2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)
4-7 March 2018
Las Vegas, Nevada, USA

Coarse-to-Fine Multi-Task Training of Convolutional Neural Networks for Automated Information Extraction from Cancer Pathology Reports

Mohammed Alawad, Hong-Jun Yoon and Georgia D. Tourassi

***Abstract*— Information extraction and coding of free-text pathology reports is an important activity for cancer registries to support national cancer surveillance. Cancer registrars must process high volumes of pathology reports on an annual basis. In this study, we investigated an automated approach using a coarse-to-fine training of convolutional neural networks (CNNs) for extracting the primary site, histological grade and laterality from unstructured cancer pathology text reports. Our proposed training scheme consists of two stages. In the first stage, the multi-task learning (MTL) with hard parameter sharing approach is used to train a multi-task MT-CNN model for all the tasks. Then, the TM-CNN model parameters are used to initialize a CNN model for each task to be fine trained individually using its corresponding dataset. The performance of our proposed approach was compared against a state-of-the-art CNN and the commonly used SVM classifier. We observed that the proposed model consistently outperformed the base line models, especially for the less prevalent classes. Specifically, the proposed training approach achieved a micro-F score of 0.7749 over 12 ICD-O-3 topography codes which is a significant improvement as compared with state-of-the-art CNN (0.7101) and the SVM (0.6019) classifiers. Also, the results demonstrate the potential of the proposed method for handling class imbalance within each task. It significantly improves macro-F score by 24% and 12% of the primary site and histology grade tasks, respectively.**

I. Introduction

Manual information extraction from pathology reports is a standard component of clinical reporting in cancer registries. With the large volumes of pathology reports generated on an annual basis, manual annotation is costly, error-prone, and not easily scalable [1]. Automating the information extraction process has attracted tremendous interest from Natural Language Processing (NLP) researchers. Conventional NLP approaches rely heavily on text matching and rule-based approaches. However, such approaches require heavy involvement of domain experts to develop text matching, task-specific dictionaries, and inference rules to handle effectively the linguistic variability that exists across pathologists describing tissue specimens. Consequently, machine learning based NLP approaches have been explored to help the national cancer surveillance program handle in an effective and scalable way the increasing information extraction challenges it faces.

Deep learning has surged in popularity and proven to be effective for various artificial intelligence applications including NLP [2]. This approach often results in state-of-the-art performance in NLP tasks by learning high-level abstractions of word-level and sentence-level representations. It extends the feature extraction methods to a vector space representation for words. Previous deep learning approaches focused on the sequential nature of text data [3] or encoding documents into individual latent feature vectors [4]. Recently, CNNs have shown unprecedented performance compared with other machine learning techniques in a variety of computer vision tasks [5], such as image classification and face recognition. Inspired by the CNNs superiority in computer vision tasks, CNNs have been extended to NLP applications, such as sentence classification [6]. They have demonstrated outstanding performance compared with traditional vector space models for document-level information extraction and classification utilizing word embeddings [6]. Recently CNNs were shown to achieve superior performance on information extraction of primary cancer topography from free-text pathology reports [7].

Multi-task learning has been proposed as an efficient scheme to improve the performance of deep neural networks and convolutional neural networks by learning related tasks using shared representations. MTL has been successfully used in various fields, such as computer vision [8] and NLP [9]. Multi-task deep neural networks were also shown to boost performance compared to single task deep neural networks for information extraction from pathology reports [10].

Although MT-CNNs have proven to be an efficient technique for solving many machine learning and computer vision problems, to the best of our knowledge, this is the first attempt to adapt MTL of CNNs to NLP tasks. This paper explores the MTL to pre-train CNNs for automated extraction of different information from cancer pathology reports. Existing methods consider each target of cancer reporting, such as primary site, as a single information extraction task. However, these are not isolated information extraction since they describe different aspects of the same tumor. In particular, our framework adopts a deep learning convolutional network that extracts primary site, histological grade and laterality of cancer in a coarse-to-fine manner.

