Exploring Cross-Institutional Recognition of Cancer Registration Items: A Case Study on Catastrophic Forgetting

You Chen Zhang[1], Chen-Kai Wang[2,3], Ming-Ju Tsai[4,5], Hong-Jie Dai[1,6]
[1]National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan
[2]Advanced Technology Laboratory Chunghwa Telecom Laboratories Taoyuan, Taiwan
[3]Department of Computer Science National Yang Ming Chiao Tung University Hsinchu, Taiwan
[4]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
[5]School of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
[6]Center for Big Data Research, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
{ uchenzhang0220, dennisckwang}@gmail.com, mjt@kmu.edu.tw, hjdai@nkust.edu.tw

Abstract

A cancer registry is a critical database for cancer research, which require diverse domain knowledge and manual extraction of vital information from patient records for surveillance. In order to building a real-time and high-quality cancer registry database, a named entity recognition (NER) model based on bidirectional long short-term memory (BiLSTM)-conditional random fields (CRFs) to automatically extract 14 cancer registry items from unstructured pathology reports was developed for five hospitals. Because not all hospitals have sufficient training data, so that we apply transfer learning to develop our models for different hospitals. However, catastrophic forgetting leads to poor performance of the transferred model on the source hospital. To address this issue, we study the effectiveness of applying the elastic weight consolidation (EWC) method for the extraction of cancer registry items from the unstructured pathology reports of colorectal cancer to mitigate the occurrence of catastrophic forgetting. In our results, we observe that effective parameter settings can reduce the impact of catastrophic forgetting.

Keywords: Electronic Medical Records, Natural Language Processing, Transfer Learning, Elastic Weight Consolidation

1 Introduction

Electronic medical records (EMR) contain large amounts of data collected during routine medical care delivery and have the potential to generate practice-based evidence, such as early diagnosis of cancer patients and improved quality of care. Cancer is one of the main causes of mortality worldwide, and it is the leading cause of death in Taiwan, and the overall incidence rate has gradually increased (Kuo et al., 2020). In recent years, domestic cancer research has continued to increase, promoting cooperation and resource integration among cancer centers to accelerate breakthroughs in cancer research bottlenecks. The Taiwan Cancer Registry (TCR), which provide a comprehensive measurement of cancer incidence, morbidity, survival, and mortality for persons with cancer in Taiwan. Unfortunately, the process of reporting cancer cases requires manual review of numerous reports, such as radiology reports and pathology reports, which is obviously labor-intensive and time-consuming. One solution to this problem currently being explored is the application of Natural Language Processing (NLP) techniques to automatically read and extract information from cancer reports.

In the field of machine learning, the quantity of the dataset has a significant impact on the performance and generalization ability of algorithms. Transfer learning has been proven to be an effective learning method to solve the problem of dataset scarcity (Hutchinson et al., 2017). It uses the knowledge gained from training a model on one task to improve the performance of another related task, which can speed up convergence, reduce data requirements and improve performance when obtaining labeled data for the new task is challenging or time-consuming. Dai et

Source	HA	HB	HC	HD	HE
# of Reports	541	1,735	965	1,732	748
Training Set	300	300	300	300	300
Test Set	100	100	100	100	100

Table 1: Datasets collected from five medical institutions.

al. (2021) demonstrated the utility of employing transfer learning for cross-corpus training in cancer registries. However, their study was limited to cases where the source hospital had same cancer registry items as the target hospital. In practical scenarios, cancer registration standards followed by different hospitals at different times may lead to different items and content of the target cancer. For example, different American Joint Committee on Cancer (AJCC) versions have different numbers of items, staging criteria, tumor descriptors and prognostic factors.

