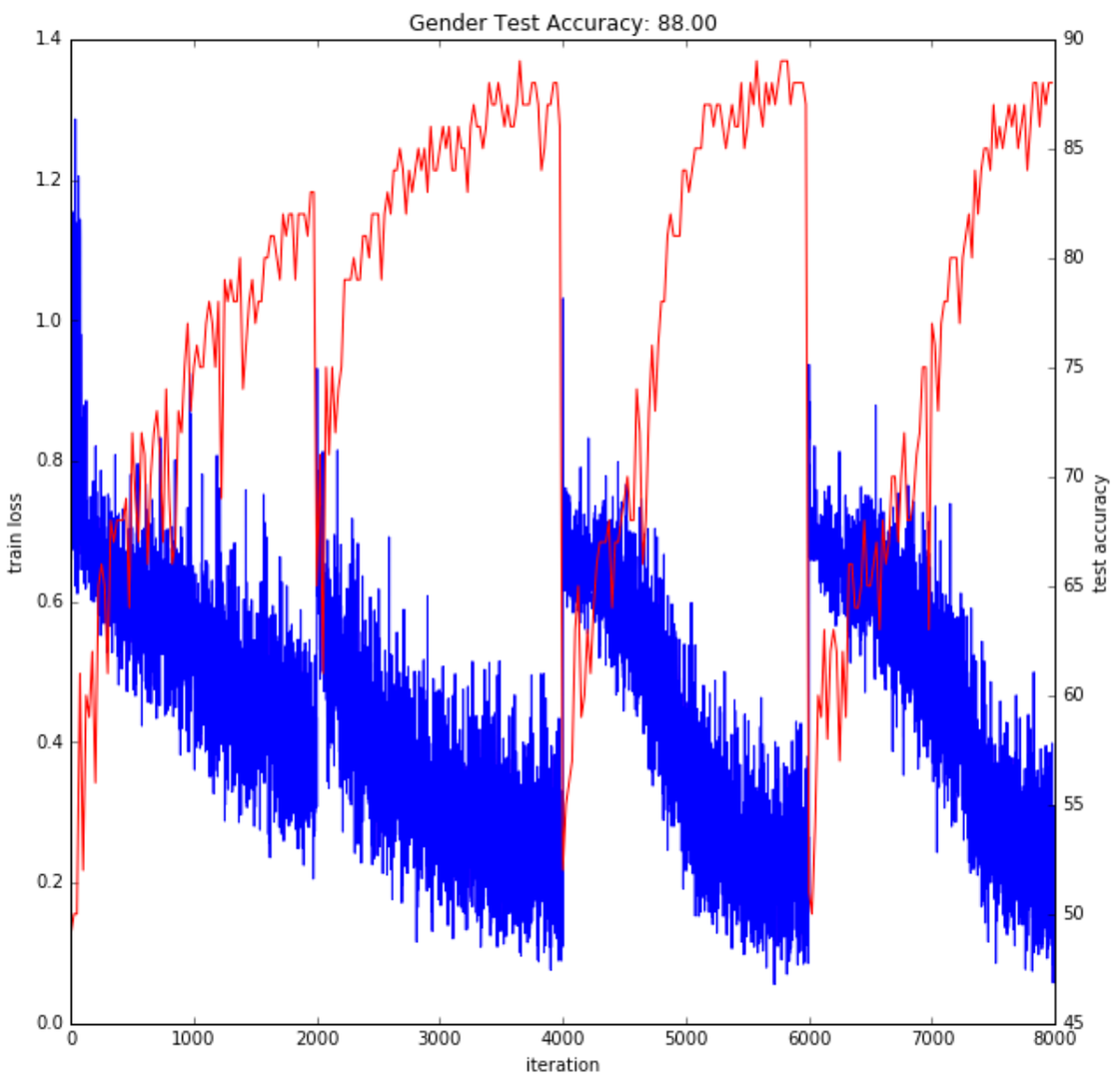
Experiment 4: [0,..1,..2,..3 sparsity]. step LR (might have had a bug in the gender sparsification)



