Artificial Intelligence In Medicine 97 (2019) 79–88

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification

Imon Banerjee[a,*], Yuan Ling[b], Matthew C. Chen[c], Sadid A. Hasan[b], Curtis P. Langlotz[c], Nathaniel Moradzadeh[c], Brian Chapman[d], Timothy Amrhein[e], David Mong[f], Daniel L. Rubin[a,c], Oladimeji Farri[b], Matthew P. Lungren[c]

[a] Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA
[b] Artificial Intelligence Laboratory, Philips Research North America, Cambridge, MA, USA
[c] Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA
[d] Department of Bioinformatics, University of Utah Medical Center, UT, USA
[e] Department of Neuroradiology, Duke University School of Medicine, NC, USA
[f] Department of Radiology, Children Hospital Colorado, CO, USA

ARTICLE INFO

Keywords:
Convolutional neural network (CNN)
Recurrent neural network (RNN)
Pulmonary embolism
Text report classification
Radiology report analysis

ABSTRACT

This paper explores cutting-edge deep learning methods for information extraction from medical imaging free text reports at a multi-institutional scale and compares them to the state-of-the-art domain-specific rule-based system – PEFinder and traditional machine learning methods – SVM and Adaboost. We proposed two distinct deep learning models – (i) CNN Word – Glove, and (ii) Domain phrase attention-based hierarchical recurrent neural network (DPA-HNN), for synthesizing information on pulmonary emboli (PE) from over 7370 clinical thoracic computed tomography (CT) free-text radiology reports collected from four major healthcare centers. Our proposed DPA-HNN model encodes domain-dependent phrases into an attention mechanism and represents a radiology report through a hierarchical RNN structure composed of word-level, sentence-level and document-level representations. Experimental results suggest that the performance of the deep learning models that are trained on a single institutional dataset, are better than rule-based PEFinder on our multi-institutional test sets. The best F1 score for the presence of PE in an adult patient population was 0.99 (DPA-HNN) and for a pediatrics population was 0.99 (HNN) which shows that the deep learning models being trained on adult data, demonstrated generalizability to pediatrics population with comparable accuracy. Our work suggests feasibility of broader usage of neural network models in automated classification of multi-institutional imaging text reports for a variety of applications including evaluation of imaging utilization, imaging yield, clinical decision support tools, and as part of automated classification of large corpus for medical imaging deep learning work.

1. Introduction

Diagnostic imaging accounts for 10 percent (100 billion dollars) of annual health care costs [2]. Due to market and patient care related pressures there is a constant demand for maintaining consistent clinical diagnostic excellence in the setting of rising imaging volumes, greater imaging complexity, and a demand by clinicians for rapid results. As a result the highest demands for new technology are around solutions that drastically improve workflow while maintaining or improving quality of care, building in efficiencies for the clinician, health system, and ultimately for the patient. Consequently, technologies that can aid in automating the medical imaging workflow are in high demand and have inspired advances in machine learning methods which show promise in assisting radiologists to analyze complex imaging and text data [1,3–6]. Yet, despite the rapid exploration around new machine learning tools for use in medical imaging diagnostic tasks, a significant barrier remains for this technology to be applied at scale: the diagnostic information for the imaging studies are contained within unstructured clinician-created free text reports. So while it may be possible to acquire millions of medical images for machine learning applications, extracting the diagnostic information in those images as structured labels for machine learning model training requires a highly specialized

* Corresponding author.
E-mail address: imonb@stanford.edu (I. Banerjee).

https://doi.org/10.1016/j.artmed.2018.11.004
Received 14 November 2017; Received in revised form 6 August 2018; Accepted 13 November 2018
0933-3657/

understanding of diagnostic imaging report context, syntax, structure, and specific terminologies all unique to the radiology enterprise. One example of an important life-threatening medical condition is pulmonary embolism (PE) which is diagnosed with computed tomography imaging (CT) and relevant details stored in the medical record as free-text. CT exam for PE is an expensive test with risks of radiation exposure, contrast injection reactions, and incidental findings leading to further testing [7]. To identify imaging studies that are positive for PE, like many other radiology diagnoses, information extraction from the free-text radiology reports is needed in order to perform large scale analyses. If the hospital was able to leverage the computerized tools, it would be possible to rapidly and accurately account for the percentage of positive and negative studies, per provider/service/specialty, and use that data over time to identify opportunities for implementing or re-educating ordering providers about decision support tools for PE evaluation and reduce the number of negative examinations in a targeted approach [8].

Natural Language Processing (NLP) techniques can be a key to successfully analyzed the radiology reports to extract clinically important findings and recommendations [9,10]. NLP has already shown potential to automate the task of classifying imaging reports in a way that could inform decisions regarding medical imaging utilization and appropriateness [11–13]. Chapman et al. [14] developed an application called PEFinder based on an extension of NegEx to detect lexical cues other than negation and defined how each cue modifies a preceding or succeeding concept. Yu et al. [15] used an NLP system called Narrative Information Linear Extraction (NILE) that combines linguistic and machine learning approaches to improve identification of pulmonary embolism location and also demonstrated promising results. Well-studied early programs that have also leveraged machine learning techniques include MedLEE (Medical Language Extraction and Encoding System), which relies on controlled vocabulary and grammatical rules in order to convert free-text into a structured database [16,17]. Dang et al. [9] processed 1059 radiology reports with Lexicon Mediated Entropy Reduction (LEXIMER) to identify the reports that include clinically important findings and recommendations for subsequent action. More recently, Sohn et al. [18] used machine learning to identify patients with abdominal aortic aneurysms. Other work has used various machine learning techniques such as support vector machines to classify MRI free text reports [11].

Recent advances in computing power and machine learning techniques have prompted the rise of a new generation of machine learning technique called convolutional neural networks (CNNs) that have successfully been applied to image recognition tasks, including facial recognition, object detection, and image classification [6,19]. CNNs and other similar deep learning architectures have significant implications for diagnostic imaging [5,20,4,21,22], but rather than being limited to image recognition tasks, it is possible that CNNs can also be applied to text data. Character-based and very deep CNN models are generally applied for text classification tasks with very large datasets [23,24], where complex neural network architectures play a major role to capture all generic features of the data while disregarding domain-dependent word-level features. Instead, we used a shallow CNN model proposed by Kim [25] as it has been shown to obtain state-of-the-art results on multiple benchmark datasets using pre-trained word vectors. The adaptation of this CNN model combined with GloVe word embeddings [26] served as a strong baseline for our experiments.

Another class of neural networks architectures, recurrent neural networks (RNNs), has recently gained a lot of attention from researchers for modeling sequences and have been shown to perform well in solving various NLP tasks because of their ability to deal with variable-length input and output [27]. RNNs have also been used to address tasks in the medical domain [28]. To the best of our knowledge, there is no existing work that applies RNN-based models to classify free text medical imaging reports based on PE categorical measures and analyze the comparative performance against the CNN model.

