# Information Extraction from Cancer Pathology Reports with Graph Convolution Networks for Natural Language Texts

Hong-Jun Yoon, John Gounley, M. Todd Young and Georgia Tourassi
Biomedical Sciences, Engineering and Computing Group, Health Data Sciences Institute
Oak Ridge National Laboratory, Oak Ridge, TN, United States

***Abstract*—Graph-of-words is a flexible a nd e fficient te xt representation which addresses well-known challenges, such as word ordering and variation of expressions, to natural language processing. In this paper, we consider the latest graph-based convolutional neural network technique, the Text Graph Convolutional Network (Text GCN), in the context of performing classification tasks on free-form natural language texts. To do this, we designed a study of multi-task information extraction from medical text documents. We implemented multi-task learning in the Text GCN, performed hyperparameter optimization, and measured the clinical task performances. We evaluated micro and macro-F1 scores of four information extraction tasks, including subsite, laterality, behavior, and histological grades from cancer pathology reports. The scores for the Text GCN significantly outperformed our previous studies with convolutional neural networks, suggesting that the Text GCN model is superior to traditional models in task performance.**

***Index Terms*—graph convolutional network, convolutional neural network, text classification, n atural l anguage processing, biomedical informatics**

## I. Introduction

Fast, reliable text analytics and information extraction from free form natural language texts are essential parts of the analysis and processing of big text data. Within this space, biomedical and health informatics are notable applications of natural language processing (NLP) and text classification. In this study, we applied machine learning techniques to perform information extraction tasks related to cancer surveillance. Cancer is one of the leading causes of death in the United States [1] [2] and monitoring cancer prevalence and outcomes with population-level statistics is an essential part of understanding the disease and establishing plans for treatment and prevention [3]. However, despite more than a million new cancer cases in the United States each year, cancer surveillance is performed manually by human observers. Such a manual process is both challenging to perform at scale and potentially error-prone. To address these challenges to manual classification, the development of mechanisms for performing automatic information extraction from cancer text data is desirable.

In our previous work, we investigated the potential of using convolutional neural networks (CNN) for information extraction from medical texts for cancer surveillance [4]. We demonstrated that the multi-task CNN (MT-CNN) is an effective and accurate means of information extraction from cancer pathology reports, when compared with traditional machine learning techniques [5]. To address the limitations resulting from the fixed filter size in the MT-CNN, we developed the multi-task hierarchical convolutional attention network (MT-HCAN) [6], which improved accuracy in our test dataset of cancer pathology reports [7].

The graph-of-words (GOW) representation was designed to mitigate complexity and vagueness caused by the variety of word orders and expressions which appear in free-form text [8]. GOW is derived from a corpus such that the nodes in the graph represent words in the corpus and the graph's edges represent word co-occurrence information. We have applied the Bag-of-Graph (BOG) technique to the classification of histological grades of cancer pathology reports with the traditional Random Forest classifier and demonstrated the feasibility of GOW in this context [9]. Yao et al. introduced text graph convolution networks (Text GCN) [10], which are an application of the GCN developed by Kipf and Welling [11], and demonstrated high classification performance on openly available datasets. In this paper, we designed a study that applies the Text GCN to a cancer pathology report dataset provided by National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) registries. The data, classification model, and the study design are described in Section II; the experimental results are listed in Section III; and discussion and follow-up research ideas are provided in Section IV.

## II. Methods

### A. Cancer Pathology Report Dataset

The dataset on which this study is based was provided by five SEER cancer registries (Connecticut, Hawaii, Kentucky, New Mexico, and Seattle) with a proper Institutional

Review Board (IRB)-approved protocol. Trained cancer registry experts annotated 942 de-identified cancer pathology reports for 4 cancer properties: subsite, laterality, behavior, and histological grade. Based on this dataset, we synthesized a data corpus of 10,000 cancer pathology reports using a long short-term memory (LSTM)-based recurrent neural network (RNN). The text synthesis code is available at https://github.com/ECP-CANDLE/Benchmarks/tree/master/Pilot3/P3B2. Information about the tasks, labels, and the numbers of actual and synthesized cancer pathology reports in the dataset are listed in Table I.