Despite transfer learning alleviates the issue of learning from small datasets in cancer registries across healthcar institutions, catastrophic forgetting may occur during the process of learning a new set of cancer registry items leading to a degradation of the model's performance on the original item set. The issue of catastrophic forgetting is paramount importance as it directly impacts the effectiveness of transfer learning and the overall performance of models. When catastrophic forgetting occurs, the learned knowledge from earlier tasks may be overwritten or weakened by the learning of subsequent tasks, leading to suboptimal performance on all tasks. McCloskey and Cohen (1989) demonstrated that interference leading to forgetting occurs whenever new knowledge could alter the weights of old knowledge. Ratcliff (1990) conducted experiments using backpropagation-based training on multi-layer models, revealing that memory and context models with pre-learned knowledge are unable to address catastrophic forgetting. Recently, Ramasesh, Dyer, and Raghu (2020) conducted experiments on the publicly available CIFAR-10 image dataset, showing that catastrophic forgetting often occurs in deep neural network layers closer to the output. Arumae, Sun, and Bhatia (2020) used the RoBERTa model pre-trained on PubMed articles by combining with the elastic weight consolidation (EWC) (Kirkpatrick et al., 2017) method to achieve better results in the i2b2 named entity recognition (NER) task than that of the original RoBERTa model alone. Arumae found that using the EWC method helped mitigate catastrophic forgetting with only a 0.33% decrease in performance across the seven general-domain tasks in the GLUE benchmark. This approach demonstrated competitive performance in biomedical tasks as well.

In this study, we focus on mitigating the adverse repercussions of catastrophic forgetting in transfer learning. To this end, we conduct experiments to study the following two interrelated research questions, each of which will be discussed and elaborated in subsequent sections, as follows:
RQ1: The effect of different transfer learning strategies.
RQ2: Extent of catastrophic forgetting in transfer learning: To illustrate the extent of catastrophic forgetting in transfer learning scenarios when the developed model learned on one additional hospital's data.

2 Method

2.1 Datasets

In this study, we used pathology reports of colorectal cancer from five medical institutes including Hospital-A (HA), Hospital-B (HB), Hospital-C (HC), Hospital-D (HD) and Hospital-E (HE) as our dataset. In order to simulate the situation of limited data, we randomly selected 300 and 100 pathology reports from each medical institution in the pre-processing stage as the training set and test set respectively. Table 1 shows the number of datasets compiled for the five medical institutions.

2.2 Corpus Construction

Due to the variations in cancer-related items of interest across different hospitals, which is owing to the adoption of different AJCC versions or other clinical research concerns, the annotation process was discussed separately. To enhance the precision of annotations, each hospital established an annotation team consisting of at least three members and utilized Fleiss' Kappa (Fleiss, Nee, & Landis, 1979) to assess annotation consistency.

Type	Description	HA	HB	HC	HD	HE
H	The structure of primary tumor cells under a microscope.	O	O	O	O	O
G	Grading/differentiation of solid tumors at the primary site after surgery.	O	O	O	O	O
NE	Total number of regional lymph nodes examined by pathologists.	O	O	O	O	O
PN	Total number of regional lymph nodes examined by pathologists that tested positive.	O	O	O	O	O
TS	Size of tumor.	O	O	O	O	O
SC	Symbols of AJCC Pathological Staging Prefixes/Roots.	O	O	O	O	O
T	Size or extent of the primary tumor.	O	O	O	O	O
N	Presence of regional lymph node metastasis and extent of metastasis.	O	O	O	O	O
M	Presence of distant metastasis of the tumor.	O	X	O	O	O
LI	Presence of lymphatic or vascular invasion in the primary site report.	O	X	X	X	O
PI	Presence of neural invasion documented in the pathology report for the primary site in the medical record.	O	X	X	X	O
ASC	AJCC Cancer Staging Edition.	O	X	X	X	O
KRAS	Normal value for KRAS testing .	O	X	O	O	O
CEA	carcinoembryonic antigen.	O	X	X	X	O

Table 2: The fourteen defined cancer registry items. If the hospital does not contain the cancer registry item, it will be noted as X.

Based on Taiwan's cancer registration reports, we focused on specific factors related to pathological examinations and colorectal cancer site-specific factors (SSFs), resulting in a total of 14 items. Table 2 presents the 14 colorectal cancer items, including histology types (H), grades (G)、stage classification (SC), pathological TNM classifications (TNM), the number of examined nodes (NE) and positive nodes (PN), tumor size (TS), lymphovascular invasion (LI), perineural invasion (PI), AJCC stage classification (ASC), carcinoembryonic antigen (CEA), and Kirsten rat sarcoma viral oncogen homolog (KRAS).