In this study, we present a comparative analysis of distinct deep learning techniques (CNN and RNN) in classifying a database of free text chest (contrast enhanced) CT reports based on the presence of pulmonary embolism. We train the models on a small sub-set of Stanford training set (2512 reports) and present the performance on four distinct institutional datasets. Our results show that deep learning models can be transferable to completely different institutional dataset only being trained on a single institutional data. The main contributions of this paper can be summarized as follows:

1. Methodological contributions
   (a) We proposed a novel domain phrase attention-based hierarchical recurrent neural network model (**DPA-HNN**) to classify the radiology reports. The model encodes radiology reports through an intuitive hierarchical structure composed of word-level, sentence-level and document-level representations. Compared to a simple word-level attention mechanism, domain phrase attention suits the radiology domain better as the radiologists traditionally follow a domain-specific note writing style and some domain phrases occur frequently in reports.
   (b) We also proposed the use of a well-known state-of-the-art convolutional neural network model [25] for classifying radiology reports corpus based on the presence of PE factors. We integrated the CNN model with pre-trained Glove vectors: (**CNN Word–Glove**) [26] for learning the semantics of the whole radiology reports from a relatively small annotated training set.
2. Comparative analysis
   (a) We analyzed the comparative performance of the two proposed deep learning architectures (**CNN Word–Glove** and **DPA-HNN**) in the context of free-text radiology reports labeling according to the presence of PE factors.
   (b) We compared the performance of the proposed deep learning models with a state-of-the-art representative of well feature engineered NLP approaches – **PEFinder** [14], as well as with traditional machine learning models – **Support Vector Machine (SVM)** and **Adaboost** with bag-of-words features which represent the class of traditional methods with limited feature engineering.
   (c) We also compared the performance of our proposed **DPA-HNN** model with two related variations of hierarchical RNN models [29,30]: (1) **HNN**: Hierarchical Neural Network, and (2) **A-HNN**: Attention-based Hierarchical Neural Network.

In the following sections, we present a detailed description of the dataset, materials and method, followed by the experimental setup, results and discussion. Finally, we discuss some limitations in Section 6 and conclude the paper in Section 7.

2. Dataset

Corpora – We obtained 227,809 radiology reports from Duke University Medical Center, 117,816 reports from Stanford University Medical Center, and 12,091 reports from Colorado Childrens Hospital for contrast-enhanced CT examinations of the chest performed between January 1, 1998 and January 1, 2016.[1]

The Impression sections were extracted from the original reports by searching for the “impression” keyword, signifying the beginning of the impression section, and using simple heuristics to determine the end of the section. We structured our heuristics to extract more details on the impressions rather than filtering out the noise, which inevitably

[1] All examination reports were de-identified in a fully HIPAA- compliant manner, and acquisition and processing of data was approved by the Institutional Review Board (IRB) of the institution where the reports were obtained.

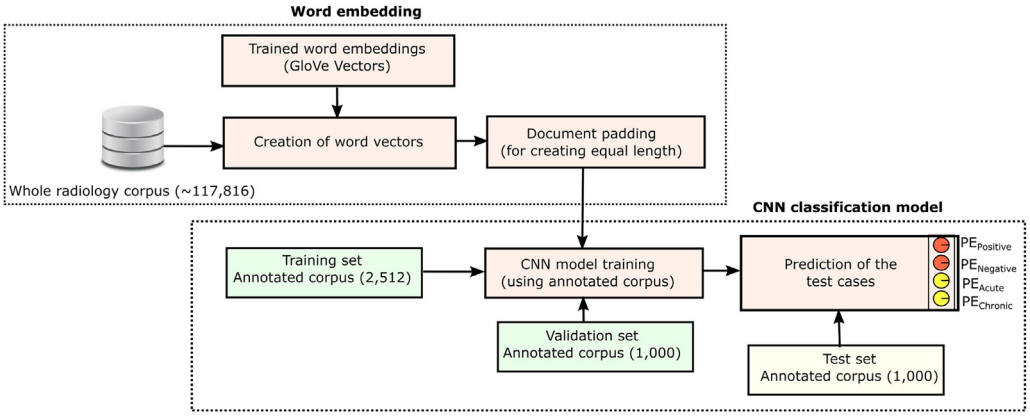

Fig. 1. CNN Word – Glove Architecture, showing the training, validation and testing situation for Stanford dataset.

resulted in some parses being more verbose than necessary.

Training Set and Test Set – The model training was performed using the Stanford University radiology report corpus of 117,816 reports of adult patients; from these, 4512 reports were randomly selected and annotated by three experienced radiologists who read the whole report and assigned binary class labels. Two binary labels were assigned to individual reports – these were defined according to two categorical measures of PE: (1) PE present/absent; (ii) PE acute/chronic. The annotators assigned binary labels based on the following assumption; if a PE was definitely present in the report it was annotated as positive, or else annotated as negative. Chronicity was labeled as either acute or chronic based on the text description. In the setting of acute on chronic, or "mixed" chronicity, the report was to be labeled as acute to reduce the false negative rate. 4512 annotated reports were randomly divided into training (2512 reports), validation (1000 reports), and testing (1000 reports) sets.

External Test Set – External test data sets were created from the Duke and Colorado reports: 1000 reports from the Duke medical center corpus and 1000 reports from the Colorado Children's corpus were randomly selected and annotated by two experienced radiologists following the same annotation protocol as adopted during the Stanford dataset annotation (above).

Inter-rater reliability among the three raters for annotation of Stanford, Duke and Colorado datasets was calculated using Fleiss' kappa [31] – a generalization of Cohen's Kappa for more than two raters. The raters were highly consistent for two categories, "PE positive" and "PE Acute", with kappa scores of 0.959 and 0.969 respectively. The kappa metric reflects the initial ratings, following which a single radiologist resolved all conflicting cases manually for preparing the ground truth labels.

External Validation Set (UPMC dataset) – We obtained 858 reports from University of Pittsburgh medical center (UPMC) that were originally used to develop PEFinder classifiers. The reports were all de-identified in a fully HIPAA-compliant manner. Three medical students independently annotated the reports with five distinct states and binary annotations for each document were obtained from the user annotations as follows: "probably positive" and "definitely positive" were collapsed to positive; and "probably negative", "indeterminate", and "definitely negative" were considered negative. After collapsing annotations to binary values, the authors generated labels for each report by a majority vote of the annotators. We refer the reader to the original paper [14] for a detailed description of the UPMC dataset.

3. Materials and methods

3.1. CNN Word – Glove Architecture

CNNs have been applied to natural language modeling by convolving a filter across a fixed length word representation as input into the model, and final classification is obtained through a series of matrix multiplications and non-linear function mappings [25]. Such a model can automatically learn the semantics of local structure of text, such as phrases, while avoiding having to keep memory over the entire sequence of words for a given sequence. However these models have empirically been shown to require larger training sets. To the downstream performance of a CNN, the efficient vectorized representation of words and/or phrases are important factors. Neural network based word embedding approach [32] can learn a vector which enables a model to perform well on the task of relating a word to its context in a given window. For example, Glove [26] combines components of skip-gram and continuous bag of words model by incorporating local window based information with document level statistics in their word vectors.

The CNN Word – Glove model used for this task was initially proposed by Kim et al. [25]. The CNN model with only one layer of convolution was trained on top of GloVe vectors obtained from an unsupervised neural network model trained on 42 billion words from text extracted from crawling web pages. Since the GloVe vector is created on very generic documents, it is expected that some specific clinical terms never occur in the vocabulary. All words that do not match the Glove corpus are converted to an unknown token, which is initialized to a random vector. The input embeddings are then padded with zeros at the end to ensure that all training examples have the same size. For this study, we choose an input length of 300 words, which is the 99.5 percentile for all impressions in our dataset. The examples containing longer sequences are truncated at the tail end. Such truncation of the report length has a minor effect on the model performance as radiologists usually document the key findings earlier in the impression section (within the first 2 or 3 sentences).

