CS512 Class Project: Exploring self-supervised learning methods for unlabeled dataset

Zong Fan
zongfan2@illinois.edu
University of Illinois at Urbana, Champaign
Urbana, Illinois, USA

ACM Reference Format:
Zong Fan. 2022. CS512 Class Project: Exploring self-supervised learning methods for unlabeled dataset. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROPOSAL CONTENT

1.1 Description of problem

1.1.1 Limitation of Supervised learning methods. Since the idea of machine learning is introduced, deep neural networks have shown outstanding performances on countless tasks in various fields, especially on supervised learning in the fields of imaging (Computer Vision), text (Natural Language Processing) and graph. In general, supervised learning methods usually need large datasets to achieve satisfying performance. Also, these methods require the fully-annotated labels to provide full supervision to optimize the model. However, there are some limitations with supervised learning. First, annotating data can be very time- and effort-consuming, especially when annotating process need expertise knowledge which can be very costly. Therefore, large-scale annotated dataset are not always available in many cases. Second, supervised learning might suffer from generalization error, spurious correlations, and adversarial attacks [5]. As a promising alternative, self-supervised learning is introduced to enable computers to label, categorize, and analyze data automatically or semi-automatically, which shows the impressive data efficiency and generalization ability in both computer vision and natural language processing community [9].

1.1.2 Potential self-supervised learning methods. Instead of using fully-annotated data during training, a lot of self-supervised learning methods have been proposed which address the limitation by obtaining the data labels in a semi-automatic process [5, 9]. The intuition of self-supervised learning is to leverage the inherent co-occurrence relationships as the self-supervision from large amounts of unlabeled data [5]. Such inherent knowledge can help the model learn the data distribution more comprehensively, significantly alleviating the out-of-distribution or generalization problem and improving the model performance [11]. Considering the difference of model architecture and objectives, most self-supervised learning methods can be grouped into three types: generative methods, contrastive methods, and adversarial methods [5]. In order to simulate input data distribution, generative methods train an encoder to encode input sample into an explicit vector and a decoder to reconstruct the sample from the vector. Contrastive methods would train an encoder to encode input sample into an explicit vector to measure feature similarity. Adversarial methods employ encoder-decoder architectures to generate fake samples and a discriminator to distinguish the fake samples from real samples. Previous study show that the performance of generative methods are less competitive than the other two methods, primary due to the inherent defects of point-wise nature of the generative objective. Therefore, in this study, we'd like to collect several either image or text datasets and implement several contrastive and adversarial methods for comparison, in order to investigate their advantages and disadvantages under different conditions.

1.2 Work plan

Here's the time line of our project:

(1) First Month(now - 10.18): Learn through the basis of self supervised learning and settled on the dataset to use. Here are several candidate datasets to be investigated:
Image dataset:
- NIH Chest X-ray Dataset [8]: This is a chest X-ray image dataset with 5606 images of 1024×1024 pixels, which can be used for classification and localization of several thorax diseases.
- STL-10 Image Recognition Dataset [2]: This is a natural image dataset with a total of 10,000 images of 96×96 pixels in 10 classes. Only 500 of them are annotated and the rest images are unlabeled, which can be used for analysis of self-supervised classification methods.

Text dataset:
- Emotion Dataset for Emotion Recognition Tasks [6]: This is a English text dataset collected on Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise.

(2) Second Month(10.19 - 11.16): Finish building the self supervised learning model for both contrastive and adversarial methods and apply on the selected datasets.
(3) Third Month(11.17 - 12.12): Fine-tune the model to increase the performance of both contrastive and adversarial methods, contrasting the two methods and analyzing the performance.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.3 Team role

- Xiaobai Li: investigating constractive self supervised learning methods
- Zong Fan: investigating adversarial self supervised learning methods

2 MIDTERM REPORT

2.1 Progress of project

After doing comprehensive literature search, we determined to focus on self-supervised learning methods in the computer vision field. For comparison, natural language processing has achieved great success by use of self-supervised learning approaches based on masked language modeling technique. Nowadays, very large-scale language models trained on huge amounts of text data, such as BERT [3] and GPT-3 [1], have achieved state-of-art performances in a broad range of natural language processing tasks, including text synthesis, question answering, language translation, common sense reasoning, text completion, etc.

