Learning about Cancer from Images and Text by Regina Barzilay: A summary

By Arshitha Basavaraj

Problem Definition: From the talk titled "Learning about Cancer from Images and Text" by Ragina Barzilay, I gathered that the higher level overview of the problem that's being (or trying to be) dealt with is to significantly reduce the loss of data and make effective use of it in the process of cancer diagnosis (particularly breast cancer) from imaging to the line of treatment decision making. To that end, two of the significant accomplishments that were discussed are: (1) Using machine learning to parse breast pathology reports. Development of a database that Dr. Hughes (one of her collaborators at MGH) and his team can use to monitor the development of atypias, which help identify which patients are at risk of developing cancer later in life. (2) Developing Machine Learning algorithms to categorise patients with HRLs (High Risk breast Lesions) based on risk levels.

Algorithms and results summary:

Algorithm in "**Using machine learning to parse breast pathology reports**": Supervised classification has been used to train a classifier with breast cancer pathology reports using electronic medical records. In particular, boosting classification was used for training the classifier. Text in the pathology report was represented using vectors that consisted of words and phrases (standard n-gram representation) and for each of the features (i.e., the words and phrases) the model learnt weights during the training phase. This training was carried out for each of the 20 categories.

Results for above: In the testing phase, the classifier was tested on 500 reports that were not contaminated with the training set. On this it achieved 90% accuracy. For each of the 20 categories, it achieved an average accuracy of 97%. The classifier was then used in creating a database of 91,505 parsed pathology reports.

Algorithm in "High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision": *Machine learning model used:* Random Forest Classifier. *My understanding of it from the paper:* From the entire feature pool in the training data set, the random forest model, picks a subset of features randomly and repeatedly. Using these random feature subsets, an ensemble of decision trees are built. And using a constructive algorithm, these decision trees allow for correct classification of that sample of the training set. The model was trained with 671 HRLs in 654 patients and tested with 335 HRLs in 332 patients.

Results for above: The model correctly diagnosed 97% of the 335 HRLs it was tested on as malignant. It reduced the number of unnecessary surgeries by more than 30%.

Critical Discussion: In the first study, one of the concerns is accuracy on the test set. Out of 500 testing samples, 90% of them were correctly classified, i.e., 50 reports were incorrectly classified which is a large number for a medical application. Also, in the collection of the training set, since three hospitals in Boston have been considered, wouldn't that introduce an unintentional bias due to demographical distribution of breast cancer patients. Can the resulting database then be a considered a comprehensive one?

Also, it has been mentioned in the paper that the patient data was anonymised and then used. In the past, there has been evidence that even anonymised data can be back tracked to a specific individual and techniques of differential privacy has proved to deal with the problem of re-identification with sufficient accuracy. It's unclear if such precautions were considered during the study.

In the second study, the results achieved are impressive. Over-treatment due to inexact diagnosis is a problem that plagues almost every field of medicine and the 30% reduced number of unnecessary surgeries due to HRLs is a significant improvement. However, it was mentioned during the talk that the human intervention from imaging to summarising the diagnosis is minimal. In my opinion this seems like a risky endeavour at this stage of study. Wouldn't having a trained radiologist in the loop during the intermediate stages from imaging to diagnosis summarisation make the outcomes more reliable? Also, bias in the training data is one of the concerns here as well.

Relevance to material discussed in class: Personally, I could relate the content discussed in this talk with variety of material discussed in class as well as addressed in the assignments. We recently covered logistic regression in the last couple of lab sessions which is basically one of the supervised classification algorithms and in both studies, a supervised classifier was trained using annotated training samples. Having studied this in a classroom environment and then being able to understand it's real world applications, I could better appreciate, for instance, the reasons behind the importance of training data selection and making sure that training and testing data aren't contaminated and so on. Also, in the first study, it was mentioned that rule based techniques were employed in building the database which is also one of the things that was covered in the lectures.

Also, on an ethical level, this is probably one of the areas of application that Ali Rahimi cautions about in his talk. How trustworthy is such a system that has minimal human intervention. To what depth is the system understood from a theoretical standpoint. Having worked on that assignment, I found myself on both sides of the argument where on one side I'm skeptical about a system that has very little interaction with humans with such a critical and significant application such as breast cancer diagnosis and on the other hand, I would argue that it would be great injustice to the society to keep such an innovation from the people simply because we are in the process of fully understanding these machine learning algorithms on theoretical level.

In conclusion, the talk has helped me evaluate gaps in my technical understanding and understand the real world problems that arise while taking on a study of this scale. It's also got me thinking about the ethical issues involved in developing such tools.