Fall 2013

Learning abstract features from images and audio with stacked denoising autoencoders

Masters Thesis Defense
Nathan H. Nifong

Dept. of Systems Science
Portland State University

Wednesday, October 16, 2013

Welcome to my thesis defense, Learning Abstract Features from Images and Audio with stacked denoising autoencoders.

Presentation Outline

Stacked Denoising Autoencoder (SDA)

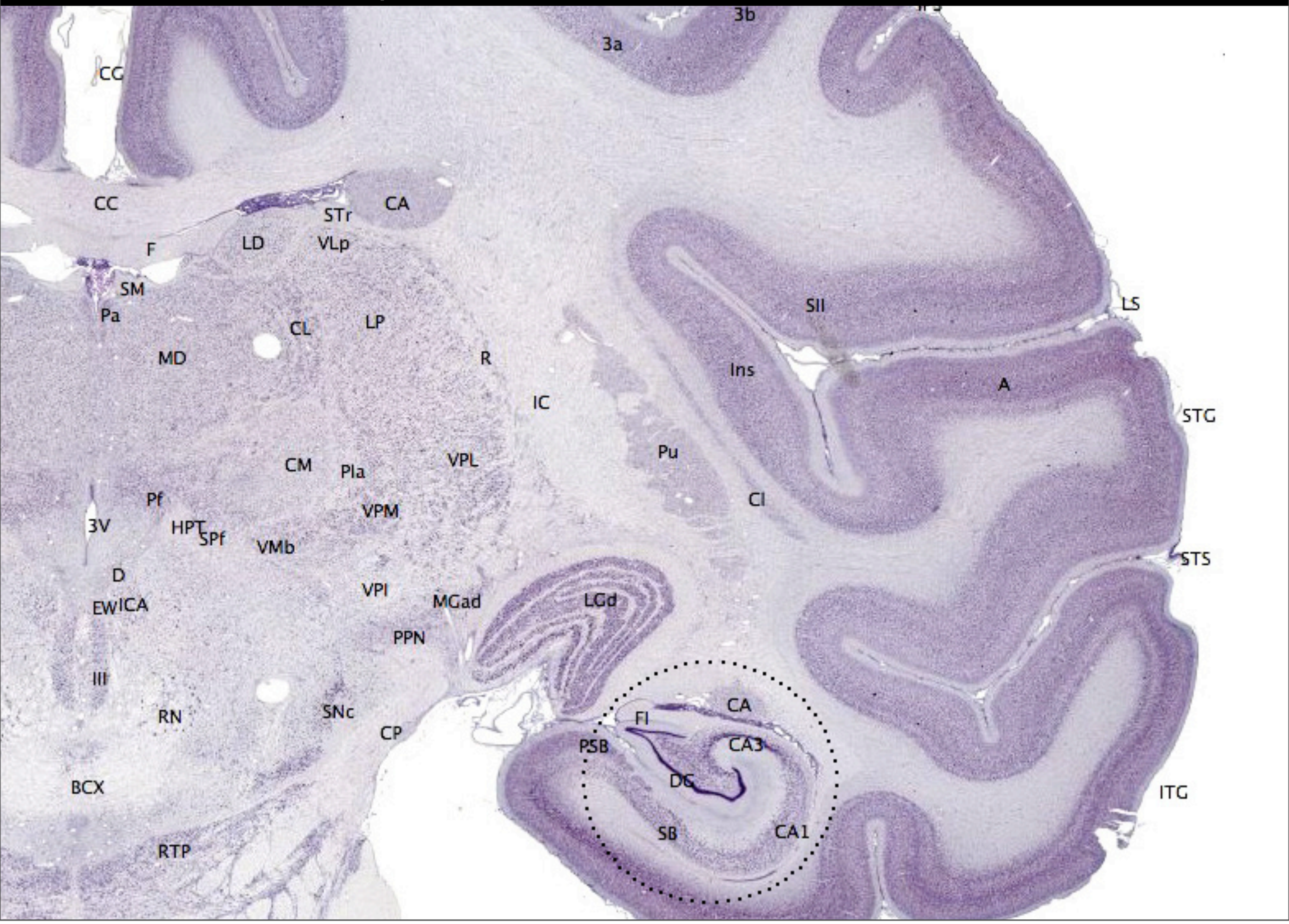
- Motivation for plasticity experiments
- SDA algorithm and Experiment Design
- Experiment results and Interpretation
- Further Study

Wednesday, October 16, 2013

I'll be showing some experiments where I've taken SDAs trained on images, and switched them to audio to see how they perform, and vice versa.

First I'll be talking about the motivation for studying plasticity in a machine learning context.
I'll then explain the SDA algorithm, and my experiment design.
Next, I'll show my results and interpretation, along with some graphs that visualize what the networks learned.
Finally, I'll talk about my future work and where I think these algorithms are going.

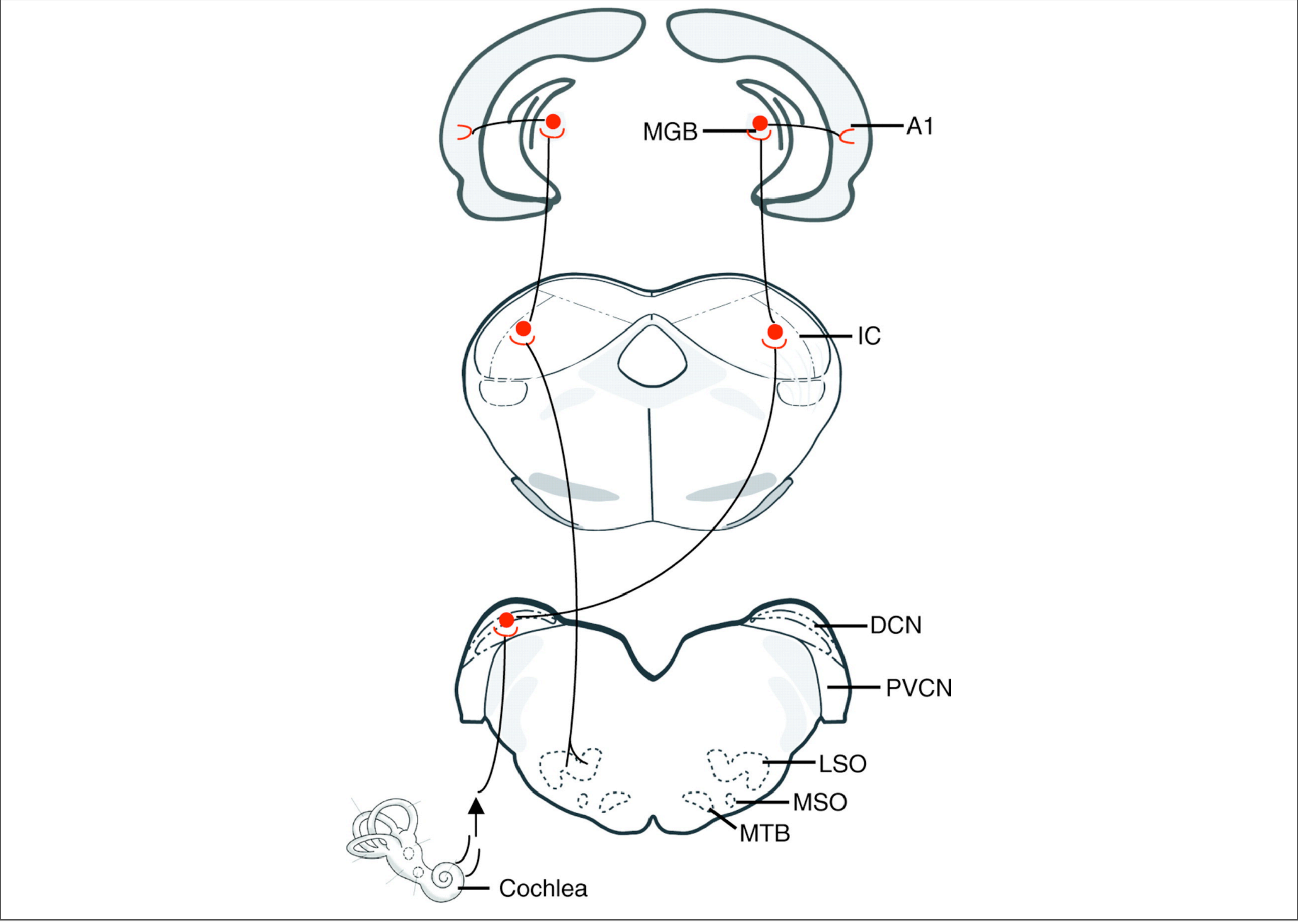
Macaque neocortex - Brainmaps.org - CC Attribution 3.0



Wednesday, October 16, 2013

The neocortex performs an incredible service to the rest of the brain. It finds patterns in both sensory inflow and motor outflow. It learns a general model of "what happens" in the very short term, to the very long term, and eventually overrides observations and instincts with predictions of it's own. The neocortex is a generalized pattern finder and controller, which can learn to make the best of anything you throw at it. It's uniform structure and the speed which with it can adapt to new tasks suggests that it is using a similar process throughout it's extent.

Auditory pathway in a rat's brain



Wednesday, October 16, 2013

The brain is said to display “plasticity”, meaning that it is flexible. If you wire up the auditory nerve from a rat's ear to it's primary visual cortex, it can learn to hear with a part of its brain that usually does vision. Plasticity experiments like this in animals were some of the first evidence that the neocortex may perform a generic uniform function.

Jan Scheuermann learning to use a neuroprosthesis

Wednesday, October 16, 2013

Communicating directly with the neocortex is a challenge, but it's still easier than researchers first thought. Connect a robotic arm to the neocortex and it will be assimilated like your own limb. The brain learns the associations between the signals it produces and the results in the world. Not only can areas of the cortex adapt to sensory and motor modalities that they are not natively connected to, they can adapt to non-biological systems so long as there is a sensory feedback loop to learn from. This supports the idea that the cortex is a general pattern finder and controller.

Tongue vision system developed by Paul Bach-y-Rita

Wednesday, October 16, 2013

If you place an array of electrodes on your tongue and connect them to a camera, you can learn to see without your eyes. This technology is one of many being used right now to assist the blind. You could also connect that sensor to virtually anything if you wanted. After a little practice, it begins to feel like another sense.

CC - Robert Wallace - Flickr

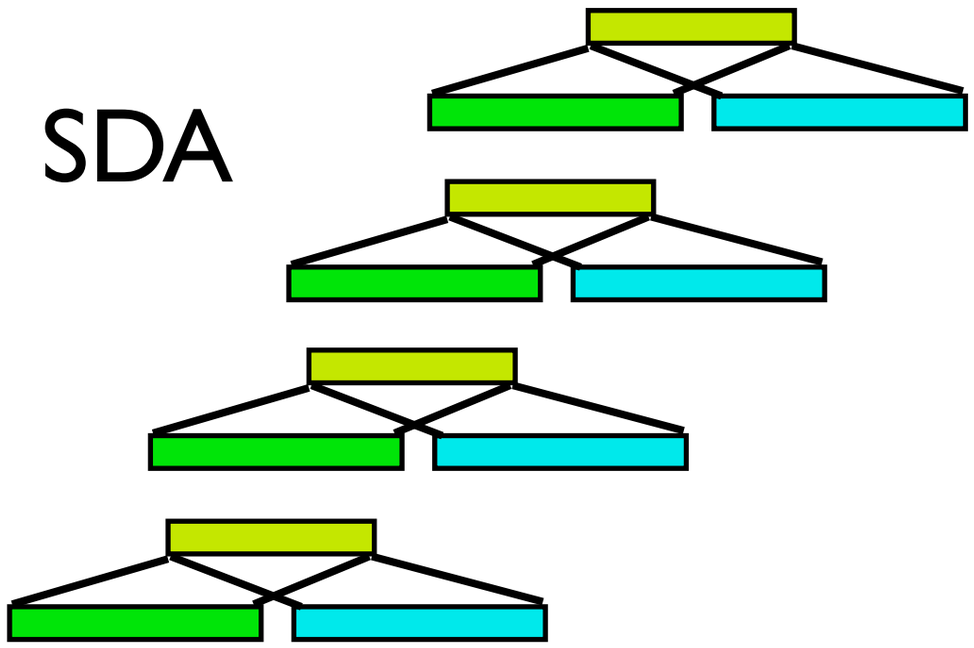
CC - Timo Newton-Syms - Flickr

Wednesday, October 16, 2013

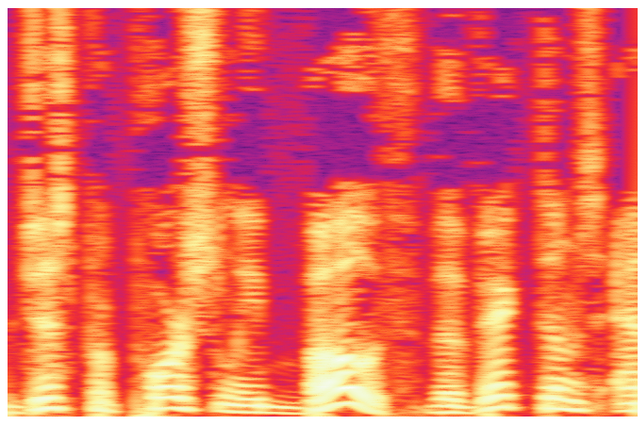
We also see plasticity at work in the way that the brain treats tools as extensions of our bodies. The purpose for which selective pressures have shaped the brain is not to learn to “control the body” it is to learn to control “anything and everything the brain is connected to, however indirectly”

Such a generic pattern-finding controller would revolutionize the way we design everything, and level the playing field of technologies. We need to find this algorithm and I believe we will find it, and deep networks are very close. We are going to find out whatever the cortex is doing, replicate it, and turn it up to eleven!

SDA



images



audio

Does an SDA learn similarly to the neocortex?
Does it share any idiosyncrasies?

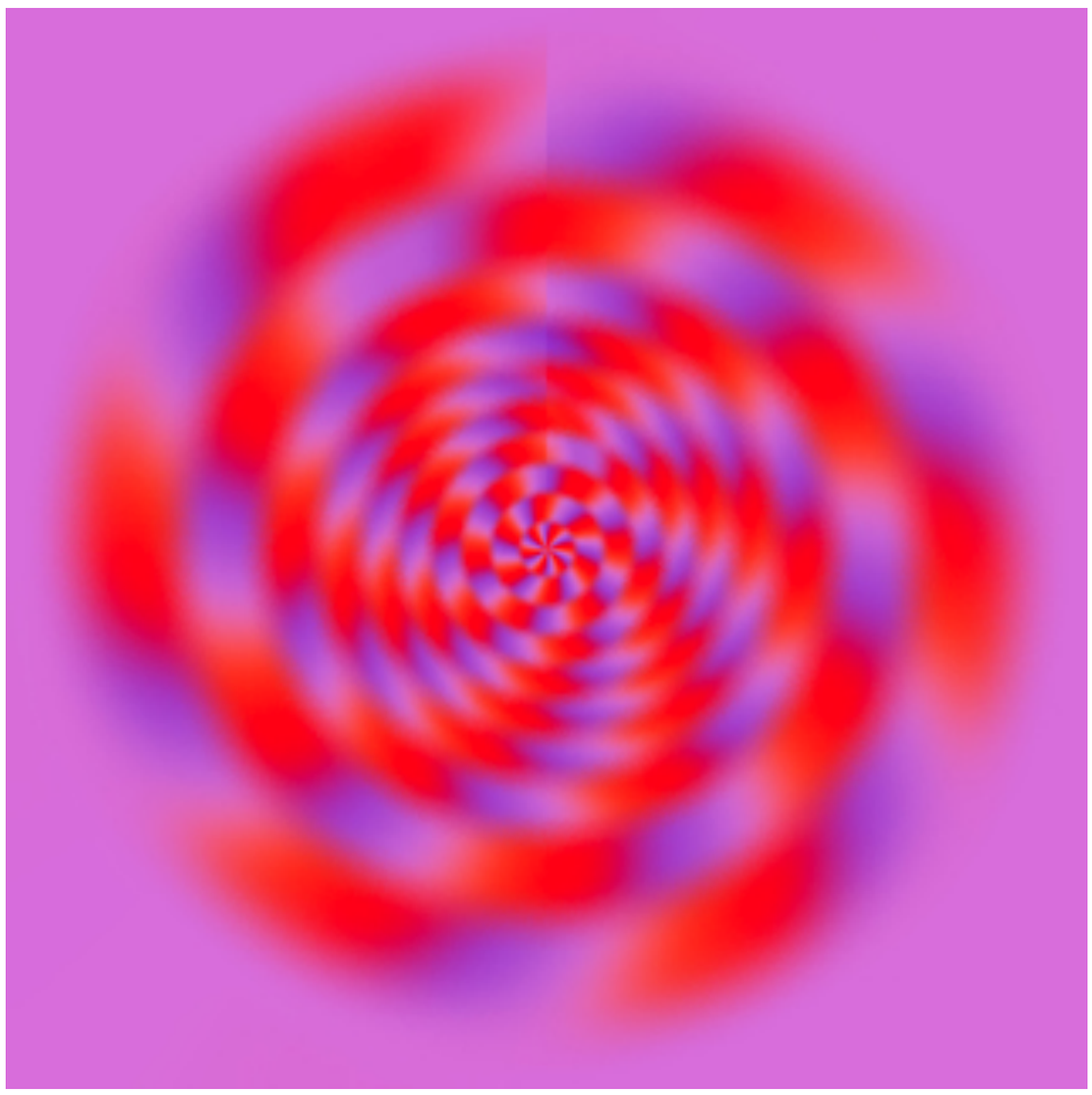
Wednesday, October 16, 2013

That's the motive for learning about stacked denoising autoencoders. I'll show you some experiments which tell us something about the plasticity of stacked denoising autoencoders by assigning them to multiple sensory modalities. I'll go ahead and tell you up front, I didn't find what I expected, but that happens whenever you try something new.

CC - DARREN FLETCHER

Wednesday, October 16, 2013

While I say we are close, Human level intelligence is still a really far-off goal, There's also no good consensus on how to define intelligence, although there are some workable definitions out there. What I find more useful in the short term are the idiosyncrasies of the brain that help us frame the problem, like what kinds of biases we have, and unique limitations that may be operationally defined by the brain’s way of learning.

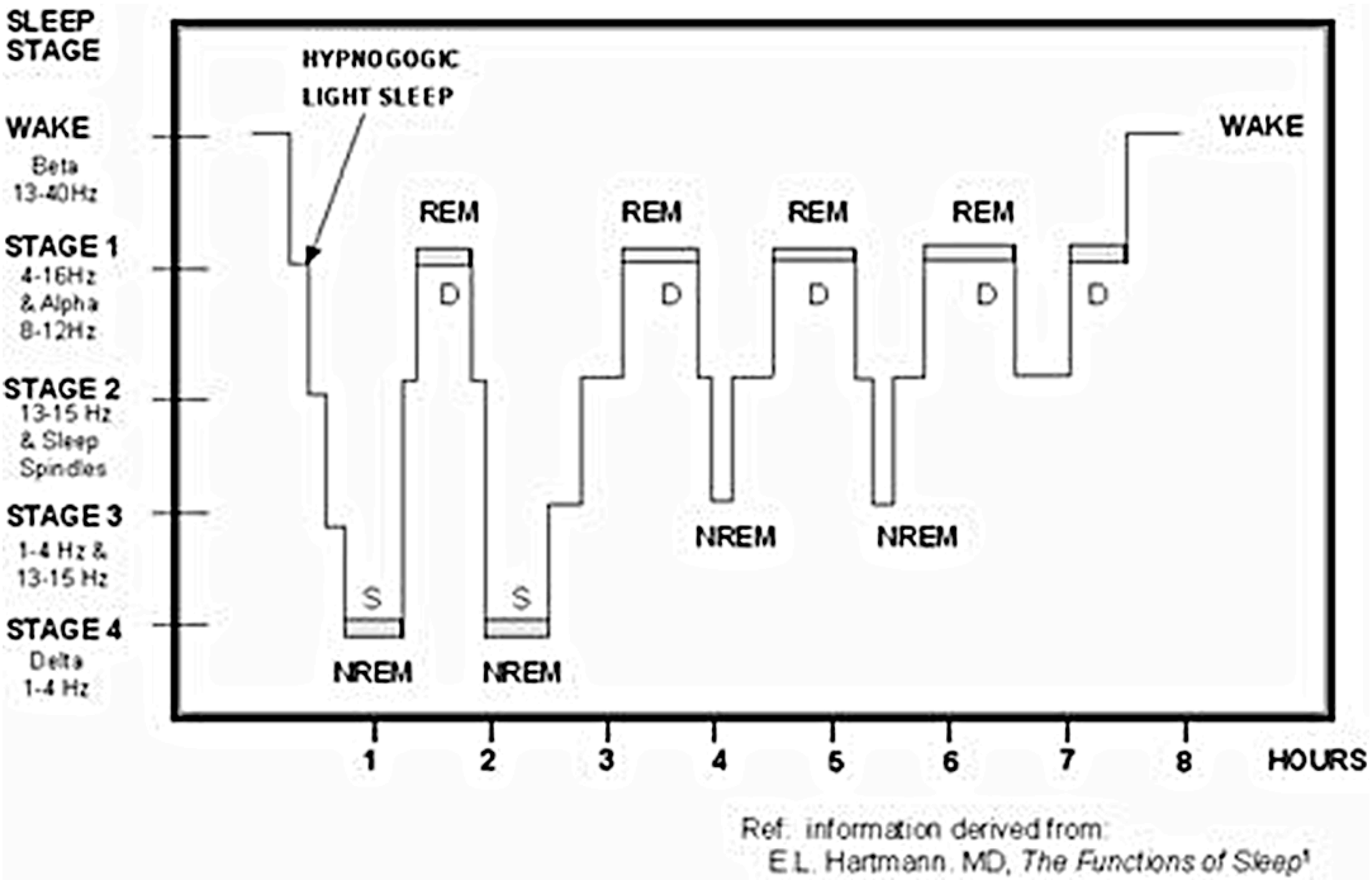


Wednesday, October 16, 2013

For example, our way of perceiving things mixes together our wishes, our predictions, and our observations, giving us a biased picture of the world that is never entirely true. The behavior of existing unsupervised learning algorithms suggest that this is not a flaw, but a fundamental part of perception. Committing to any interpretation of the data allows you to compress it and present a more information-rich stream to higher level processes.