TABLE I
NUMBER OF CANCER PATHOLOGY REPORTS BELONGING TO EACH CLASS IN BOTH THE ACTUAL AND SYNTHESIZED DATASETS.

| Codes | Actual | Synthesized |
|---|---|---|
| Site | | |
| C34.1 | 73 | 2200 |
| C34.3 | 23 | 700 |
| C34.9 | 24 | 520 |
| C50.4 | 13 | 400 |
| C50.8 | 22 | 680 |
| C50.9 | 181 | 5500 |
| Laterality | | |
| right | 141 | 4280 |
| left | 195 | 5720 |
| Behavior | | |
| malignant | 300 | 8900 |
| in situ | 36 | 1100 |
| Grade | | |
| 1 | 42 | 1280 |
| 2 | 135 | 4120 |
| 3 | 159 | 4600 |

We randomly split the dataset for training (8,000 reports) and model validation (the remaining 2,000). The tokenized, synthesized dataset is available to download at http://ftp.mcs.anl.gov/pub/candle/public/benchmarks/Pilot3/.

### B. Text Graph Convolutional Networks

One of the advantages of the MT-CNN model is its capability for finding optimal features of keywords and expressions by itself. The model training determines the key phrases and patterns in the latent representations of word embedding and convolution layers; thus, we do not need to perform manual curation of features. The MT-CNN model can handle the ambiguity of expression variation if we supply a sufficient number samples of the variability to train the network. However, such a dataset is not always available, especially in the medical text corpus that may contain rare disease types. Conversely, the GOW representation can easily handle ambiguity of expression variations, such as "word inversion" (e.g., "RIGHT BREAST CORE BIOPSY" and "BREAST, RIGHT, CORE BIOPSY") and "subset matching" (e.g., "RIGHT LOBE" and "RIGHT UPPER LOBE").

A GCN is a multilayer neural network that operates directly on a graph and induces embedding vectors of nodes based on the properties of their neighbors. Suppose we have a graph $G = (V, E)$, where $V$ is the set of edges and $E$ is the set of nodes of the graph. Let $X \in \mathbb{R}^{n \times m}$ be a matrix containing all $n$ nodes with their features, where $m$ is the dimension of the feature vectors. The GCN layer is described as

$$L^{(j+1)} = \rho(\tilde{A} L^{(j)} W_j)$$

where $W$ is a weight matrix, $\tilde{A} = D^{-1/2} A D^{-1/2}$ is the normalized symmetric adjacency matrix, $D_{ii} = \sum_j A_{ij}$ is a degree matrix, and $\rho(\cdot)$ is an activation function. If $j = 0$ (corresponding to the input layer), $L^{(0)} = X$.

In the implementation of the Text GCN, documents in the corpus and unique words in the vocabulary become nodes of the graph. Based on these nodes, we build edges based on word occurrence and word co-occurrence in the documents. The weight of the edge between a document node and a word node is the term frequency-inverse document frequency (TF-IDF). Likewise, the weight of an edge between two words is the point-wise mutual information (PMI) of the two words. In the Text GCN, we set the feature matrix $X = I$, which means that the words and documents in the corpus are represented as a one-hot vector as the input to the Text GCN.

To implement multi-task learning in the Text GCN, we locate multiple fully-connected layers at the end of the graph layer with a final softmax output layer. Performance is measured with cumulative training and validation losses across tasks. Figure 1 illustrates the multi-task Text GCN model implemented in this study.

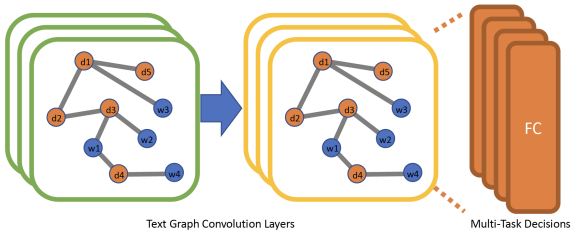


Fig. 1. Illustration of the multi-task Text GCN model.