The annotation process of the dataset was carried out independently by the annotation teams in the five medical institutes. They followed a consistent annotation guideline when the cancer registry items were shared among them. Initially, the annotators annotated a set of 100 randomly sampled pathology reports according to the annotation guidelines to estimate the Kappa value, which is interpreted as follows: value ≤ 0 as no agreement, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect agreement. If the kappa value did not exceed 0.85, further discussions and criteria modifications were carried out iteratively. Once the consistency criterion was met, the remaining reports were evenly distributed among the annotators for individual annotation.

2.3 Network Architecture for Cancer Registry Information Extraction

To process pathology reports, we first de-identify the unstructured reports and then apply the sentence segmentation. Subsequently, the task is formulated as a sequence labeling task by using the IOB2 encoding. We utilize a neural network architecture that combines bidirectional long short-term memory (BiLSTM) with conditional random fields (CRFs) as depicted in Figure 1.

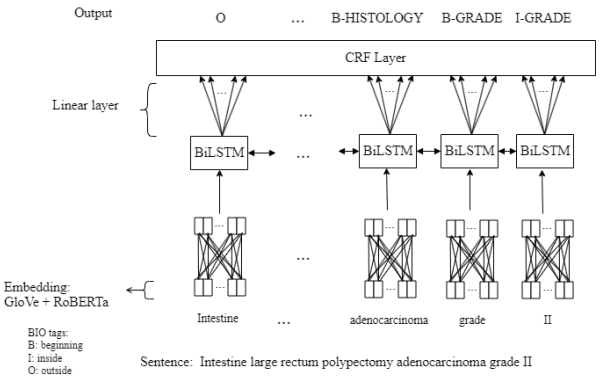

Figure 1: BiLSTM-CRF Network Architecture.

2.4 Fine-tuning with EWC

EWC employs a penalty mechanism in updating model parameters based on their importance. The

Fisher information matrix ($F_{i,i}$) is utilized to identify significant parameters. During EWC fine-tuning, the Fisher information matrix serves as a criterion to slow down the decrease of loss, scaling the cost of the original parameters θ_i^* to the updating parameters θ_i. The following equation is the lose function defined for the model with the parameter set θ.

$$L(\theta) = L_{FT}(\theta) + \sum_i \frac{\lambda}{2} F_{i,i}(\theta_i - \theta_i^*)^2 \quad (1)$$

Here, λ is a controllable hyperparameter. $L_{FT}(\theta)$ is the loss of target domain.

2.5 Transfer Learning among Different Hospitals

Previous studies have observed that transfering the parameters of all layers of the BiLSTM-CRF model for the recognition of cancer registry items achieve the best scores even with a small amount of data. However, those works only focuses on the transfer learning of the same recognition task. In this study, the number of cancer registry items can be different as shown in Table 2, which can be summarized as the following three types:

1. The numbers and types of items are the same.
2. Transfer from more items to fewer items: In this case, the set of the types of the source domain items is the superset of the target hospital's items.
3. Transfer from fewer items to more items: In this case, the number of the types of the target domain items is the superset of the source hospital's items.

Due to the fact that the number of the target hospital's items surpassed that in the source domain, it is necessary to modify the last linear layer shown in Figure 1 to align with the target domain. In our implementation for the first and second cases, the parameters of all layers of the developed models were directly transferred to the new models. For the third case, we migrated the trained parameters from the source hospital to the target hospital for the matched registry items. For new items not present in the source hospital, random initialization was applied to set the initial weights for the corresponding node in the last linear layer.

2.6 Experiment Configurations

We conduct experiments to study the effectiveness of applying EWC in the aforementioned scenarios to mitigate catastrophic forgetting. For comparison purpose, we developed models followed the conventional transfer learning methods. Furthermore, the following two methods were developed, which are served as the upper and lower bounds respectively:

- Merged corpus: Models trained on the merged training sets of the source and target hospitals. The configuration is served as an upper bound.
- Direct Prediction: Making predictions directly by using the source model. The configuration serves as a lower bound.

