Automatic Diagnosis Coding of Radiology Reports: A Comparison of Deep Learning and Conventional Classification Methods

Sarvnaz Karimi[1], Xiang Dai[1], Hamed Hassanzadeh[2], Anthony Nguyen[2]
[1] CSIRO Data61, [2] CSIRO The Australian e-Health Research Centre

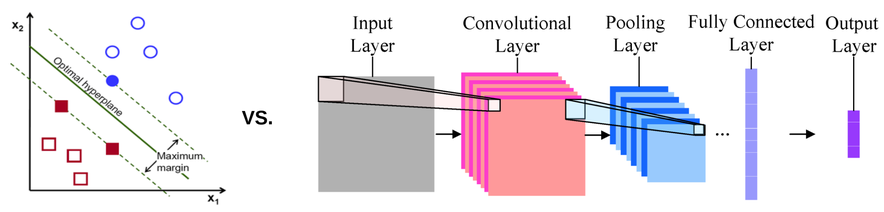
www.data61.csiro.au

Diagnosis autocoding improves the productivity of clinical coders and the accuracy of the coding. We investigate the applicability of deep learning methods at autocoding of radiology reports using International Classification of Diseases (ICD). We explore how to use these methods when training data is sparse, skewed and relatively small.

Deep Learning vs Conventional Text Classification

Clinical coding is a specialised skill requiring excellent knowledge of medical terminology, disease processes, and coding rules. Problems with the manual process are high costs of labour and human errors resulting in misleading statistics. Literature reports automatic methods for this problem using conventional machine learning methods, such as SVM classifiers. We investigate the following research questions:

- Can deep learning-based text classification compete or improve the state-of-the-art clinical autocoding?
- In what setting is Convolutional Neural Networks (CNNs) effective for autocoding of radiology reports?



Method

We implemented text classifiers as below:

- CNN text classifier - one convolutional layer using multiple filters and filter sizes followed by max pooling land fully-connected layers to assign a label.
- Input to CNN: (1) a matrix of random vectors representing all the words in a document; or (2) word embeddings.
- Word embedding: (1) In-domain created using Medline; or (2) Out-of-domain using Wikipedia.
- Static versus dynamic embeddings: pre-fixed input vector values based on the collection versus allowing to adjust during the training stage.
- Hyperparameter tuning to find the optimal values: (1) Batch size, (2) Number of epochs, (3,4) activation function on convolution layer and fully connected layers, (5) dropout rate, (6) filter size, depth, (7) learning rate, (8) word representation, (9) vector size, and (10) stride.

Dataset

We used one in-domain, and one out-of-domain dataset:

- ICD9 shared task: 978 anonymised radiology reports and their corresponding ICD-9-CM codes (38 unique codes). A subset (rICD9) was created by removing reports with codes of less than 15 instances. rICD9 had 894 documents.
- IMDB movie reviews: 100,000 documents, with two classes (positive and negative).

Experimental Setup

- A multi-class classification
- CNN using Tensorflow, SVM using Scikit-learn, embeddings using Word2Vec
- For ICD9 and rICD9 we used 10-fold stratified cross-validation
- IMDB dataset had a fixed training and testing data
- Measures: Accuracy, Precision, Recall, F-score (macro-averaged)
- All the experiments were repeated 50 times and averaged
- Hyper-parameter tuning using grid search

Experiments and Results

- CNN versus conventional classifiers
  - SVM Features: normalised tf-idf
- Effect of pre-trained word vectors

Table 1: CNN with random vectors versus conventional classifiers on ICD9 dataset.

Classifier	Accuracy	Precision	Recall	F1-score
SVM	80.52	66.02	67.69	65.63
Random Forests	68.22	50.85	49.38	48.66
Logistic Regression	79.43	66.08	66.15	64.63
CNN (default)	81.55	78.93	81.55	79.05
CNN (optimal)	**83.84**	**81.44**	**83.84**	**81.55**

Findings:
- For all three datasets (ICD9, rICD9, IMDB), CNN achieved comparable or better results than conventional classifiers.