### C. Hyperparameter Optimization

A hyperparameter is a parameter that defines the characteristics of an artificial intelligence (AI) model. While there exists no analytical solution to determine the optimal hyperparameters, iterative algorithms have been developed which optimize these values by successively training a number of models with different sets of hyperparameters. In this study, we performed hyperparameter optimization with the CANcer Distributed Learning Environment (CANDLE) framework [12]. Developed by the Exascale Computing Project (ECP), the CANDLE framework aims to facilitate the deployment and scalability of deep learning on high performance computing resources. CANDLE configured for various types of computing environments, including Department of Energy (DOE)-operated leadership computing facilities. We performed the hyperparameter optimization of the Text GCN model

TABLE II
HYPERPARAMETERS OF THE MULTI-TASK TEXT GCN AND THEIR SWEEPING RANGES.

| Parameter | Description | Bounds |
|---|---|---|
| lr | Learning Rate | 0.005∼0.1 |
| n_batch | Batch Size | 16∼128 |
| dropout | Dropout | 0.0∼0.95 |
| n_layers | Number of GCN layers | 1∼5 |
| n_hidden | Number of output nodes of GCN | 100∼1000 |
| min_df | Minimum # of document frequency | 1∼5 |
| w_cooc | Window size of word co-occurrence | 5∼30 |

on the Summit supercomputer at the Oak Ridge Leadership Computing Facility (OLCF). Summit has IBM Power System AC922 nodes with 6 Nvidia V100 GPUs per node.

We employed the model-based optimization (MBO) hyperparameter optimization workflow in CANDLE to determine the seven hyperparameters in this study. The minimum document frequency of the word and word co-occurrence window size define the vocabulary size and the edges between words, respectively, and thereby define properties of the text graph. The number of GCN layers and the number of hidden nodes describe the topology of the Text GCN itself. Finally, the learning rate, mini-batch size, and a dropout rate of the GCN output define the characteristics of network optimization. Table II lists each of the hyperparameters to be determined and their sweeping ranges during optimization.

## III. RESULTS

The Text GCN is implemented with version 0.4 of the Deep Graph Library [13] and PyTorch 1.0 [14] using Python 3.6. We utilized the mlrMBO [15] workflow with 22 proposed points and 20 MBO search iterations, corresponding to a total design size of 440 points. We implemented an early stopping mechanism that monitored validation loss and stopped the training procedure if there were no improvement in accuracy for more than five consecutive epochs. If early stopping was not triggered, training was performed up to a maximum number of 50 training epochs.

### A. Optimal Hyperparameters

Validation losses of the 440 trials during the MBO runs are plotted in Figure 2. Note that the validation losses were consistently lower for successive MBO iterations, which is an indication that the MBO process converges toward an optimal solution.

We chose the five most optimal solutions from the MBO search and performed further analyses, as shown in Table III. We determined the 95% confidence intervals of the minimum validation losses for the five models with bootstrapping of 200 samples. We observe that the five optimal models do not differ significantly with respect to clinical performance; however, we can observe the trend that the 50th percentiles of minimum validation losses were increasing as the rank order increases.

### B. Clinical Task Performance

We examined the information extraction of the four tasks (subsite, laterality, behavior, and histological grade) from the 2,000 validation samples. Their micro and macro-averaged F1 scores, along with the 95% confidence intervals, are reported in Table IV. Consistent with the cumulative validation loss noted above, we cannot observe significance differences in task performance among the models trained by the five optimal hyperparameter sets. However, when compared to our previous study [7], in which we optimized and evaluated for the same four tasks with the MT-CNN and multi-task hierarchical convolutional networks (MT-HCAN), our achievements from this study with the Text GCN are reflected in the significantly higher F1 scores than found for the MT-CNN and MT-HCAN.