On the left is a picture of some furniture, but our expectations make an entirely different image pop out at us. On the right, the appearance of motion you see is an artifact of the way your visual system has adapted to temporal regularities in colors caused by the slow response time of the retina. Even though the image you see is not moving, your brain is trying to compensate for the usual delay, and *predict* where the shapes will be in a few more milliseconds.

We are not rational observers who only infer upwards from the data, we also infer downwards from our expectations to get the most out of potentially relevant stimuli. This “omni-directional inference” ought to be a requirement for state-of-the-art machine learning algorithms.

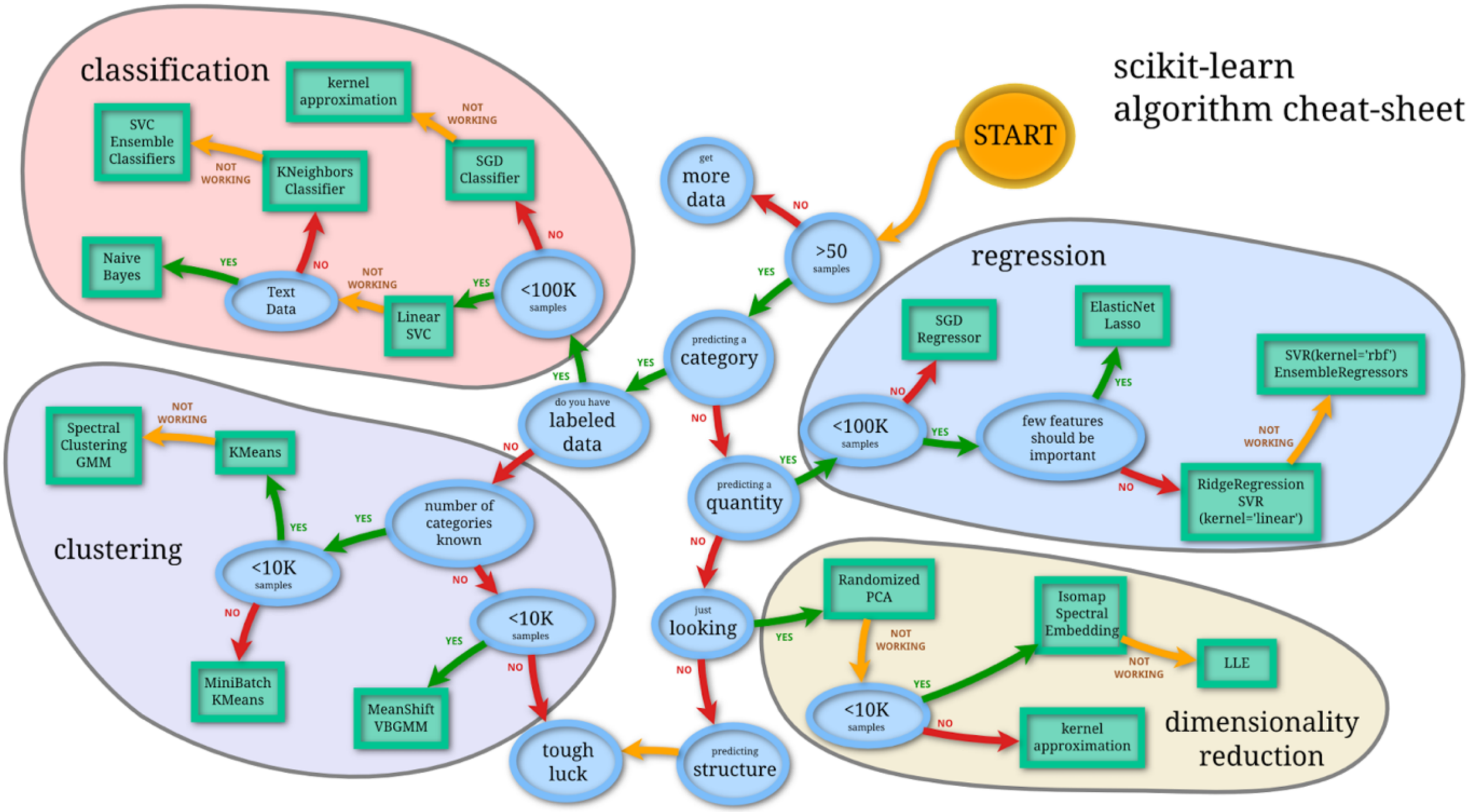


Ref. information derived from:
E.L. Hartmann, MD, *The Functions of Sleep*[1]

Wednesday, October 16, 2013

Another idiosyncrasy is that we dream. For reasons not clearly understood, our brain needs to periodically change modes. From a machine learning point of view this may be an insightful design constraint which leads us in the right direction, or it may be a dead end, but at least it's an easier constraint to meet than "having intelligence".

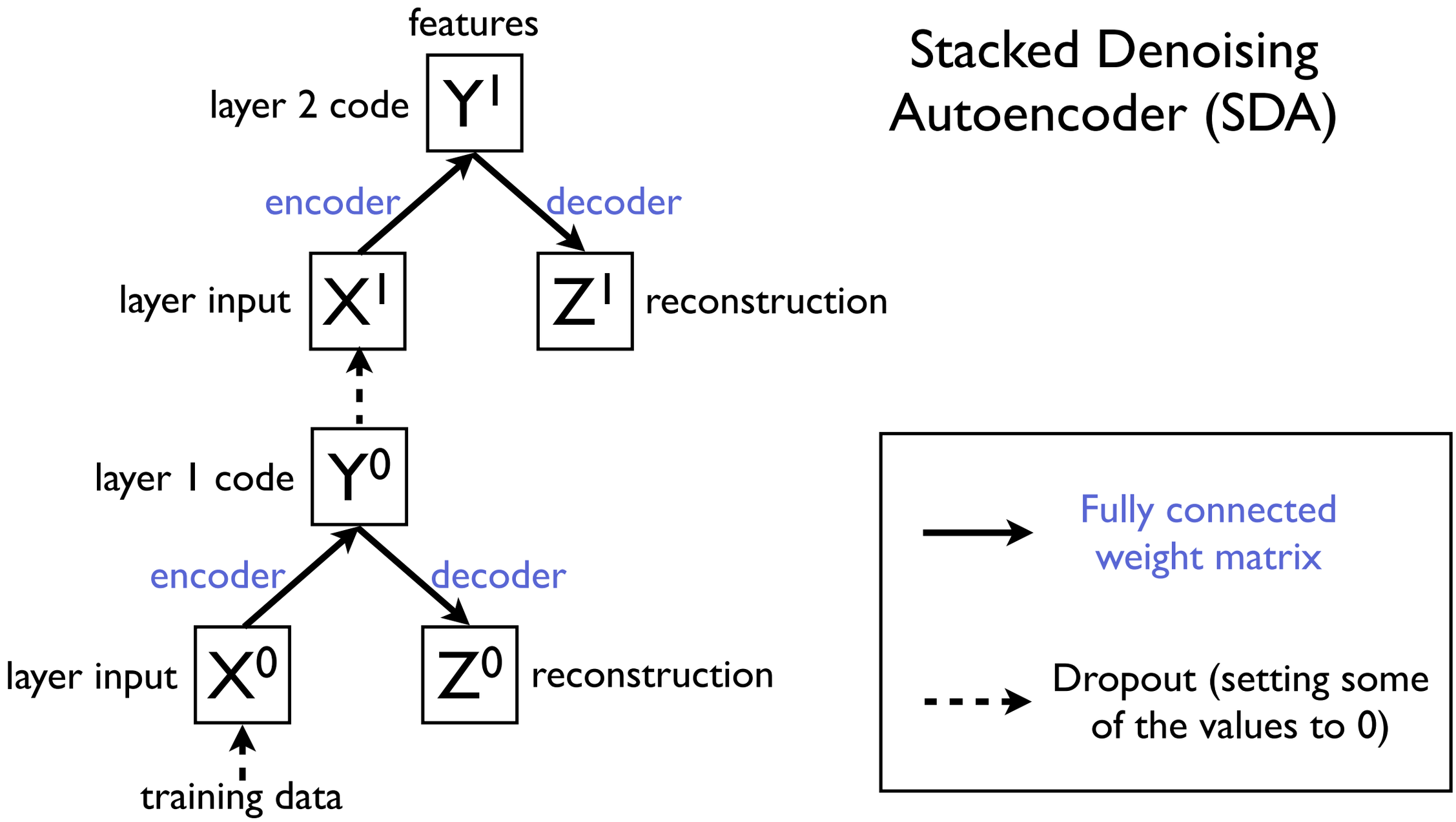
Many machine learning algorithms can operate in a “generative” mode, in which they infer sensory data from labels, or generate a time-series from conditional probabilities, like in the case of hidden Markov models. Many researchers have likened this mode to dreaming, but it’s not clear whether it is useful or just an interesting quirk of these algorithms.



Wednesday, October 16, 2013

I am considering the plasticity of the neocortex with regards to it's behavior when switched from one sensory modality to another to be an idiosyncrasy worth studying. I believe that a learning algorithm which is mediocre at a variety of tasks is more brain-like than a specialized system with near human-level performance on only one task.

This is a diagram of the learning algorithms available the scikit-learn python library. But when I first saw it, I followed along with what I wanted to do and found myself down here in this “tough luck” section. These algorithms are all very good at what they do, but what we really need is a killer algorithm that goes right there, in the "tough luck" section, that can find the hidden structure in data of arbitrary complexity, given sufficient resources and time, and without requiring a googol of training examples.

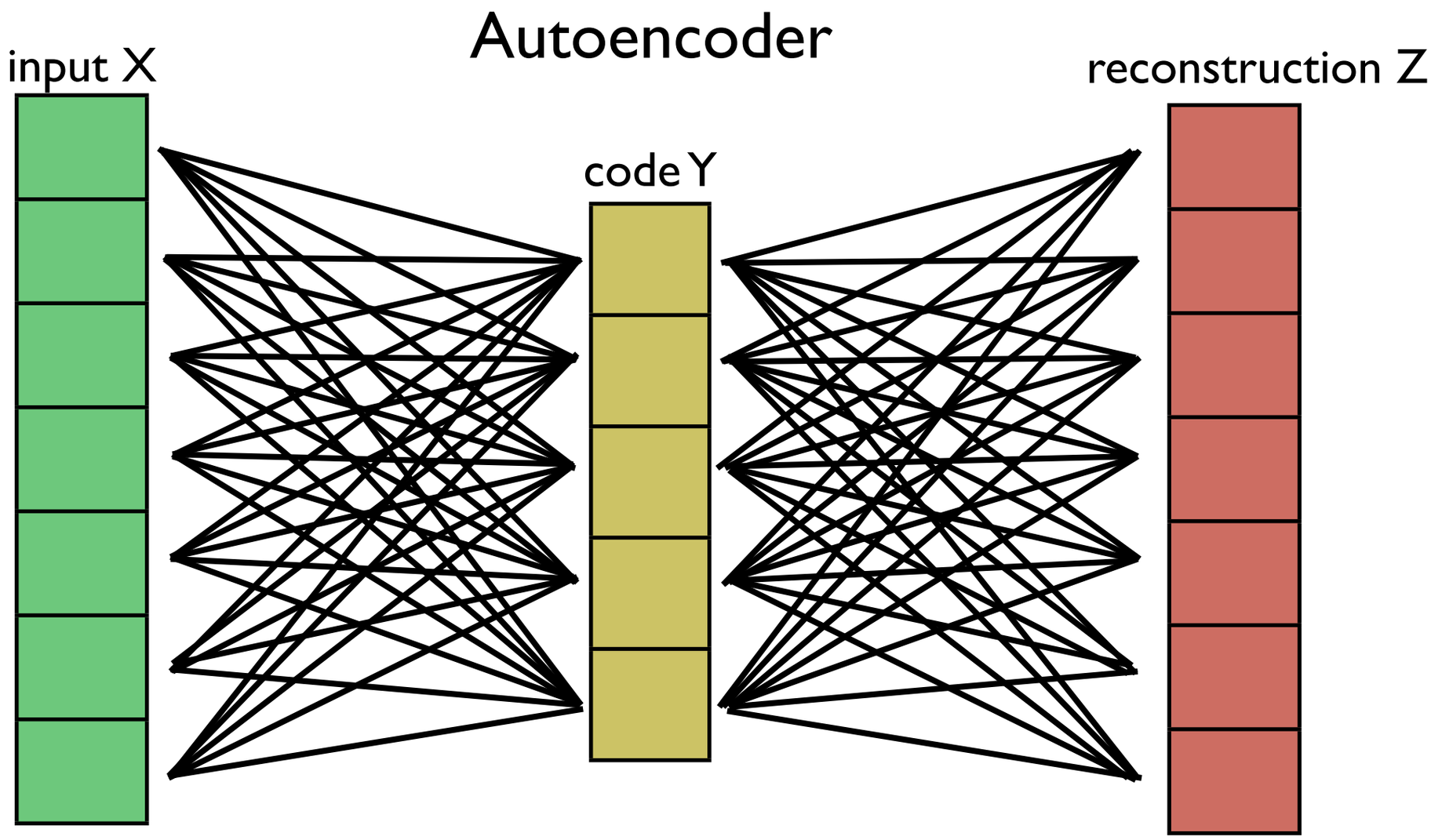


Wednesday, October 16, 2013

I'm using an unsupervised learning algorithm called the stacked denoising autoencoder, recently brought into the spotlight and improved upon by Geoff Hinton at the LISA lab of the University of Montreal, and by Andrew Ng at Google. Deep networks in general have been making huge gains thanks to these researchers and their laboratories.

The SDA's purpose is to learn useful features which can compactly represent examples from the input distribution. Ideally, each layer learns a more abstract representation than the one below it, but in practice, extra measures are needed to ensure this result.

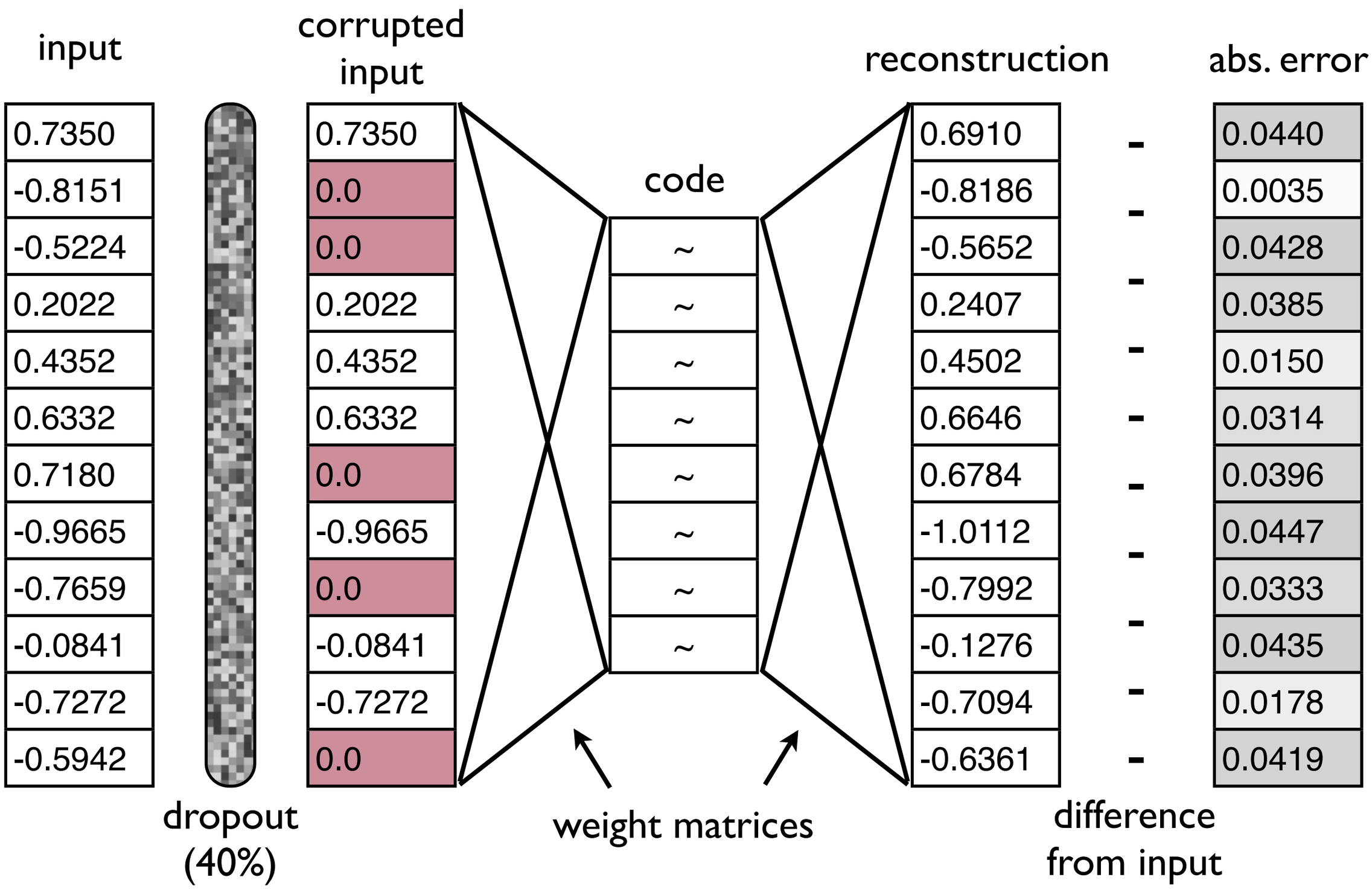
The SDA can be understood in three parts, I'll talk about them in the reverse order they are referred to in the acronym: Autoencoder, Denoising, and Stacked.