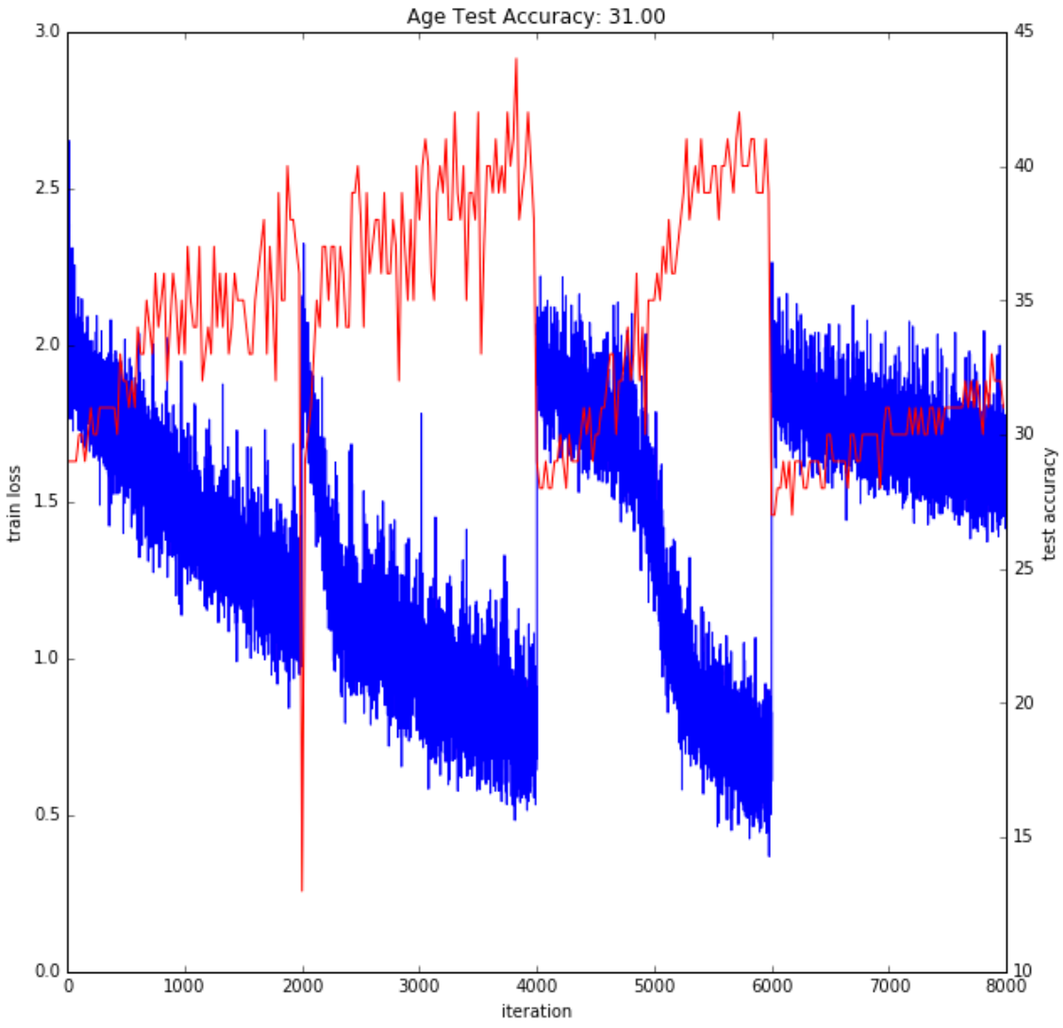


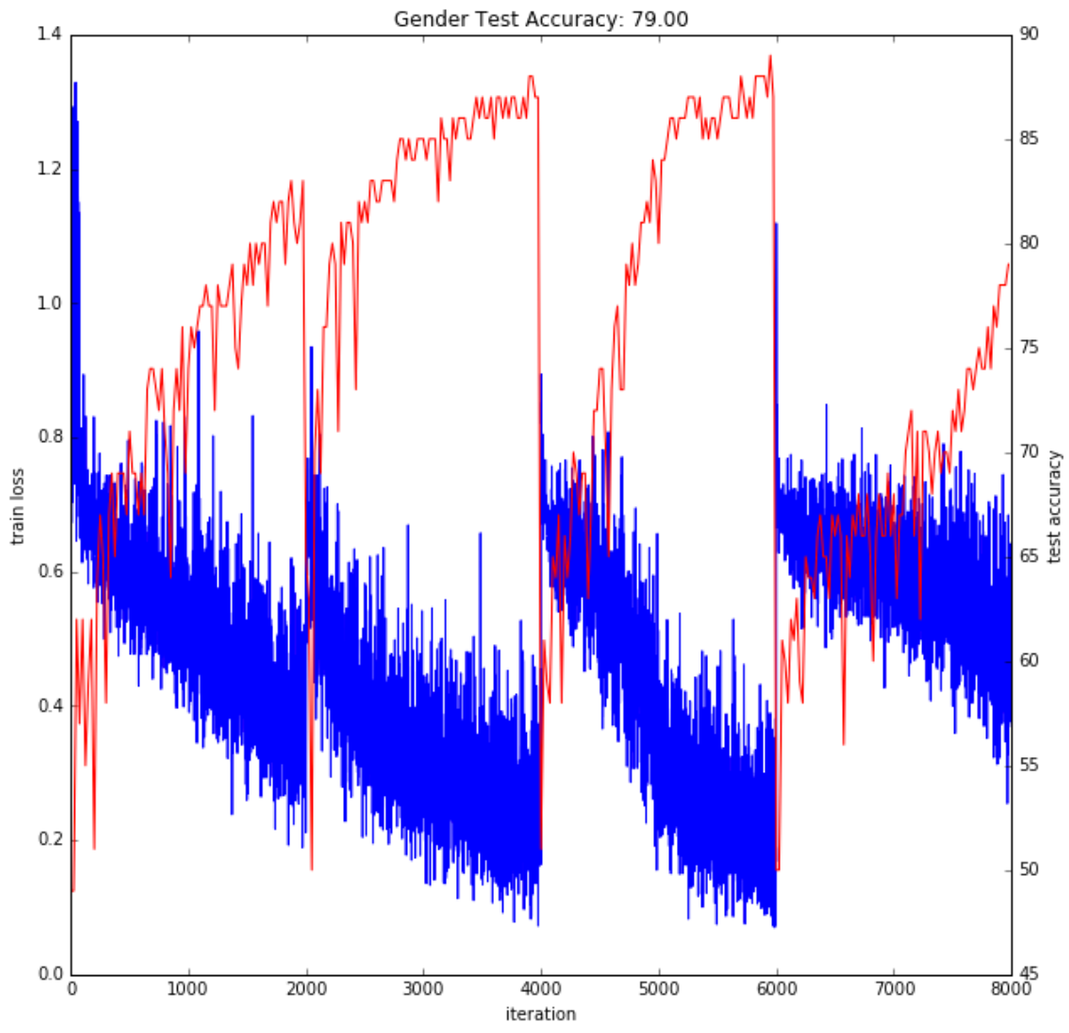
Experiment 5: [0,.1,.2,.3 