**NN-course**

Taking the Stanford course CS231N: <https://www.youtube.com/watch?v=vT1JzLTH4G4&list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv>

Class 1. Basic introduction to NN

Class 2. How NN have developed.

* Focus on vision
* How CNN have dramatically improved vision.

Class 3. The loss function - <https://www.youtube.com/watch?v=h7iBpEHGVNc&index=3&list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv>

* Assignment 1 -
* Loss Function - based on output, what needs to change to adapt the W matrix.
* Hinge Loss -
  + Multi-class SVM Loss @15:45 -> @18:30 - good example
  + Initialized with small weights initially
    - Initial loss should be C-1.
    - Minimum loss is 0
    - Maximum loss is technically infinite
  + Consider square hinge loss function - emphasize different types of errors. Can increase the impact of larger errors.
  + When W matrix is found, there are actually multiple W's, e.g. 2\*W. \*\* How does the classifier choose the correct W if there are multiple? Our error function doesn't push the selection one direction or another. As a result, the classifier may minimize the training data points and Overfits. Unfortunately, we want something more general rather than a complex W that matches.
    - Add Regulariztion - this attempts to make W "simpler". The regularization attempts to make the W solution more general. The regularization strength is lambda. A really important hyper-parameter. \*\* You can imagine the regularization adds a penalty for higher order polynomials.

L2 regularization - euclidean norm - W^2 L1 regularization - | W | Elastic Net - combinations of L1 + L2

All regularization attempts to reduce the complexity of a model. However, they each have different effects on the selected W.

* The linear SVM loss – there is no real meaning to the values, just that there is a difference.
* In SoftMax – we try to provide additional information. We effectively normalize the value outputs of the classes to provide a probability distribution for each class output.
* SoftMax debugging – when starting out with many small values in W – the output should be log C.
* In SVM – the classifier wants a certain difference between desired class and other classes. In SoftMax, the classifier will always push the desired class toward infinite while the wrong classes are pushed toward negative infinite.

* Once we have the loss function, how is it applied to obtain W?
* Optimization!
  + Think of walking through valley and try to minimize the W.
  + 1. Random search – generally bad algorithm.
  + 2. Gradient Descent – generally descend based on the slope, i.e. derivative of the function.
    - Can test it practically by taking small steps instead of taking actual derivatives.
    - Generate the gradient, use this information to adjust the parameter vector, W.
    - Practically – do not calculate the gradient by finite differences. Instead calculate the gradient based on **analytic expression** of the gradient, dW.
    - But, still test the analytic gradient with the numeric gradient as a cross check.
    - The **step\_size (learning rate)** determines how strongly the gradient is applied to W.
  + Another challenge is the actual calculation of the loss. When the training data set is large, calculating the loss could take a long time. **Stochastic Gradient Descent** (SGD) takes a sample of the training data set and calculates the loss and gradient based on the mini-batch. These estimates are then used to update W.
* SVM looks at data as data/pixels. However, this does not work very well. Another option is to feed in feature quantities into the SVM.
  + One common option is to obtain the colour histogram. This provides the colour distribution of the image.
  + Another option is histogram of oriented gradients (HoG)
  + Another option is to extract random patches. This effectively uses the bag of words option. We would need to define a “codebook” of “visual words”. We can then use the codebook to describe the image.
* CNN changes everything.
  + Rather than defining the codebook or HoG, the CNN learns that information from the data. Therefore, no explicit programming required.