Efficient Medical Image Assessment via Self-supervised Learning

Chun-Yin Huang[1], Qi Lei[2], and Xiaoxiao Li[1]

[1] University of British Columbia
[2] Princeton University
{chunyinh, xiaoxiao.li}@ece.ubc.ca, qilei@princeton.edu

Abstract. High-performance deep learning methods typically rely on large annotated training datasets, which are difficult to obtain in many clinical applications due to the high cost of medical image labeling. Existing data assessment methods commonly require knowing the labels in advance, which are not feasible to achieve our goal of *'knowing which data to label.'* To this end, we formulate and propose a novel and efficient data assessment strategy, **EX**ponenti**A**l **M**arginal s**IN**gular valu**E** (EXAMINE) score, to rank the quality of unlabeled medical image data based on their useful latent representations extracted via Self-supervised Learning (SSL) networks. Motivated by theoretical implication of SSL embedding space, we leverage a Masked Autoencoder [8] for feature extraction. Furthermore, we evaluate data quality based on the marginal change of the largest singular value after excluding the data point in the dataset. We conduct extensive experiments on a pathology dataset. Our results indicate the effectiveness and efficiency of our proposed methods for selecting the most valuable data to label.

1 Introduction

Artificial intelligence (AI) such as deep learning has became a powerful tool for medical image analysis. Its success relies on the availability of abundant high quality dataset. However, medical images collected from different sources vary in their quality due to the various imaging devices, protocols and techniques. When trained with low-quality data, AI models can be compromised. Furthermore, labeling medical images for AI training requires domain experts and is usually costly and time consuming. Therefore, it is demanding to have an automated framework to effectively assess and screen data quality before data labeling and model training.

There are numerous definitions of data quality. Data is generally considered to be of high quality if "fit for [its] intended uses in operations, decision making and planning." [4, 5, 16]. In the context of training an AI predictive model, good data are the fuel of AI. Namely, data with better quality can help obtain higher prediction accuracy. However, how to quantitatively assess data's quality for AI tasks is under-explored. Previous works [6, 10] mainly propose to estimate data values in the context of supervised machine learning, which requires knowledge

of labels and repeated training of a target utility. Such setting lacks practical value as data labels are typically not available at the data preparation stage for data privacy, labeling cost, and computational efficiency concerns. Differently, we aim to develop a cost-effective scheme for data assessment in the context of unsupervised learning to tackle the limitations of the existing methods, in which no labeling is required during assessment.

A trending and powerful unsupervised representation learning strategy is self-supervised learning (SSL). SSL solves auxiliary pretext tasks without requiring labeled data to learn useful semantic representations. These pretext tasks are created solely using the input features, such as predicting a missing image patch [8], recovering the color channels of an image from context [19], predicting missing words in texts [12], forcing the similarity of the different views of images [1, 7], etc. Motivated by the recent discovery that SSL could embed data into linearly separable representations under proper data assumptions [13, 17], we show that 'good' and 'bad' data can be distinguished by examining the change of the data representation matrices' singular value by removing a certain data point.

In this work, we tackle a practically demanding yet challenging problem — medical image assessment (also referred as data assessment in this work). To this end, we develop a novel and efficient pipeline for medical image assessment *without knowing data labels*. As shown in Fig 1, we propose a new metric, **EX**ponenti**A**l **M**arginal s**IN**ular valu**E** (EXAMINE) score to evaluate the value (or referred as quality) of individual data by first using SSL to extract the features, and then calculate the value of the data using Singular Value Decomposition (SVD). EXAMINE scores are useful in indicating the essential data to be annotated, which can not only abundantly reduce the effort in manual labeling but also mitigate the negative effect of mislabeled data, and further improve the target model. Our chief contributions are summarized as follows:
- We are the first to show the feasibility of using an unsupervised learning framework to assess medical data by utilizing SSL and SVD, which is a more cost-efficient and practical method to evaluate data compared to previous work.
- EXAMINE can assess data *without* knowing the label, which reduces annotation efforts and the chance of mislabeling.
- We conduct experiments on the simulated medical dataset to demonstrate the feasibility of using EXAMINE scores to distinguish data with different qualities and show comparable performance to previous supervised learning based works.

2 Preliminaries

2.1 Supervised-learning-based Data Assessment

The goal of data assessment is using a valuation function to map an input data to a single value that indicate its quality. Supervised-learning-based data assessments assume knowing a labeled dataset $\mathcal{N}^l = \{(x_i, y_i)|i \in [N], x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ where N is the number of the labeled data, an utility model $f : \mathcal{N}^l \mapsto \mathcal{Y}$, a held-out labeled testing set $\mathcal{N}^t = \{(x'_i, y'_i)|i \in [M], x'_i \in \mathcal{X}, y'_i \in \mathcal{Y}\}$ where M is the number of the testing data, and a value function $V : (f, \mathcal{N}^l, \mathcal{N}^t) \mapsto \mathbb{R}$