sparsity]. fixed LR



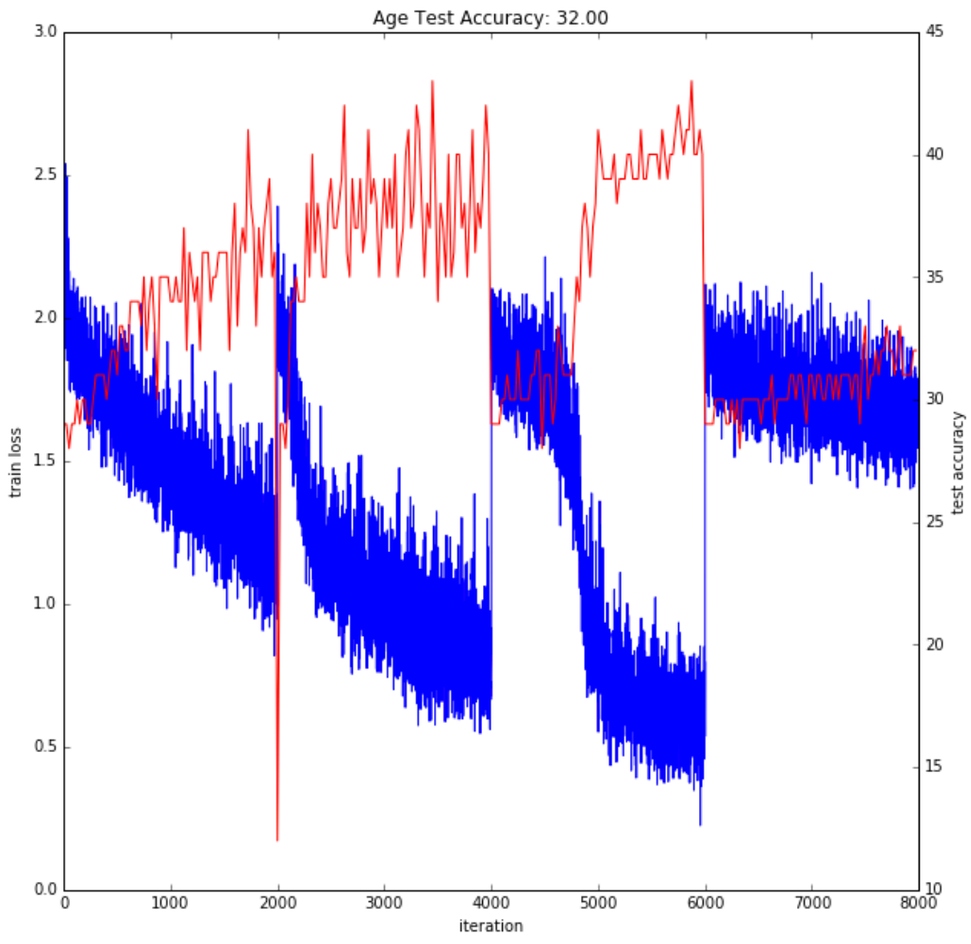


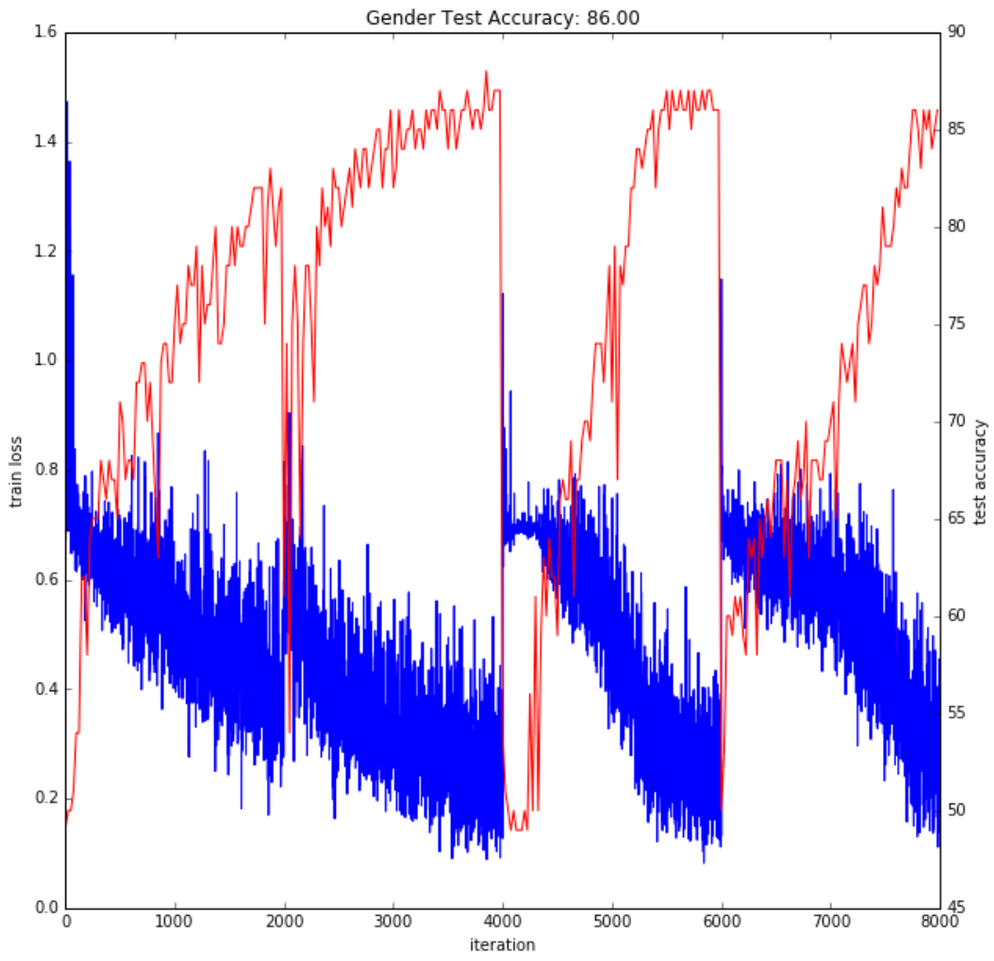
Experiment 6: [0,.05,.1,.15 