FEEDBACK

by the official opponent on the dissertation work of

Xu Jiashu

on the topic «Research and Development of Self-Supervised Visual Feature Learning Based on Neural Networks»,

submitted for the degree of Doctor of Philosophy

in the field of knowledge 12 Information Technologies

with the specialization 121 Software Engineering

Relevance of the dissertation topic.

The research theme of this dissertation, "Self-Supervised Visual Feature Learning", represents a current hotspot in the field of computer vision. This thesis primarily addresses the challenge of effectively training deep learning algorithms without reliance on extensively labeled datasets. By exploring self-supervised learning algorithms, it introduces two novel approaches: 'Mixup Feature' and 'Denoising Self-Distillation Masked Autoencoder'. These methods constitute significant additions to the broader family of self-supervised learning algorithms. Notably, they reduce dependence on labeled data while achieving accuracy comparable to supervised visual feature learning methods. In practical applications, particularly in medical image analysis, where acquiring a large volume of labeled medical imagery is often challenging, these algorithms hold immense potential. Given the global shortage of medical experts and the escalating demand for diagnostic services, there is an urgent need for accurately interpreting medical images with minimal human intervention. This work could facilitate the development of cost-effective pre-training schemes that obviate the need for expensive and time-consuming data labeling.

Evaluation of the Justification of the Dissertation's Scientific Results, Their Reliability, and Novelty.

The scientific novelty of the dissertation research results lies in the following:

A novel self-supervised learning methodology leveraging the Mixup Feature function is introduced. This method involves the pre-training of visual representations by predicting Mixup features from masked images, which stand in as proxies for advanced semantic information. The approach is poised to potentially booster the aggregate efficacy of the model.

A masked autoencoder model for self-supervised learning is proposed, featuring novel mechanisms for noise suppression and self-distillation. The architecture utilizes

a masked autoencoder in conjunction with a teacher network to facilitate the reconstruction of corrupted image segments afflicted with random Gaussian noise. This model extends the utility of self-supervised techniques in restoring visual data.

For the first time, a model is proposed that, by combining losses at the pixel level and feature level, enables the extraction of deep semantic characteristics of the image. This complements existing techniques for modeling masked images and also increases the robustness of self-supervised learning models when working with unbalanced data sets.

Thus, in the dissertation work, the scientific task has been fully accomplished, and the candidate has fully mastered the methodology of scientific activity.

Evaluation of the content of the dissertation, its completeness, and adherence to the principles of academic integrity.

The content of the dissertation work by Xu Jiashu fully meets the Higher Education Standard for the specialization 121 Software Engineering and research directions according to the Software Engineering educational program.

The dissertation is a completed scientific work and demonstrates the presence of the candidate's personal contribution to the scientific direction of Self-Supervised Visual Feature Learning.

Having examined the similarity report based on the check of the dissertation for text matches, it can be concluded that the dissertation work of Xu Jiashu is the result of the candidate's independent research and does not contain elements of falsification, compilation, fabrication, plagiarism, and borrowing. The ideas, results, and texts of other authors used have proper references to the relevant source.

Language and Style of Presentation of Results.

The dissertation is written in English.

It comprises an introduction, 4 chapters, conclusions, a reference, and appendices. The total length of the dissertation is 168 pages.

In the introduction, the dissertation underscores the increasingly pivotal role of Self-Supervised Learning (SSL) in deep learning, particularly for tasks where labeled data is scarce, such as medical image analysis. This research is dedicated to advancing SSL for image understanding, a domain within computer vision that poses significant challenges and remains underexplored. The introduction initially highlights the significance of SSL and its applications in both natural language processing and computer vision. The principal aim is to establish a self-supervised learning framework designed to extract visual features in the absence of labeled data, with a special focus on applications in medical imaging. Furthermore, it delineates

key tasks, including the development of innovative SSL algorithms, the creation of effective model architectures, the definition of self-supervised tasks and loss functions, the application of transfer learning, and the evaluation of model performance. The introduction culminates with a summary of its contributions and the scientific novelty it presents. It especially emphasizes the potential applications of these advancements in medical image classification, showcasing how this research contributes new insights and methodologies in the realm of computer vision and self-supervised learning.