This paper is organized as follows: Section II describes the

This manuscript has been authored by UT - Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid- up, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public- access-plan).

M. Alawad, H.J. Yoon, G.D. Tourassi are with the Health Data Sciences Institute and the Biomedical Sciences, Engineering, and Computing Group in the Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 {alawadmm,yoonh,tourassig}@ornl.gov

cancer pathology reports and the fundamental three targets that are considered in this paper. In Section III, we present the CNN and MT-CNN models for automatic information extraction from pathology reports and our proposed coarse-to-fine training approach. In Section IV we discuss the performance evaluation and present the experimental results. Finally, we conclude the paper in Section V.

II. Cancer Pathology Reports

Cancer pathology reports are unstructured free-text documents containing highly descriptive and specific observations of cells and tissues. These reports are a primary information source for cancer registries, and the Surveillance, Epidemiology, and End Results (SEER) program which covers 30% of the US population. In this paper, we used de-identified pathology reports of breast and lung cancers, provided from five different SEER cancer registries (CT, HI, KY, NM, Seattle), with the proper IRB-approved protocol. Cancer registry experts manually annotated all pathology reports based on standard guidelines and coding instructions used in cancer surveillance. Their annotations served as the gold standard. Below we describe the tasks with the total number of annotated reports available per information extraction task.

1) Task-1: Primary Site: We used a corpus of 942 de-identified pathology reports. The corpus matched to 12 ICD-O3 topography codes representing 7 breast and 5 lung primary cancer topographies. For 6 ICD- O-3 codes the dataset included at least 50 reports per code but the remaining 6 ICD-O-3 codes were minimally populated with at least 10 but less than 50 pathology reports per code. Table I describes the 12 classes and the corresponding number of annotated reports per code included in the database. Since report sections available in each pathology report vary across pathology laboratories and registries, we aggregated the text content of every section in the pre-processing phase. The average length of the reports was 469 words.

TABLE I
ICD-O-3 TOPOGRAPHICAL CODES AND NUMBER OF CASES EMPLOYED IN THE STUDY.

Code	Cases	Description	Code	Cases	Description
C34.0	26	Main Bronchus	C50.2	36	U-inner quadrant of breast
C34.1	139	Upper lobe, lung	C50.3	10	L-inner quadrant of breast
C34.2	11	Middle lobe, lung	C50.4	63	U-outer quadrant of breast
C34.3	78	Lower lobe, lung	C50.5	21	L-outer quadrant of breast
C34.9	191	Lung, NOS	C50.8	62	Overlapping lesion of breast
C50.1	13	Central portion of breast	C50.9	292	Breast NOS

2) Task-2: Histological Grade: Cancer histological grade is determined by examining the cells and their patterns under a microscope and observing three features: frequency of cell mitosis, tubule formation, and nuclear pleomorphism. It is important to determine how quickly the cells are growing and spreading. In general, a lower grade indicates slower-growing cancer and a higher grade indicates a faster-growing and spreading one. Depending on the extent of the malignancy, a tumor can be graded as 1,2,3, and 4. Our corpus included 645 de-identified pathology reports with ground truth for histologic grade. Definitions of cancer histologic grades and the number of reports available for this dataset are listed in Table II.

TABLE II
DEFINITIONS OF CANCER HISTOLOGY GRADES AND NUMBER OF CASES EMPLOYED IN THE STUDY.

Grade	Cases	Description	Grade	Cases	Description
1	124	Well differentiated	3	271	Poorly differentiated
2	233	Moderately differentiated	4	17	Undifferentiated / Anaplastic

3) Task-3: Tumor Laterality: Laterality in cancer describes which side of a paired organ, in this study breast or lung, is the origin of the primary cancer. Our analysis used a corpus of 815 de-identified pathology reports for which the annotations of 400 reports are clearly stated the cancer laterality is right and the laterality of 415 reports is left.