sparsity]. step LR



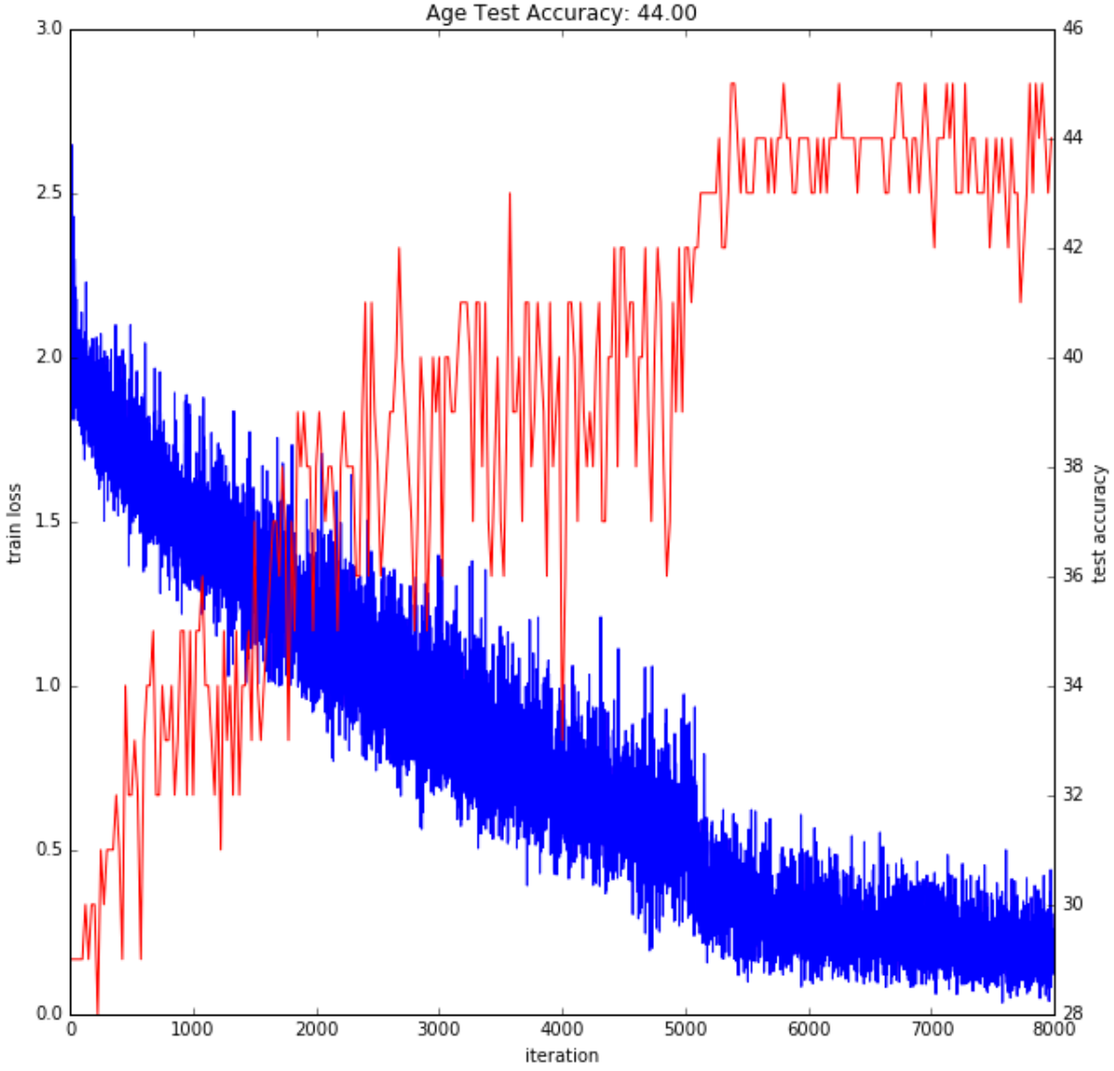


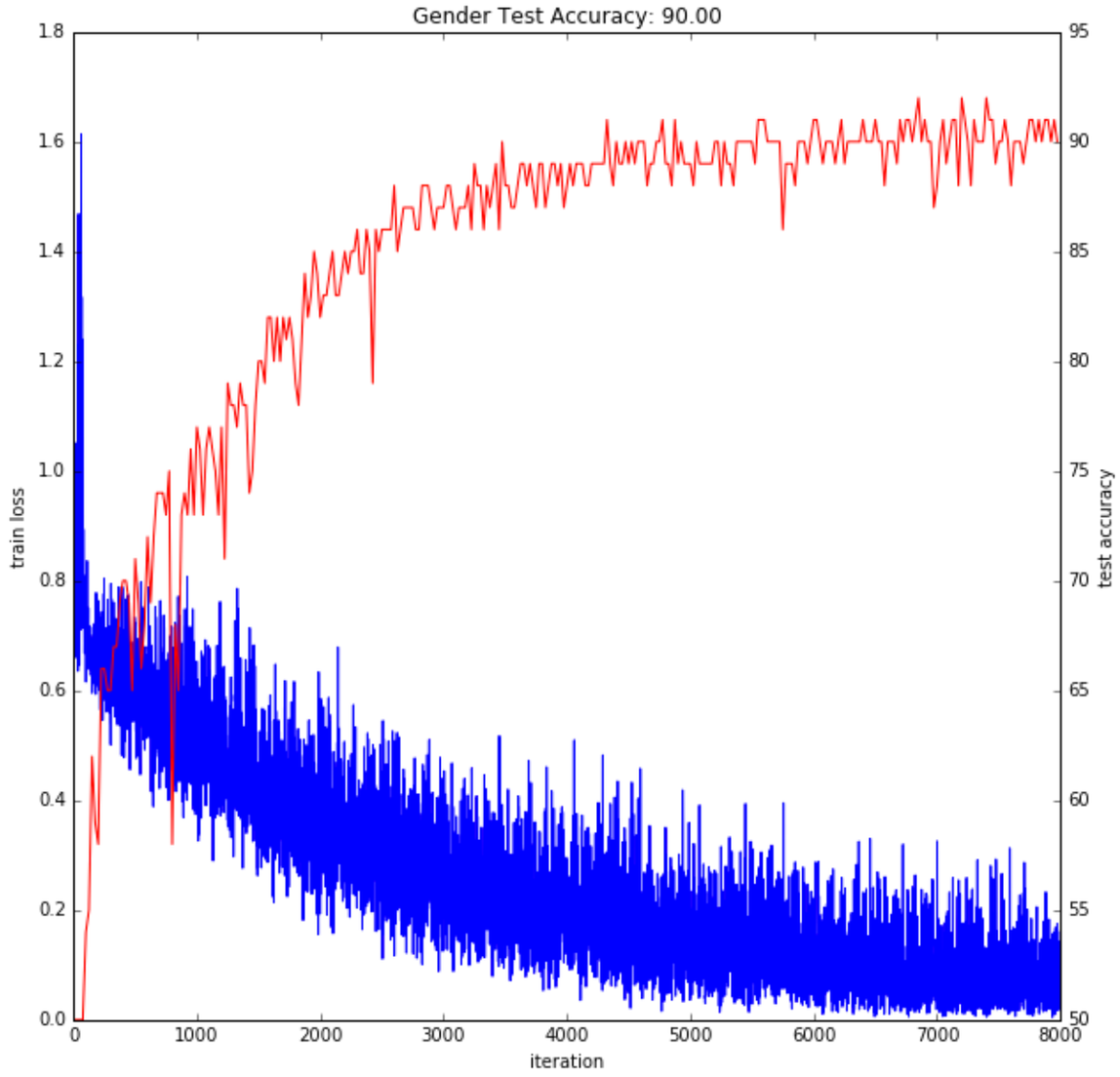
Experiment7: [0,.05,.1,.15 sparsity].
Adding noise (.2) to just conv1
and conv2.





Experiment8:
[0,.05,.1,.15 sparsity] to conv1 conv2.
Adding noise (.2) to conv1, conv2, conv3.





Experiment 9:
no sparsity
Adding noise (.2) to conv1, conv2, conv3.
Sharing conv1, conv2