An autoencoder is a classical neural network which learns the identity function

Wednesday, October 16, 2013

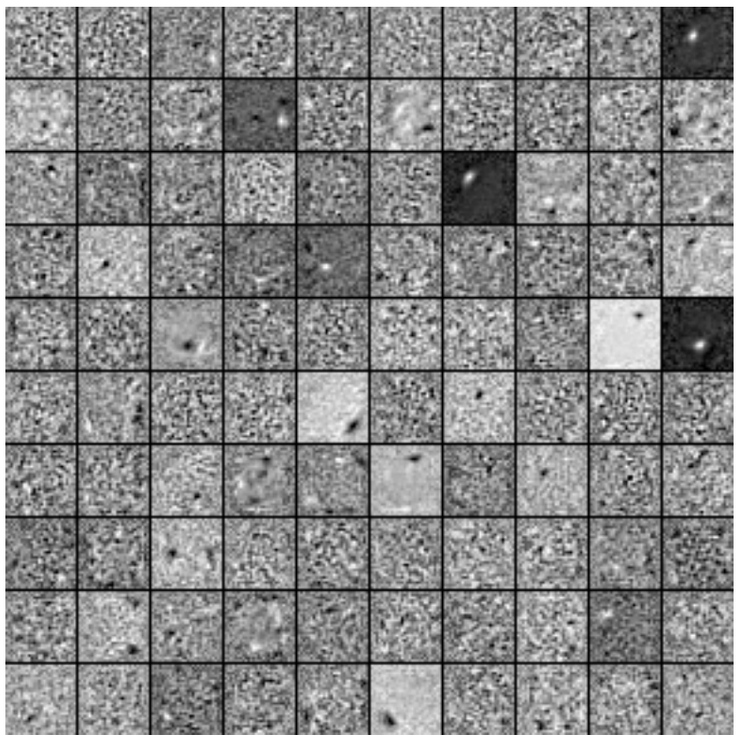
An autoencoder is a classical three-layer neural network which learns the identity function. This means that it is given a vector on its input and the weights are optimized to produce the same vector on the output. The reason this is non-trivial and useful is that the number of hidden nodes is less than the size of the input and output, so the autoencoder must perform some lossy compression. Alternatively, more hidden nodes can be used if a sparsity term is used in the cost function. Because an autoencoder is just a neural network, it can take advantage of the GPU-accelerated linear algebra libraries and specialized hardware that has been designed for neural networks if it is available.



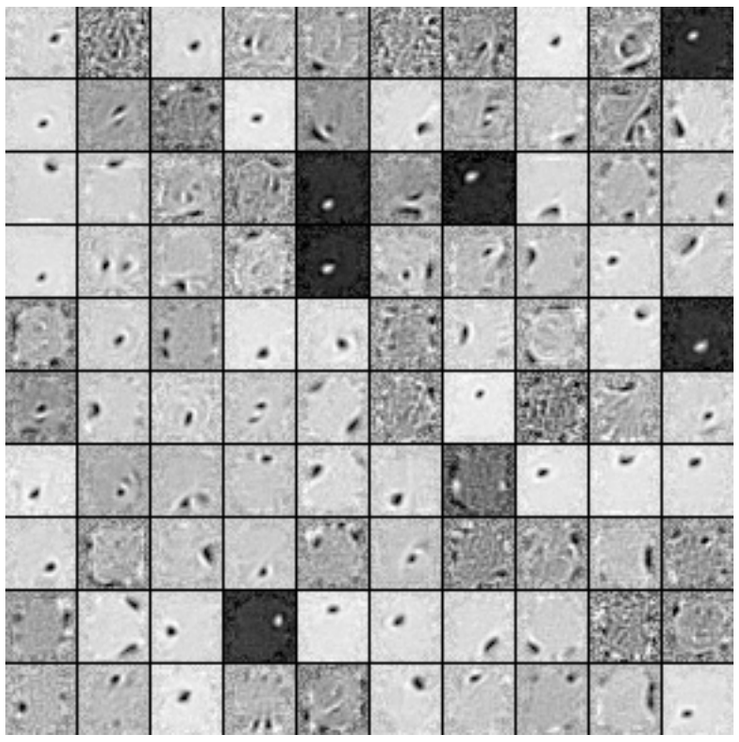
Wednesday, October 16, 2013

A denoising autoencoder is a slight improvement on an autoencoder that generalizes better. By randomly setting some of the values in each input vector to zero, it reconstructs corrupted versions of training examples. To do this, it must learn the correlations that exist between the values in the input distribution. The code is expected to converge on a sparse distributed representation of any given input vector. The weights are optimized to minimize the error between the original uncorrupted input and the reconstruction.

Effect of dropout on the features discovered by a denoising autoencoder on MNIST



without dropout

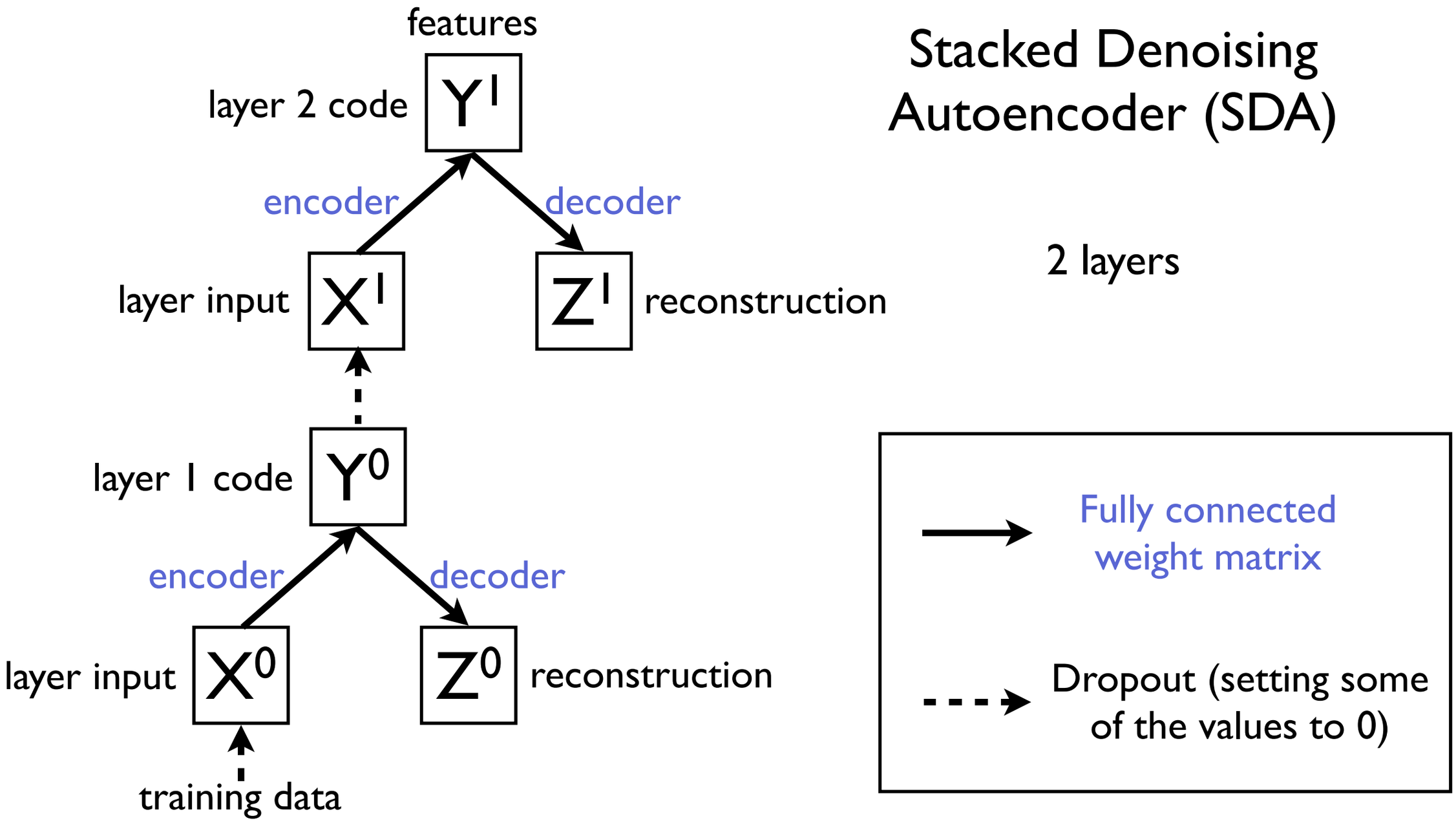


with 30% dropout

Dropout is the process of setting some fraction of the numbers in each input vector to zero.

Wednesday, October 16, 2013

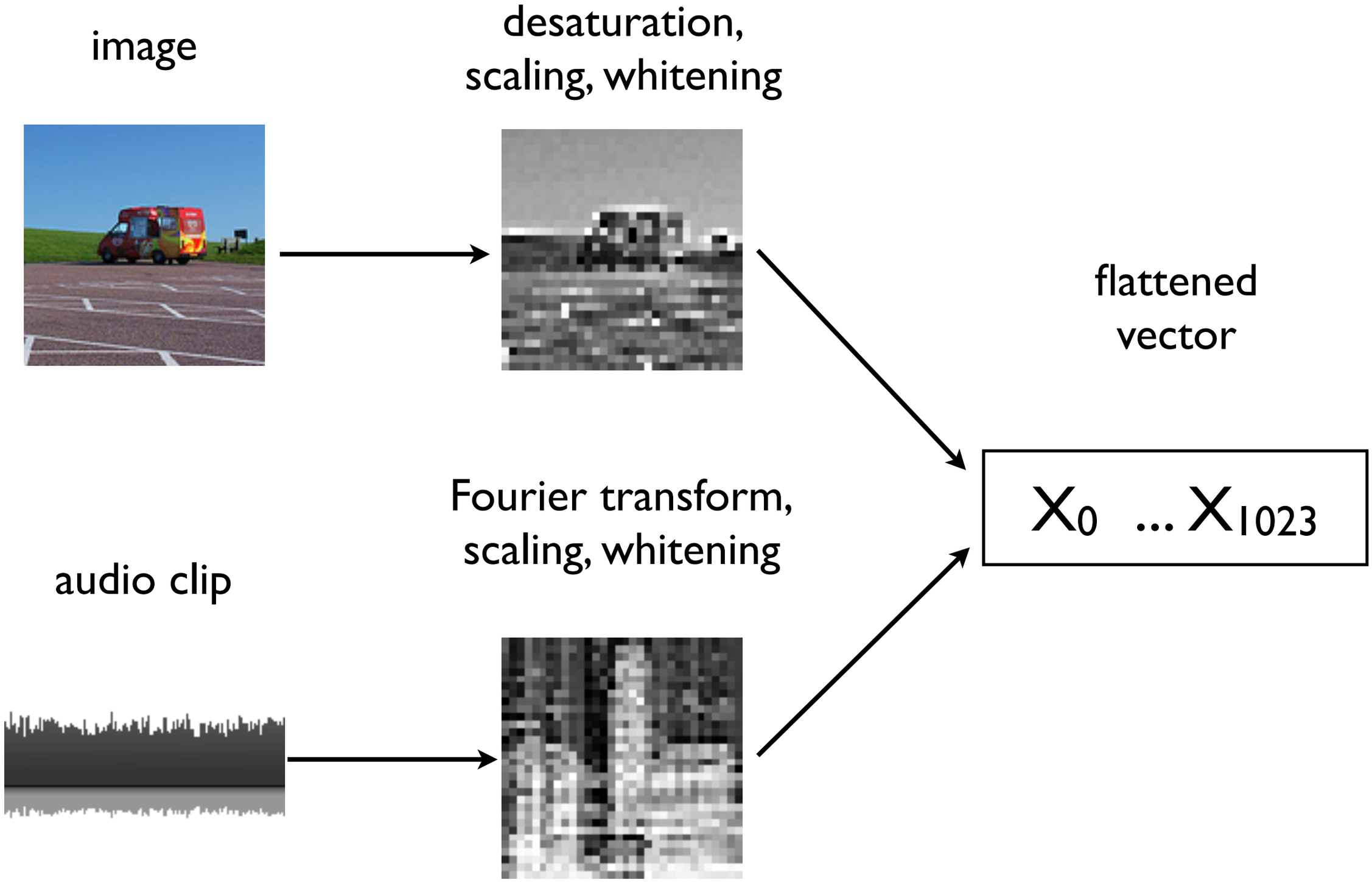
Here you can see what dropout does to the smoothness, sparsity, and generality of the features learned by a denoising autoencoder. These are visualizations of the hidden nodes in a network trained on the MNIST digit recognition dataset. This technique is likened to stochastic resonance.



Wednesday, October 16, 2013

Finally, A stacked denoising autoencoder, is a stack of denoising autoencoders where each one takes as input the activations of the hidden nodes from the one below. The innovation that allows these deep architectures to be trained is called greedy layer-wise pre-training. If you train the layers one at a time from the bottom up, you can create a deep network that finds a much better solution than you could get if you started with a classical neural network of the same topology.

Data set preparation



Wednesday, October 16, 2013

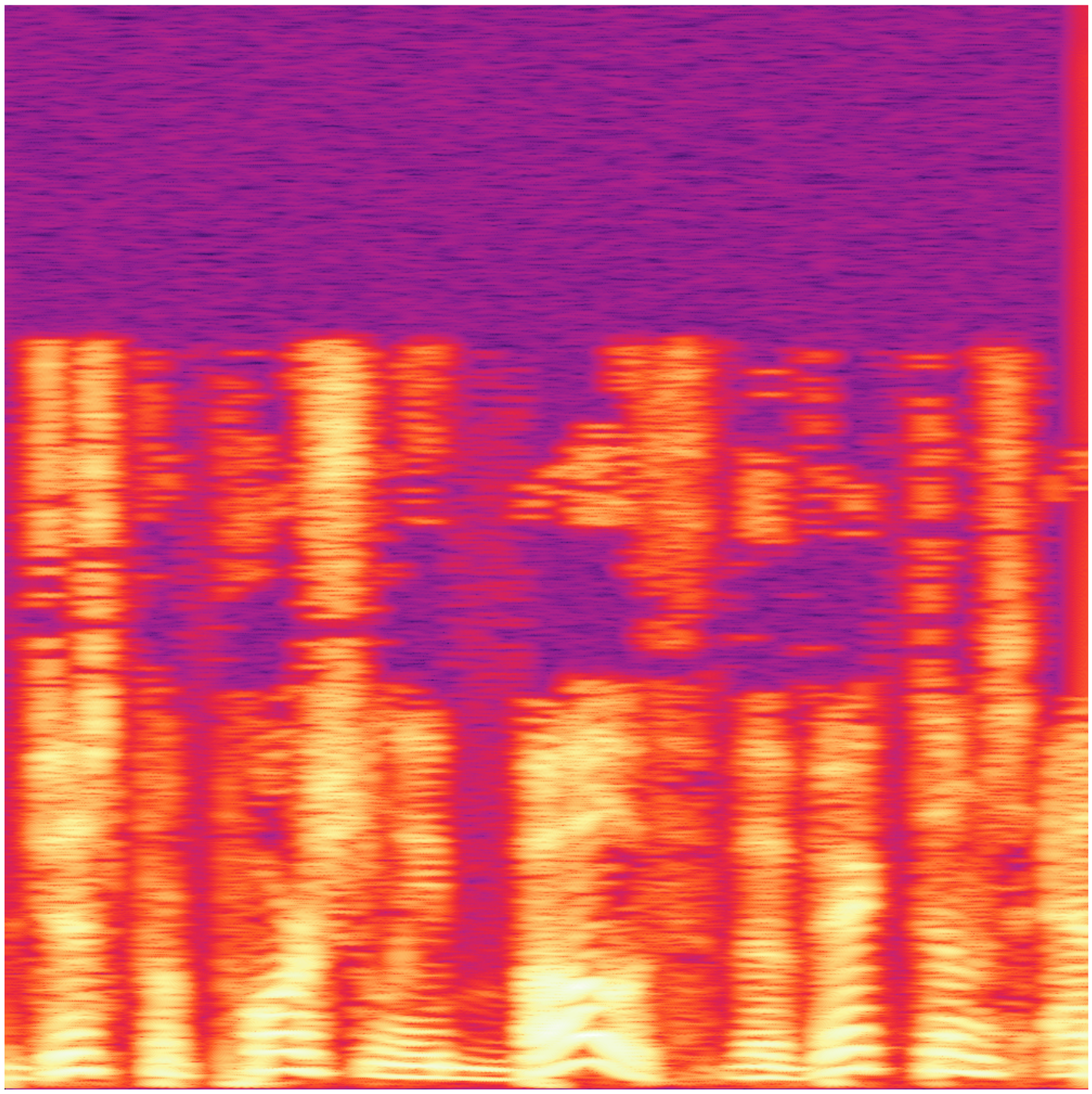
Now that I've covered the algorithm I'll be using, I'll explain how I'm going to explore it's plasticity. I wanted to switch the between different sensory modalities mid-training, so I prepared two datasets of identical size and dimensionality. One of images from flickr, and one of audio from NPR and several college radio stations.

Example images from MFLICKR-1M

Wednesday, October 16, 2013

The Image dataset is a collection of 1 million images from Flickr released under a Creative Commons license by their original authors. I am using a set of ten-thousand patches that I sampled from a subset of this dataset. I don't have enough GPUs to use the whole thing, but I'd love to give it a try eventually!

Example spectrogram from NPR 10 seconds

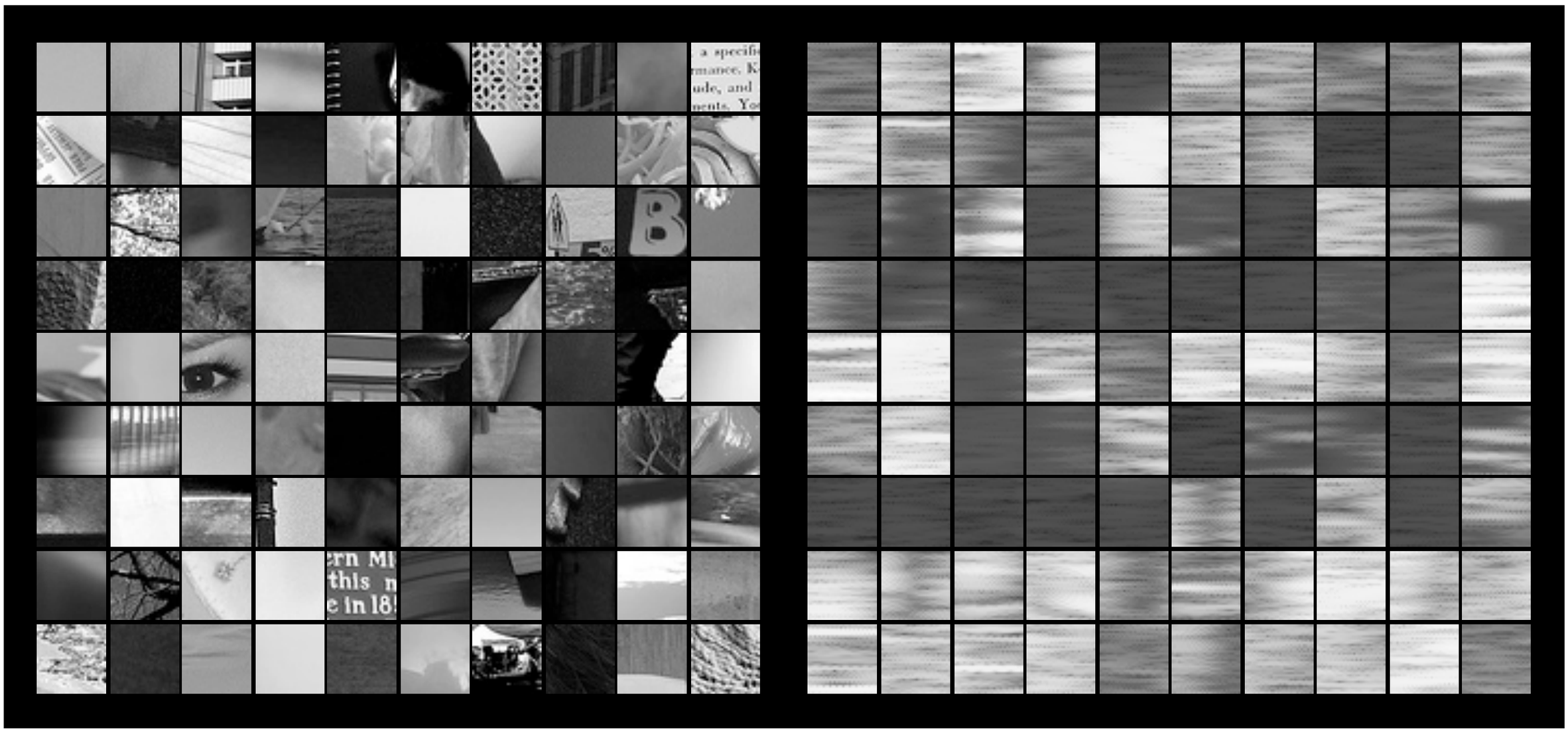


Wednesday, October 16, 2013

The audio comes from NPR talk shows, and a favorite college radio station of mine called WKNC. periodically, two-second audio clips are captured from these streams, and transformed into spectrograms like this using open source software available on Ubuntu. I collected a few thousand of these spectrograms, and then sampled patches from them, just like the images.