Fig. 1 shows the high level schema of the CNN model. First, we generated the vector representations of all the words that belong to the Impression section of the Stanford corpus using the pre-trained GloVe model. The 115,816 Stanford report (excluding 2000 validation and test reports) impressions each contained 124 words and symbols on average with 31,470 unique tokens across all impressions. A preprocessing step was included to treat all punctuation as separate word tokens. The vector representation of the words are processed through a convolutional layer with 200 filters, a window size of 10 words, and a rectified linear unit as the non-linearity. This layer is followed by a max pooling layer which takes the max across all values of a given filter. Dropout, a regularization method to prevent overfitting of the data, is then used on the output. The result is fed into a fully connected output layer followed by a sigmoid layer to convert the raw scores to probabilities. The two prediction tasks are trained on two different CNN models – (1) $\{PE_{Positive}, PE_{Negative}\}$ and (2) $\{PE_{Acute}, PE_{Chronic}\}$.

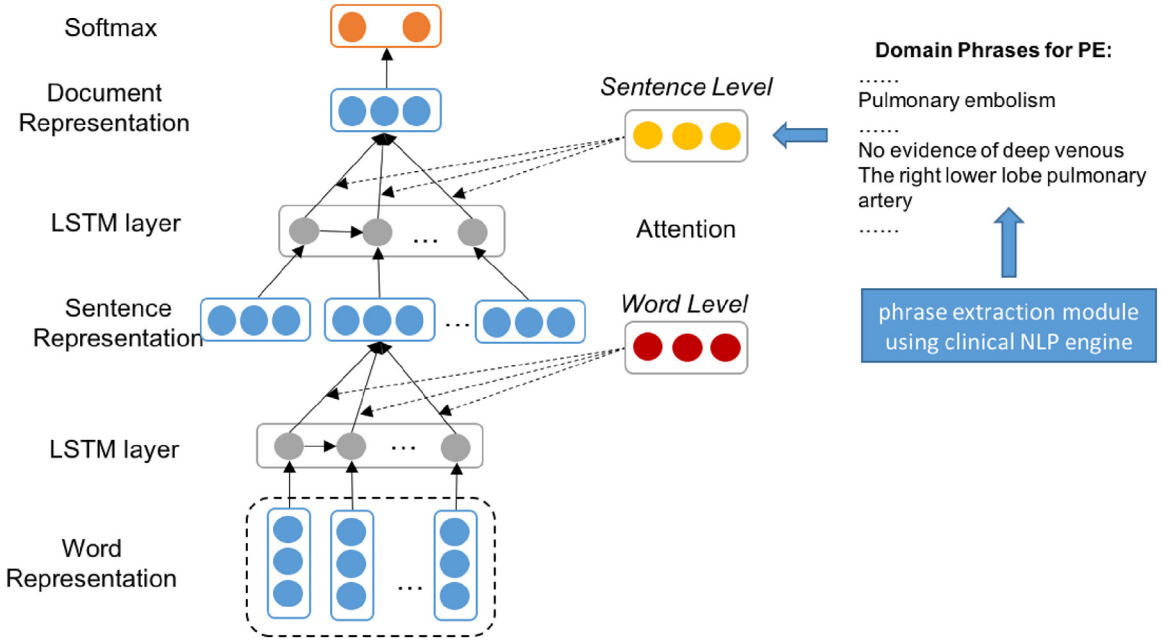

Fig. 2. Domain Phrase Attention-based Hierarchical Neural Network (DPA-HNN) architecture.

3.2. Domain Phrase Attention-based Hierarchical Recurrent Neural Network (DPA-HNN) model

The semantics of a radiology report can be modeled through a hierarchical structure composed of word-level, sentence-level and document-level representations. Recent work on document-level sentiment classification showed promising performance by taking similar hierarchical structures into consideration [29,30].

In contrast to the existing work [30] that encodes global user preference and product characteristics via a word-level user-product attention mechanism for a product-oriented sentiment classification task, we propose a novel Domain Phrase Attention-based Hierarchical Neural Network (DPA-HNN) model (Fig. 2) by encoding clinical domain-dependent phrases into a sentence-level attention mechanism and representing a radiology report through a hierarchical structure composed of word-level, sentence-level and document-level representations. Compared to a general word-level attention mechanism, our domain phrase attention applied at the sentence-level plays a more important role in classifying radiology reports as radiologists traditionally follow a domain-specific note writing style. Moreover, some domain phrases occur frequently in radiology documents, justifying the need to propose our DPA-HNN model with a domain phrase attention mechanism.

Although RNN is theoretically a powerful model to encode sequential information, in practice it often suffers from the vanishing/exploding gradient problems while learning long-range dependencies [33]. LSTM [34] and GRU [35] networks are known to be successful remedies to these problems. We chose LSTMs for our experiments because there had not been shown any distinguishable difference between the performance of a LSTM unit and a GRU unit in the literature [36,37]. We use LSTM as our hidden layer activation unit to model the semantic representations of sentences and documents. Typically, each cell in a LSTM is computed as follows:

$$X = \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{1}$$

$$f_t = \sigma(W_f \cdot X + b_f) \tag{2}$$

$$i_t = \sigma(W_i \cdot X + b_i) \tag{3}$$

$$o_t = \sigma(W_o \cdot X + b_o) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot X + b_c) \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where W_i, W_f, $W_o \in \mathbb{R}^{d \times 2d}$ are the weight matrices and b_i, b_f, $b_o \in \mathbb{R}^d$ are the biases of LSTM to be learned during training, parameterization, and transformations of the input, forget and output gates, respectively. σ is the sigmoid function and $\odot$ stands for element-wise multiplication. x_t is the input of a LSTM cell unit and h_t represents the hidden state at time t.

Assume that a document has L number of sentences, where each sentence s_i contains T_i words. w_{it} with $t \in [1, T]$ represents the words in the ith sentence. For word-level computations, x_t represents the word embedding vectors w_t. The first hidden layer vectors h_{it} with $t \in [1, T]$ are used to represent a sentence. For sentence-level computations, x_t represents the sentence embedding vectors s_i. The hidden layer vectors h_i with $i \in [1, L]$ are used to represent a document in this case.

We regard the last hidden layer as the representation of a document and place a softmax layer on top of it to predict the class labels e.g. $\{PE_{Positive}, PE_{Negative}\}$ or $\{PE_{Acute}, PE_{Chronic}\}$ for the radiology report. In other words, the output layer (softmax) of DPA-HNN considers two possible labels. As such, we build two separate models to classify PE Positive/Negative and PE acute/chronic. Considering h^* as the final representation of a radiology report, the softmax layer can be formulated as:

$$y = \text{softmax}(W_s h^* + b_s), \tag{7}$$

where W_s and b_s are the parameters of the softmax layer. We use the negative log likelihood of the correct labels as our training loss function:

$$L = -\sum_d \log y_{\text{dj}}, \tag{8}$$

where j is the label of a document d.