4) Pre-processing: After extracting the pathology report text content, we first removed all non-alphabetical and numeric characters and stop words, and setting all alphabetical characters to lowercase. For the conventional SVM classifier, we pre-processed the pathology report data using the term frequency-inverse document frequency (TF-IDF) vector space model, which is considered the standard approach of document feature representation in NLP [11]. Specifically, we tokenized the processed pathology reports into n-grams of up to length 2 and generated an n-gram based term-frequency vector for each report while aggregating a training corpus document-frequency dictionary for each n-gram. We removed all but the top 400 document-occurring n-grams from our corpus vocabulary and finalized the pathology report feature vector by dividing each report term frequency vectors by the corpus document frequency vector. However, for the CNN and MT-CNN models, we used the word embedding method described in [7]. According to this method, word representations are learned instead of document representations. Specifically, a new word embedding vector is initialized, in the pre-trained embedding's word vector, for each tokenized word if no embedding exists in the embeddings vocabulary and if the frequency of the word in the pathology report is at least 2.

III. Network Architecture

In this section, we present the architectures of the CNN and MT-CNN models, and our proposed coarse-to-fine training approach.

A. Convolutional Neural Networks

CNNs can be applied to the document matrix, as in image processing, by using linear filters with region size l that corresponds to a context length of l word vectors. Convolution layer generates feature maps which are the representation of every context window over the document matrix. Max pooling layer captures the most important features by taking

the max value from each feature map as the extracted feature from a particular filter. Then, aggregating the selected contexts by concatenation. To extract multiple features, we use multiple filters with variable window sizes. The output is connected to a soft-max fully connected layer to produce a rank for each label. For the CNN architecture used in this paper, we applied the hyper parameters presented in [7], [6]. The document length input is defined as 1500 words vectors. The word vector space is 300. The window sizes l of the convolutional filters are 3,4, and 5 with 100 feature maps each. Rectified Linear Unit ReLU is used as the activation function. Finally, to account for class imbalance, we weighed our error costs inversely proportional to a class prevalence in the dataset. Figure 1 illustrates the architecture diagram of the CNN.

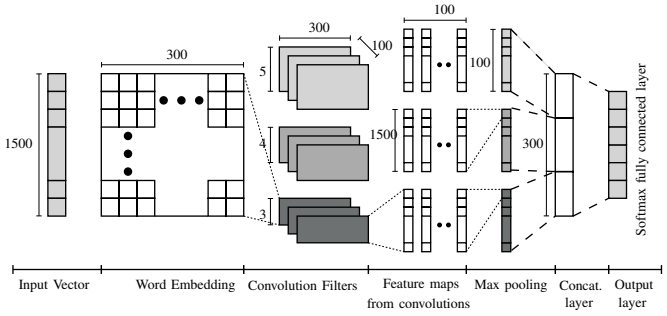


Fig. 1. Architecture diagram of a conventional CNN.

B. Coarse-to-Fine Training of CNNs

Due to the imbalance in the number of cases for classes within each task in the pathology reports, we propose a coarse-to-fine training technique for the CNN model. The proposed method consists of two stages. In the first stage, we use the MTL scheme with hard parameter sharing approach to train a MT-CNN model to learn the shared features. The network architecture consists of a shared convolutional layer, as the one presented in the previous subsection. The shared layer is followed by multiple fully connected layers. Each task has a separate fully connected layer and its size is determined by the number of classes for that task, for example, primary site task has a fully connected layer with 12 neurons. We leveraged three tasks to train our MT-CNN: cancer primary site, histological grade, and laterality. We treated the loss weight for all tasks equally as in [8]. The reports that have available labels for all tasks are chosen and used to coarse train the entire network jointly. The diagram of the multi-task CNN used in this paper is shown in Figure 2. In the second stage, the MT-CNN parameters, obtained from the first training stage, are used to initialize a CNN model for each individual task. Each CNN model has the same shared layer structure, the box shown in Figure 2, and keeps only its corresponding fully connected layer. Then, the complete set of cases for each task is used for task-wise fine training its corresponding CNN network. At the end, we will have three different CNN models, one for each task.