Image patch samples

Audio patch samples



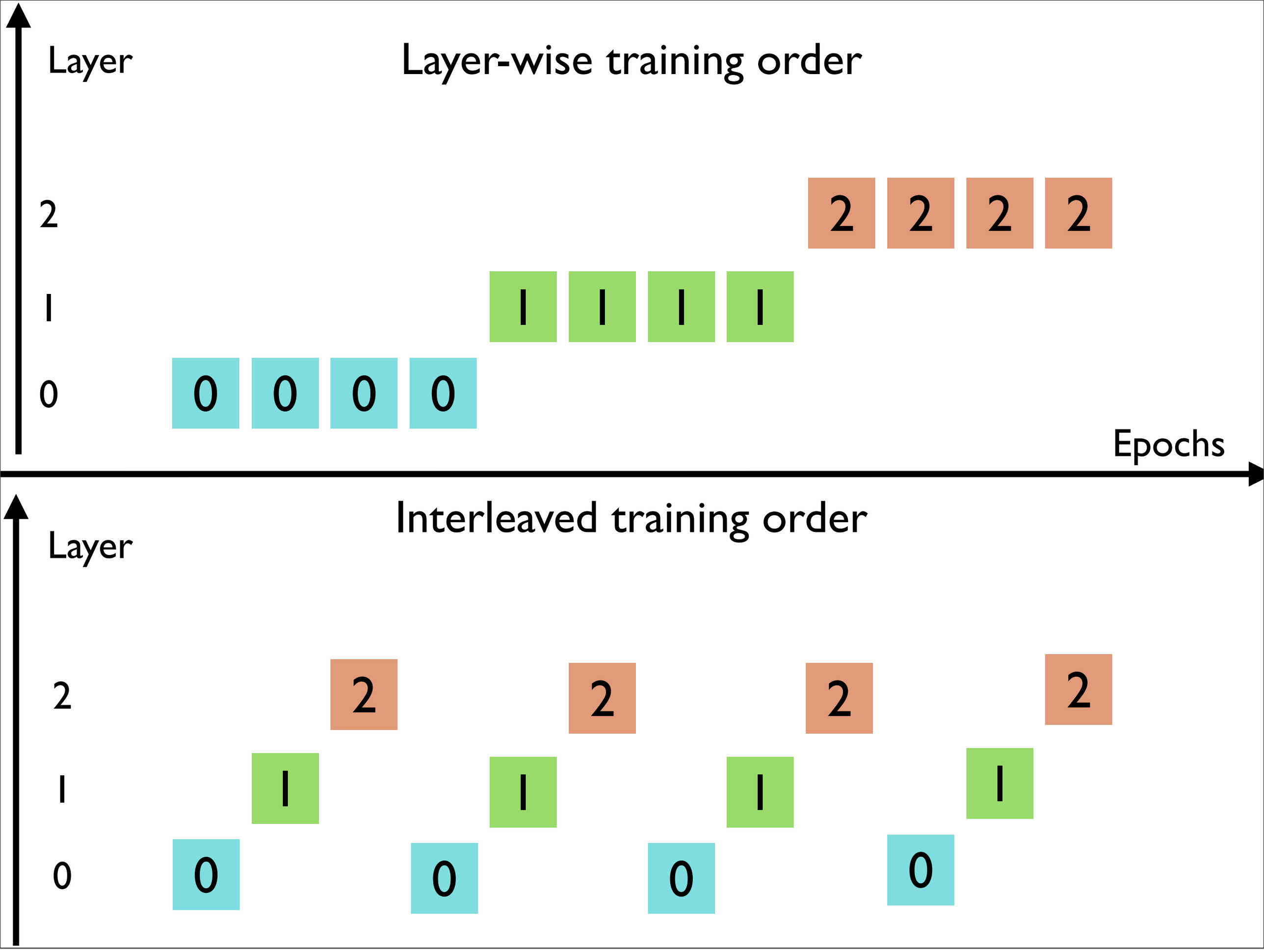
Wednesday, October 16, 2013

Here are some examples of those patches. They undergo one more step after this called whitening, which is effectively a form of local contrast enhancement.

Summary of Experiments	Interleaved or Layer-wise	Single or multi-modal	Audio or Images
Interleaved, single-modal, audio	0	0	0
Interleaved, single-modal, images	0	0	1
Interleaved, multi-modal, audio last	0	1	0
Interleaved, multi-modal, images last	0	1	1
Layer-wise, single-modal, audio	1	0	0
Layer-wise, single-modal, images	1	0	1
Layer-wise, multi-modal, audio last	1	1	0
Layer-wise, multi-modal, images last	1	1	1

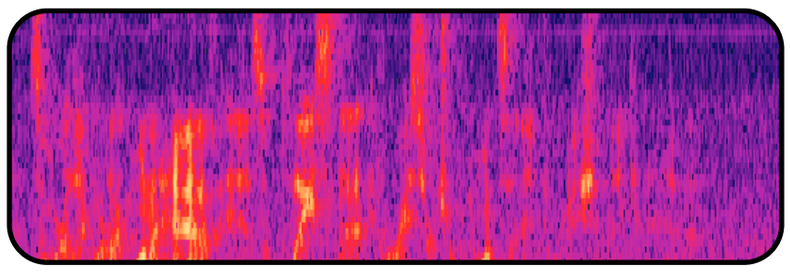
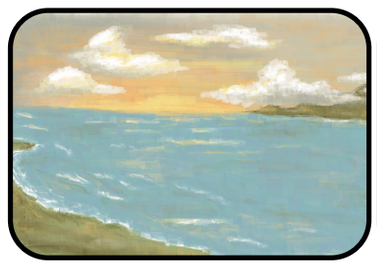
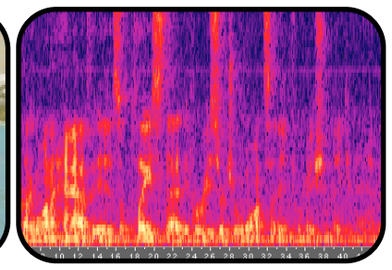
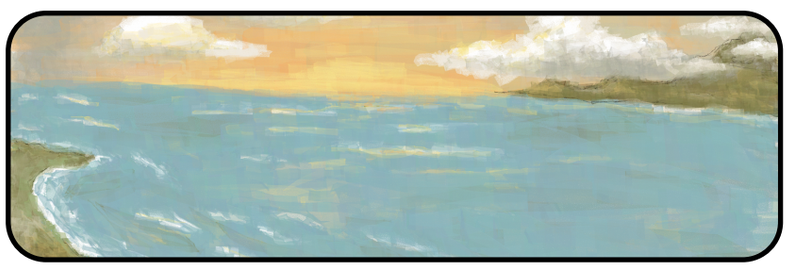
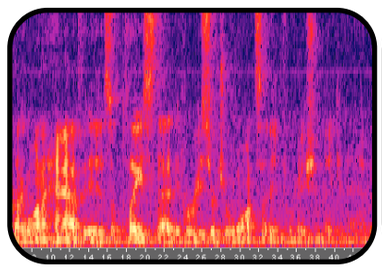
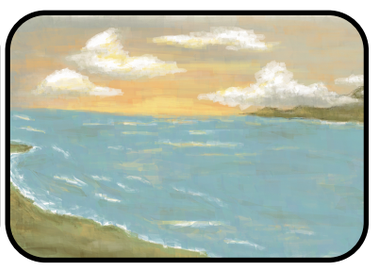
Wednesday, October 16, 2013

I set up 8 different experiments to fully explore the combinations of three different binary variables. These are: interleaved or layer-wise training, single-modal or multi-modal data sources, and audio or images as the test data. I'll explain each of these binary variables in the next two slides.



Wednesday, October 16, 2013

Layer-wise vs. Interleaved training is the difference between training the individual DAs in order, vs starting with all of them present and training them in a round-robin fashion. Since layer-wise training is known to help the higher layers of SDAs converge on better features, I hypothesized that it would improve adaptability in the case of multi-modal data as well.

	Single-modal	Multi-modal
Audio (last)	audio  1000 epochs	images  audio  500 epochs 500 epochs
Images (last)	images  1000 epochs	audio  images  500 epochs 500 epochs

Wednesday, October 16, 2013

Single-modal vs. multi-modal refers to whether I use the same dataset thru-out the experiment or swap out the dataset half-way. the term "modal" comes from the term "sensory modality". To control for the advantage given by training order, I run multi modal experiments with audio first, and images first. I run single-modal experiments with either just audio or just images.

Experiments outline

- 8 models trained & tested
- 20 random initializations
- 1000 training epochs per run
- 10000 data points viewed every epoch

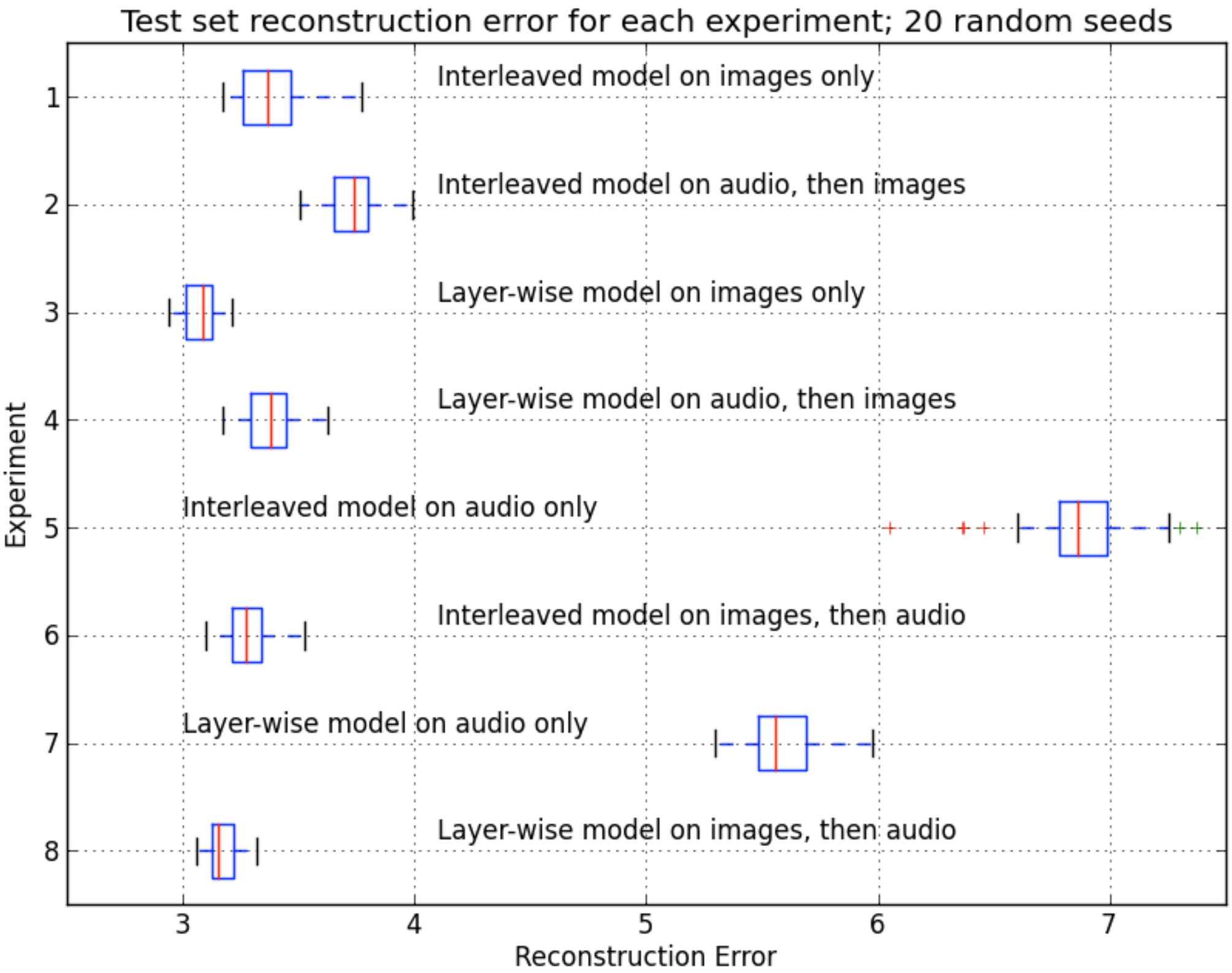
Approximate run time: 4 days at 1.4 GFLOPS

Data collected

Training error over time for each layer
Weights of finished models
Reconstruction error on test set for each model

Wednesday, October 16, 2013

Each of these 8 experiments is run 20 times, using random seeds, to obtain a distribution of results. I compare the reconstruction error on the test data from each dataset to observe the effect of multi-modal initialization, and layer-wise training

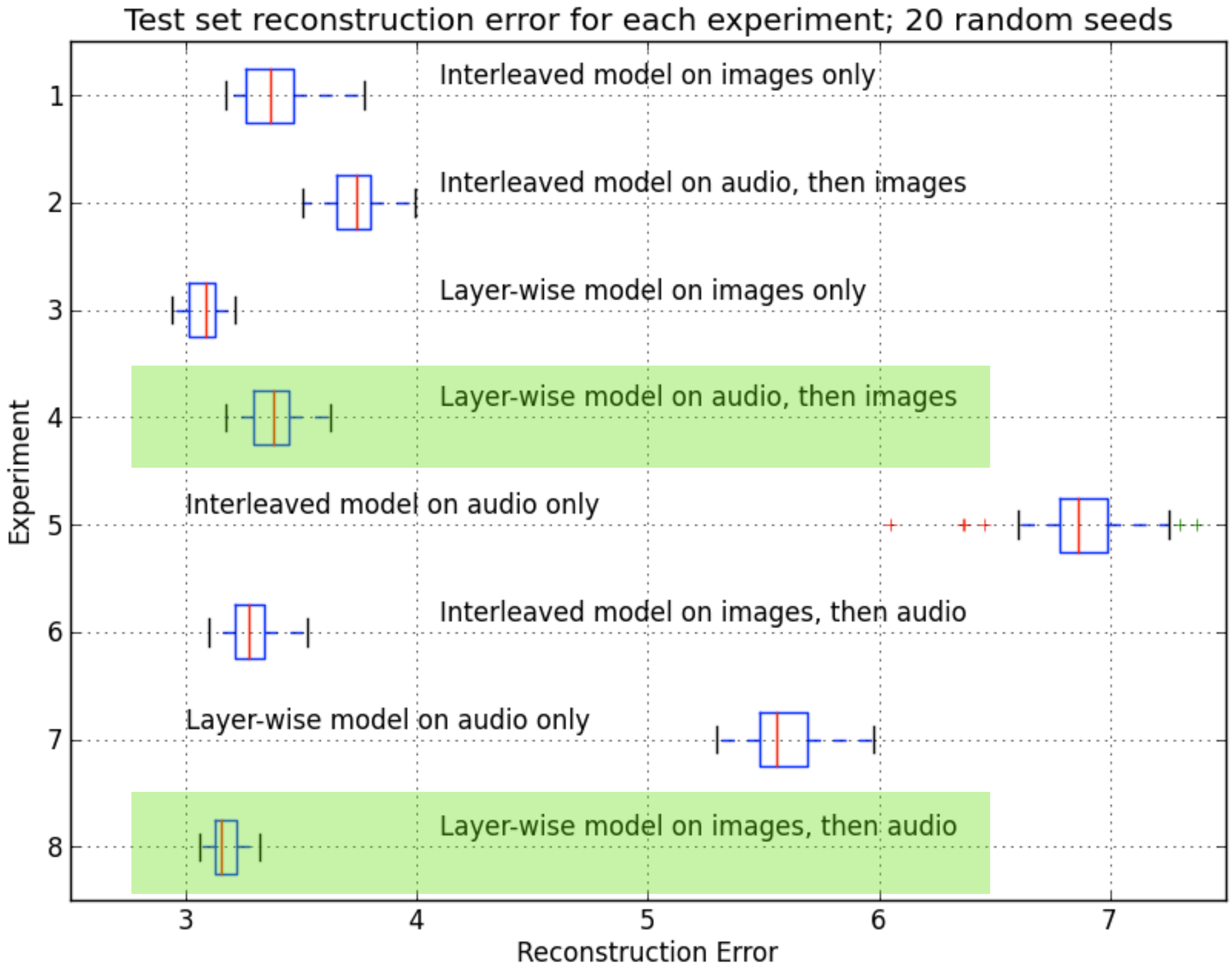


Wednesday, October 16, 2013

Here are the reconstruction error distributions for each experiment. Better performance is to the left. By comparing these values, we can learn whether layer-wise training, the supposed magic bullet of deep networks, has an effect on the plasticity of SDAs. In other words, do the layer-wise multi-modal experiments (highlight) have a lower reconstruction error than their interleaved counterparts. (highlight, indicate association) An improved (smaller) reconstruction error would indicate that the model is successfully adapting to the new sensory modality.

The answer is that yes, it did help a little bit for both modality orders. The distribution for experiment 4 lies entirely to the left of experiment 2, and the distribution for experiment 8 lies not entirely, but significantly to the left of experiment 6. (point) (clear)

There are some unexpected differences though. The error distributions for models trained on audio only are way over on the right. Something went wrong with those models, and in the next few slides I'll show some graphs that explain that a little better, but the truth is I don't know why they do that, all I can say is that it reliably happens every time and we're talking about the same code that worked fine on images. If the audio-only models were like the image-only models, we would expect them to be here. (highlight)

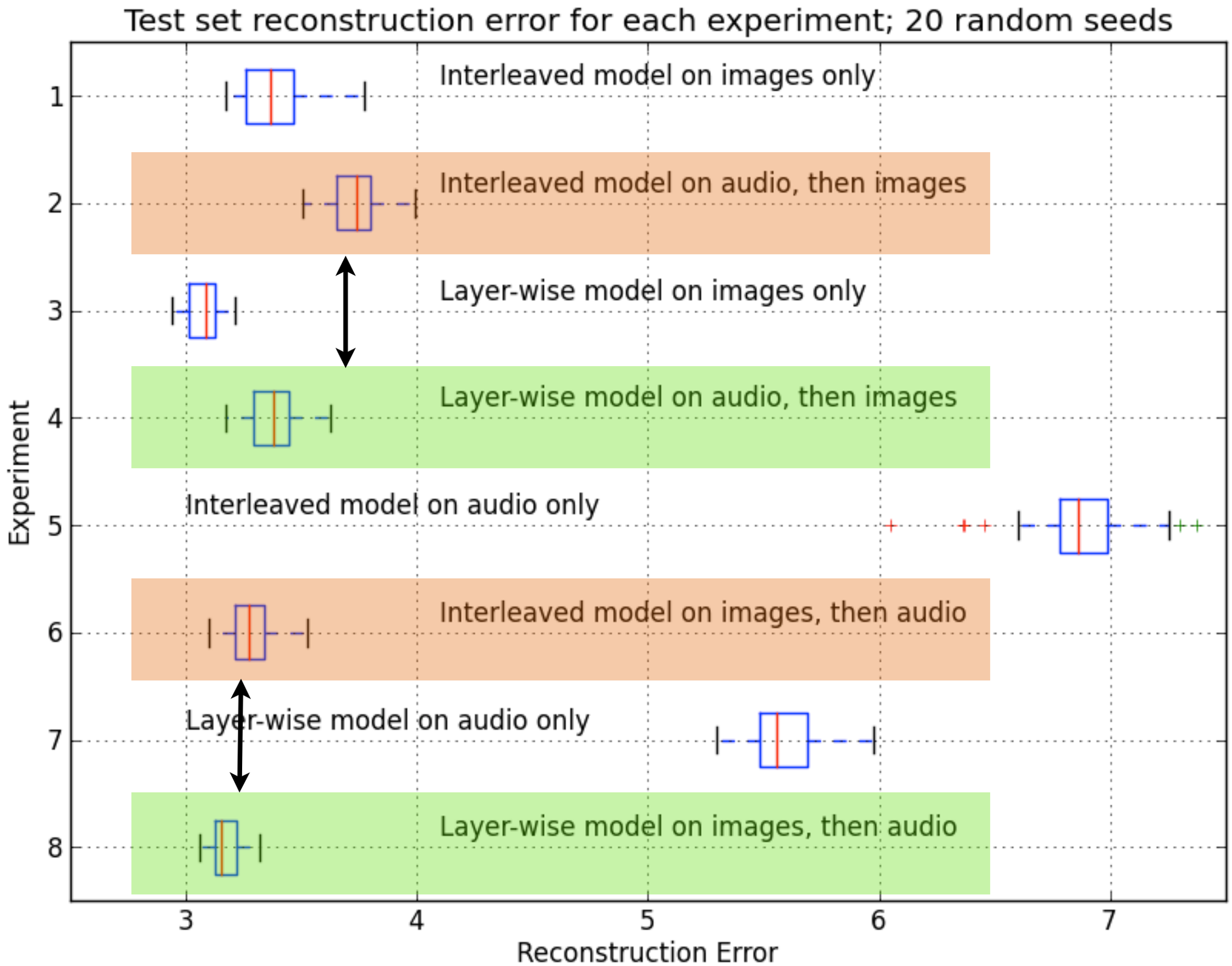


Wednesday, October 16, 2013

Here are the reconstruction error distributions for each experiment. Better performance is to the left. By comparing these values, we can learn whether layer-wise training, the supposed magic bullet of deep networks, has an effect on the plasticity of SDAs. In other words, do the layer-wise multi-modal experiments (highlight) have a lower reconstruction error than their interleaved counterparts. (highlight, indicate association) An improved (smaller) reconstruction error would indicate that the model is successfully adapting to the new sensory modality.