The neural networks were implemented using PyTorch and trained with a Nvidia GeForce RTX 2080 Ti GPU with 11GB of memory.

In the following experiments, the number of epochs was set to 150 with a batch size of 256 and the learning rate was set to 1×10^{-1} . We used cross entropy as the loss function and employed stochastic gradient descent as the optimizer. The λ of EWC was set to 400, same as Kirkpatrick et al. (2017).

3 Results

3.1 Statistics of the Experimental Datasets and the Evaluation Results

We collected a total of 5,721 pathology reports from five hospitals. In this study, the corpora from each hospital (shown in Table 3) were further randomly sampled to extract 300 reports as the training set, ensuring no overlap with the 100 reports in the test set. The training set was then divided proportionally into subsets of 15, 60, 120, 180, and 240 reports each. This process aimed to simulate scenarios of learning with limited data. The Kappa values for each hospital are detailed in Table 4. As HE did not undergo Kappa consistency testing, the table does not include its Kappa score.

For the collected data, we notice that each hospital has its unique way of releasing the pathology reports, leading to variations in the amount of information included. For instance, the reports for each patient are created separately at

Type	HA	HB	HC	HD	HE
H	539	948	911	537	2,097
G	436	908	852	695	919
NE	584	450	1,046	710	1,148
PN	516	450	770	714	920
TS	1,119	350	1,671	1,272	727
SC	534	320	629	275	1,273
T	364	319	352	275	785
N	366	198	337	275	682
M	364	1	41	84	214
LI	303	N/A	N/A	N/A	294
PI	298	N/A	N/A	N/A	252
ASC	316	N/A	N/A	N/A	298
KRAS	8	N/A	1	312	256
Numbers of reports	300	300	300	300	300
Numbers of sentences	18,544	14,054	29,877	39,794	31,913
Numbers of annotations	2,039	1,928	3,236	2,759	5,507

Table 3: Corpus statistics for the compiled corpora of train sets.

Hospital	**Kappa Value**
HA	0.802 (substantial)
HB	0.914 (almost perfect)
HC	0.955 (almost perfect)
HD	0.819 (substantial)
HE	N/A

Table 4: Kappa values of the compiled dataset.

HA, but HD consolidates diagnostic reports for the same patient and clinical pathology number into a single report. Table 3 shows the performance of the developed models evaluated on their test sets respectively. The models were then served as the pre-trained models for transferring the learned parameters to the model for other target hospitals in the following experiments.

While this practice can save time in case finding, it may introduce uniqueness to the labeling process. Taking HD's corpus as an example, a single report could contain multiple diagnostic reports with the same writing style. However, the annotators only label the grade based on the last diagnostic report in that combined report.

The varying annotation styles across different hospitals pose a challenge for transferring learning from one hospital to another in this study.

3.2 RQ1: The Effect of Different Transfer Learning Strategies

To investigate RQ1, this experiment is divided into three configurations based on whether to inherit the parameters of the last layer:

- Non-inherit: Not inheriting the parameters of the last layer, and initializing all parameters of that layer randomly (while still inheriting parameters of other layers).
- Inherited: Based on the "Non-inherit", the configuration further inherits the parameters of the last layer matched with the output nodes of the source model.
- EWC: Based on the "Inherited", this configuration further apply the EWC method during the training phase.

The datasets compiled for all of the five hospitals were used in this experiment, and transfer learning was conducted between each pair of hospitals. The results are presented according to the task types described in Section 2.5 which can be divided as follows:

- Type 1: The number and types of items are the same. The model was first pre-trained on the full source dataset and then transferred to the target training dataset.

The evaluation results on the target test set was presented in Figure 2.

- Type 2: The model was transferred from the source dataset with more item types to the target dataset with less item types. The test set results for the target hospital is presented in Figure 3.

- Type 3: The model was trained with less item types but transferred to the target hospital with more item types. The evaluation results on the target test set is illustrated in Figure 4.