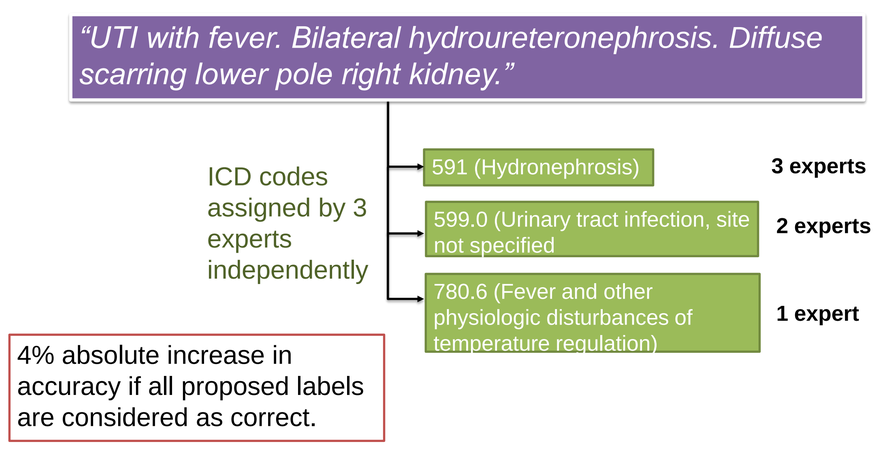
Table 2: Impact of methods of generating input vector on accuracy.

Word embedding	Vector Size	Dynamic	ICD9	rICD9	IMDB
Random embedding			81.93	86.69	87.75
Word2Vec Wikipedia	40	Yes	81.03	86.24	88.79
		No	69.75	74.90	85.02
	100	Yes	82.04	86.86	88.55
		No	75.93	81.40	86.98
	300	Yes	82.41	87.22	88.21
		No	79.34	84.94	88.14
	400	Yes	82.60	87.24	88.10
		No	80.03	85.53	88.19
Word2Vec Medline	40	Yes	81.59	87.06	**89.00**
		No	72.31	78.05	82.11
	100	Yes	82.55	**87.76**	89.00
		No	78.66	84.06	85.70
	300	Yes	**83.84**	87.45	88.58
		No	80.88	86.30	87.10
	400	Yes	82.55	87.56	88.62
		No	81.39	86.66	87.21

Findings:
- pre-trained word vectors significantly improve classification accuracy
- Dynamic word vectors almost always beat static ones on all datasets.
- For ICD9 and rICD9 word vectors using Medline were better. Domain did matter to capture the meaning of medical terminology.

Error Analysis: There were two sources of classification mistakes:

1. Multi-class classification did not take into account of multiple labels in the gold standard.
2. Companion diseases: when there are common symptoms



Conclusions

- Some of the hyperparameters such as depth are specific to a dataset and task and should be tuned each time using a CNN network. Some however, such as vector size and learning rate, are less sensitive and can be set in advance.
- In-domain embeddings (using Medline) were better than out-of-domain.
- Dynamic word embeddings worked better than static ones.
- CNNs can get comparable or in some situations better results than conventional state-of-the-art methods.

FOR FURTHER INFORMATION
Sarvnaz Karimi
e sarvnaz.karimi@csiro.au
Anthony Nguyen
e Anthony.nguyen@csiro.au

REFERENCES
1. J. Pestian, C. Brew, P. Matykiewicz, D. Hovermale, N. Johnson, K.B. Cohen, and W. Duch. 2007. A shared task involving multi-label classification of clinical free text. In Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Prague, Czech Republic, pages 97–104.
2. Y. Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, pages 1746–1751.
3. CNN figure: https://github.com/pacocp/Convolutional-Neural-Net-Tutorial

ACKNOWLEDGEMENTS
Xiang Dai was a summer student funded by Data61 Decision Sciences Program while working on this project.

THE AUSTRALIAN E-HEALTH RESEARCH CENTRE