The answer is that yes, it did help a little bit for both modality orders. The distribution for experiment 4 lies entirely to the left of experiment 2, and the distribution for experiment 8 lies not entirely, but significantly to the left of experiment 6. (point) (clear)

There are some unexpected differences though. The error distributions for models trained on audio only are way over on the right. Something went wrong with those models, and in the next few slides I'll show some graphs that explain that a little better, but the truth is I don't know why they do that, all I can say is that it reliably happens every time and we're talking about the same code that worked fine on images. If the audio-only models were like the image-only models, we would expect them to be here. (highlight)

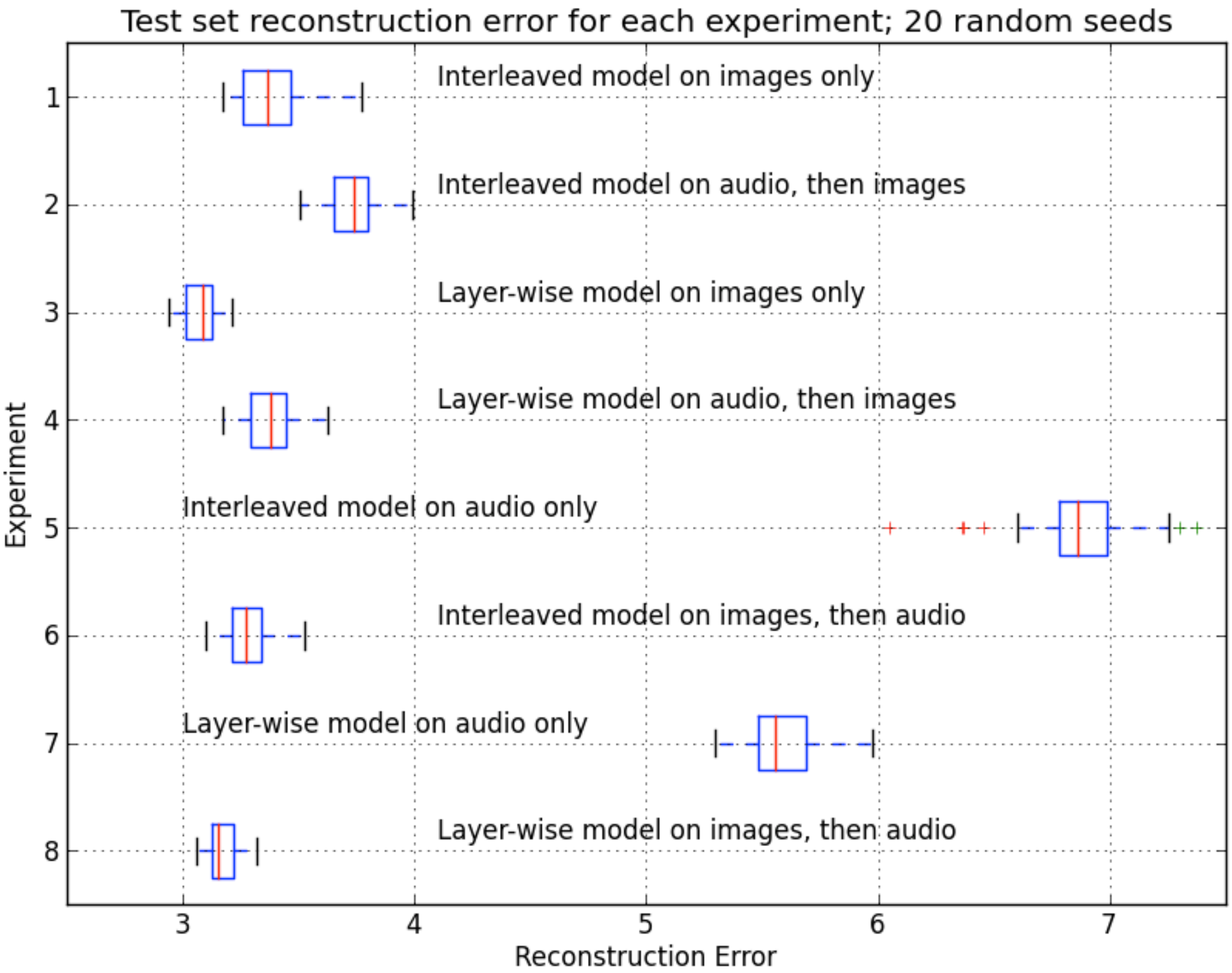


Wednesday, October 16, 2013

Here are the reconstruction error distributions for each experiment. Better performance is to the left. By comparing these values, we can learn whether layer-wise training, the supposed magic bullet of deep networks, has an effect on the plasticity of SDAs. In other words, do the layer-wise multi-modal experiments (highlight) have a lower reconstruction error than their interleaved counterparts. (highlight, indicate association) An improved (smaller) reconstruction error would indicate that the model is successfully adapting to the new sensory modality.

The answer is that yes, it did help a little bit for both modality orders. The distribution for experiment 4 lies entirely to the left of experiment 2, and the distribution for experiment 8 lies not entirely, but significantly to the left of experiment 6. (point) (clear)

There are some unexpected differences though. The error distributions for models trained on audio only are way over on the right. Something went wrong with those models, and in the next few slides I'll show some graphs that explain that a little better, but the truth is I don't know why they do that, all I can say is that it reliably happens every time and we're talking about the same code that worked fine on images. If the audio-only models were like the image-only models, we would expect them to be here. (highlight)

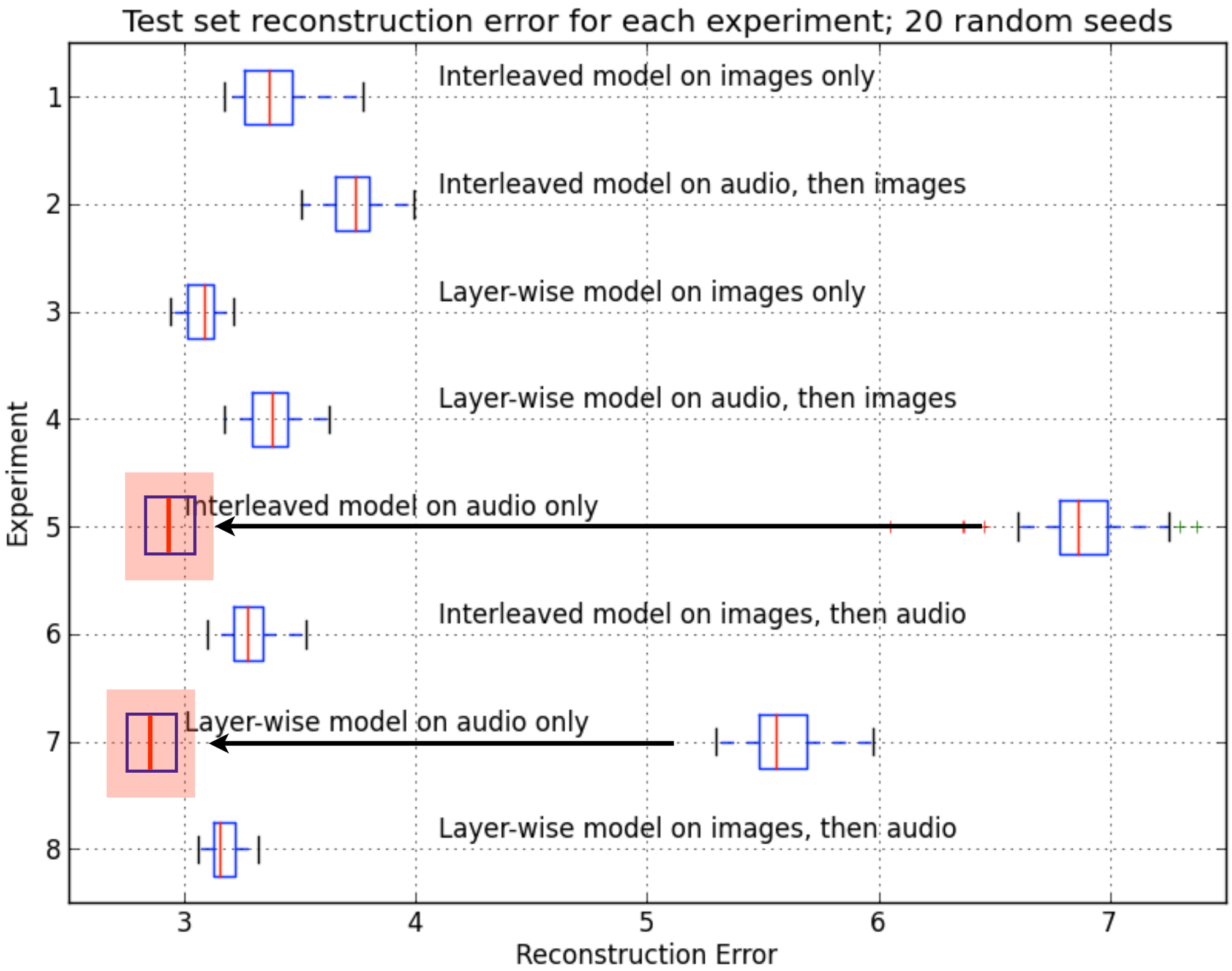


Wednesday, October 16, 2013

Here are the reconstruction error distributions for each experiment. Better performance is to the left. By comparing these values, we can learn whether layer-wise training, the supposed magic bullet of deep networks, has an effect on the plasticity of SDAs. In other words, do the layer-wise multi-modal experiments (highlight) have a lower reconstruction error than their interleaved counterparts. (highlight, indicate association) An improved (smaller) reconstruction error would indicate that the model is successfully adapting to the new sensory modality.

The answer is that yes, it did help a little bit for both modality orders. The distribution for experiment 4 lies entirely to the left of experiment 2, and the distribution for experiment 8 lies not entirely, but significantly to the left of experiment 6. (point) (clear)

There are some unexpected differences though. The error distributions for models trained on audio only are way over on the right. Something went wrong with those models, and in the next few slides I'll show some graphs that explain that a little better, but the truth is I don't know why they do that, all I can say is that it reliably happens every time and we're talking about the same code that worked fine on images. If the audio-only models were like the image-only models, we would expect them to be here. (highlight)



Wednesday, October 16, 2013

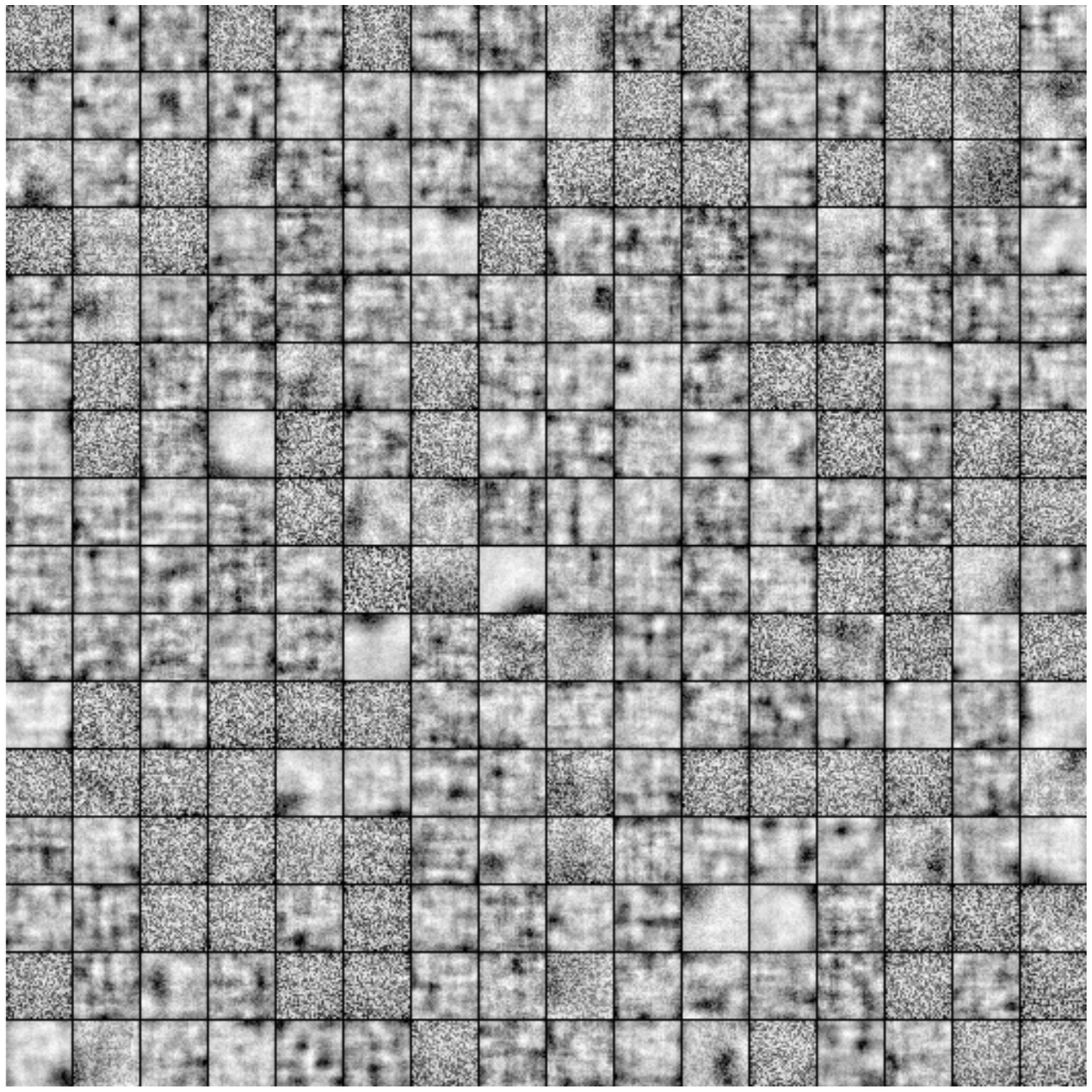
Here are the reconstruction error distributions for each experiment. Better performance is to the left. By comparing these values, we can learn whether layer-wise training, the supposed magic bullet of deep networks, has an effect on the plasticity of SDAs. In other words, do the layer-wise multi-modal experiments (highlight) have a lower reconstruction error than their interleaved counterparts. (highlight, indicate association) An improved (smaller) reconstruction error would indicate that the model is successfully adapting to the new sensory modality.

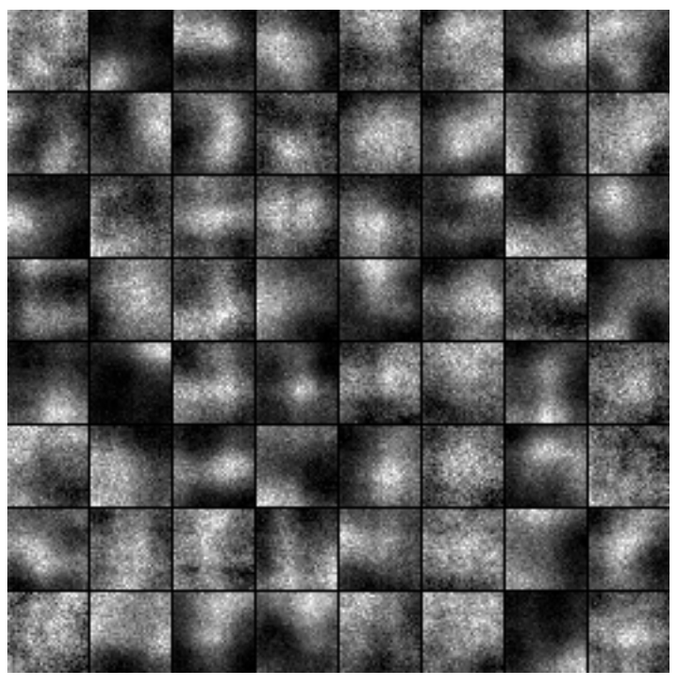
The answer is that yes, it did help a little bit for both modality orders. The distribution for experiment 4 lies entirely to the left of experiment 2, and the distribution for experiment 8 lies not entirely, but significantly to the left of experiment 6. (point) (clear)

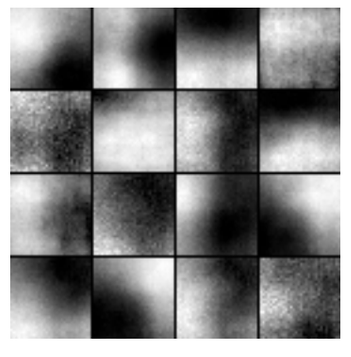
There are some unexpected differences though. The error distributions for models trained on audio only are way over on the right. Something went wrong with those models, and in the next few slides I'll show some graphs that explain that a little better, but the truth is I don't know why they do that, all I can say is that it reliably happens every time and we're talking about the same code that worked fine on images. If the audio-only models were like the image-only models, we would expect them to be here. (highlight)