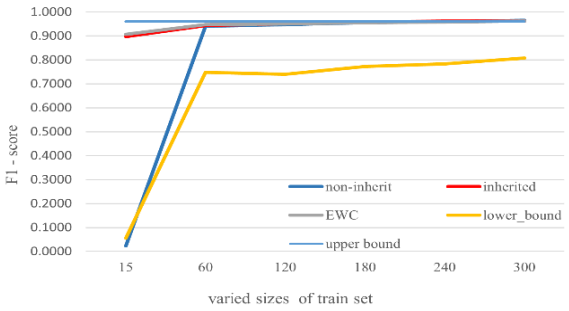

Figure 2: Type 1 results for the HC test set; the model was transferred from HD (10) to HC (10).

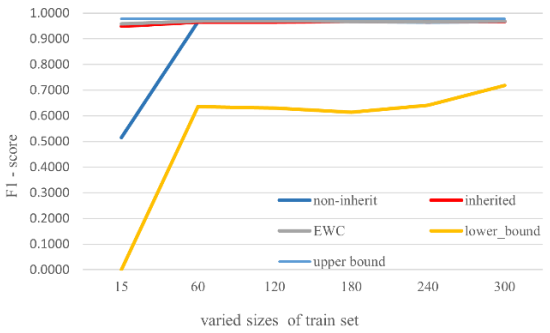

Figure 3: Type 2 results for the HB test set; the model was transferred from HA (13) to HB (9).

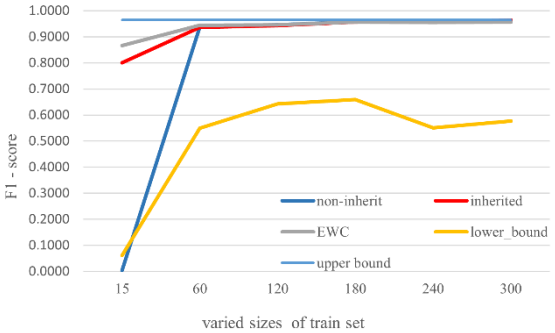

Figure 4: Type 3 results for the HC test set; the model was transferred from HB (9) to HC (10).

We only select three results with different types as a result of the gr eat mass of data. In general, the outcomes are mostly consistent. Take Figure 2 as an example. We fine-tuned the models pre-trained with the HD training set on the varied sizes of the HC training set (ranged from 15 reports to 300 reports as depicted in the x-axis). It's worth noting that the performance of the configurations of all inherited approaches among all of the three types achieved above 0.9 scores when the target hospital only provides 15 reports. The configurations trained with more than 15 reports achieved an F-score of 0.9 or higher, except for the lower-bound configuration. Consistent with the observations of other related configuration results, the inclusion of EWC during the training phase results in a model with a better F-score than that of the model trained with the conventional transfer learning. On the other hand, we can observe that the performance of the non-inherited configurations is significantly lower when the training set size is limited. Some of them even underperform the lower bound model. We will discuss it later in the Error Analysis section.

3.3 RQ2: Extent of Catastrophic Forgetting in Transfer Learning

In this section we study the extent of catastrophic forgetting following the same type definitions used in the RQ1. The results are depicted in Figures 5-7 in which we report the performance of the transferred models evaluated on the original source test sets. Take Figure 5 as an example. We fine-tuned the models pre-trained with the HD training set on the sampled HC training set ranging from 15 reports to 300 reports. We then plot the fine-tuned models' performance on the HD test set.

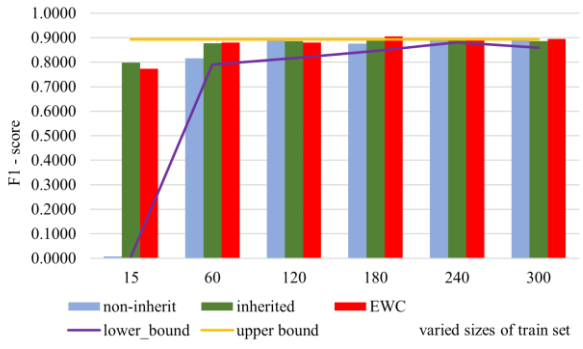

Figure 5: The HD test set performance of the HD (10) model fine-tune on the corresponding HC (10) training set with varied sizes.