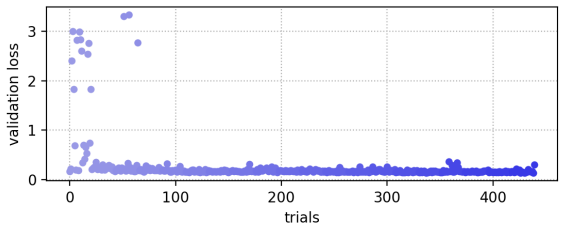


Fig. 2. Validation losses of the hyperparameter optimization of multi-task Text GCN model, showing MBO algorithm is converging to the optimal hyperparameters.

TABLE III
THE FIVE OPTIMAL HYPERPARAMETERS OF THE MULTI-TASK TEXT GCN DETERMINED BY CANDLE WORKFLOW AND THE CORRESPONDING VALIDATION LOSS. NOTE THAT THE LB AND UB ARE THE LOWER AND UPPER BOUNDS OF 95% CONFIDENCE INTERVALS OF THE VALIDATION LOSSES DETERMINED BY BOOTSTRAPPING OF 200 SAMPLES.

| Parameters | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| lr | 0.0406 | 0.0406 | 0.0493 | 0.0406 | 0.0644 |
| n_batch | 16 | 16 | 16 | 27 | 16 |
| dropout | 0.3712 | 0.6188 | 0.0000 | 0.0264 | 0.3712 |
| n_layers | 2 | 2 | 2 | 2 | 2 |
| n_hidden | 662 | 431 | 393 | 1000 | 438 |
| min_df | 1 | 3 | 1 | 4 | 5 |
| w_cooc | 5 | 5 | 5 | 5 | 30 |
| Val Loss | 0.1288 | 0.1292 | 0.1306 | 0.1322 | 0.1331 |
| LB | 0.1163 | 0.1159 | 0.1197 | 0.1188 | 0.1195 |
| UB | 0.1462 | 0.1508 | 0.1545 | 0.1555 | 0.1553 |

## IV. DISCUSSION

In this paper, we briefly reviewed the Text GCN, implemented the multi-task learning of the GCN, and evaluated the model for four information extraction tasks from a cancer pathology report dataset. We ran hyperparameter optimization procedures with the CANDLE workflow engine on the Summit supercomputer. The clinical task performance scores from the models trained by the optimal hyperparameters were substantially higher than the scores we have achieved with the MT-CNN and MT-HCAN models. Our results suggest that the Text GCN is a superior method for performing document classification and natural language processing, and that is suited to performing information extraction tasks from medical texts. The identification of optimal hyperparameters from this study will inform the design of an in-depth study a larger dataset and a greater number of cancer phenotypes.

TABLE IV
THE FOUR CLINICAL TASK PERFORMANCES (SUBSITE, LATERALITY, BEHAVIOR AND HISTOLOGICAL GRADE) FROM THE FIVE MODELS TRAINED BY THE FIVE OPTIMAL HYPERPARAMETERS. THE CLINICAL TASK PERFORMANCE SCORES WERE MEASURED BY MICRO AND MACRO-AVERAGED F1 SCORES. NOTE THAT THE 95% CONFIDENCE INTERVALS OF THE F1 SCORES WERE DETERMINED BY BOOTSTRAPPING OF 200 SAMPLES.