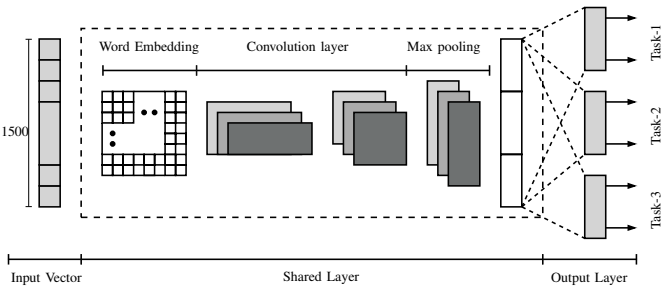


Fig. 2. Architecture diagram of a multi-task CNN.

TABLE III
IN-HOUSE CLINICAL DATASET RESULTS COMPARISON, WHERE LB/UB ARE LOWER/UPPER BOUNDS.

Model	Task	Micro-F	LB	UB	Macro-F	LB	UB
Primary Site	SVM	0.6019	0.5711	0.6337	0.3334	0.3017	0.3612
	CNN	0.7101	0.6804	0.7377	0.3367	0.3190	0.3537
	Proposed	**0.7749**	0.7473	0.8014	**0.5792**	0.5221	0.6275
Hist-grade	SVM	0.6759	0.6403	0.7116	0.6083	0.5392	0.6675
	CNN	0.7751	0.7426	0.8062	0.6036	0.5588	0.6571
	Proposed	**0.7906**	0.7596	0.8201	**0.7213**	0.6522	0.7796
Laterality	SVM	0.9354	0.9174	0.9521	0.9354	0.9174	0.9520
	CNN	0.9447	0.9276	0.9595	0.9447	0.9276	0.9595
	Proposed	**0.9558**	0.9423	0.9693	**0.9558**	0.9418	0.9693

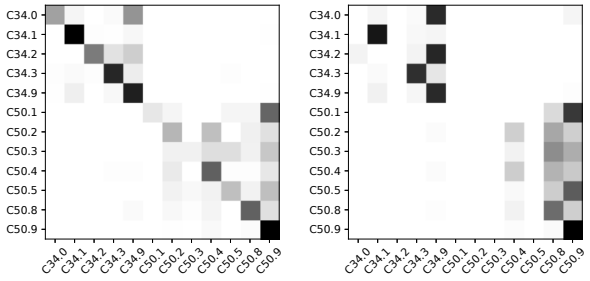
IV. Performance Evaluation and Experimental Results

In our experiments, we compare the proposed coarse-to-fine trained CNN scheme with two baseline architectures, the conventional single-task CNN [7] and the common SVM classifiers.

We implemented a balanced tenfold cross validation scheme by randomly partitioning the dataset into ten parts with near balanced label distributions. For each fold we used one partition once for testing and combined the rest for our training set. We evaluated model performance by aggregating the predicted responses from each test fold. To evaluate the performance accuracy of the three models, we used the standard NLP metrics of micro and macro averaged F-scores. The micro-averaged metrics have class representation roughly proportional to their test set representation, whereas macro-averaged metrics are averaged by class without weighing by class prevalence [12].

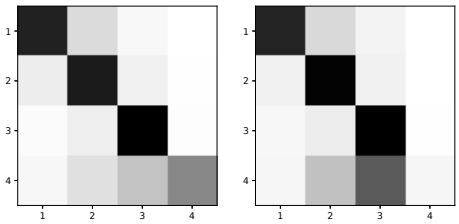
A. Experimental Results

Table III summarizes the performance of the proposed approach (coarse-to-fine trained CNN) benchmarked against the conventional deep learning approach (CNN) and the state-of-the-art shallow learning approach (SVM) for each one of the three information extraction tasks at hand (primary cancer site, laterality, and histologic grade). The table shows the micro-F and macro-F scores for each classifier, along with the respective 95% confidence intervals for each performance metric, derived using bootstrapping [13]. The