However, although the idea of masked autoencoding in these NLP models has attracted a lot of research interest, the research progress of this technique in computer vision still lags behind the NLP, which may be explained by difficulties of transferring the method to different data modalities. Recently, several approaches have been proposed in computer vision to address these difficulties, such as masked autoencoder [4] and simple masked image modeling framework (SimMIM) [10]. In this study, we will mainly explore the SimMIM method as the self-supervised representation learning method and investigate its effect in computer vision tasks, especially the classification performance on unlabeled data samples.

For this goal, an medical image dataset with breast tumor histopathological images called BreakHis [7] was also determined for testing the method. We will discuss them in detail as follows.

2.1.1 Selected Method: SimMIM. SimMIM was proposed by Xie et al [10], which masks random patches from the input image and reconstructs the missing patches in the pixel space by use of transformer architecture via regression. The framework architecture of SimMIM is shown in Figure. 1. To understand the philosophy of framework design, it's important to know the difference between the task of natural language processing and computer vision. First, the information density is different between language and vision. Images are raw and low-level signals which has high spatial redundancy and relatively low information density. Language are high-level human-generated data which is highly semantic and has high information density. The visual signals are more continuous than the discretization of language tokens. Second, the images have a strong locality where neighboring pixels usually share highly-correlated information, while the language tokens don't show such a strong locality. Third, the decoder of the autoencoder architecture maps the latent representation back to the input, which is important for reconstructing texts and images. In vision, the decoder reconstructs pixels that are lower semantic level. In contrast, the decoder predicts missing words that contain rich semantic information in language. Therefore, SiMMIM proposed several innovative designs to address these differences as follows.

- The input image was first divided into regular non-overlapping patches. Random masking was applied on image patches. Random masking with a high masking ratio or larger patch size eliminates potential redundancy and makes the reconstruction task unable to be solved easily by extrapolating from visible neighboring patches. Such highly sparse input makes it possible for the encoder to efficiently learn the meaningful representation from the input image for reconstruction.
- A raw pixel regression task was designed to predict the value of missing pixels. This design aligns well with the continuous nature of visual signals. It was conducted by a lightweight prediction head which can significantly speed up training.

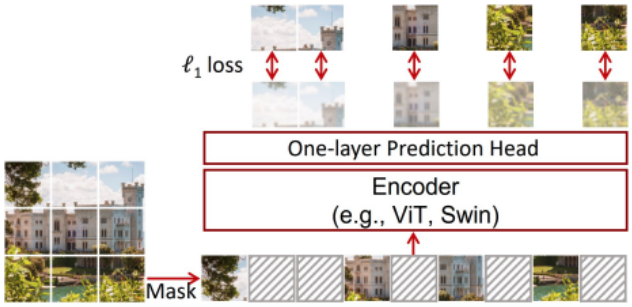

Figure 1: Framework architecture of SimMIM. After the encoder extracts representations of each patch, a one-layer lightweight head is designed to predict the raw pixel values of the randomly masked patches. A simple L1 loss was employed to supervise the model training. Credit: Xie et al. 2022 [10]

The code of this approach has been released publicly on GitHub (https://github.com/microsoft/SimMIM). We are currently reading the code and will try to start running it with a toy dataset soon. There are four major components of this method, including masking strategy, encoder architecture, prediction head, and prediction target. We will probably adjust some of them on our data in order to achieve the optimal learning effect.

2.1.2 Datasets. The public breast tumor histopathological image database, BreaKHis (http://web.inf.ufpr.br/vri/breast-cancer-database), is utilized to demonstrate the performance of the selected method. The whole-slide images of HE-strained histopathological images provide rich texture and pattern information of breast tumor lesions. This can support the analysis of tumor samples and help pathologists make clinical decisions on lesion diagnosis, treatment strategy design, and other clinical applications. From the perspective of clinical diagnosis, digital pathology images can be employed to differentiate lesion subtypes or binary classification between malignancy and benign. However, annotating the histopathological images usually needs expertise knowledge and large amounts of time, which is very expensive and time-consuming in obtain large-scale fully annotated dataset. Therefore, it's always desired to utilize the unlabeled image to facilitate the prediction accuracy.