Interleaved
Images only

Reconstructed
features for each
layer

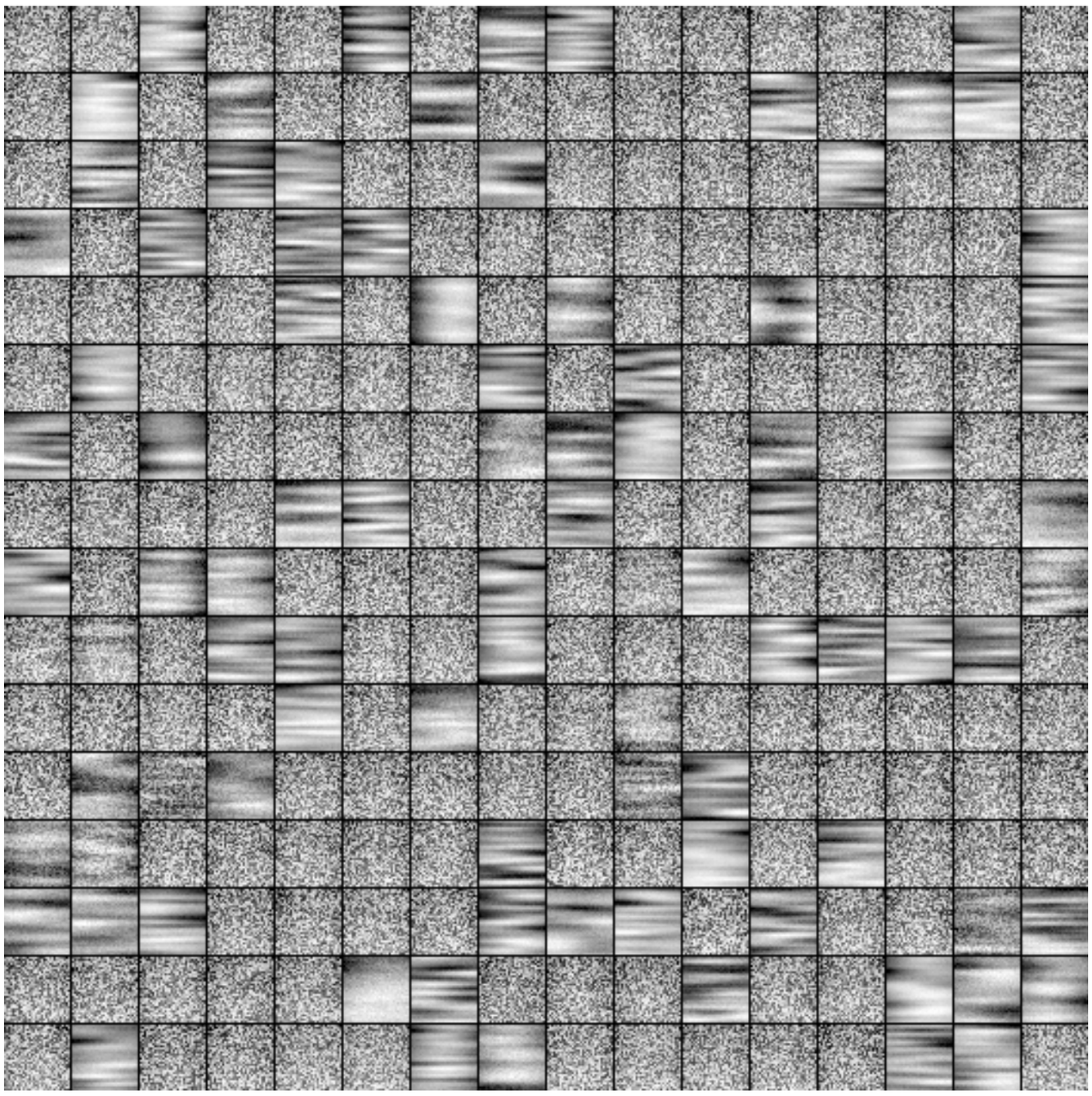






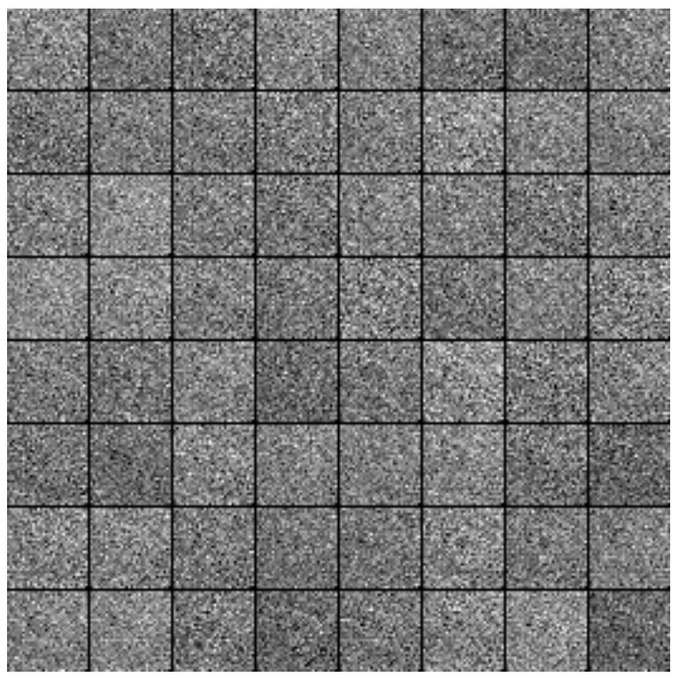
Wednesday, October 16, 2013

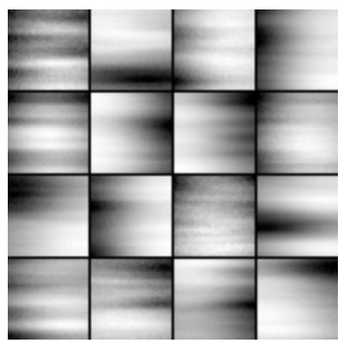
Another way to get an idea of what a network has learned is to visualize the features in each layer. These are the hypothetical input examples that would maximally activate each node. This gives an approximation of what the node is sensitive to. In this visualization of the interleaved image-only model, the 1st layer's 256 hidden nodes are on the left, the second layer's 64 hidden nodes are in the middle, and the third layer's 16 hidden nodes are at the bottom right. Note what fraction of the available nodes have converged on a smooth, spatially correlated pattern, vs. nodes which still display uniformly distributed noise. Also, note how the smoothness, the size, and the redundancy of the features changes from layer one to three.



Interleaved
Audio only

Reconstructed
features for each
layer

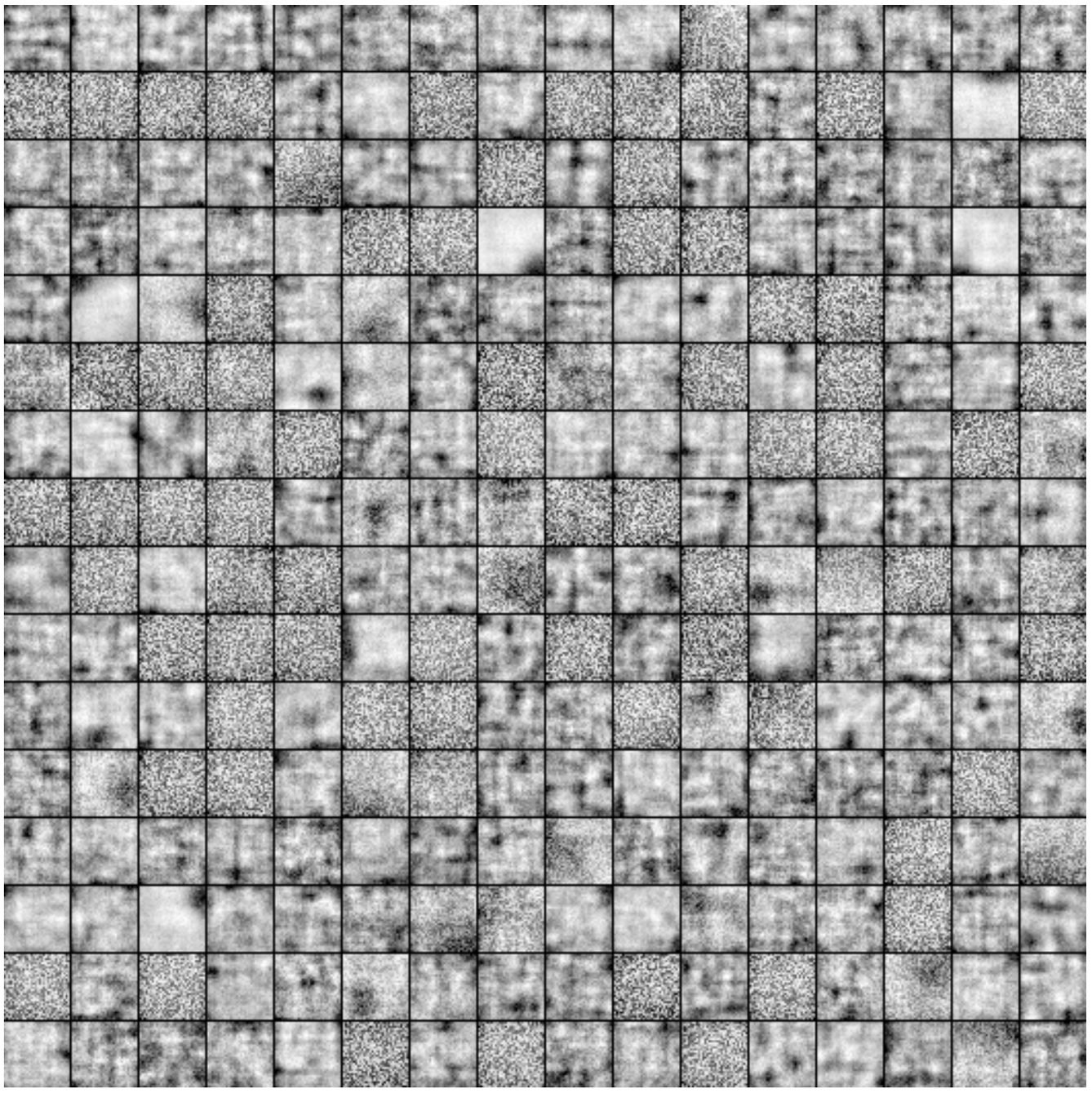




Wednesday, October 16, 2013

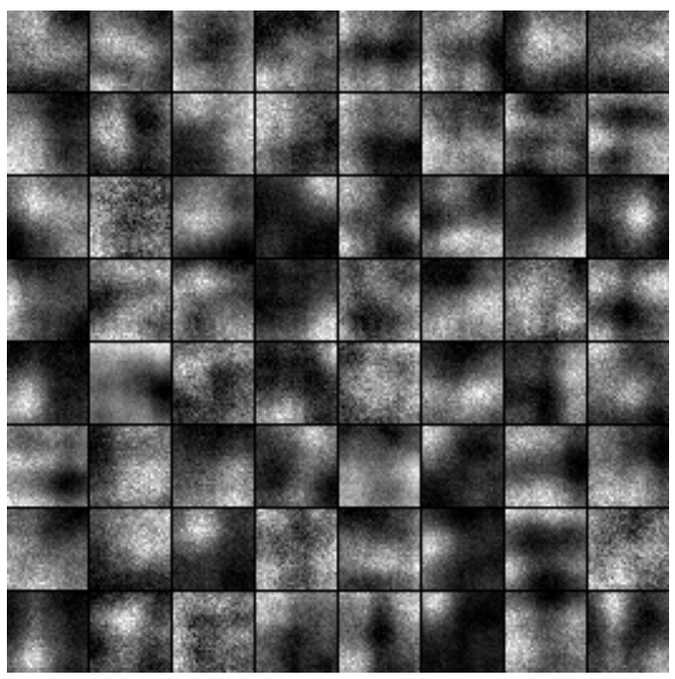
In the next model, interleaved audio-only, we get a clue as to why the reconstruction error for audio only models was worse than expected. Layer two has not converged on spatially correlated patterns. But even more strangely, layer three has. At first I looked at this and thought, "that's impossible, something is broken." But there is nothing in the cost function of the interleaved training method that demands spatially correlated patterns in individual layers (only layer-wise training demands that) the interleaved training method only asks for a decent reconstruction from the whole network.

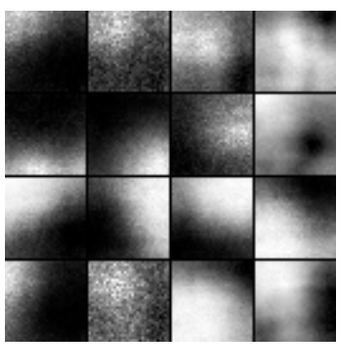
Additionally, this type of visualization has a shortcoming that may explain why nodes may not appear coherent. It does not give the whole story of what a node is coding for, only what the hypothetical input that would produce an activation of 1 for that node and an activation of 0 for all others in the same layer, a condition that is unlikely to ever happen in practice.



Layer-wise
Images only

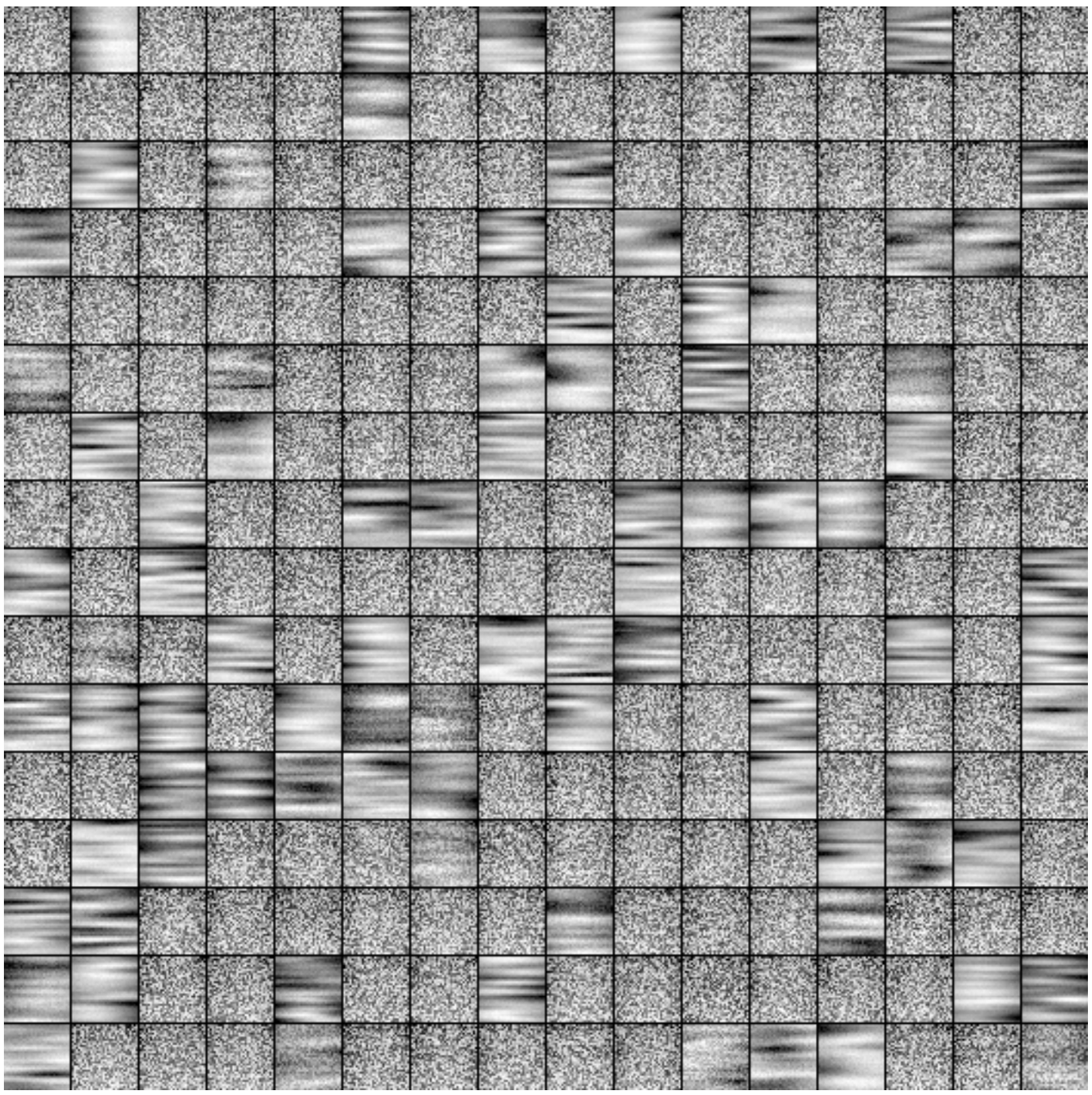
Reconstructed
features for each
layer





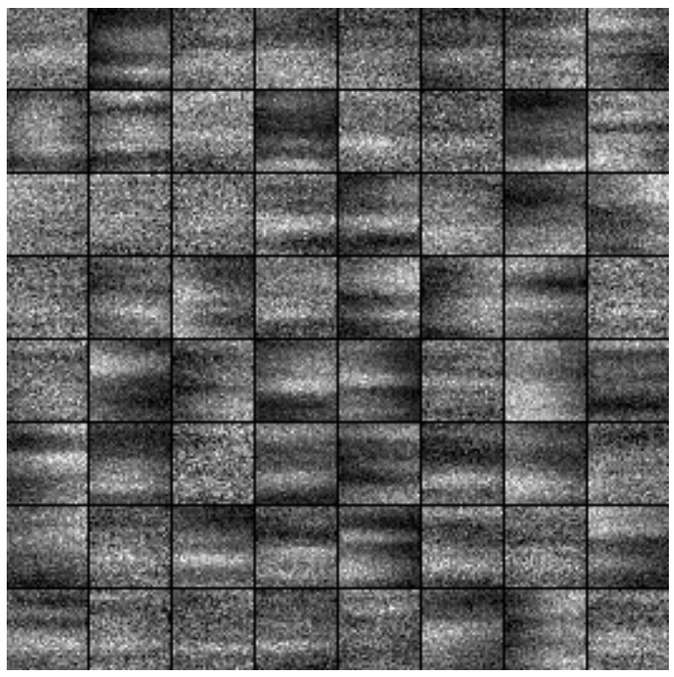
Wednesday, October 16, 2013

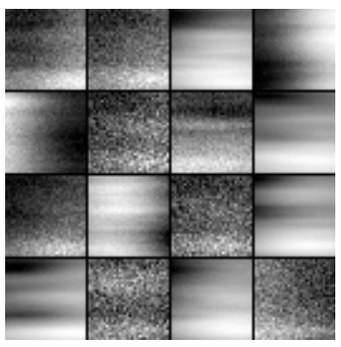
In the next two slides we can see layer-wise pre-training doing its job: enforcing every layer to converge on a sparse distributed representation of the one below. The resulting features are more spatially correlated, and allow the network to generalize better and thus achieve better reconstruction error on the test set.



Layer-wise
Audio only

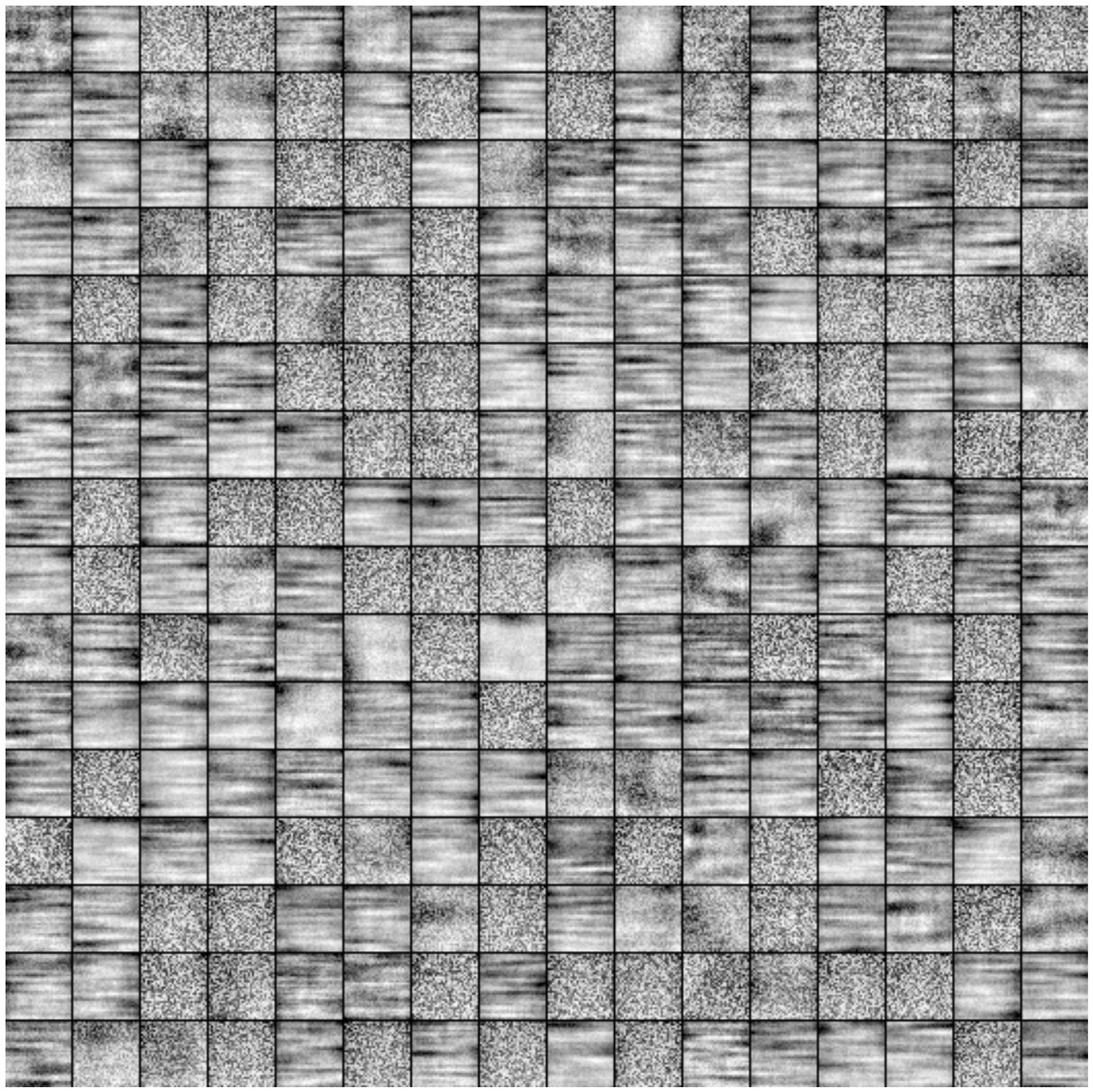
Reconstructed features for each layer





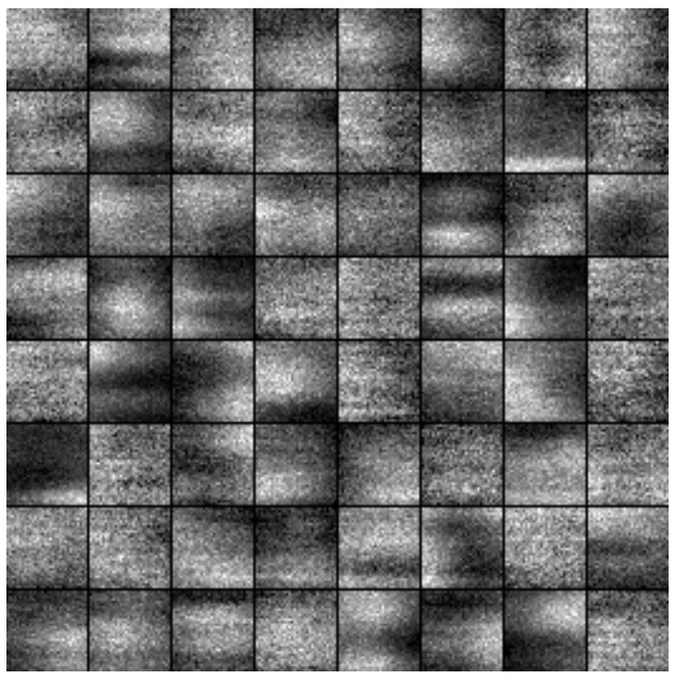
Wednesday, October 16, 2013

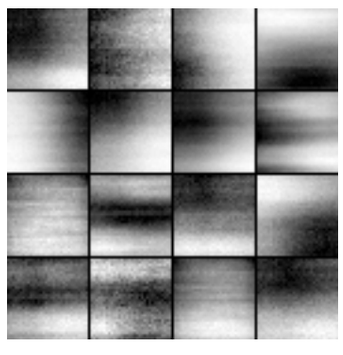
Back on the audio dataset, layer-wise pre-training has helped with the lack of spatially correlated patterns in layer two, but it still displays a strange tendency for that second layer to be noisier than usual.