| Parameter | Task | Micro F1 | Macro F1 |
|---|---|---|---|
| **1** | Subsite | 0.9895 (0.9865∼0.9920) | 0.9861 (0.9801∼0.9905) |
| | Laterality | 0.9845 (0.9825∼0.9870) | 0.9839 (0.9818∼0.9865) |
| | Behavior | 0.9940 (0.9920∼0.9950) | 0.9856 (0.9806∼0.9880) |
| | Grade | 0.9795 (0.9755∼0.9825) | 0.9786 (0.9742∼0.9826) |
| **2** | Subsite | 0.9890 (0.9860∼0.9920) | 0.9851 (0.9801∼0.9898) |
| | Laterality | 0.9850 (0.9825∼0.9870) | 0.9844 (0.9819∼0.9865) |
| | Behavior | 0.9940 (0.9920∼0.9950) | 0.9855 (0.9806∼0.9880) |
| | Grade | 0.9805 (0.9755∼0.9840) | 0.9798 (0.9747∼0.9839) |
| **3** | Subsite | 0.9890 (0.9855∼0.9910) | 0.9850 (0.9801∼0.9893) |
| | Laterality | 0.9845 (0.9810∼0.9870) | 0.9839 (0.9803∼0.9865) |
| | Behavior | 0.9935 (0.9915∼0.9945) | 0.9844 (0.9794∼0.9870) |
| | Grade | 0.9790 (0.9750∼0.9820) | 0.9786 (0.9734∼0.9822) |
| **4** | Subsite | 0.9895 (0.9865∼0.9915) | 0.9860 (0.9810∼0.9899) |
| | Laterality | 0.9845 (0.9815∼0.9865) | 0.9839 (0.9807∼0.9860) |
| | Behavior | 0.9935 (0.9915∼0.9950) | 0.9845 (0.9794∼0.9881) |
| | Grade | 0.9785 (0.9750∼0.9825) | 0.9779 (0.9729∼0.9819) |
| **5** | Subsite | 0.9885 (0.9860∼0.9915) | 0.9838 (0.9792∼0.9897) |
| | Laterality | 0.9850 (0.9820∼0.9870) | 0.9844 (0.9813∼0.9865) |
| | Behavior | 0.9935 (0.9910∼0.9950) | 0.9844 (0.9782∼0.9881) |
| | Grade | 0.9775 (0.9730∼0.9810) | 0.9763 (0.9718∼0.9806) |

## ACKNOWLEDGMENT

This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725.

The authors wish to thank Valentina Petkov of the Surveillance Research Program from the National Cancer Institute and the SEER registry at Connecticut, Hawaii, Kentucky, New Mexico and Seattle for the 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] Sherry L Murphy, Jiaquan Xu, Kenneth D Kochanek, and Elizabeth Arias, "Mortality in the united states, 2017," 2018.

[2] American Cancer Society, *Cancer facts & figures 2019*, American Cancer Society, 2019.

[3] Matthew J Hayat, Nadia Howlader, Marsha E Reichman, and Brenda K Edwards, "Cancer statistics, trends, and multiple primary cancer analyses from the surveillance, epidemiology, and end results (seer) program," *The oncologist*, vol. 12, no. 1, pp. 20–37, 2007.

[4] John X Qiu, Hong-Jun Yoon, Paul A Fearn, and Georgia 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, 2018.

[5] Mohammed Alawad, Hong-Jun Yoon, and Georgia D Tourassi, "Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports," in *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*. IEEE, 2018, pp. 218–221.

[6] Shang Gao, Arvind Ramanathan, and Georgia Tourassi, "Hierarchical convolutional attention networks for text classification," Tech. Rep., Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2018.

[7] Hong-Jun Yoon, John Gounley, Shang Gao, Mohammed M Alawad, Arvind Ramanathan, and Georgia Tourassi, "Model-based hyperparameter optimization of convolutional neural networks for information extraction from cancer pathology reports on hpc," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019.

[8] François Rousseau and Michalis Vazirgiannis, "Graph-of-word and tw-idf: new approach to ad hoc ir," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 2013, pp. 59–68.

[9] Hong-Jun Yoon, Larry Roberts, and Georgia Tourassi, "Automated histologic grading from free-text pathology reports using graph-of-words features and machine learning," in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017, pp. 369–372.

[10] Liang Yao, Chengsheng Mao, and Yuan Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 7370–7377.

[11] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[12] Justin M Wozniak, Rajeev Jain, Prasanna Balaprakash, Jonathan Ozik, Nicholson Collier, John Bauer, Fangfang Xia, Thomas Brettin, Rick Stevens, Jamaludin Mohd-Yusof, et al., "Candle/supervisor: A workflow framework for machine learning applied to cancer research," 2017.

[13] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al., "Deep graph library: Towards efficient and scalable deep learning on graphs," *arXiv preprint arXiv:1909.01315*, 2019.

[14] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[15] Bernd Bischl, Jakob Richter, Jakob Bossek, Daniel Horn, Janek Thomas, and Michel Lang, *mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions*, 2017.