Machine Intelligence and Reliable AI Lab

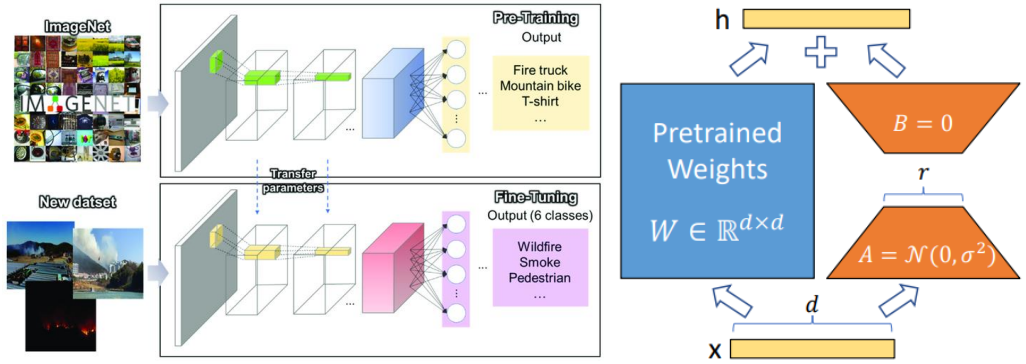
School of Software Convergence, Myongji University

Machine Intelligence and Reliable AI Lab

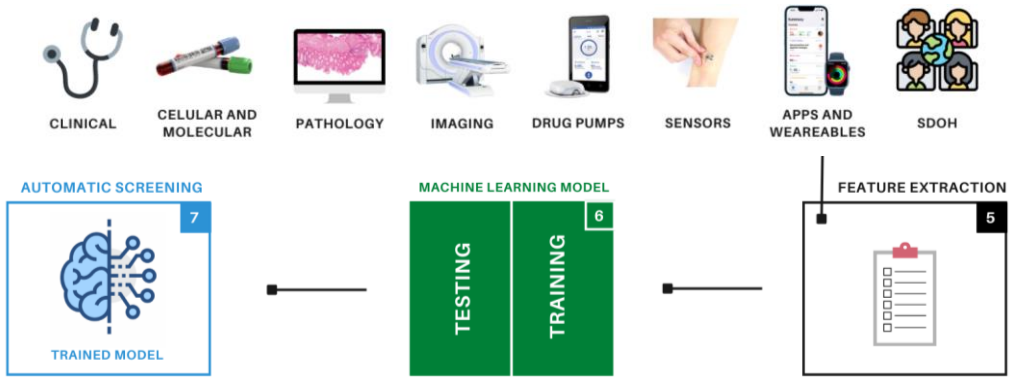
- Professor – Daeyoung Hong (홍대영)
  - Yonsei University, Seoul, South Korea (Mar 2010 – Feb 2016)
    B.S. in Electrical & Electronic Engineering
  - Seoul National University, Seoul, South Korea (Mar 2016 – Feb 2024)
    Ph.D. in Electrical & Computer Engineering
  - Seoul National University, Seoul, South Korea (Mar 2024 – Aug 2024)
    Postdoctoral Researcher in Institute of Computer Technology
  - Myongji University, Seoul, South Korea (Sep 2024 – Present)
    Assistant Professor in School of Software Convergence