In the first chapter, there is a comprehensive review of the evolution of self-supervised learning algorithms used in visual feature learning. While the methodologies selected may no longer be considered competitive in the contemporary context, these early self-supervision techniques have indelibly shaped the foundation upon which current strategies are built. These algorithms are broadly categorized into four main types: contrastive learning, masked image modeling, self-distillation, and canonical correlation analysis. The chapter provides an extensive overview of various methodologies, ranging from early techniques to recent approaches based on Vision Transformers (ViT). It meticulously examines how these methods collectively contribute to the advancement of self-supervised visual representation learning. This thorough analysis not only highlights the historical progress in the field but also sets the stage for understanding current and future trends in self-supervised learning within computer vision.

In Chapter 2, the text presents two self-supervised learning algorithms that are centered around masked image modeling, offering substantial advancements in the field of visual feature learning. The first algorithm, Mixup Feature, proposes a new pretext task of reconstructing a Mixup of traditional image features like Sobel, HOG, and LBP as the target for a masked autoencoder. This unique mixed feature target provides more complex visual representations for pre-training. The second algorithm, Denoising Self-Distillation Masked Autoencoder, combines self-distillation with masked autoencoders for robust denoising. It considers both pixel-level image restoration and feature-level regression. Gaussian noise is randomly added to image patches as a pretext task for the student network to denoise. The teacher network guides the student through exponential moving average parameter updates. An asymmetric decoder further enhances efficiency. The loss function balances pixel reconstruction loss and feature alignment loss.

In Chapter 3, comprehensive experiments were conducted to verify the two proposed self-supervised learning algorithms - Mixup Feature and Denoising Distillation Masked Autoencoder. For Mixup Feature, the model was pretrained on the CIFAR-10, CIFAR-100, and STL-10 datasets. Downstream evaluation involves linear probing and full fine-tuning. Results showed that mixing traditional features like Sobel, HOG, LBP improved performance compared to single features. The

unnormalized combination of features delivered the most optimal performance when set as targets. Ideally, a masking ratio ranging from 40-60% is preferred. The mixup factor λ of 0.2-0.3 yielded the best results for the HOG-Sobel mixup. This method matches or surpasses MAE and is superior to contrastive learning. For the Denoising Distillation Masked Autoencoder, the model adopted ViT as the backbone network and was pretrained on the same three datasets. Removing randomly added Gaussian noise blocks served as the pretext task. Full fine-tuning and linear probing benchmark tests exhibited comparable performance to MAE and MaskFeat within 500 epochs, faster than the 1600 epochs of MAE. This strategy outperforms the seminal contrastive learning approach of MoCo v3. The ablation study confirmed the combined effect of pixel reconstruction loss and feature regression loss, demonstrating the effectiveness of the proposed framework. In conclusion, comprehensive experiments demonstrated the effectiveness of the two proposed self-supervised learning algorithms on three different datasets compared to existing methods. The results verified the innovation of advancing masked image modeling techniques through mixed feature objectives and denoising distillation MAE. Future work can evaluate extending these methods to larger datasets.

