CSC 550 notes:

Definitions:

* Linear filters: **Linear filters** process time-varying input signals to produce output signals, subject to the constraint of [linearity](https://en.wikipedia.org/wiki/Linearity)
* **Linearity:** is the property of a mathematical relationship or function which means that it can be graphically represented as a straight [line](https://en.wikipedia.org/wiki/Line_(geometry))
* Natural scenes:
* Receptive fields:
* Simple cells:
* Complex cells:
* Sparse code: Sparse coding is a class of unsupervised methods for learning sets of over-complete bases to represent data efficiently. The aim of sparse coding is to find a set of basis vectors \mathbf{\phi}_i such that we can represent an input vector \mathbf{x} as a linear combination of these basis vectors:
* **Universal function approximator:** model family that can approximate any function that maps input patterns to output patterns (with arbitrary precision when allowed enough parameters)
* **Deep neural network:** network with more than one hidden layer between the input and output layers; more loosely, a network with many hidden layers
* **Supervised learning:** a learning process requiring input patterns and additional information about the desired representation or the outputs (e.g., category labels)
* **Unsupervised learning:** a learning process that requires only a set of input patterns and captures aspects of their probability distribution
* **Convolutional network:**network in which the preactivation of a layer (before the nonlinearity) implements convolutions of the previous layer with a number of weight-template patterns
* **Receptive field modeling:** predictive modeling of the response to arbitrary sensory inputs of neurons (or measured channels of brain activity)
* **Max-pooling:** summary operation implementing invariances by retaining only the maxima of sets of detectors differing in irrelevant properties (e.g., local position)

From Visual Cortex: The continuing puzzle of area V2

* Little is known about role of V2 in visual processing
* Ito and Komatsu 🡪 responses of V2 neurons to pairs of angled lines could be predicted from their responses to the individual line components. They might be simply the sum of responses from one or more V1 neurons (orientation selective)
* V1 neurons driven by orientation and essentially deconstruct a scene based on spatially localized oriented components. Other area of the visual cortex also clearly displays their role in visual processing by the way they fire, given motion in a specific direction or speed? V2 neuron specificity, however, is not clearly defined, but it is known that it is essential to vision, given that an injury in that area significantly affects the ability to perform complex spatial tasks [1]
* V2 cells are selective to orientation, color, stereoscopic disparity and motion
* So far, no excitatory stimulus has been found that reveals their role in vision
* Ito and Komatsu studied the selectivity of V2 neurons by placing two lines in different angles in the center of each cells receptive field
* They chose angled lines because angles are a basic component for determining contours and co-occurring lines are common in natural scenes
* Main results 🡪 presented as a pattern of responses from a given neuron across the chosen stimulus space
* Their results evidence that the neurons respond primarily to the components of the angels, but not really to size or orientation
* They then compared the V2 responses of the angled stimuli to the responses to the components alone. They found that the responses to an angle stimulus can be predicted by the responses to the individual components
* This gave rise to the thought that V2 works based on V1 level inputs. Perhaps V2 neurons with two preferred orientations simply receives direct inputs from two orientation-selective V1 neurons
* Hubel and Wiesel [11] proposed a V1 simple cell might be constructed from a series of center-surround neurons in the lateral geniculate nucleus (LGN). Maybe V2 is constructed within a similar structure
* A V2 cell can then be constructed by summing the responses of 2 hypothetical V1 simple cells, each modeled as oriented linear filters with excitatory centers and suppressive flanking surrounds.
* The simple linear models cannot explain these results, which show some V2 cells respond to particular combinations of line components that form angles, but not necessarily to each of the components alone, implying that there is nonlinearity
* They show a very simple model of V2 but it is evident that a lot more is involved between V1 and V2.
* It is difficult to choose stimuli to study unknown areas like V2. They are mostly based on guesses that make up fundamental components of a visual scene to perform tasks such as object recognition and contour segregation
* One advantage of V2 is that V1 is well understood, and it is the area from which V2 receives its predominant input.
* It makes sense to wonder how a V2 neuron can be constructed based on V1 inputs.
* From Ito and Komatsu studies, we can see that the seemingly complex patterns of results can be partially explained with a simple model in which V2 neurons are summing the response from two orientation selective V1 simple cells.

