Machine Learning and Medical Imaging

Computer Science Senior Seminar April 18, 2020

By: Trent Merkins

Outline

Background

CSAL Model

CNN-Seq2Seq-Attention model

AliPy Active Learning

Experiment

Results

Background

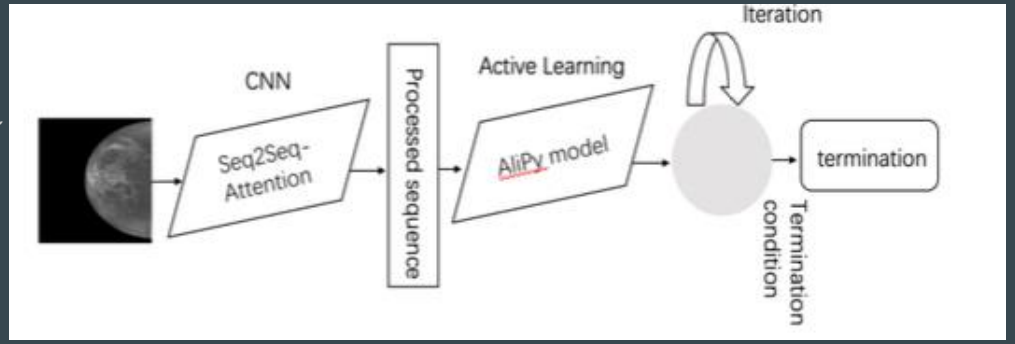
Medical imaging is frequently used in the diagnosis and treatment of patient conditions.

The scarcity of medical image standard data sets and the high cost of manually labeling images means that it is necessary to use active learning methods to select the most valuable data.

The purpose of these learning algorithms is to select the most valuable data to reduce labeling cost, reduce pressure on physicians and make clinical decision making easier.

CSAL Model

CSAL is a machine learning model that integrates CNN-based Seq2Seq-Attention and active learning (AliPy model).



CNN-Based Seq2Seq-Attention Model

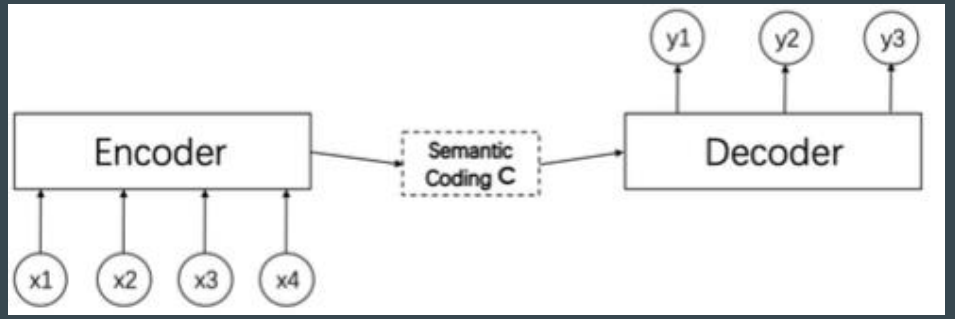
Widely used sequence-to-sequence deep learning algorithm, used in image voice, etc.

Used for image processing, machine translation, image recognition and others.

Traditional encoder-decoder model

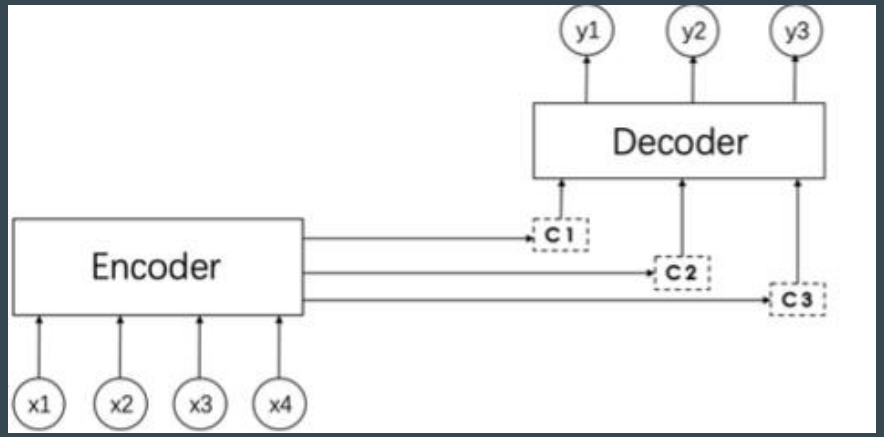
Limited and the only connection between encoding and decoding is a fixed vector length

- Two drawbacks
  - These vectors cannot fully express the input information
  - As the input length increases it needs to overwrite previous data



Attention model solution

This solution allows for multiple intermediate C vectors to provide a better transformation between the input and output.