(a) Proposed CNN (b) Conventional CNN

Fig. 3. Confusion Matrices for the primary site task



(a) Proposed CNN (b) Conventional CNN

Fig. 4. Confusion Matrices for the histology grades task

proposed coarse-to-fine trained CNN classifier consistently performed better than the conventional CNN and SVM classifiers for the primary site task and the improvement is statistically significant. For the histological grade, it shows a statistically significant improvement of macro-F which demonstrates the effectiveness of our approach in handling low prevalence classes. However, the performance improvement with respect to laterality was marginal. Figures 3 and 4 compare the normalized confusion matrices for the proposed model and the conventional CNN model across all 12 classes for the primary cancer site and 4 classes for the histology grade classification tasks, respectively. The true-positives are indicated on the diagonal, while false-positives are indicated on the vertical axis and false-negatives on the horizontal. The figures clearly show that the proposed CNN outperforms the conventional CNN for the less prevalent classes, suggesting better handling of class imbalance in the data.

V. Conclusion

This work presents a new approach for using coarse-to-fine training of CNNs for automated information extraction from cancer pathology reports. This approach can significantly improve the performance of CNNs, particularly for low prevalence classes in imbalanced datasets. The proposed training scheme guarantees attaining at least the same performance as the conventional CNN since it uses fine training for each individual task. Experimental results demonstrate that our method consistently outperformed the state-of-the-art single task CNN and SVM classifiers. The (micro, macro) -F scores of the proposed scheme are (0.7749, 0.5792), (0.7906, 0.7213), and (0.9558, 0.9558) for primary cancer site, laterality, and histologic grade tasks, respectively. Future direction of this work includes mapping the proposed model to a bigger dataset, analyze computational efficiency, and perform comparative analysis with different MTL implementations for information extraction from unstructured pathology reports.

Acknowledgment

This work has been supported in part by the Joint Design of Advanced Computing Solutions (JDASC4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. The authors wish to thank Valentina Petkov of the Surveillance Research Program from the National Cancer Institute and the SEER registries at HI, KY, CT, NM and Seattle for the de-identified pathology reports used in this investigation. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S., Department of Energy under Contract No. DE-AC05- 00OR22725.

References

[1] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research." *Yearbook of medical informatics*, pp. 128–44, Jan. 2008.

[2] M. R. Vargas, B. S. L. P. de Lima, and A. G. Evsukoff, "Deep learning for stock market prediction from financial news articles," in *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, June 2017, pp. 60–65.

[3] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model." in *INTERSPEECH*, T. Kobayashi, K. Hirose, and S. Nakamura, Eds. ISCA, 2010.

[4] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML14. JMLR.org, 2014, pp. 1188–1196.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ser. NIPS12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[6] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014.

[7] J. X. Qiu, H. J. Yoon, P. A. Fearn, and G. D. Tourassi, "Deep learning for automated extraction of primary sites from cancer pathology reports," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 244–251, Jan 2018.

[8] J. Yim, H. Jung, B. Yoo, C. Choi, D.-S. Park, and J. Kim, "Rotating your face using multi-task deep neural network." in *CVPR*. IEEE Computer Society, 2015, pp. 676–684.

[9] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML 08. New York, NY, USA: ACM, 2008, pp. 160–167.

[10] H.-J. Yoon, A. Ramanathan, and G. Tourassi, *Multi-task Deep Neural Networks for Automated Extraction of Primary Site and Laterality Information from Cancer Pathology Reports*. Springer International Publishing, 2017, pp. 195–204.

[11] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Int. Res.*, vol. 37, no. 1, Jan. 2010.

[12] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, "Extracting information from the text of electronic medical records to improve case detection: a systematic review," *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 1007–1015, 2016.

[13] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, ser. Chapman and Hall/CRC Monographs on Statistics and Applied Probability. Taylor and Francis, 1994.