How do you answer all these question?

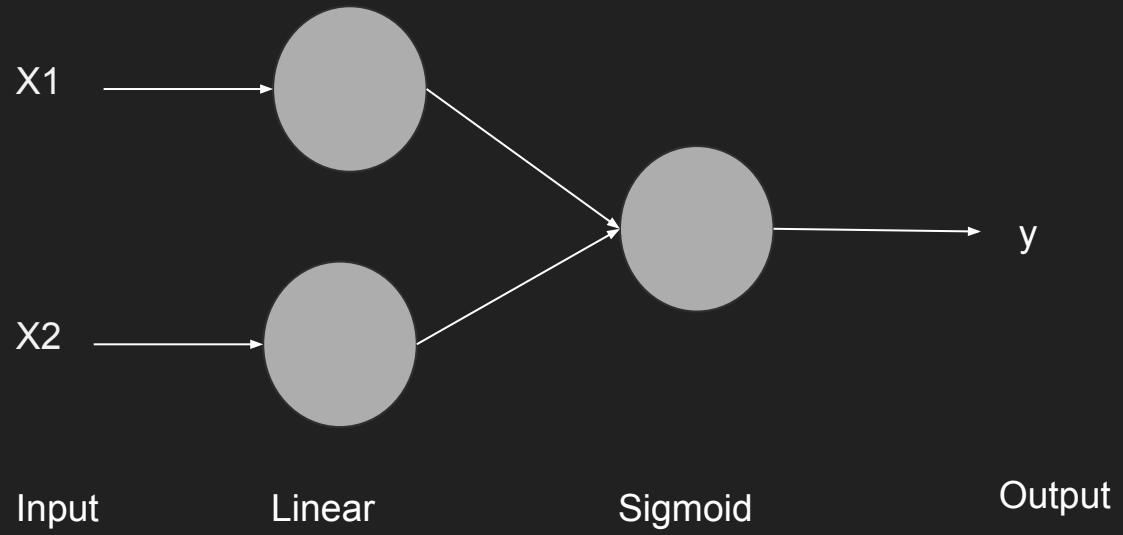
CS 434 - Machine Learning & Data Mining Final Project

Quora Questions

Miles A. Curry - Christian Armatas

Approach - Neural Network Design

- Siamese Two Layer Design
  - Inputs and recognized features are captured and categorized separately in Layer 1
  - Then, inputs are reduced down into Layer 2 to determine if the two questions were similar
- Similar in design as Homma et al’
- Using Keras Library



1. Data Processing
    a. Removed entries that had encoding errors or special characters unrecognizable by python
    b. Tokenized sentences into words and then encode them as a ***bag-of-words***
        i. Each question was added together to be its own bag-of-words instead of a bag-of-words of the entire question set
            1. We reasoned that data in other questions is not relevant in determining whether or not two questions were directly related.
            2. This would also reduce the algorithm's amount of data consumption
        ii. Questions were encoded as ID's and their count
        iii. Parts of speech (POS) for each word was also recorded
            1. We figured POS would give the Neural Network more data to analyze
            2. Difficult to encode this number effectively
            3. Thought about giving some POS higher priority than others
                a. See if added weights to specific parts of speech affected performance

```
What is the step by step guide to invest in share market in india?
What is the step by step guide to invest in share market?
{u'what': 1, u'invest': 2, u'is': 3, u'india': 4, u'by': 10, u'to': 6, u'step': 7, u'in': 8, u'the': 0, u'share':
 9, u'guide': 5, u'market': 11}
[  1.   2.   3.   4.   5.   6.   7.   8.   9.  10.  11.  12.   0.   0.   0.
   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.]
[  1.   2.   3.   4.   6.   7.   8.   9.  10.  11.  12.   0.   0.   0.   0.
   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.]
```

Data Representation of the first question pair

NN Design cont.

- Distance calculated by Euclidian Distance
  - Gave the best accuracy ~ 50-54%
  - Cosine accuracy gave ~ 15%
- Contrastive Loss as Loss Function
  - Function given with Keras' simease example (https://github.com/fchollet/keras/blob/master/examples/mnist_siamese_graph.py)
  - Not sure what this is… (We pulled it from the internet)
- Dropout = 0.1
  - No Dropout gave us a higher percent correct… but it's probably overfitting

Results Thus Far

- Around 50% Accuracy on training set
  - Depending on number of epochs and input size
- N = 395,548 , Epochs = 20 , Accuracy = 54.84% (Entire Testing Set)
  - ~8000 Entries Removed for encoding errors
- N = 10,000, Epochs = 20, Accuracy = 52.81%

```
Epoch 19/20
395548/395548 [==============================] - 3s - loss: 0.2096
Epoch 20/20
395548/395548 [==============================] - 3s - loss: 0.2094
* Accuracy on training set: 54.84%
```

```
Epoch 20/20
10000/10000 [==============================] - 0s - loss: 0.2231
* Accuracy on training set: 52.81%
```

Current Problems and Lessons Learned

1. Data Processing takes a long time and needs to be optimized
    a. Its around cubic time currently…
2. See if we can encode POS into data
3. Accuracy is bad.. Very baaad..
    a. Need to try
        i. different activation functions
        ii. different difference functions
        iii. see what other groups have done for there siamese nn
        iv. Verify if our nn is predicting correctly or not.
4. Lessons Learned
    a. Inputs of differing sizes
        i. Can’t take dot product of different size 1D arrays
        ii. Compensated by padding the small inputs with zeros and truncating large inputs
    b. Verification - Do n-fold testing on training set
    c. Experiment more with Keras!
    d. Do our data encoding ideas work