Domain Phrase Attention Mechanism: In the settings of hierarchical neural networks without attention, the hidden states are fed to an average pooling layer to obtain the sentence representation and the final document representation. For example, the final feature representation of a radiology report can be computed as:

$$h^* = \sum_{i \in [1,L]}^{i} h_i \tag{9}$$

We propose a domain phrase attention mechanism to capture the most important part of a document by considering domain phrases at the sentence level. It is reasonable to reward sentences that are clues to correctly classify a document as indicated in [29]. Hence, we pay extra attention to domain phrases if they are present in a sentence. We encode each domain phrase as continuous and real-valued vectors $p \in \mathbb{R}^d$, which are randomly initialized. This yields:

$$u_i = \tanh(W_s h_i + W_{\text{dp}} p + b_s) \tag{10}$$

$$\alpha_i = \frac{\exp(u_i^T u_s)}{\sum_i \exp(u_i^T u_s)} \tag{11}$$

$$h^* = \sum_{i \in [1,L]}^{i} \alpha_i h_i \tag{12}$$

where W_s and W_{dp} are projection parameters and b_s is the bias parameter to be learned during training. Our domain phrase (DP) generation algorithm works as follows. First, we extract clinical concepts from the radiology reports based on SNOMED-CT ontology [38] using a hybrid clinical NLP engine [39]. Then, our algorithm uses the extracted clinical concepts to form the DPs based on two heuristics: (1) we combine the consecutive clinical concepts occurred in the same sentence as one domain phrase. For example, in the sentence, *"three small low attenuation lesions within the liver, which are too small to characterize"*, "low attenuation" and "lesions" are tagged as two separate clinical concepts by the clinical NLP engine. Since they are consecutive words in the sentence, we regard them as the domain phrase "low attenuation lesions". In other words, the clinical concepts should be consecutively present in a sentence in order to be the part of a domain phrase, and (2) we obtain shallow parsing annotation for the text. If one token is tagged as "B-NP", then together with its following tokens tagged as "I-NP", "B-PP", and "B-NP", the entire chunk will be considered as one phrase. If any token in this phrase is annotated as a clinical concept by the clinical NLP engine, we regard the phrase as a domain phrase. For example, the shallow parsing annotation for "The right lower lobe pulmonary artery" is "The[B-NP] right[I-NP] lower[I-NP] lobe[I-NP] pulmonary[I-NP] artery[I-NP]". Here, the token "The" is tagged as "B-NP", and its following tokens are tagged as "I-NP" while "right lower lobe pulmonary artery" is tagged as a clinical concept. Therefore, the whole chunk "The right lower lobe pulmonary artery" is considered as one domain phrase. The list of Domain Phrases (DPs) is generated from the Stanford training set. The total number of DPs in the list is 343. The average number of words in DPs is ≈4. We display the frequency distribution of word counts in DPs with few examples in Fig. 3.

4. Experiments and evaluation

4.1. Models for comparison

For legitimate comparison, we experimented with the following seven models.

1. **PEFinder**: a state-of-the-art rule-based method for PE radiology text report classification [14].
2. **Support Vector Machine with radial basis kernel function (SVM)**: a popular machine learning method for binary classification; we used sparse bag-of-words as feature vector.
3. **Adaptive Boosting (Adaboost)**: a machine learning algorithm that uses selective boosting on top of the bag-of-words feature vector.
4. **CNN Word-Glove**: CNN model built on top of GloVe vectors (Section 3.1).
5. **Hierarchical Recurrent Neural Network (HNN)**: RNN baseline model without any attention mechanism (Section 3.2).
6. **Attention-based Hierarchical Neural Network (A-HNN)**: RNN-based model with the general word-level attention mechanism using random initialization (Section 3.2).
7. **Domain Phrase Attention-based Hierarchical Neural Network (DPA-HNN)**: our proposed model with Domain Phrase (DP) attention mechanism (Section 3.2).

4.2. Model training

The CNN model has been trained on the vector embedding of 2512 reports from the Stanford corpus. We used dropout rate (p) of 0.5, l_2 constraint (s) of 3, and mini-batch size of 50. The network weight update has been done through stochastic gradient descent (SGD) over shuffled mini-batches with the Adadelta update rule. Weights for the model were initialized randomly using the Xavier initialization method to roughly keep the scale of the gradients the same across all layers.

For RNN-based models, we used the same pre-trained word embeddings as used by CNN model to represent the words. Each RNN was built with 300 hidden units (i.e. LSTMs as discussed in Section 3.2). Similar to the CNN models, our RNN-based models were also trained with the SGD algorithm with update direction computed using a batch size of 32 via the Adadelta update rule. We trained these models for 200 epochs and 400 iterations of validation. The vector size for each granularity i.e. word, sentence, and document were 300 dimensions.

Using the Stanford validation set (1000 reports), we optimized the hyperparameters such as dropout rate, number of epochs, and batch size by selecting a best trade-off between the validation accuracy and the computational memory requirement.

4.3. Classification performance

We evaluate the performance of our models on all the data sets (Stanford, UPMC, Colorado Childrens and Duke) using the following metrics: Precision, Recall, F1 value, and Area Under the Curve (AUC).

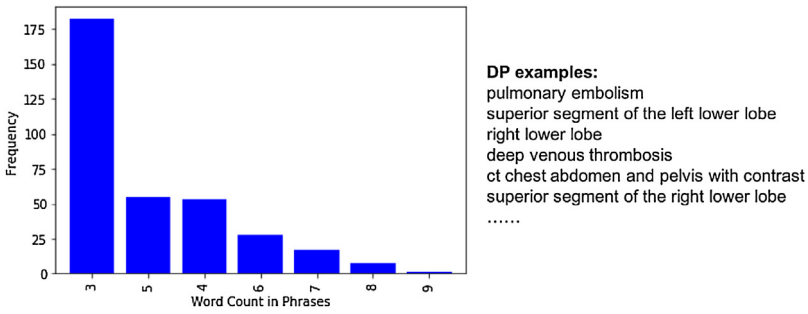

Fig. 3. The frequency distribution of word count in DPs and DP examples.

In order to convert the predicted probability values of neural models to binary class labels, we determine an optimal cutoff threshold of the probability of the positive class by maximizing: Precision(t_i) + Recall (t_i) for all the thresholds (t_i) between 0 and 1. The classification results from all methods are displayed in Table 2 .

For Stanford test set, DPA-HNN has the best scores on all evaluation metrics for both PE Positive/Negative and PE Acute/Chronic classifications. Compared with HNN and A-HNN, DPA-HNN encodes domain phrase attention and improves the performance. All the improvements of DPA-HNN model over HNN model and A-HNN model were found to be statistically significant ($p < 0.05$). From the results we can see that, overall, neural network-based methods have better performance than the classic PEFinder, SVM and Adaboost methods in term of F1 and AUC scores on Stanford test set.

On UPMC dataset, DPA-HNN has the best precision scores for both tasks, while CNN has the best AUC scores. On Duke test set, DPA-HNN has the best AUC scores for both tasks, while CNN has the best precision and F1 scores. On Colorado Childrens test set, HNN has the best scores on all evaluation metrics for PE Positive/Negative classification, while not performing well on PE Acute/Chronic classification.