In Chapter 4, the focus is on investigating the application of self-supervised pre-trained models in the classification of CT scans. This chapter examines how self-supervised pre-training learns representations from unlabeled CT scans, addressing key challenges in this area. Three self-supervised learning methods - MAE, Mixup Feature, and Denoising Self-Distillation MAE - were applied to pre-train on the unlabeled COVID-CTset dataset. Reconstruction images on held-out data indicated the model captured global anatomical structure but faced challenges with fine tissue textures. However, reconstruction performance does not necessarily correlate with transfer learning ability. The pre-trained models were fine-tuned on two smaller labeled datasets for classification. All self-supervised methods outperformed direct training, demonstrating enhanced performance through self-supervised pre-training. Mixup Feature achieved results comparable to MAE, highlighting its potential in medical domains. To evaluate robustness under real-world class imbalances, models were tested with varying positive and negative sample ratios. Self-supervised pre-training stabilized performance as imbalance increased, whereas direct training showed steeper declines. This suggests self-supervised learning provides a more robust feature representation foundation. Comparison to CNNs pre-trained on ImageNet revealed self-supervised learning maintained higher accuracy even at extreme 8:1 imbalance. While supervised pre-training performed slightly better with balanced data, self-supervised learning exhibited stronger generalization to highly imbalanced scenarios. In conclusion, this work demonstrated the promise of self-supervised pre- training for enhancing CT scan analysis and robust medical image classification.

In the conclusions section, the dissertation summarizes the results that hold both academic value and practical applicability.

The dissertation is formatted in accordance with the requirements of the Order of the Ministry of Education and Science of Ukraine dated January 12, 2017, No. 40 "On Approval of the Requirements for Dissertation Formatting".

Publication of the results of the dissertation work.

The scientific results of the dissertation are presented in 4 scientific publications by the candidate, including: 4 articles in periodic scientific journals indexed in the Web of Science Core Collection and Scopus databases, of which 1 article is in journals ranked in the first quartile (Q1) according to the SCImago Journal and Country Rank or Journal Citation Reports classification;

Additionally, the results of the dissertation were presented at 3 scientific professional conferences.

Thus, the scientific results described in the dissertation are fully reflected in the candidate's scientific publications.

Shortcomings and Comments to the dissertation work.

1. The dissertation presents two innovative methods that demonstrate promise through positive results in experiments across three datasets. However, it is recommended that the authors compare these methods more thoroughly with current state-of-the-art techniques.
2. While the proposed methodologies show enhancements in performance metrics, the degree of improvement appears limited when benchmarked against the state-of-the-art in the field.
3. A notable limitation of the study is its lack of experimental validation on large-scale datasets, such as ImageNet. Given the importance of such datasets in assessing scalability and generalizability, the authors are advised to extend their evaluations to include these or similar comprehensive datasets.
4. The dissertation contains some instances of imprecise language that need to be addressed for improvement.

I believe that the expressed remarks are not decisive and do not diminish the overall scientific novelty and practical significance of the results and do not affect the positive evaluation of the dissertation work.

Conclusion about the dissertation work.

I consider that the dissertation work of the candidate for the Doctor of Philosophy degree, Xu Jiashu, on the topic "Research and Development of Self-Supervised Visual Feature Learning Based on Neural Networks" is conducted at a

high scientific level, does not violate the principles of academic integrity, and is a completed scientific research. The collective theoretical and practical results solve a scientific task of significant importance for the field of Self-Supervised Visual Feature Learning. The dissertation work, in terms of relevance, practical value, and scientific novelty, fully meets the requirements of the current legislation of Ukraine as provided in paragraphs 6 – 9 of the "Procedure for awarding the degree of Doctor of Philosophy and cancellation of the decision of a one-time specialized academic council of a higher education institution, scientific institution on awarding the degree of Doctor of Philosophy," approved by the Resolution of the Cabinet of Ministers of Ukraine dated January 12, 2022, No. 44.

The candidate Xu Jiashu deserves to be awarded the degree of Doctor of Philosophy in the field of knowledge 12 Information Technologies with the specialization 121 Software Engineering.

Official opponent:

Professor of the Department of Mathematical Informatics, the Faculty of Computer Sciences and Informatics, Taras Shevchenko Kyiv National University, Doctor of Sciences in Physics and Mathematics, Professor Oleksandr MARCHENKO

Підпис Марченка засвідчую
Заст. декана О.А.Капустян

Seal

«21» February 2024