Research Areas

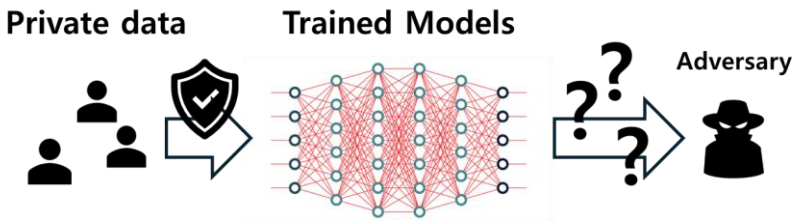
Leveraging Large Models

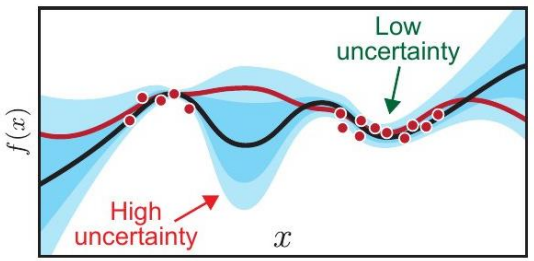


Medical AI

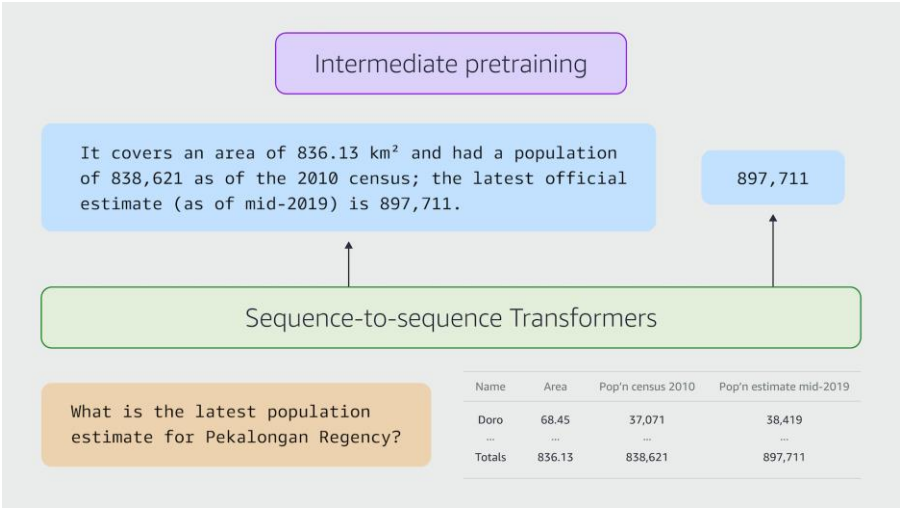


Reliable & Trustworthy AI





NLP / Recommendation System / ...



Recent Papers

Large Deep Learning 모델 활용 & Reliable AI (2025)

International Conference on Computational Linguistics

DP-FROST: Differentially Private Fine-tuning of Pre-trained Models with Freezing Model Parameters

Daeyoung Hong
Myongji University
South Korea
dyhong@mju.ac.kr

Woohwan Jung
Hanyang University
South Korea
whjung@hanyang.ac.kr

Kyuseok Shim*
Seoul National University
South Korea
kshim@snu.ac.kr

Abstract

Training models with differential privacy has received a lot of attentions since differential privacy provides theoretical guarantee of privacy preservation. For a task in a specific domain, since a large-scale pre-trained model in the same domain contains general knowledge of the task, using such a model requires less effort in designing and training the model. However, differentially privately fine-tuning such models having a large number of trainable parameters results in large degradation of utility. Thus, we propose methods that effectively fine-tune the large-scale pre-trained models with freezing unimportant parameters for downstream tasks while satisfying differential privacy. To select the parameters to be fine-tuned, we propose several efficient methods based on the gradients of model parameters. We show the effectiveness of the proposed method by performing experiments with real datasets[1].

1 Introduction

Deep learning models conventionally trained on private user data have risks of leaking sensitive information (Shokri et al., 2017; Carlini et al., 2019, 2021). To protect sensitive statistical data, the differential privacy (Dwork and Roth, 2013) is widely used due to theoretical guarantee of privacy preservation by limiting the influence of any individual user's data and injecting noise into the results computed from the sensitive data. The differential privacy has also attracted attentions in the machine learning community to build a differentially private model (Abadi et al., 2016).

To satisfy differential privacy, deep learning et al., 2021), fine-tuning large pre-trained models may significantly degrade their accuracies. Thus, LoRA (Li et al., 2022) adds a small-sized low-rank matrix to each existing pre-trained weight matrix in a model and fine-tunes only the low-rank matrices. Similarly, Adapter (Yu et al., 2022) adds a small-sized intermediate layer after each existing layer of the pre-trained model, and fine-tunes only the intermediate layers. However, both methods fine-tune only added parameters without considering the impact of parameters to improve the accuracy. On the other hand, the method in (Luo et al., 2021) freezes the pre-trained model parameters with small absolute values and fine-tune only the other parameters. But updating such freezed parameters may significantly improve the accuracy of the trained model.