Overall, DPA-HNN model shows performance improvement on Stanford test set, and partially on UPMC dataset and Duke test set. However, the performance on Colorado Childrens test set is not the best, which is reasonable because DPA-HNN and other neural network-based methods are trained on Stanford dataset which are mostly adult patients compared to the specific pediatrics population of Colorado Childrens. Further analyses revealed that the external datasets (UPMC dataset, Duke test set, Colorado Childrens test set) have varying distributions of average number of sentences and domain phrases in a document. The distributions and statistics are displayed in Fig. 4 and Table 1 . DPs play an important role in our proposed DPA-HNN model. For example, the Colorado data has an average number of 1.2 DPs in a document, which is much lower than the average number of 3.5 in Stanford test data, while the percentage of documents without DPs for Colorado data is much lower than the Stanford test data – that could be the reason why the DPA-HNN model trained on Stanford dataset does not work equally well on Colorado data. However, the average number of sentences in a document for this dataset is 6.2, which is very close to Stanford data of 6.4. Since the HNN model does not rely on DPs, it performed well on the Colorado test set, but the DPA-HNN model suffered due to lack of DPs in this test set.

From the results, we can also observe that in general the evaluation scores of the PE Acute/Chronic classification task are lower than the PE Positive/Negative classification task denoting the complexity of the former compared to the later task.

4.4. Qualitative analysis

To better understand what information a deep learning model in a natural language task is using to make its decisions, methods have been developed to visualize the impact of input words on the output decision [40]. For the CNN model, we used one such method called sensitivity analysis, which takes the partial derivative of the loss function with respect to each input variable. Since our input variables represent word vectors in this case, in order to get a importance score for a particular word we take the L1 norm of the vector of partial derivatives.

The heat maps (in Fig. 5) represent result of sensitivity analysis of input for a positive example (left) and a negative example (right) where the CNN model predicted correctly from the Stanford test set. The result on the left shows the text of a report that is positive for all two prediction classes. We can see that there is little importance (light colored) placed on the long document with the exception of the relevant phrase (dark colored), which contains information on the PE status. Similarly on the right is an example of a report that is negative for PE and all other prediction classes. We can see in this case that the model places the most emphasis (dark color) on the first sentence which clearly states that there is no evidence of PE. From these qualitative results we note that the network is able to parse through large sequences of text to focus on phrases that are relevant to classification simply from the document level annotation.

In Fig. 6, we present two examples of misclassified reports using the CNN model. In the first case (left) the model predicted PE negative, but the ground truth was PE positive. In the second case (right) the model predicted a PE positive, but the ground truth was PE negative. It is apparent in the example on the left that the CNN model was unable to infer that PE was present with the statement that there was decreased volume of clot. In the example on the right, the word “no” seemed to be separated by an abnormally long string of words prior to “pulmonary embolism” leading to a classification error. Such errors are reasonable as CNNs are only able to capture local information from the data to perform well in classification tasks [25].

However, our RNN based models A-HNN and DPA-HNN were able to predict the classes correctly for the same reports where CNN failed to classify (shown in Fig. 6). This is shown via the hierarchical attention-

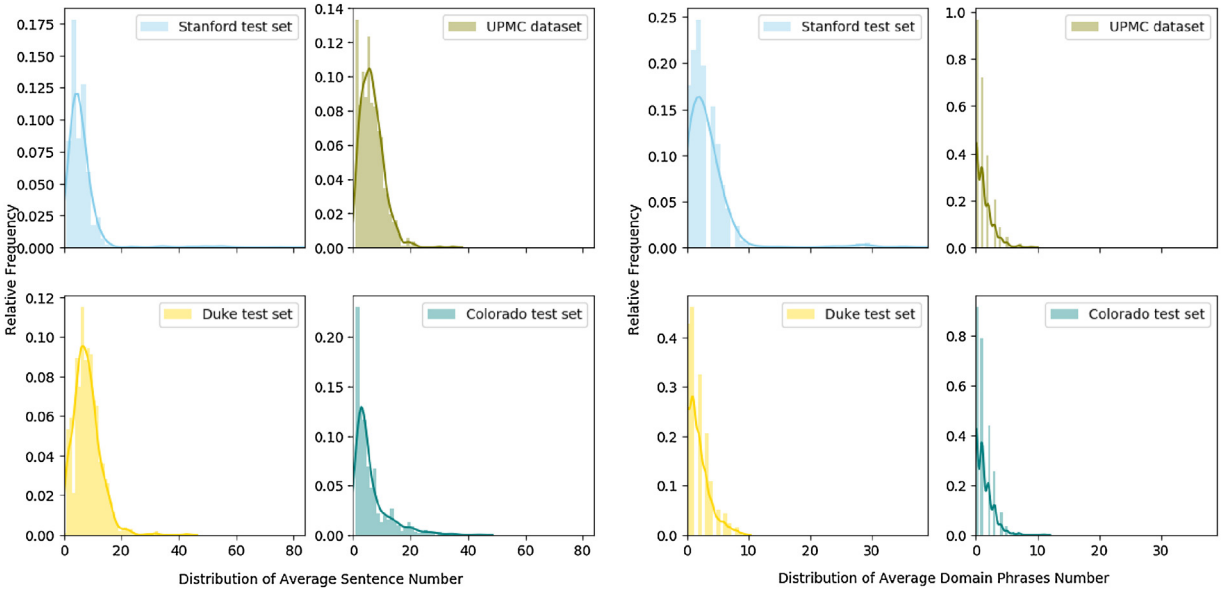

Fig. 4. Distribution of average number of sentences in a document (left) and distribution of average number of domain phrases in a document (right).

Table 1
Statistics of average number of sentences/DPs in a document across 4 datasets.

Dataset	Stanford test set	UPMC dataset	Duke test set	Colorado test set
Average number of sentences in document	6.4	6.7	8.0	6.2
Average number of DPs in document	3.5	1.2	1.8	1.2
Percentage of Documents without DPs	13.7%	39.6%	25.7%	35.9%

Table 2
Comparative performance measures. Boldface numbers represent column-wise superior performance achievement on a particular dataset.

	PE Positive/Negative				PE Acute/Chronic			
	P	R	F1	AUC	P	R	F1	AUC
Stanford test set								
PEFinder	0.87	0.90	0.89	N/A	0.91	0.91	0.91	N/A
SVM	0.96	0.96	0.95	0.85	0.93	**0.97**	0.95	0.93
Adaboost	0.98	0.98	0.98	0.95	0.97	**0.97**	0.97	0.92
CNN	0.92	0.97	0.95	**0.99**	0.91	0.91	0.91	0.99
HNN	0.94	0.96	0.95	0.98	0.92	**0.97**	0.94	0.95
A-HNN	**0.99**	0.96	0.97	0.98	0.97	0.96	0.97	0.99
DPA-HNN	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.97**	**0.98**	**0.99**
UPMC dataset								
PEFinder	**0.87**	**0.96**	**0.91**	N/A	**0.91**	**0.95**	**0.93**	N/A
SVM	0.71	0.70	0.58	0.83	0.49	0.70	0.57	0.86
Adaboost	0.83	0.84	0.83	0.90	0.84	0.83	0.82	0.91
CNN	0.76	0.95	0.85	**0.97**	0.82	0.92	0.87	**0.97**
HNN	0.82	0.74	0.77	0.90	0.88	0.73	0.75	0.88
A-HNN	0.82	0.74	0.77	0.90	0.88	0.72	0.75	0.88
DPA-HNN	**0.87**	0.87	0.87	0.95	**0.91**	0.90	0.90	0.94
Duke test set								
PEFinder	0.84	0.99	0.90	N/A	0.86	**0.99**	0.92	N/A
SVM	0.95	**0.98**	0.96	**0.94**	0.96	0.98	0.97	0.95
Adaboost	0.97	0.96	**0.97**	0.85	0.97	0.97	0.97	0.77
CNN	**0.98**	0.97	**0.97**	0.90	**0.98**	0.98	**0.98**	0.91
HNN	0.93	0.79	0.85	0.92	0.95	0.74	0.81	0.91
A-HNN	0.90	0.83	0.86	0.91	0.89	0.79	0.83	0.88
DPA-HNN	0.94	0.81	0.86	**0.94**	0.90	0.95	0.92	**0.99**
Colorado Childrens test set								
PEFinder	0.69	0.79	0.73	N/A	0.66	0.87	0.72	N/A
SVM	0.98	**0.99**	**0.99**	0.91	**0.99**	**0.99**	**0.99**	0.94
Adaboost	**0.99**	0.98	**0.99**	0.97	**0.99**	0.99	**0.99**	0.98
CNN	**0.99**	0.95	0.96	0.97	**0.99**	0.98	0.98	**0.99**
HNN	**0.99**	**0.99**	**0.99**	**0.99**	0.56	0.74	0.60	0.81
A-HNN	0.87	0.80	0.83	0.98	0.60	0.87	0.66	0.77
DPA-HNN	0.80	0.80	0.80	0.93	0.71	0.87	0.77	0.98