This database includes a total of 7909 breast cancer histopathology images acquired on 82 different patients. I n this database, four benign breast tumor subtypes: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); and four malignant breast tumor subtypes: ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC) are available. Also, these images have four kinds of magnifications, including 40X, 100X, 200X, and 400X. In this study, 40X images would be first investigated since 40X images usually achieved better performance using the supervised learning method [7]. The detail of the 40x BreakHis data used in experiments is shown in Table 1 and several image samples are shown in Figure 2.

Tumor	type	Number of images
Adenosis	benign	114
Fibroadenoma	benign	253
Phyllodes tumor	benign	109
Tubular adenoma	benign	149
Ductal carcinoma	malignant	864
Lobular carcinoma	malignant	156
Mucinous carcinoma	malignancy	205
Papillary carcinoma	malignancy	145

Table 1: Distribution of 40X images

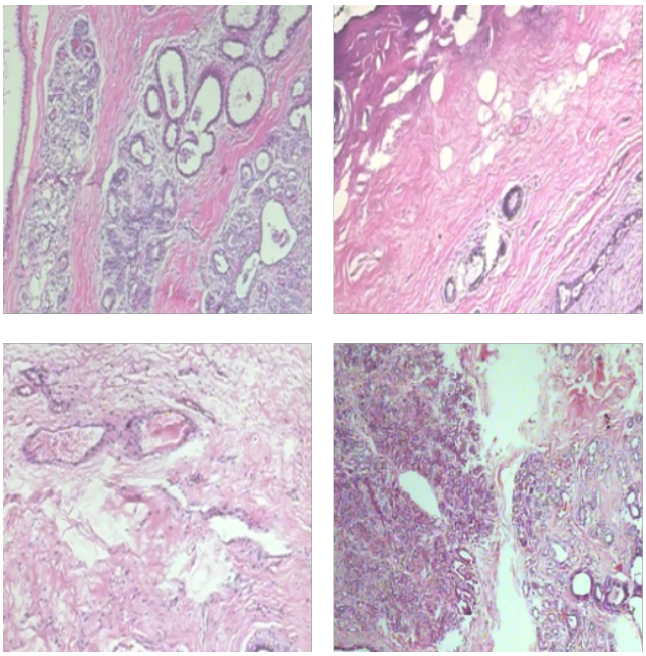

Figure 2: Examples of 40x BresKHis histopathological images

2.2 Plans of rest milestones

Zong Fan will do the rest experiments, since his roommate dropped the class. So the goal of this project is modified to only investigate a generative-based self-supervised learning strategy, SimMIM, as discussed above. Here are the rest milestones.

(1) First half month (10.19 - 11.2): Preprocess the histopathological image dataset. Read the SimMIM code and start training a toy model by resizing the images with input size as 32×32.
(2) Second half month (11.2-11.16): Debug and train the model with full image size (512×512). Investigate the factors that may affect the model performance, such as random masking ratio and patch size.
(3) Third half month (11.16-12.1): train a fully-supervised learning model and semi-supervised learning model on the unlabeled images for comparison. Compare the learned feature representation of each model and analyze the interesting findings
(4) Last half month (12.1-12.12): Summarize experiment results and write the final report.

REFERENCES

[1] Tom B. Brown et al. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.
[2] Adam Coates, A. Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*.
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. 2022. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.
[5] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. 2020. Self-supervised learning: generative or contrastive. *ArXiv*, abs/2006.08218.
[6] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: contextualized affect representations for emotion recognition. In *EMNLP*.
[7] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. 2015. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63, 7, 1455–1462.
[8] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462–3471.
[9] Lirong Wu, Haitao Lin, Zhangyang Gao, Cheng Tan, and Stan.Z.Li. 2021. Self-supervised learning on graphs: contrastive, generative,or predictive. *IEEE Transactions on Knowledge and Data Engineering*.
[10] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: a simple framework for masked image modeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9643–9653.
[11] Keyulu Xu, Jingling Li, Mozhi Zhang, Simon Shaolei Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2021. How neural networks extrapolate: from feedforward to graph neural networks. *ArXiv*, abs/2009.11848.