Loss Function

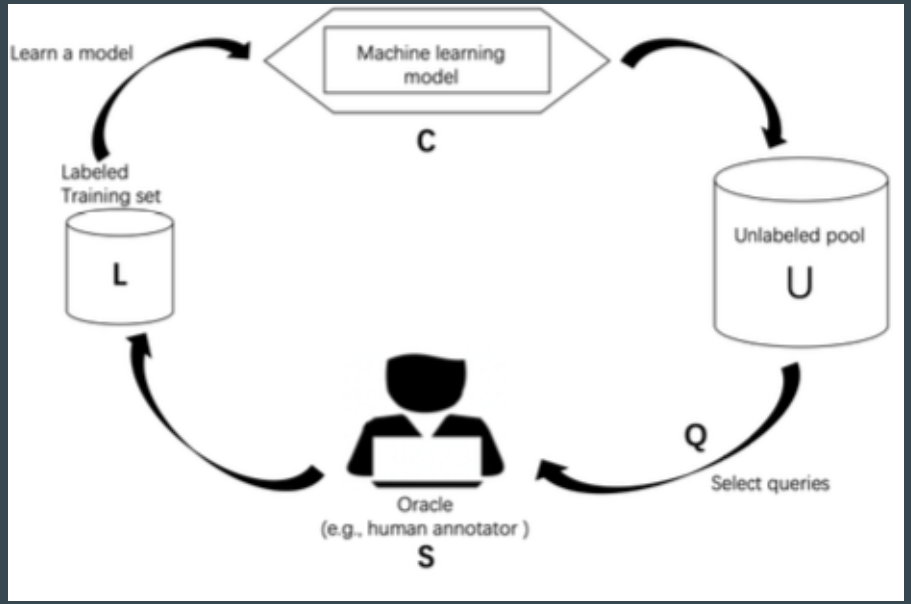
This considers how to update parameters based on this loss function.

The goal of the Seq2Seq model is to maximize the probability of the target output.

$$\text{loss} = -\frac{1}{\text{N}}\sum_{n=1}^{N} logP(Y_n|X_n,\theta)$$

Active Learning

SupposeA=(C,L,S,Q,U),where C is one or a group of classifiers, L is a set of labeled training samples, Q is a query strategy function for querying a large number of samples in unlabeled samples, U is the entire unlabeled sample set, S is the supervisor which can mark the unlabeled samples.



Active Learning Basic steps

1) Select appropriate machine learning model, actively select strategy Q, and divide the data (Training Set, Testing Set, Unlabeled Data Set).
2) Initialization: model C is initialized by training set.
3) Use the trained model C to predict U, obtain the prediction result of each sample, and query the sample n1 with large amount of information through Q.
4) Label n1 by the oracle, update the label set n=n+n1.
5) Train and update model C based on training set n.
6) Use model C to verify on the test set, stop the iteration in accordance with the convergence condition, otherwise loop execution 3-5.

AliPy Active Learning Model

Provides an implementation of a module-based active learning framework

In the setting of the uncertainty standard, there are generally two choices: 1) Use probability to indicate the degree of uncertainty. 2) Use distance to indicate the degree of uncertainty

$$x_{LC}^{*} = \underset{x}{argmax}\left(1 - p_{\theta}(\hat{y}|x)\right)$$

$$\hat{y} = \underset{x}{argmax}\, p_{\theta}(y|x)$$

$$f(x_i) = \sum_{j=1}^{n} a_i y_i K(x_j, x_i) + b$$

Experiment

Data-sets Used

Dataset 1 comes from the breast_cancer standard dataset of sklearn.datasets, its target is divided into two categories, the total sample size was 569, of which 212 were benign samples, 357 were malignant samples and has 6 types of sample label

Dataset 2 comes from INbreast (public dataset for Portuguese bcct.plan), includes 115 cases, 410 images and corresponding 117 case reports.

