HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

GRADUATION THESIS

Continual Relation Extraction with Feature Decorrelation

NGUYỄN HỮU HUY
huy.nh183553@sis.hust.edu.vn

Major: Computer Science
Specialization: Computer Science

Supervisor: MSc. Ngô Văn Linh ______________

Signature

Department: Computer Science

School: School of Information and Communications Technology

HANOI, 03/2023

ACKNOWLEDGMENT

I would like to thank Mr. Ngo Van Linh, my thesis advisor, for his guidance and assistance during my research process. His astute comments, unwavering support, and depth of research-related expertise were extremely useful when establishing the direction of my research.

I especially appreciate the lecturers at the Hanoi University of Science and Technology's School of Information and Communications Technology, whose knowledge and constructive criticism have improved my study ideas and approaches.

I am deeply grateful to VinAI Research for providing computational resources.

I want to express my sincere gratitude to my friends and family for their constant support and inspiration throughout this trying road. Their concern and love have given me courage and inspiration.

Finally, I would like to express my gratitude toward my best friend, Nguyen Van Chien. The completion of this work would not have been possible without his love and support.

ABSTRACT

Continual Relation Extraction (CRE) aims to continuously train a model to learn new relations while preserving its ability on old learned relations. To handle classification tasks, CRE models often include a softmax classifier that is trained using Cross-Entropy (CE) loss. The Nearest Class Mean (NCM) classifier, however, has recently been favored over the traditional Softmax classifier with Cross-Entropy loss in Continual Learning (CL) studies because of its higher performance. One flaw in their research is that they failed to provide an explanation for why the NCM classifier performs better than the Softmax classifier when CE Loss is present. In contrast to those works, I investigate how CE loss degrades performance by impairing the transferability of the learned features. In detail, I revisit CE loss by plugging CE loss into the state-of-the-art framework CRL and note that CRL does not use CE loss. I then give a concrete investigation of the impact of CE loss on the CRE problem through experiments using spectral analysis. In light of this analysis, I propose a simple yet effective class-wise regularization that improves the transferability of the representations. With my regularization, we can take advantage of both mechanisms Supervised Contrastive Learning (SCL) from CRL and Softmax classifier. I observe that my proposed regularization boosts the transferability of the representations and outperforms state-of-the-art CRE methods by a significant margin on the FewRel and TACRED datasets.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
BERT	Bidirectional Encoder Representation from Transformer
CE	Cross-Entropy
CL	Continual Learning
CR	Continual Replay
CRE	Continual Relation Extraction
CRL	Consistent Representation Learning
LSTM	Long-Short Term Memory
NCM	Nearest Mean Classifier
NLP	Natural Language Processing
PCA	Principal Component Analysis
RE	Relation Extraction
RNN	Recurrent Neural Network
SCL	Supervised Contrastive Learning

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

The Relation Extraction (RE) problem is a very important task in Information Extraction. Because of its effectiveness in extracting information from unstructured text, it is the key component of many NLP tasks, such as Information Retrieval [1], Question Answering [2] and Knowledge Graph Construction [3]. In particular, a relation extraction system is expected to classify the semantic relation between two entity mentions in the given context. To deal with this problem, many methods have been proposed and achieved remarkable results [4]–[6]. Nevertheless, a majority of previous RE studies only considered the traditional setting where the set of relations is pre-defined and fixed during the training and testing phases. This setting is not practical as new relations of interest might emerge during the deployment time of RE systems in practice, requiring the models to adapt their operation to accommodate new types. Therefore, Continual Relation Extraction was proposed, aiming to learn new relations from new coming data. Recently, CRE has attracted considerable attention in the literature [7]–[10] because of not only its appealing practical applications but also the challenging problems which come from both fields, Continual Learning, and Relation Extraction.

1.2 Background and Problems of Research

Compared to conventional RE, CRE has to face the stability-plasticity trade-off. Generally speaking, this is a fundamental problem in the continual learning paradigm, where stability relates to maintaining accuracy on the previous tasks, while plasticity relates to learning emerging tasks effectively. Modern deep learning models adapt to new knowledge quickly while lacking stability, this phenomenon is called Catastrophic Forgetting (CF).

Based on how models mitigate Catastrophic Forgetting, existing approaches can be categorized into one of the following three groups: (1) Regularization-based methods [11]–[13] mitigate catastrophic forgetting by constraining the updates of some network parameters depending on their importance. (2) Dynamic architecture methods [14], [15] adjust network capacity dynamically to learn emerging tasks effectively. (3) Replay-based methods [16]–[19] deploy a memory buffer to save a small number of samples from old tasks for later replay. Among three major approaches, the replay-based method has shown the best results for NLP tasks, including CRE.

The current state-of-the-art of the CRE task is Contrastive Replay (CRL) [10] which is a replay-based method. Different from previous work, CRL maintains learned knowledge by introducing a contrastive replay mechanism that removes the uses of Softmax classifier and Cross-Entropy loss. It instead proposes to use Supervised Contrastive loss and Nearest Class Mean classifier to make the representation space well-separated. Although CRL outperforms older approaches that use traditional Cross-Entropy loss, it did not learn this pattern in-depth. As a result, it begs the question of why models with Cross-Entropy loss perform so poorly.

1.3 Research Objectives and Conceptual Framework

The goal of my research is to investigate the poor performance in the CRE task caused by Cross-Entropy loss. From the investigation, I will propose solutions to improve the performance of the CRE task.

Existing works in Continual Learning mainly focus on the Catastrophic Forgetting problem, while overlooking other factors. In this thesis, instead of traditional approaches where many methods were proposed to preserve old knowledge, I give a concrete analysis of the transferability of the representations in the CRE setting. Especially, I focus on the bad impact of traditional Cross-Entropy loss on the transferability and how to improve it.

In an earlier work, [20] stated an issue in representation learning for image classification, representation bias. First, if the encoder is fixed after learning old tasks, it can preserve the learned representation space but the learned features are only helpful for the old tasks and not for the new tasks to classify. In contrast, if the encoder is updated with new knowledge, the updated representations would be not suitable for previous tasks. For better understanding, I give an example of CRE. First, a classifier is trained to classify between two relations $city\_of\_birth$ and $parents$, but I assume that all the cities from $city\_of\_birth$ samples are located in China. Then the learned feature extractor, by chance, can only classify texts if it mentions China or not. In the next task, it has to classify between $capital\_of$ and $siblings$, I see that the learned feature cannot help if $capital\_of$ samples mention cities around the world. This learned feature is considered not transferable. While if the feature extractor can learn a feature that distinguishes between people and places, it can still perform well on classifying between $capital\_of$ and $siblings$. This feature is transferable because, without updating the feature extractor, the representation space is separated to some extent. If the feature extractor learns non-transferable features in the previous task, these features can be changed considerably to adapt to the current task. Consequently, old learned features for previous tasks can be forgotten, leading to the performance reduction on those tasks [20]. They