based visualization results (Fig. 7) where the weight value for each word in the sentence is obtained from the weights learned from word-level attention for the first LSTM layer. The weight value for each sentence in the report is obtained from the weights learned from sentence-level attention for the second LSTM layer. With the weights for both word-level and sentence-level attention, we can see that different sentences play different roles in a report and different words play different roles in each sentence toward the final classification of a radiology report. In the negative example (right) of Fig. 7, the sentence 3: "*there is no filling defect in the central pulmonary artery to indicate a pulmonary embolism*" has the highest weight. The word "*embolism*" has the highest weight in this sentence, and the word "*no*" has the second highest weight in this sentence. In the positive example (left), the sentence 1: "*impression 1 interval decrease in the volume of clot within the pulmonary arterial system*" has the highest weight. The word "*clot*" has the highest weight in this sentence. The success of RNNs in classifying these reports correctly is understandable. Because, unlike CNNs, RNNs are able to capture global context from the data by considering long-term dependency among words [33].

We further contextualize the results as confusion matrix for DPA-HNN on the Stanford test set in Fig. 8. The confusion matrix is normalized class-wise to meaningfully represent the performance of the model. The *X*-axis represents gold-standard results, and the *Y*-axis represents predicted results from our DPA-HNN model. We can see that the false positive (top right part) and the false negative (down left part) rates are very low for both PE Positive/Negative classification and PE Acute/Chronic classification. In fact, for both classifications, only two cases are misclassified (sample shown in Fig. 9). These errors were evaluated and found to be related to conflicting and skeptical language in the impression. For example, in Fig. 9 (left), the impression clearly states "no definite pulmonary embolus", however, shortly thereafter the report went on to suggest "artifact vs possible pulmonary embolus in the right upper lobe" and recommended an additional imaging test. In the other example in Fig. 9 (right), the model focused on the word "subacute" to predict the report as chronic.

5. Discussion

The increasing availability of computational methods to process vast amounts of unstructured information makes it possible to derive insights from large repositories of narrative medical data [41,42]. However, the current state-of-the-art method to classify Radiology reports based on the presence or absence of a diagnosis of PE, called PEFinder, is a purely rule-based system and may not be easily scaled to classify/annotate multi-institutional free-text reports with varying narrative styles. This restricts automated interpretation of large volumes of free text reports for creating solutions to support clinical processes such as image interpretation and auto-reporting in radiology. We report on the performance of machine learning models leveraging several neural network architectures to determine the PE status from

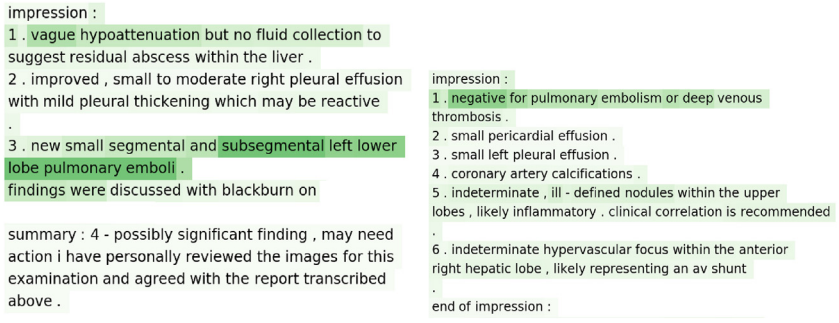

Fig. 5. Results of sensitivity analysis of CNN model: cases with correct prediction – positive (left) and negative (right). The darker the color, the more weight there is.

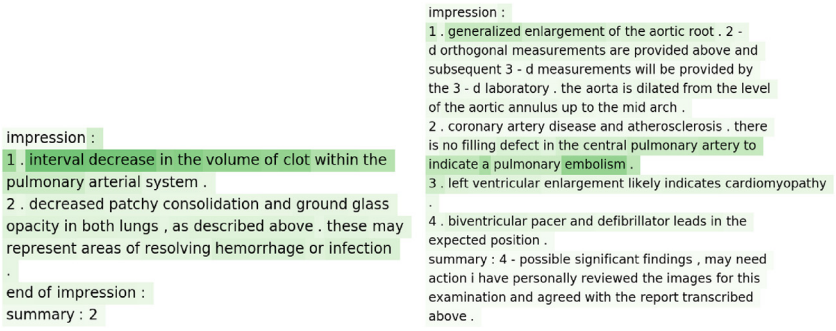

Fig. 6. Results of sensitivity analysis of CNN model: cases with incorrect prediction – positive (left, CNN model predicted negative) and negative (right, CNN model predicted positive). The darker the color, the more weight there is.

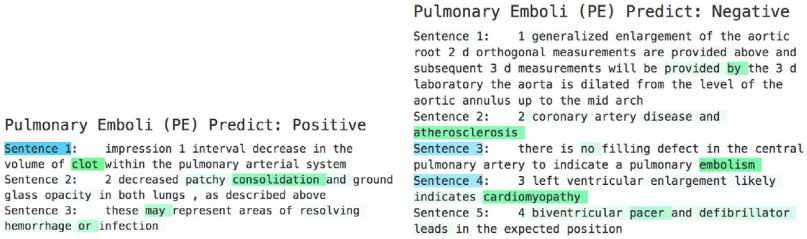

Fig. 7. Sentence-level and word-level attention of the RNN model: cases where CNN failed to classify as shown in Fig. 6. The **blue color** represents sentence-level importance and the green color stands for token-level importance; the darker the color, the more weight there is.

unstructured CT reports. We compared the neural network models with the best available published feature-engineered model. To the best of our knowledge, our work is the first to compare several neural network techniques to evaluate classification performance on free text medical imaging reports on both intra- and inter-institutional data, and compare to the current NLP rule-based gold standard, PEFinder.

Our neural network models performed equivalently or better than PEFinder and existing machine learning models. Models such as MedLEE, cTAKES, NILE, and others require named entity recognition libraries, term definitions, parsed phrases, matching, etc., and a great deal of effort is required to create these resources – representing, in certain cases, decades of prior work. By contrast, our CNN and RNN based neural network models were mainly developed *without* hand-curating any semantic input. Instead, our neural network-based models were simply trained on a small sample of document-level classified reports and *rapidly* achieved the performance levels of the best available NLP tools in the field. This suggests that neural networks may be extremely powerful for classification/ annotation of large volumes of free text reports and can achieve optimal performance without any of the previously relied upon grammatical feature definitions, concept codes, or pre-defined terms [14,15,43].