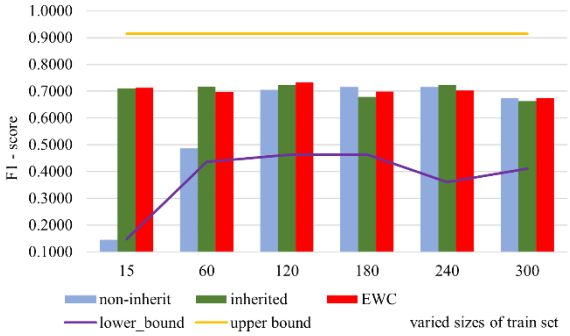

Figure 6: The HA test set performance of the HA (13) model fine-tuned on the HB (9) training sets with varied sizes.

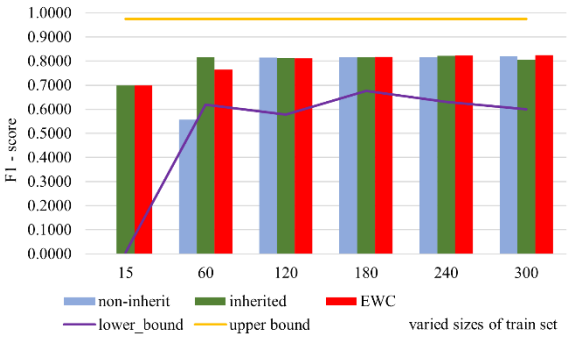

Figure 7: The HB test set performance of the HB (9) model fine-tuned on the corresponding HC (10) training sets with varied sizes.

As Figure 5-7 presented, when the target domain has less than 120 reports, the configuration of non-inherited has more serious extent of catastrophic forgetting than the inherited one. Furthermore, we observe that when the item types between the target and source domain are consistent, the extent of catastrophic forgetting for the inherited configuration is minor. As shown in Figure 5, when the size of the target domain's dataset increases, the performance of the source domain approaches the upper bound and even surpasses the models trained solely on the dataset of source domain.

With regard to the performance of EWC method, it was evident that EWC can mitigate forgetting more effectively. However, in some case EWC method perform worse than the inherited configuration when the amount of data is limited. One potential explanation for this phenomenon is that EWC's regularization of initially important parameters might lead to a slower learning rate.

It was noticed that some non-inherited configurations perform worse than the lower bound when the amount of target domain training set less than 60 reports. These cases occur when transferring from the source domain with fewer item types to a target domain with more item types. With respect to these errors, we will discuss them in following section.

4 Error Analysis

As mentioned in the previous chapter, this section focuses on the error analysis of the prominent discrepancies. First, as the result of RQ1 presented, we find that some of the non-inherited configurations underperformed the lower bound in case when they were fine-tuned on a limited training set like 15 reports. The error analysis demonstrates that fine-tuning the transferred model on such a limited dataset can enhances its recall on the target dataset, but its precision diminishes significantly, resulting in a reduced overall F-score. In contrast, the model without transfer learning struggles to recognize registry items such as G, NE, PN, TS, SC, and TNM. Nonetheless, it maintains the ability to recognize H (histology) across most cases, owing to this study only focus on the colorectal cancer type, thereby yielding a slightly higher F-score. Additionally, we notice that some histology terms like "Mucinous adenocarcinoma" appeared in one hospital's reports, does not appear in the other hospitals' reports. The counts for lymph node examination (NE) and positive nodes (NP) are typically denoted as integers in most hospitals. However, our investigation has revealed that, in the case of HC, some counts are directly expressed in English. For example, the sentence "Twelve dissected lymph nodes have no evidence of tumor metastasis" labels "Twelve" as "NE." As discussed above, directly predicting for unfamiliar

	Lower-bound	**Non-inherit**
H	**0.2985**	**0.0761**
G	**0.0000**	**0.0000**
NE	0.9748	0.9812
PN	0.9969	0.9969
TS	**0.0303**	**0.0435**
SC	0.9872	0.9829
T	0.9741	0.9697
N	0.9343	0.9343
Overall	0.6182	0.5573

Table 5: At 60 instances, when transferring from HB (9 categories) to HC (10 categories), and predicting the detailed NER performance of HB (bold scores are those below the micro-average).

knowledge can disregard the variations in labeling styles across target domains, resulting in higher accuracy compared to the transfer effect with randomly initialized parameters. This is also due to the combined impact of transfer and the random initialization of the linear layer.