Interleaved
Images, then Audio

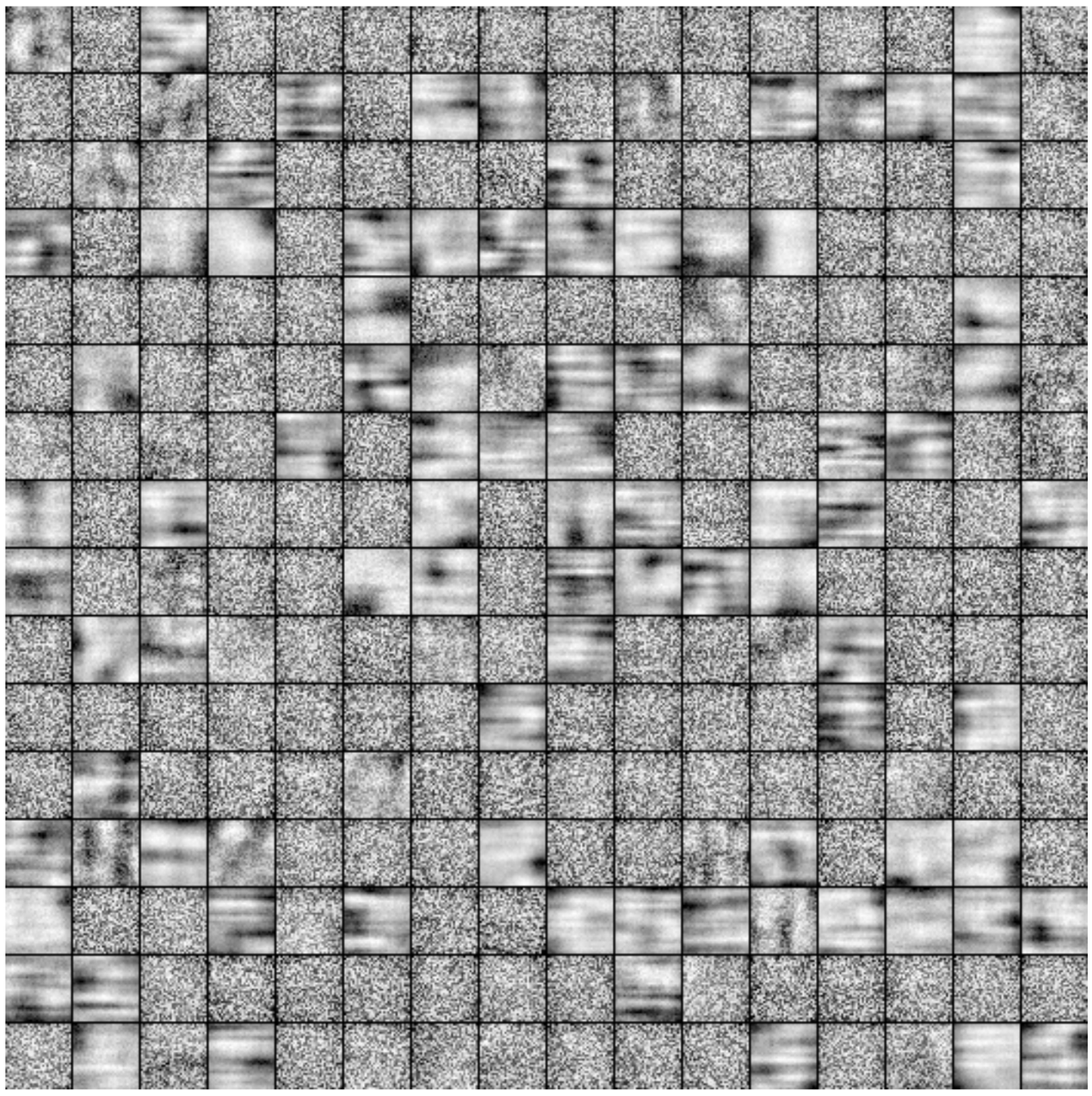
Reconstructed
features for each
layer





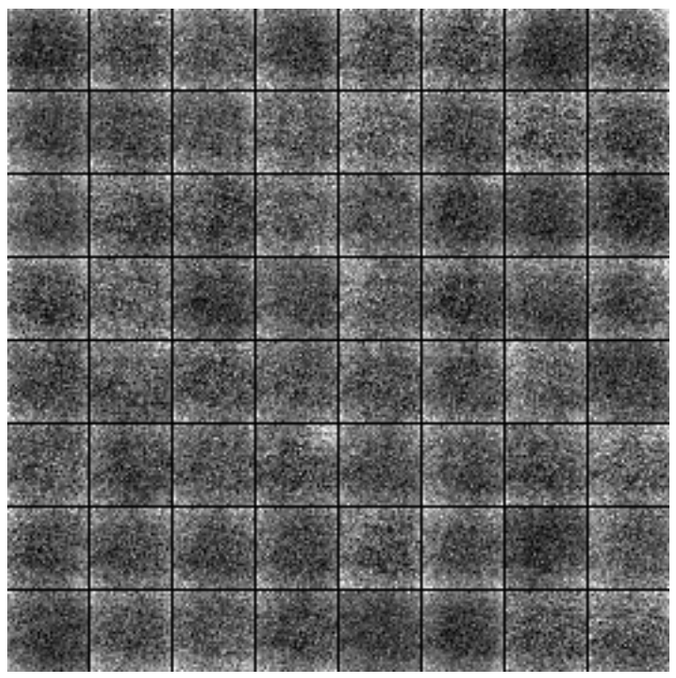
Wednesday, October 16, 2013

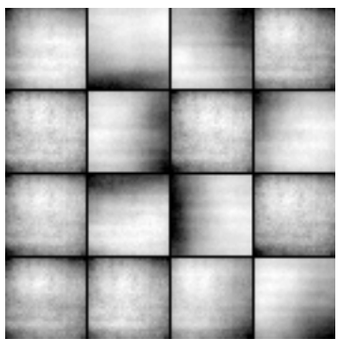
Next we have the four multi-modal models that were trained on one dataset, and switched to another one halfway through. This is the Interleaved model trained on images, then switched to audio. It retains a high ratio of smooth nodes to noisy nodes, and this probably contributes to it's higher performance. Note that the features have begun to take on the appearance of the spectrogram patches.



Interleaved
Audio, then Images

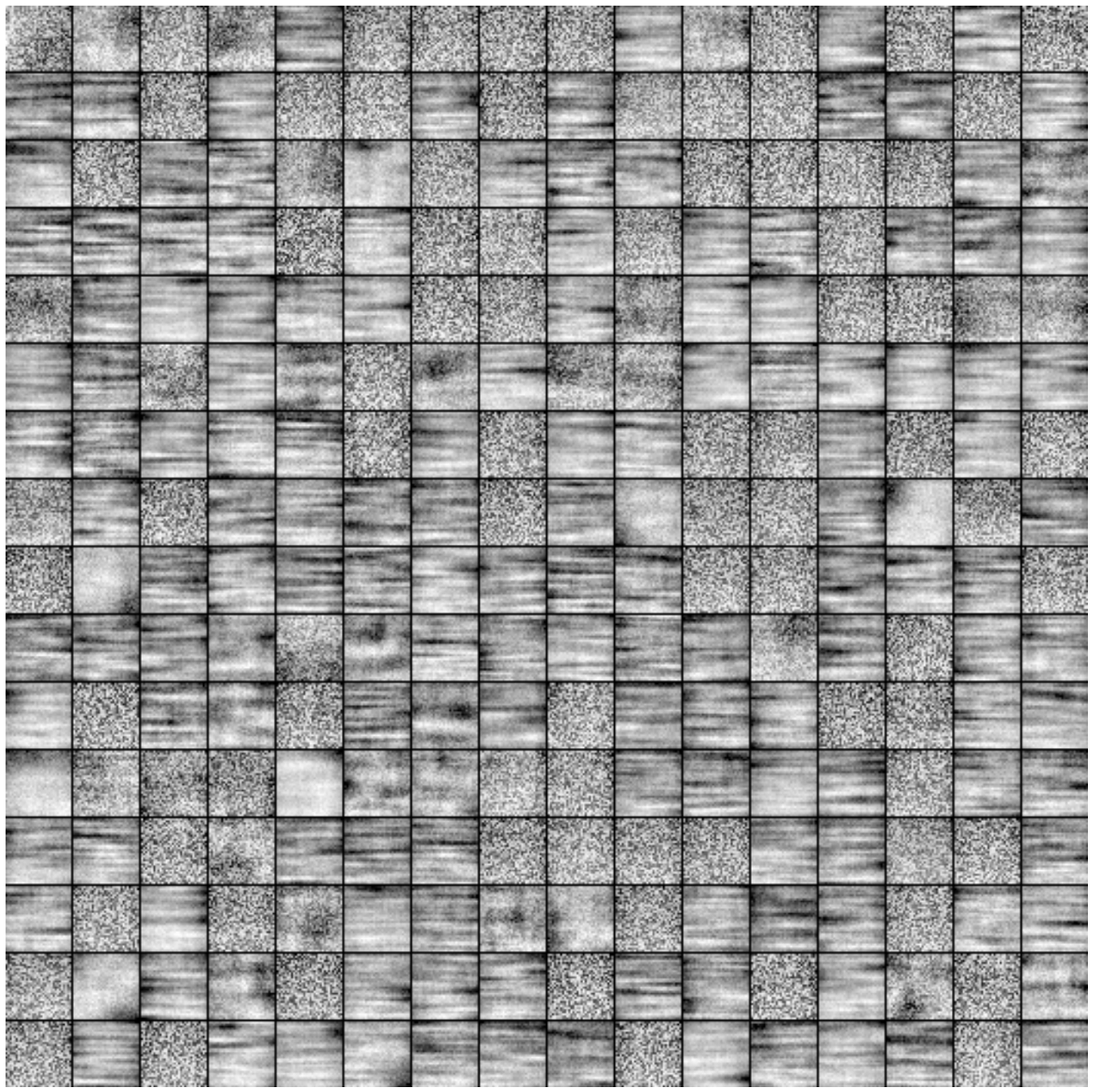
Reconstructed
features for each
layer





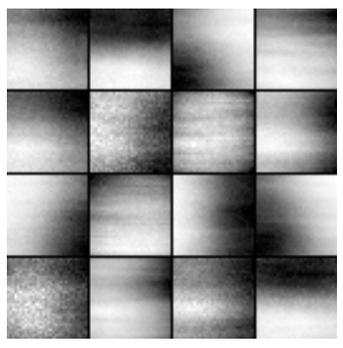
Wednesday, October 16, 2013

Here we have the interleaved model that was trained on audio, then switched to images. It also retains the ratio of smooth to noisy nodes that it originally had, and it has mostly retained the lack of spatially correlated second layer features that afflicts the interleaved audio only model, except that they have all begun to start moving towards the same "vignette" appearance.



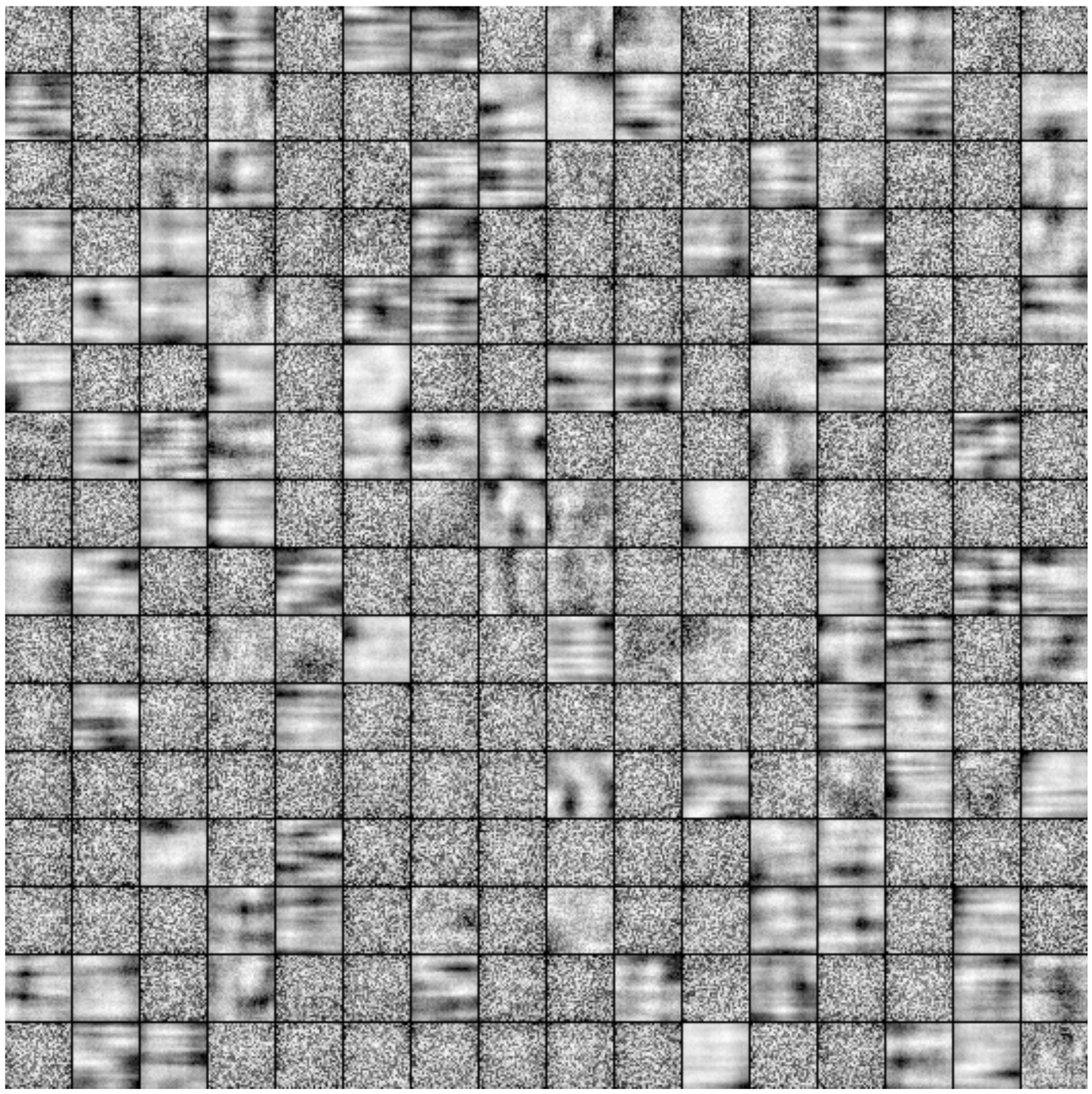
Layer-wise Images, then Audio

Reconstructed features for each layer



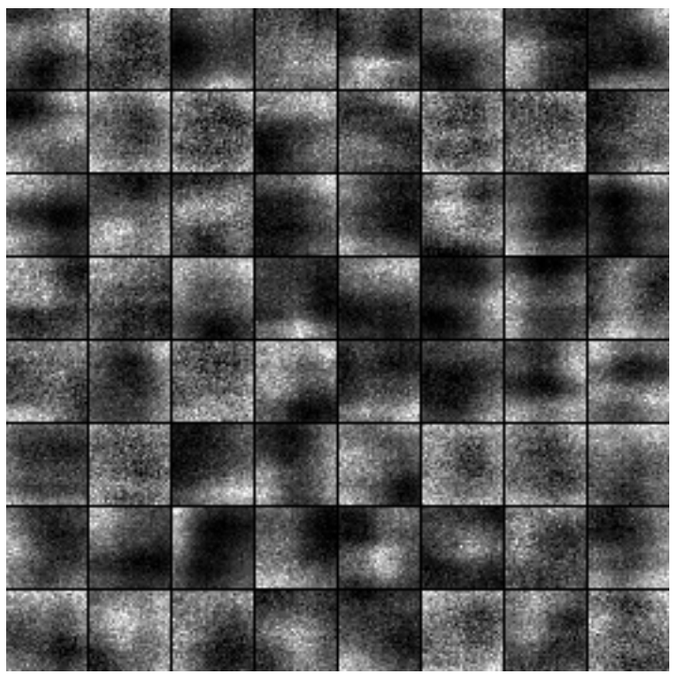
Wednesday, October 16, 2013

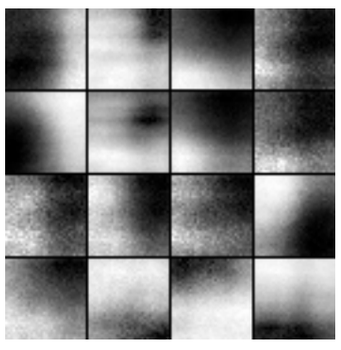
The next model show the layer-wise model trained on images, then audio. It is very similar to the interleaved model two slides back. It has a bit higher performance, but the improvement is not visually apparent just by looking at the features.



Layer-wise
Audio, then Images

Reconstructed
features for each
layer





Wednesday, October 16, 2013

However, over on the layer-wise audio then images model, we can see that the layer-wise training has helped the second and third layers start to converge on sparse distributed representations, which is a healthier thing to see. The performance of this model is better than it's interleaved counterpart which is what I would expect just looking at these features.