From Hosoya:

* Previous studies, either experimental or theoretical, demonstrated tight relationships between natural image statistics and neural representations in V1
* Sparse coding of natural imaging has presented a way to infer receptive fields of simple and complex cells, but the relationships in higher areas has not been clarified
* Their experiment consisted of a sparse coding algorithm that took the output of a V1 like model as input. The V1 filter-model was previously trained with a large variety of natural image patches as input
* The model exhibited response properties that were consistent with three neurophysiological results (quantitatively and qualitatively):
  + Homogeneous and heterogenous integration of local orientations (Anzai et al., 2007)
  + A wide range of angle selectivities with biased sensitivities to one component orientation (Ito and Komatsu, 2004)
  + Exclusive length and width suppression (Schmidt et al., 2014)
* The study proposes novel type of model cell that could detect a combination of local orientations converging toward a single spatial point, potentially related to corner-like features played an important role in reproducing tuning properties compatible with V2
* Compatible with idea that V2 uses a sparse code of natural images
* Sparse coding importance:
  + Sparse coding has successfully explained many receptive field properties in V1, but nothing is concluded about V2
  + A variety of properties distinct from V1 have been discovered in V2, and thus a more integrative understanding is called for.
  + They propose a hierarchical sparse coding model of natural images to explain three major responses known (above)

From Kriegeskorte:

* Unlinke the brain, the models of information processing that have dominated computational neuroscience are largely shallow and can only perform simple computations
* Recent advances, however, could have possibly triggered a new era, where real world tasks engages with computing power
* An early mathematical model of a single neuron was suggested by McCulloch & Pitts (1943)
  + Binary threshold unit took a number of inputs, computed a weighted sum and imposed a threshold, implementing a linear discriminant
  + Discriminating categories that are not linearly separable in the input requires an intervening layer of nonlinear transformations between the input and output units
* The problem now is how to automatically train the multilayer network with input-output pairs
* So far the most influential solution is the backpropagation algorithm, a gradient-descent method that makes iterative small adjustments to the weights in order to reduce the errors of the outputs
* Neural network models worked well on “toy” problems whereas backpropagation models didn’t. With advances in machine learning techniques and hand-engineered representations, neural networks fell out of favor in the 1990s
* Between 1990-2000s, neural nets were studied by a smaller number of scientists, who realized the limitations were due to the extreme complexity of the problem and were to be overcome through a combination of better learning algorithms, better regularization and larger training sets
* In 2012, the neural network model built by Krizhevsky et al. marked the beginning of the dominance of neural networks in computer vision
* In the past 6 years, error rates have dropped further, and now roughly match human performance in the domain of visual object classification
* Primer on neural networks:
  + A unit computes a weighted sum of its inputs and activates according to a nonlinear function
  + Model neurons 🡪 used as a distinction between biological reality and highly abstracted models
  + Linear unit 🡪 simplest, outputs a linear combination of its inputs. The units combine to form networks but never go beyond a linear combination
  + Multilayer network of linear units = single-layer network whose weights matrix W’ is the product of the weights matrices Wi of the multilayer network
  + Nonlinear units 🡪 essential since their outputs provide building blocks whose linear combinations one level up enables approximations of any desired mapping from inputs to outputs
  + The unit uses its input weights w to compute a weighted sum of its activities and passes the result through a nonlinear function to generate its activation
  + In early models, the nonlinearity was simply a step function
  + The problem with this method (hard thresholding) is that small changes to the weights don’t reflect any differences in the output, which makes it difficult to learn the weights for a multilayer network by gradient descent, in which small adjustments are made to the weights to reduce errors iteratively.
  + If you have a soft threshold that varies constantly, the gradient descent can be used for learning
* Feedforward networks with at least one layer of hidden units intervening between input and output layers are universal function approximators
* A feedforward network is composed of a sequence of layers of unit, with each unit sending output only to higher layer units
* A feedforward network is deep when it has more than one hidden layer
* Deep learning 🡪 strategy of using architectures with many hidden layers to tackle difficult problems
* Why does deep help 🡪 even though shallow networks are already very efficient in approximating functions, deep networks help because they can represent many complex functions more concisely, meaning with fewer units and weights than shallow nets and support vector machines. They have additional nonzero weights and enable the reuse of previous computations and extend the power of the network
* Even though neural networks have high representational powers given sufficient numbers of units, setting the weights of the connections is tricky. Models that can solve real world problems will have large numbers of units and weights, and global optimization techniques are not available for this nonconvex problem.
* The total error is a locally smooth function of the weights. With a gradient descent, the iterative reduction of the errors through small adjustments to the weights is made
  + Gradient descent starts with random initial weights and performs slight changes to reduce error
  + Gradient = derivative of the error with respect to the weight
  + An efficient way to compute the error derivatives is to propagate them backwards through the network
* Gradient descent sees only the local neighborhood in weight space and is not guaranteed to find globally optimal solutions
* Gradient descent only sees the local neighborhood in weight space and is not guaranteed to find globally optimal solutions, but it can help local optimization find good solutions. The gradient descent does not get stuck in the local minima, where error increases in all directions and no further progress is possible.
* Backpropagation 🡪 each time point of the recurrent computation corresponds to a layer of the feedforward net, each of which is connected to the next by the same weights’ matrix, of the recurrent network.
* By backpropagation through time, a recurrent network can learn weights that enable it to store short-term memories in its dynamics, thus relating events temporarily separated as needed in order to achieve desired classifications or predictions of temporal sequences
* Problem: Propagating error derivatives far enough backwards through time enough for the network to be able to learn makes the gradient vanish or explode
  + Since the error derivative is the product of multiple terms,
  + One solution: long short-term memory architecture, where special gated units can store short-term memories for extended periods, thus the network can remember information that will be relevant many time steps later in a sequential prediction, classification, or control task.
* The details of the gradient-descent algorithm, regularization, and weight initialization all matter to making supervised learning by backpropagation work well
* Supervised learning: data = input patterns and the associated desired outputs. A supervision of this type of signal is often unavailable in the real world. There is often unlabeled input patterns and only a smaller number of labeled input patterns
* Unsupervised techniques:
  + Unsupervised learning does not require labels for a network to learn a representation that is optimized for natural input patterns and potentially useful for a variety of tasks. Natural images form a very small subset of all possible images, enabling unsupervised learning to find compressed representations.
  + Autoencoders are a good example of unsupervised learning. They are a feedforward neural network with a central code layer that has fewer units than the input. It is trained with backpropagation to reconstruct its input in the output layer
  + It is unsupervised because it requires no other labels, or separate supervision information besides the set o input patterns.
  + *Encoder 🡪* The layers from the input to the code layer
  + *Decoder* 🡪 the layers from the code layer to the output
  + If both are linear, the network learns the linear subspace spanned by the first k (num. of layers) principal components
  + With nonlinearity, nonlinear compressed representations can be learned
  + Unsupervised learning can help pretrain a feedforward network when insufficient labeled training data are available for purely supervised learning
  + Ex: network of visual recognition can be trained layer by layer using a large set of unlabeled images. Once it has learned a reasonable representation of natural images, it can more easily be trained with backpropagation to predict the correct image labels
* Deep feedforward convolutional network currently dominates computer vision
* Deep hierarchy of representations: typically 5 to 20 layers gradually **transforming a visual representation**, whose spatial layout matches the image, **to a semantic representation that enables the recognition of object categories**
* Convolution: a feature useful in one position might be also be useful in another position. The lower layers contain local visual feature detectors with small receptive fields. Each detector is replicated across the 2-D image, forming a feature map. This amounts to a convolution of the image with each feature pattern, followed by a static nonlinearity
* In between the convolutional stages, local pooling stages are inserted. It combines the outputs of a local set of units by taking the maximum or the average in order to implement local tolerances.
* In the highest layers, units have global RFs, receiving inputs from all units of the previous layer
* Final layer 🡪 one unit/category and implements a softmax (normalized exponential), which strongly reduces all but the very highest responses and ensures that the outputs add up to 1. Output can be interpreted as a probability distribution over the categories when the training procedure is set up to minimize the crossentropy error
* Machines that are trained to categorize natural images discover features that are qualitatively similar to those found in biological visual systems
* Early layers 🡪 develop Gabor-like features similar to the V1 neurons characteristics
  + Followed by units that are selective for slightly more complex features, including curve segments.
  + Higher layers 🡪 selective for parts of objects and for entire objects, such as faces and bodies of humans and animals
* The idea is that units learn selectivities to natural image features that increase in visual complexity along the hierarchy.
* 2 things to consider:
  + In order for the selectivity to be confirmed, many images need to be considered. The response shows what drives in the context of a particular image.
  + Network by krizhevsky 🡪 appears to contain units selective for text and faces, although they were not among the trained categories
* Krizhevsky: 5 convolutional layers and three fully connected layers
  + Reducing the number of convolutional layers hurts performance
  + Uses rectified linear units, max-pooling, and local normalization
  + The network was trained through backpropagation to recognize which of 1000 object categories was shown in the input image
  + Training relied on dropout regularization, a technique in which each unit is “dropped” with a probability of 0.5 on each training trial. On a given trial, a random set of approximately half of the units is used in both the forward pass computing the output and the backpropagation pass adjusting the weights
  + The network has a total of 650,000 units and 60 million parameters. The convolutional layers are defined by their small local weight templates, which constitute less than 5% of the parameters in total
  + Training was performed over 6 days on a single workstation with 2 GPUs, parallelizing and accelerating the computations