To overcome the drawbacks of existing works, we propose the DP-FROST (Differentially Private Fine-tuning of pRe-trained mOdelS with freezing model parameTers) method. The proposed method (1) splits the model parameters into disjoint partitions, (2) adds noise to the partition-wise gradient magnitude, (3) selects partitions with their largest noisy partition-wise gradient magnitudes and (4) updates the parameters in the selected partitions.

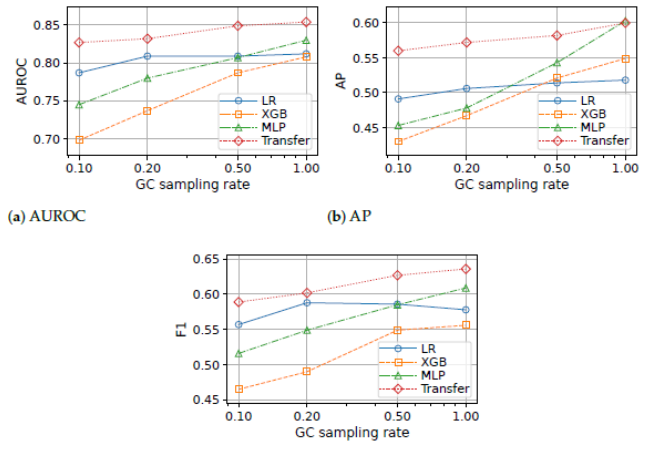
Our contributions: The contributions of this paper are summarized below.

- We prove that the model parameters with large gradient magnitudes contribute more to its accuracy than those with small gradient magnitudes.
- To reduce the amount of inserted noise, instead of selecting the parameters in the granularity of parameters by adding noise to their gradient

Medical AI (2025)

SCI급 국제 논문 (학부생 인턴과 공동으로 연구)

3.3. Model performance under reduced GC training labels



(a) AUROC (b) AP

(c) F1

Figure 1. Performance across the GC sampling rate r. Lower r simulates fewer labeled GC training cases.

As summarized in Figure 1 (AUROC, AP, and F1 panels), we varied the GC sampling rate $r \in \{1.0, 0.5, 0.2, 0.1\}$ to quantify robustness under limited GC labels. For each r, we applied stratified sampling at proportion r on the GC training partition (preserving the case:control ratio), while keeping the GC validation and test partitions fixed. Non-GC source datasets for pretraining and all preprocessing were unchanged across r. All models (LR, XGB, MLP-from-scratch, and Transfer) were then trained on the resulting r-subsampled GC training set for that rate. The Transfer model always used the same pretraining configuration on non-GC cancers and was fine-tuned on the r-subsampled GC training set.

Recent Participation in R&D Projects

- Development of an AI-Based Decision Support System for Full-Lifecycle Gastric Cancer Management, National Cancer Center, Korea (Apr. 2025 — Present)
- Development of Lightweight Multimodal Anti-Phishing Models and Split-Learning Techniques for Privacy-Preserving Anti-Phishing, Institute of Information & communications Technology Planning & Evaluation (IITP), Korea (Jan. 2024 — Present)
- Development of Web Table Question Answering System using Pre-trained Language Model, National Research Foundation of Korea (NRF), Korea (Jun. 2023 — Present)
- Derivation of a Differential Privacy Concept Applicable to National Statistics Data While Guaranteeing the Utility of Statistical Analysis, Institute of Information & communications Technology Planning & Evaluation (IITP), Korea (Apr. 2022 — Dec. 2024)

Collaborating Institutions

고려대학교 KOREA UNIVERSITY

고려대학교안산병원 KOREA UNIVERSITY ANSAN HOSPITAL

부산대학교병원 Pusan National University Hospital

서울대학교 SEOUL NATIONAL UNIVERSITY

Join Us

- Opportunities for Students
  - Undergraduate Research Internships
  - Graduate (M.S./Ph.D.) Programs

- If you are interested in joining our lab (as an ***intern***, ***M.S.***, or ***Ph.D.*** student), please email me directly:

dyhong@mju.ac.kr