The effect of multi-modal training on reconstruction

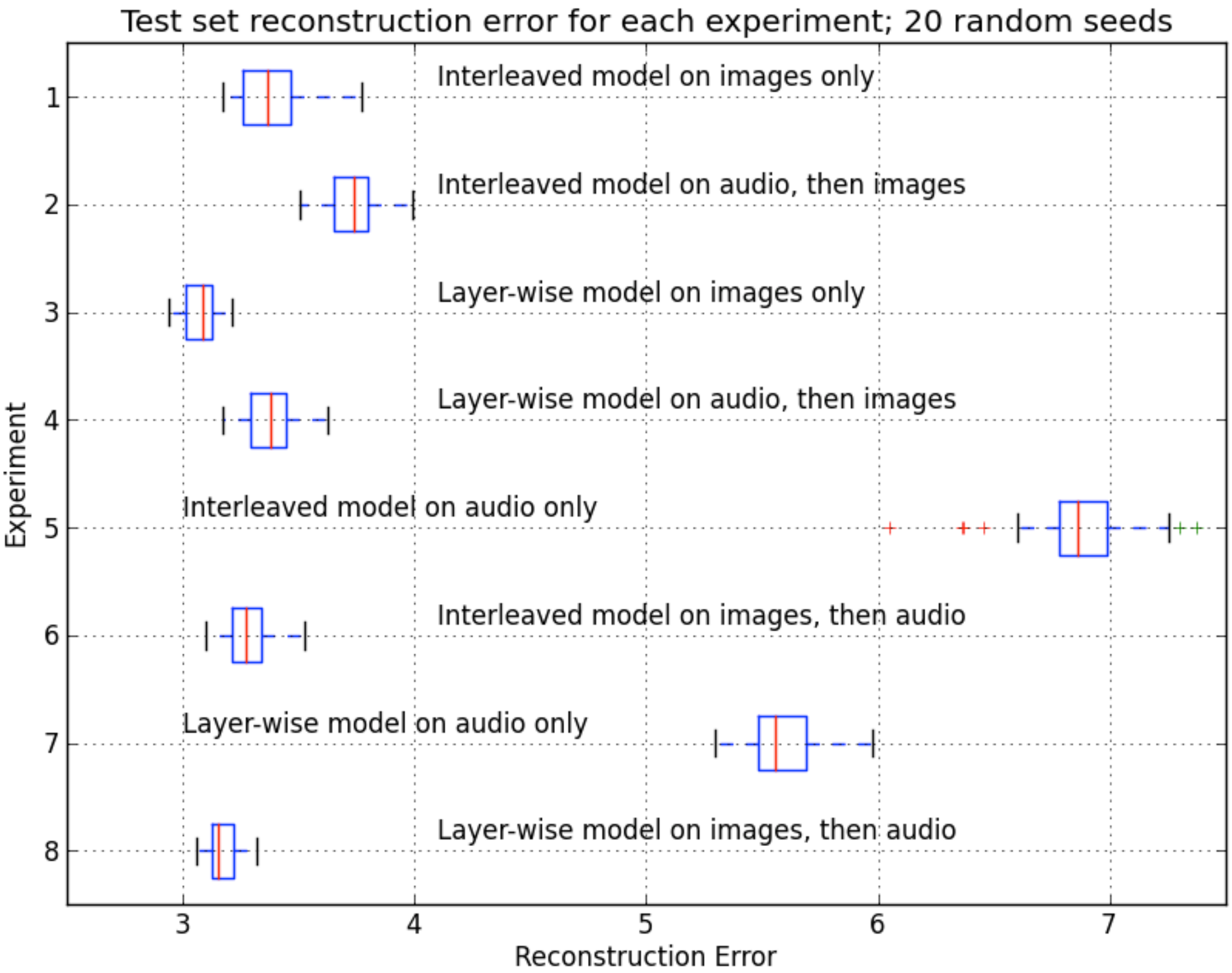
	Experiment Pair (single-modal, multi-modal)	Difference in mean reconstruction error	Difference (effect of GLWPT)
Interleaved, Tested on images	(1, 2)	-0.36072 (p-value <.0001)	-0.06051 (p-value <.0001)
Layer-wise, Tested on images	(3, 4)	-0.30021 (p-value <.0001)	
Interleaved, Tested on audio	(5, 6)	3.44677 (p-value <.0001)	1.05709 (p-value <.0001)
Layer-wise, Tested on audio	(7, 8)	2.38968 (p-value <.0001)	

Wednesday, October 16, 2013

At the outset of the experiment, I made two distinct hypotheses.

First, SDAs trained on one sensory modality and then switched to a second sensory modality half-way through training will have a higher reconstruction accuracy on test data from the second sensory modality than randomly initialized networks trained for an equal number of epochs. This hypothesis was not confirmed in all cases. (some of the differences in mean reconstruction error here are positive (green) and some are negative (red). This means that multi-modal initialization only helped in the case of images first, audio second.

The second hypothesis is that the aforementioned improvement would be amplified by greedy layer-wise pre-training since it is known to help networks discover better abstract features. On the right, I have computed the difference in the improvement due to layer-wise pre-training. Both of these numbers are expected to be negative.



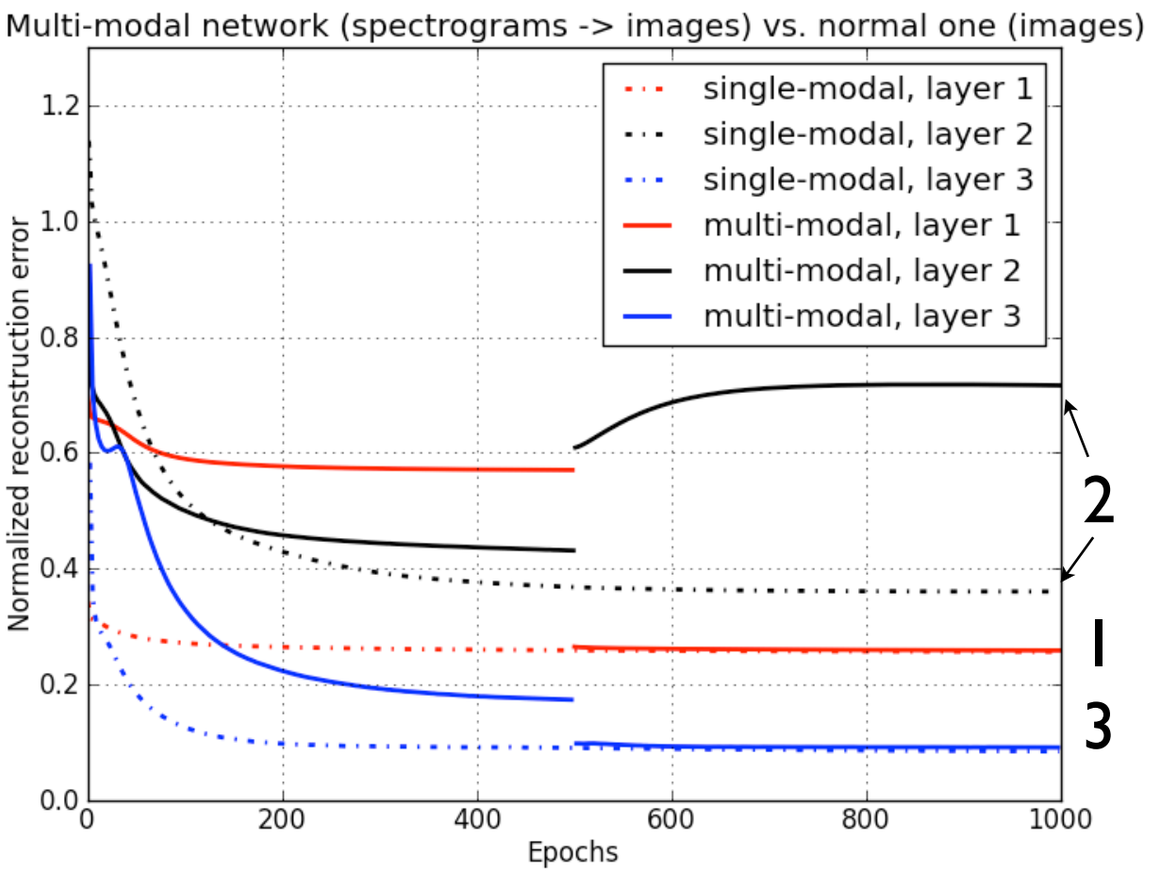
Wednesday, October 16, 2013

Neither of those hypothesis were confirmed, but as you may recall from the reconstruction error distributions, layer-wise pre-training did help multi-modal networks generalize to new data, which is a useful finding.

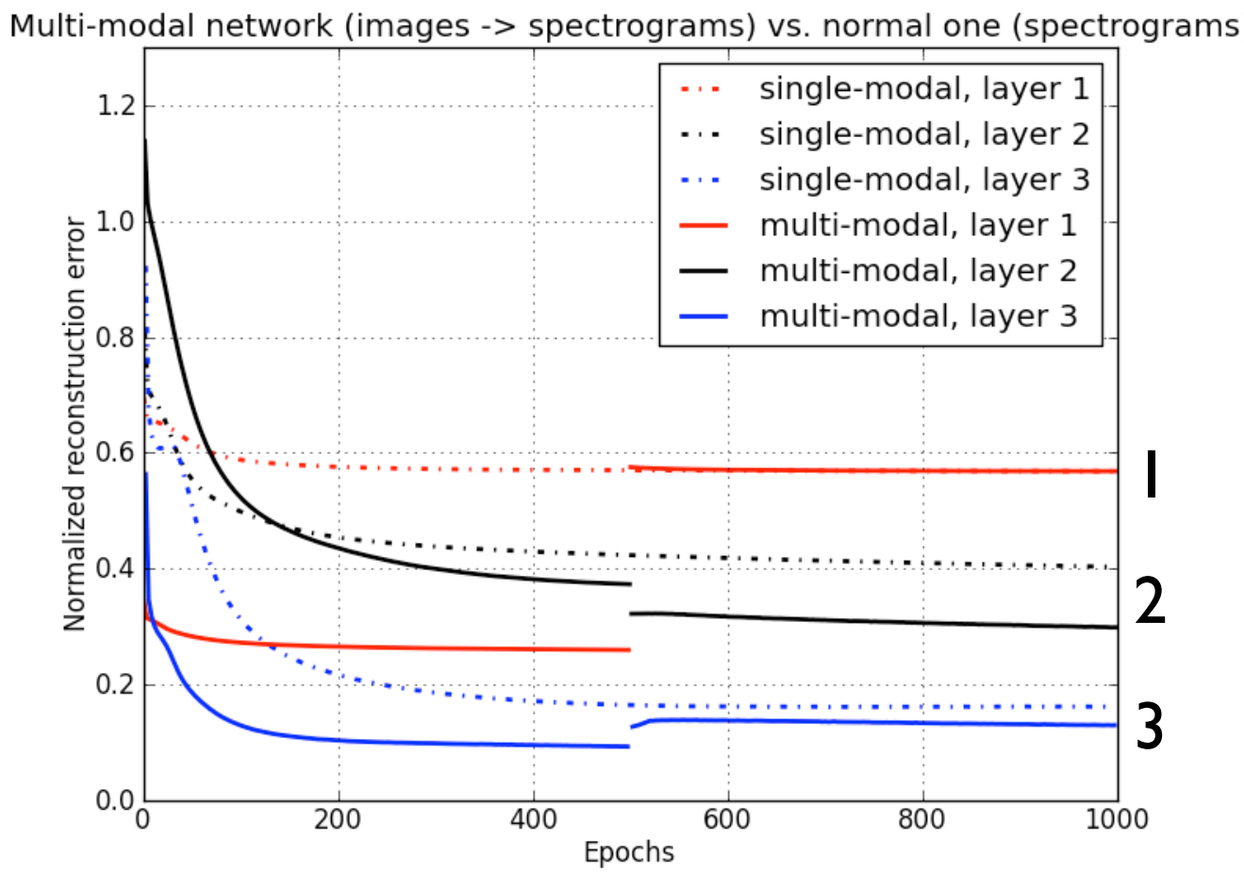
In conclusion of the results, I found SDA's display some amount of plasticity through generalization, and that it is improved by layer-wise pre-training. Unexpectedly, I found that my datasets were not as interchangeable as I thought, and that modality-order played a huge role. Natural images serve to create an excellent prior distribution for learning on spectrograms, but the reverse is not true. Perhaps the spectrograms could have been pre-processed differently to mitigate this, but that remains untested.

Training error over time

Audio first



Images first



Wednesday, October 16, 2013

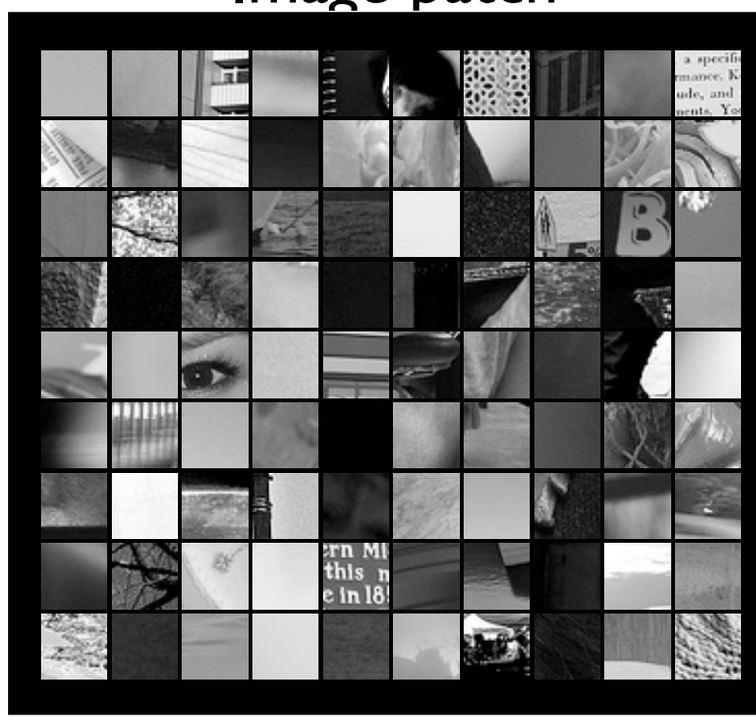
To show the effect of modality order another way, I have prepared graphs of the training error of each layer over time. All networks displayed here use interleaved layer training order. Multi-modal networks are displayed in solid lines, and single-modal networks in dotted lines. In both graphs a multi modal network is compared to a single modal network tested on the same type of data, or in other words, the second dataset of the multi-modal network is the only dataset used by the corresponding single-modal network.

The only difference between the two graphs is modality order. On the left we have audio first, and on the right, images first. The discontinuity in the solid lines in the middle of each graph is the point where the dataset is switched.

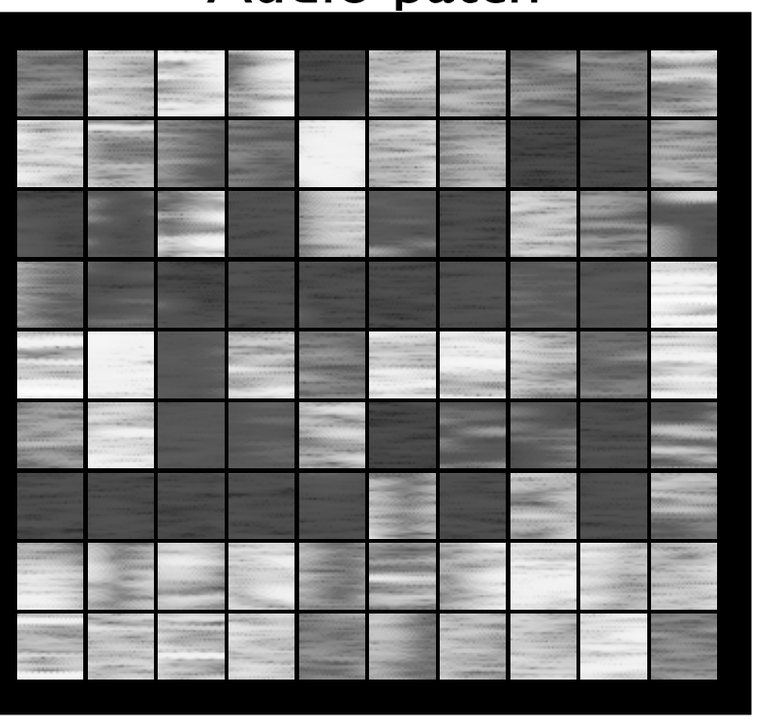
The first interesting point is that any time a model is training on audio, the reconstruction error decreases from layer 1 to layer 3 (red, black, blue from top to bottom), while models training on images are in a different order, the first layer has a lower reconstruction error than the second layer (black, red, blue from top to bottom)

At the moment the dataset is switched, the reconstruction errors for each layer immediately jump to roughly the value that they would have if the network had been training on that type

Image patch

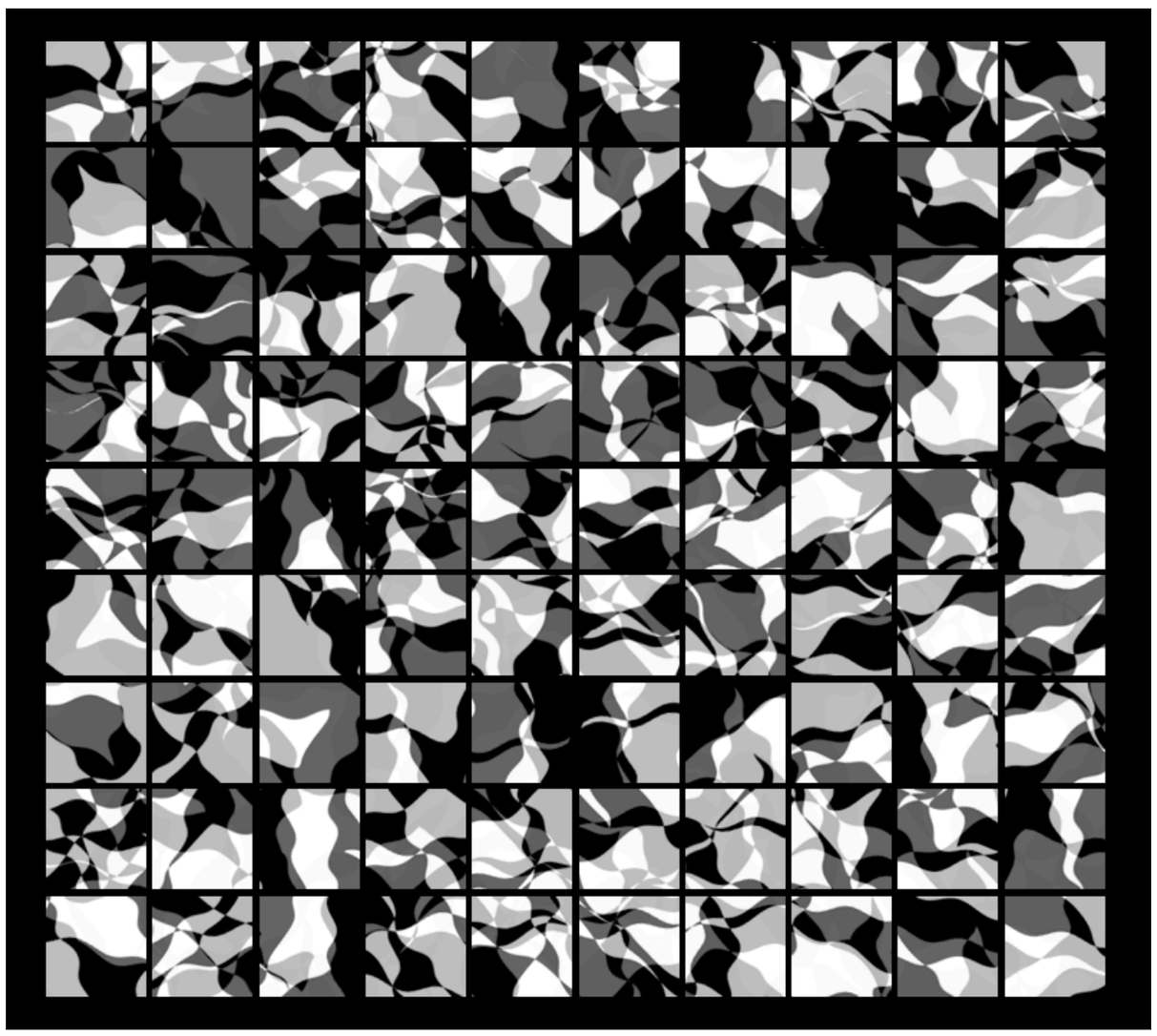


Audio patch



Future Work

Synthetic dataset for optimal one-time pre-training.



Wednesday, October 16, 2013

If images serve to create a better prior distribution for later training on images or audio,perhaps data used for pre-training does not need to be from the same distribution as the data that the network is intended for. There may be a dataset that serves as a good pre-training set for a variety of problems.

Future Work

Further experimentation with the generative capabilities of SDAs

Thorough exploration of hyper-parameters.

What variables effect the ratio of smooth nodes to noisy nodes, and how is that ratio related to performance?

What variables affect the normalized reconstruction error order of layers? Is it an important indicator of good performance?

What other metrics besides performance could we use to evaluate how brain-like new algorithms are?

Wednesday, October 16, 2013

Much more research on SDAs is needed, and much more will be done in the near future with deep networks being currently the largest focus of attention at machine learning laboratories around the world. Here are just a few things that could prove interesting.

SDA's can function in a generative mode. How can we use this ability? How is it similar to dreaming?

There are a number of hyper-parameters that can be adjusted with SDAs. I did not have time to explore these fully in my experiments, but I chose values that worked well for similarly sized problems in the literature. A more thorough exploration of those hyper-parameters would be an informative exercise.

The ratio of smooth nodes to noisy nodes in the feature visualizations seems to correlate with performance. It would be useful to have a computable metric of smoothness so that it could be experimented with, and even used to optimize or guide the network during training.

In my graphs of training error over time, I noticed that the order of layers when sorted by their normalized reconstruction error was not the same between sensory modalities. Future

Wednesday, October 16, 2013

Finally, I would like to leave you with a parting thought. Many people's gut reaction to artificial intelligence is fear. They think it will be just like the movies and it's going to be malevolent and take over the world and kill us all.

That's not even close to realistic, given the direction of machine learning research and the current and historical geopolitical climate. I'm excited about the future of AI because I think it will have an effect on the human experience similar to the Internet. The Internet changed the requirements for success from being about "who you know" to "what you know". AI will change it from "what you know" to "what you want". We have always been at the pinnacle of our world, and it will remain our responsibility indefinitely to decide where it goes.