More recently, the surge in medical imaging classification and computer vision in radiology has led to a demand for high-fidelity labeled medical images; neural network models, such as the ones demonstrated here, can serve to provide highly structured labels for medical images useful for deep learning techniques. One of the most exciting opportunities that deep learning classification of imaging text offers is the ability to aid in training and modeling in computer vision projects for which a large corpus of annotated imaging data is needed (and the free-text narrative report style is poorly suited) [44,45]. Medical images are usually saved with accompanying radiology reports, and leveraging the natural language information for image analysis has great potential. For example, imaging studies from a clinical Picture Archiving and Communication Systems (PACS) can be automatically annotated by using our models to analyze the corresponding radiology reports. We can unleash the full capacity of deep learning for analyzing large volumes of medical images by automating the data collection and annotation process. Moreover, a sustainable system can be developed which allows the annotated image data to be continuously updated, shared, and integrated within a learning health care system.

Application of the neural network methodologies described in this work to other imaging free-text report annotation tasks would be both rapid and scalable using domain adaptation and transfer learning algorithms as shown in the literature [46–48]. To examine the

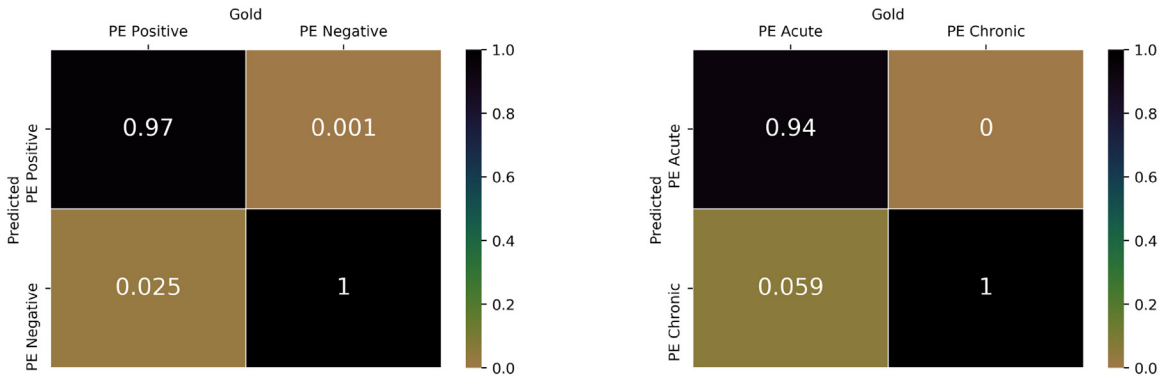

Fig. 8. Normalized confusion matrices for results from DPA-HNN model. Darker color represents a higher value.

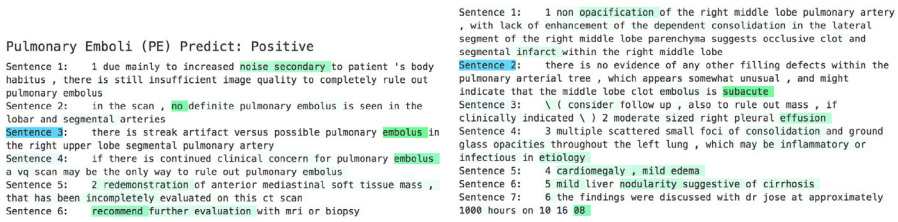

Fig. 9. DPA-HNN misclassified cases – PE negative but predicted positive (left) and PE Acute but predicted Chronic (right). The **blue color** represents sentence-level importance and the **green color** stands for token-level importance; the darker the color, the more weight there is.

generalization of our neural network models we also used imaging reports from other institutions as test sets. Our neural network models (trained on Stanford data) were tested on four different test sets. The rules of PEFinder were developed using reports from the UPMC test set, and so this provided an optimal test case. There is no significant difference in performance between our neural network models and PEFinder on the UPMC dataset. Our models were able to generalize to new reports not included in the training data and on reports from independent institutions, our models performed better than a rule-based NLP system and traditional machine learning models. The mild performance degradations noticed for the models applied on other institutional datasets except Stanford is still clinically acceptable with the lowest AUC of 0.93 for DPA-HNN. However, to achieve additional performance improvements, institutions can tune our trained models further based on their local datasets. Moreover, the hospitals can leverage the scalable deep learning tools demonstrated in this work on large volumes of data to identify trends and provide metrics around which to build performance and quality improvement programs, including the use of clinical decision support systems for PE imaging assessment, something that is mandated by congressional legislation beginning next year [49]. Moreover, the neural networks used for NLP tasks have tremendous value in many applications such as generating systems for radiology case prioritization based on report analysis, patient cohort generation, eligibility screening for clinical trials, and in radiology clinical decision support to manage imaging utilization or use yield as a metric. For example, understanding the rate of negative studies could serve as an indirect marker of utilization appropriateness and guide clinical decision making [50,51,8,52–55].

Overall, our work contributes to advancements in automated radiology report classification for PE leveraging a minimal amount of labeled examples, and offers opportunity to generate large volumes of annotated reports that would drive the development of clinical decision support systems towards improved imaging utilization, clinical research and automatic medical image interpretation for PE and other critical conditions.

6. Limitations

We acknowledge that our neural network classifiers are yet to achieve perfect scores. However, the accuracy measures are either superior or comparable to other published NLP studies. For instance, although there were 10 misclassifications out of the 1000 test cases by the CNN model, the classification errors are rather difficult to explore in neural networks, as they are often a “black box”. However, the visualization we generated helped us understand the source of errors (Fig. 6). The most common source of errors was the lack of a direct mention of the existence or absence of PE and limited documentation due to insufficient image quality; instead, inference was needed based on context. As shown in Fig. 7, our RNN based model correctly predicted the classes and located the most important sentences in the reports, but it is still hard to generalize how the model made the inference. We found only one example case for PE Positive/Negative classification, where CNN correctly predicted it positive, but RNN predicted it negative. However, it was not apparent how CNN was able to predict this case correctly based on the generated heat map. As such, many of these errors require subtle reasoning to reach the correct conclusion, which may be a limitation to the architecture of our models, in addition to training constraints posed by the size of our datasets.

7. Conclusion

We proposed a novel domain phrase attention-based hierarchical recurrent neural network model (**DPA-HNN**) that can accurately classify free text chest CT reports into pre-defined PE related criteria. We adopted a well-known state-of-the-art convolutional neural network model [25] for radiology domain and integrated the model with pre-trained Glove vectors. For both models, we demonstrated both intra- and cross-institutional fidelity when compared to other more laborious NLP methods. Our results suggest feasibility of CNNs and RNNs in automated classification of imaging text reports and support the application of these techniques at scale in classifying free text imaging reports for various use cases including radiology patient prioritization, cohort generation for clinical research, eligibility screening for clinical trials, and assessing imaging utilization. These approaches may also have impact on other important research areas such as large scale generation of labeled data for generating computer vision models for automatic medical image interpretation.

Acknowledgements

Financial support for this project was provided by grants from Philips Healthcare, Stanford Child Health Research Institute (Stanford NIH-NCATS-CTSA Grant #UL1 TR001085) and National Library of Medicine #R01LM01296601.