Next, RQ2 discuss the extent of catastrophic forgetting, and the comparison table of HB fine-tuning result presented in Table 5. Additionally, during the examination of the original training data, it was noticed that a few annotation errors which may causing the confusion during the training phase and prediction confusion. For instance, "Grade 1 (moderately differentiated)" was entirely labeled as Histology, when in reality, this annotation should be "Grade". The above observation highlight the potential for annotation errors can contribute to inaccurate predictions and confusion in the training and prediction phases. We discovered that in the non-inherited setting, there are instances where “NOS” is wrongly predicted as Path_N, resulting in the frequent occurrence of “NOS” and the subsequent decrease in accuracy.

In conclusion, based on the observations from the results of RQ1 and RQ2, it's evident that the inherited approach indeed outperforms the non-inherited approach, and the EWC method exactly perform well when the target domain have more than 120 reports.

5 Conclusions

In this study, we aimed to mitigate catastrophic forgetting under transferring learning. The total of five different hospitals provided the unstructured reports of colorectal cancer. We utilized manually annotated pathology reports to create datasets which including 14 items of cancer registry. Our research method explored the importance of inherited parameters and the EWC method under various transfer learning scenarios with different labeling quantities and transfer orders. In RQ1, we arrive at the conclusion that regardless of the amount of target domain item, inheriting the parameters in the last linear layer with little training data leads to better performance. Besides, we also demonstrating that EWC doesn't negatively affect the training of the original model and that it effectively mitigates forgetting. The transfer order between unequal label types doesn't significantly impact the effectiveness of the approach. In RQ2, we demonstrated that EWC method can mitigate the extent of forgetting whether the quantities of transferring labels were consistent or not. The configuration of inheriting parameters cause the lower catastrophic forgetting when the target hospital had limited data.

The error analysis explained that the mislabeling led to the worse performance and the stylish of labeling cause the knowledge transferring problem. In the future work, we prefer to the integration of the labeling golden standard, and try more deep learning algorithm and regularization method on transferring to avoid forgetting.

Acknowledgments

The support for the research of this work from National Science and Technology Council [MOST 109-2221-E-992-074-MY3, NSTC 112-2221-E-992 -056 -MY3], and NKUST-KMU Research and Development and Industry-Academia Collaboration Grant Program [#NKUSTKMU-111-KK-027].

References

Arumae, K., Sun, Q., & Bhatia, P. J. a. p. a. (2020). An empirical investigation towards efficient multi-domain language model pre-training.

Dai, H.-J., Yang, Y.-H., Wang, T.-H., Lin, Y.-J., Lu, P.-J., Wu, C.-Y., . . . Hsu, Y.-C. J. I. A. (2021). Cancer registry coding via hybrid neural symbolic systems in the cross-hospital setting. *9*, 112081-112096.

Fleiss, J. L., Nee, J. C., & Landis, J. R. J. P. b. (1979). Large sample variance of kappa in the case of different sets of raters. *86*(5), 974.

Hutchinson, M. L., Antono, E., Gibbons, B. M., Paradiso, S., Ling, J., & Meredig, B. J. a. p. a. (2017). Overcoming data scarcity with transfer learning.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., . . . Grabska-Barwinska, A. J. P. o. t. n. a. o. s. (2017). Overcoming catastrophic forgetting in neural networks. *114*(13), 3521-3526.

Kuo, C.-N., Liao, Y.-M., Kuo, L.-N., Tsai, H.-J., Chang, W.-C., & Yen, Y. J. J. o. t. F. M. A. (2020). Cancers in Taiwan: Practical insight from epidemiology, treatments, biomarkers, and cost. *119*(12), 1731-1741.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109-165): Elsevier.

Ramasesh, V. V., Dyer, E., & Raghu, M. J. a. p. a. (2020). Anatomy of catastrophic forgetting: Hidden representations and task semantics.

Ratcliff, R. J. P. r. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *97*(2), 285.