Improvements to the NICE Machine Learning Interview problem

November 14, 2018

1. *Term-Weighting:* In general, term weighting consists of two parts: the one that corresponds to statistics of term occurrences in the given document (dialogue in our case) and the part that estimates statistics of term occurrences in the whole collection. As a rule, term weighting is a multiplication of these two parts. While computing the first part is usually the same, there are several techniques to compute the latter one. We have used the popular Inverse Document Frequency (IDF) (unsupervised approach) for this problem but there exists the following supervised alternatives 1) Gain ratio (GR): Gain ratio measures how much a particular term is useful in predicting the classes on similar lines as features are assigned importance by a Decision Tree 2) Confident Weights (CW) implements the main idea that a term has a non-zero weight in a particular class only if the frequency of the term in that class is higher than the frequency of the term in all other classes. CW has been shown to outperform TF-IDF and GR based features on a number of benchmark corpora but on the other hand, this method is more computationally expensive. 3) Term Second Moment (TM2), Relevance frequency (RF) and Novel Term Weighting (NTW) are other supervised term-weighting techniques that are modifications of the same idea as in CW. A straightforward way to improve the performance of our classifier is to try features based on one or combination of all of these term weighting techniques.

2. *Additional Classification Model:* In this exercise, I have shown the performance of two deep learning architectures namely Convolutional Neural Network(CNN) and a Bidirectional LSTM both of which failed to perform as well as other simpler classifiers. The appalling performance of CNN on this problem can be attributed to the fact that CNN are specially designed for images where the local spatial information is important whereas clearly long-distance term dependencies are important in this classification problem. A recent advancement in deep learning models is attention mechanism where the basic idea is to focus on a subset of the information they are given. This idea has gained prominence because attention based neural networks have performed better than state of the art for sequence to sequence translation. Most recently, a paper named ”Hierarchical Attention Networks for Document Classification” by Yang et. al. was published which proposes a novel neural network architecture which the authors call Hierarchical Attention Network (HAN) for document classification. This paper includes two levels of attention mechanism at the word level and sentence level respectively while creating the representation of the document. While the sentence level attention mechanism is probably not necessary for our problem (since our dataset mostly consists of small dialogues), implementing HAN should lead to a significant improvement over the current results.

3. *Hyperparameter Tuning:* The hyperparameters in ANN (which has been shown to be the best model for this classification) should be tuned using K-fold cross-validation where the best parameters should be chosen by a random search of parameters. Since the number of hyperparameters is small, random search significantly outperforms grid search. More complicated techniques for hyperparameter tuning include Bayesian optimization which tries to balance exploration (hyperparameters for which the outcome is most uncertain) and exploitation (hyperparameters expected close to the optimum).