References

[1] Pons E, Braun LM, Hunink MM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279(2):329–43.
[2] Imaging utilization trends and reimbursement. http://www.diagnosticimaging.com/reimbursement/imaging-utilization-trends-and-reimbursement [accessed 30.03.18].
[3] Xu Y, Tsujii J, Chang EI-C. Named entity recognition of follow-up and time information in 20 000 radiology reports. J Am Med Inform Assoc 2012;19(5):792–9.
[4] Hua K-L, Hsu C-H, Hidayati SC, Cheng W-H, Chen Y-J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets Ther 2015;8.
[5] Cho J, Lee K, Shin E, Choy G, Do S. Medical image deep learning with hospital pacs dataset. arXiv preprint. arXiv:1511.06348.
[6] Bakthula R, Agarwal S. Automated human bone age assessment using image processing methods-survey. Int J Comput Appl 2014;104(13).
[7] Lang K, Huang H, Lee DW, Federico V, Menzin J. National trends in advanced outpatient diagnostic imaging utilization: an analysis of the medical expenditure panel survey, 2000–2009. BMC Med Imaging 2013;13(1):40.
[8] Kilani RK, Paxton BE, Stinnett SS, Barnhart HX, Bindal V, Lungren MP. Self-referral in medical imaging: a meta-analysis of the literature. J Am Coll Radiol 2011;8(7):469–76.
[9] Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study 1. Radiology 2005;234(2):323–9.
[10] Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform 2013;46(2):354–62.
[11] Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP. Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield. Am J Roentgenol 2017:1–4.
[12] Hassanpour S, Langlotz CP. Predicting high imaging utilization based on initial radiology reports:: a feasibility study of machine learning. Acad Radiol 2016;23(1):84–9.
[13] Banerjee I, Madhavan S, Goldman RE, Rubin DL. Intelligent word embeddings of free-text radiology reports. AMIA 2017 annual symposium, Washington, DC 2017.

(in press).
[14] Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of ct pulmonary angiography reports based on an extension of the context algorithm. J Biomed Inform 2011;44(5):728–37.
[15] Yu S, Kumamaru KK, George E, Dunne RM, Bedayat A, Neykov M, et al. Classification of ct pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. J Biomed Inform 2014;52:386–93.
[16] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–74.
[17] Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. Radiology 2002;224(1):157–63.
[18] Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. AMIA summits on translational science proceedings, 2013 2013:249.
[19] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[20] Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35(5):1285–98.
[21] Anavi Y, Kogan I, Gelbart E, Geva O, Greenspan H. A comparative study for chest radiograph image retrieval using binary texture and deep learning classification. Engineering in Medicine and Biology Society (EMBC), 2015 37th annual international conference of the IEEE, IEEE 2015:2940–3.
[22] Banerjee I, Crawley A, Bhethanabotla M, Daldrup-Link HE, Rubin DL. Transfer learning on fused multiparametric MR images for classifying histopathological subtypes of rhabdomyosarcoma. Comput Med Imaging Graph 2017.
[23] Conneau A, Schwenk H, Barrault L, LeCun Y. Very deep convolutional networks for text classification. Proceedings of the 15th conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 2017:1107–16.
[24] Shin B, Chokshi FH, Lee T, Choi JD. Classification of radiology reports using neural attention models. International joint conference on neural networks (IJCNN) 2017:4363–70.
[25] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint. arXiv:1408.5882.
[26] Pennington J, Socher R, Manning CD. Glove Global vectors for word representation. EMNLP 2014;14:1532–43.
[27] Sutskever I, Martens J, Hinton GE. generating text with recurrent neural networks. Proceedings of ICML 2011:1017–24.
[28] Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2016;24(2):361–70.
[29] Yang Z, Yang D, Dyer C, He X, Smola AJ, Hovy EH. Hierarchical attention networks for document classification. 2016.
[30] Chen H, Sun M, Tu C, Lin Y, Liu Z. Neural sentiment classification with user and product attention. EMNLP 2016:1650–9.
[31] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378.
[32] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint. arXiv:1301.3781.
[33] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 1994;5(2):157–66.
[34] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[35] Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: encoder-decoder approaches. Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation 2014:103–11.
[36] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. Proceedings of the deep learning and representation learning workshop: NIPS 2014.
[37] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures. Proceedings of the 32nd international conference on international conference on machine learning – volume 37 2015:2342–50.
[38] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proceedings of the AMIA symposium, American Medical Informatics Association 2001:662.
[39] Datla VV, Hasan SA, Qadir A, Lee K, Ling Y, Liu J, et al. Automated clinical diagnosis: the role of content in various sections of a clinical document. IEEE international conference on bioinformatics and biomedicine, BIBM 2017:1004–11.
[40] Arras L, Horn F, Montavon G, Müller K-R, Samek W. Explaining predictions of non-linear classifiers in NLP. arXiv preprint. arXiv:1606.07298.
[41] Leeper NJ, Bauer-Mehren A, Iyer SV, LePendu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLOS ONE 2013;8(5):e63499.
[42] Gallego B, Dunn AG, Coiera E. Role of electronic health records in comparative effectiveness research. J Comp Eff Res 2013;2(6):529–32.
[43] Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural language processing technologies in radiology research and clinical applications. RadioGraphics 2016;36(1):176–91.
[44] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19(1).
[45] Chen C-h. Computer vision in medical imaging vol. 2. World Scientific; 2014.
[46] Long M, Cao Y, Wang J, Jordan MI. Learning transferable features with deep adaptation networks. Proceedings of the 32nd international conference on machine learning, ICML 2015:97–105.
[47] Long M, Zhu H, Wang J, Jordan MI. Unsupervised domain adaptation with residual transfer networks. Adv Neural Inf Process Syst 2016:136–44.
[48] Long M, Zhu H, Wang J, Jordan MI. Deep transfer learning with joint adaptation networks. Proceedings of the 34th international conference on machine learning, ICML 2017:2208–17.
[49] Powell DK, Silberzweig JE. The use of ACR appropriateness criteria: a survey of radiology residents and program directors. Clin Imaging 2015;39(2):334–8.
[50] Lungren MP, Amrhein TJ, Paxton BE, Srinivasan RC, Collins HR, Eastwood JD, et al. Physician self-referral: frequency of negative findings at MR imaging of the knee as a marker of appropriate utilization. Radiology 2013;269(3):810–5.
[51] Lungren MP, Paxton BE, Kilani RK. Imaging self-referral: here we go again. Am J Roentgenol 2013;201(4):W658.
[52] Paxton BE, Lungren MP, Srinivasan RC, Jung S-H, Yu M, Eastwood JD, et al. Physician self-referral of lumbar spine MRI with comparative analysis of negative study rates as a marker of utilization appropriateness. Am J Roentgenol 2012;198(6):1375–9.
[53] Amrhein T, Paxton B, Lungren M, Befera N, Collins H, Yurko C, et al. Physician self-referral and imaging use appropriateness: negative cervical spine MRI frequency as an assessment metric. Am J Neuroradiol 2014;35(12):2248–53.
[54] Raja AS, Ip IK, Prevedello LM, Sodickson AD, Farkas C, Zane RD, et al. Effect of computerized clinical decision support on the use and yield of ct pulmonary angiography in the emergency department. Radiology 2012;262(2):468–74.
[55] Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med 2011;365(19):1758–9.