Preprocess

Training Set	Test Set	Unlabeled Set
20%	20%	60%

Results

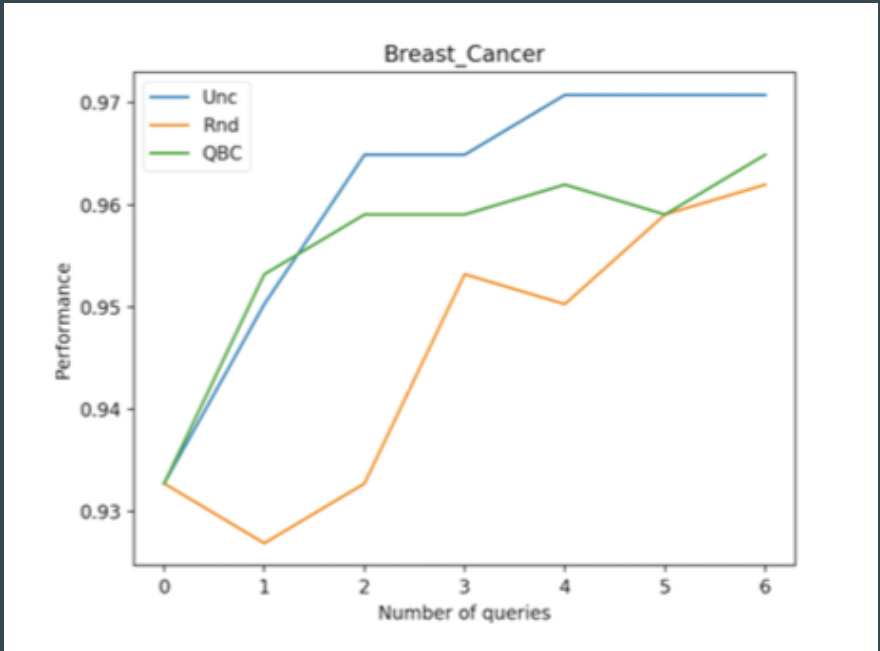
Table 4 Experimental Results Based on Data Set 1

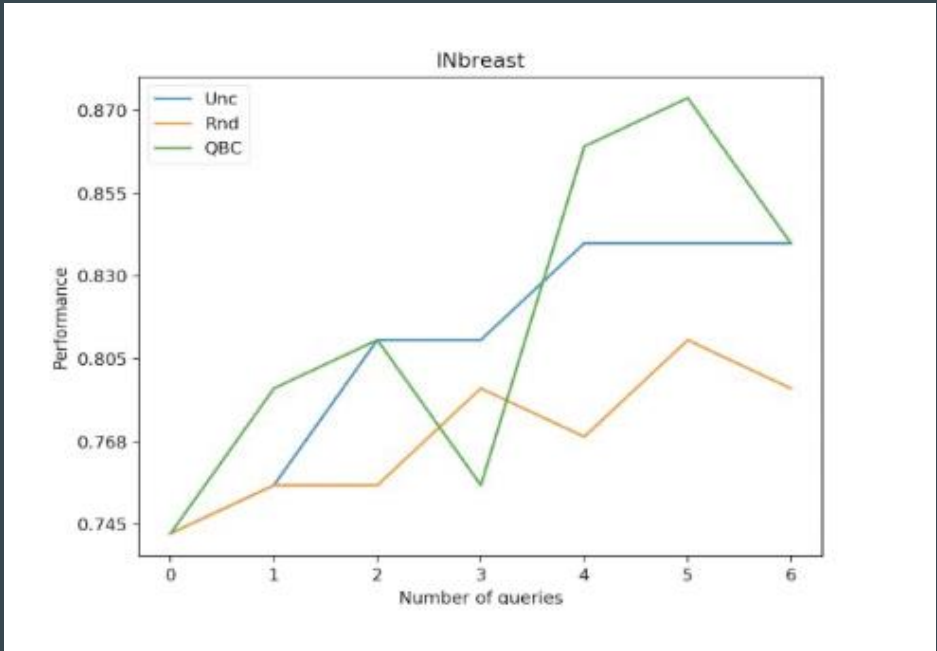
Methods	Number_of _quiries	Number_of_ different_split	Run time	Performance
Unc	6	3	1min	0.962±0.00
QBC	6	3	1.2mins	0.955±0.00
Rnd	6	3	3mins	0.943±0.01

Table 5 Experimental Results Based on Data Set 2

Methods	Number_of _quiries	Number_of_ different_split	Run time	Performance
Unc	6	3	2.2mins	0.802±0.02
QBC	6	3	3.1mins	0.809±0.01
Rnd	6	3	5.5mins	0.754±0.02

Results (cont.)





Conclusion

This is how active learning can be used to filter our more valuable samples from the data pool. These can then be labeled by the professionals in order to expand the medical image data set at a reduced cost.

Reference

J. Li and W. Mi. Study of mammography medical imaging sample selection based on csal. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, ACAI 2019, page 485–490, New York, NY, USA, 2019