(e.g., the accuracy of $\mathcal{N}^t$ evaluated by f that is trained on $\mathcal{N}^l$). The simplest assessment metric is by performing leave-one-out (LOO) on the training set and calculating the performance differences on the testing set. The i-th data samples value is defined as:

$$\phi_i^{\mathrm{LOO}} = V_f(\mathcal{N}^l) - V_f(\mathcal{N}^l \backslash \{i\}). \tag{1}$$

A more advanced but computational costly approach is Data Shapley [6]. Shapley value for data valuation resembles a game where training data points are the players and the payoff is defined by the goodness of fit achieved by a model on the testing data. Given a subset S, let $f_S(\cdot)$ be a model trained on $\mathcal{S}$. Then Shapley value of a data point $(x_i, y_i) \in \mathcal{N}$ is defined as:

$$\phi_i^{\mathrm{SHAP}} = \sum_{S \subseteq \mathcal{N} \backslash \{x_i\}} \frac{V_f\left(S \cup \{x_i\}\right) - V_f\left(S\right)}{\binom{\left|\mathcal{N}^l\right| - 1}{|S|}}, \tag{2}$$

where $V_f(S)$ is the performance of the utility model f trained on subset S of the data. Suppose each training of f takes time T, the computational complexity of Eq (1) and Eq (2) is $\mathcal{O}(TN)$ and $\mathcal{O}(T2^N)$[3], respectively. Also, training a deep utility function (*e.g.,* neural networks) leads to a large T.

2.2 Formulation of Unsupervised-learning-based Data Assessment

Motivation story Labeling is costly and time consuming in many medical imaging tasks. AI developers may want to pay for labeling some data points to train a particular machine learning model. In such a scenario, supervised-learning-based methods (Sec 2.1) cannot fulfill the aim. Therefore, algorithms that can automatically identify low quality data before labeling data are highly desired.

To address computational issue and demand for task/label-agnostic data quality, we propose to conduct quantitative data quality assessment via unsupervised learning. Different from the formulations of LOO (Eq. (1)) and Shapley value (Eq. (2)), here we propose a new problem formulation to infer i-th data's quality by assigning it a value $\phi_i : (\mathcal{X}, i) \mapsto \mathbb{R}$ using unlabeled data only.

3 Our Method

3.1 Theoretical Implication

Our proposed EXAMINE is well motivated by representation theory of SSL. We start by restating Theorem 1 proved in [13] that under proper assumptions, the embedded space obtained by the reconstruction-based SSL strategy forms a linearly separable space of the embedded feature and a related task. Then, *Remark* 1 presents how we use Theorem 1 to guide the design of EXAMINE.

[3] In practice, there are approximation methods for calculating Shapley value, but the it still requires around $\mathcal{O}(T\mathrm{poly(N)})$ [11].

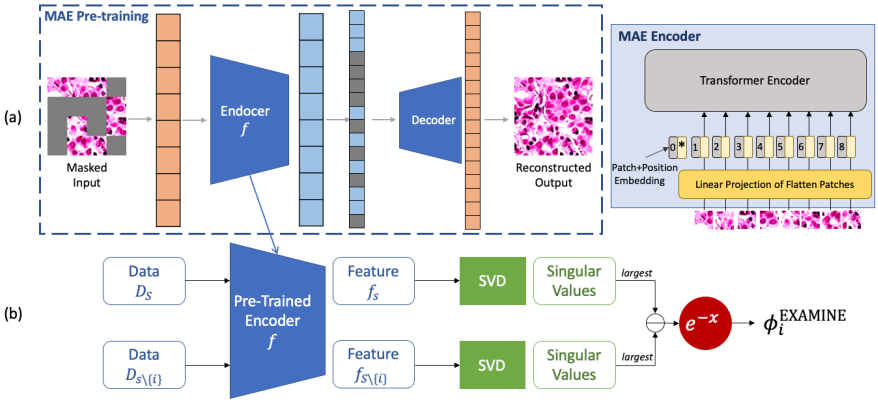


Fig. 1: Proposed pipeline for EXAMINE data assessment. (a) Using the state-of-the art reconstruction-based SSL strategy, MAE [8] architecture for pre-training an representation extractor (encoder). (b) EXAMINE first utilizes the pre-trained encoder to extract semantic features f_S and $f_{S\setminus\{i\}}$ from input data D_S and $D_{S\setminus\{i\}}$, where $D_{S\setminus\{i\}}$ denotes input data D_S *without* data point i. The features then pass the SVD module to find the largest singular values λ_S and $\lambda_{S\setminus\{i\}}$. The EXAMINE score of data point i is defined as Eq. (3).

Theorem 1 (informal [13]). *For two views of a data $X_1, X_2 \in \mathcal{X}$ and their classification label $Y \in \mathbb{R}^k$. Under the class conditional independence assumption,* i.e., *$X_1 \perp X_2 | Y$, for some $w \in \mathbb{R}^{m\times k}$ the representation $\psi^* : \mathcal{X} \mapsto \mathbb{R}^m$ that minimizes a reconstruction loss $\mathcal{L}(\psi) = \mathbb{E}_{(X_1,X_2)}\left[\|X_1 - \psi(X_2)\|^2\right]$ satisfies*

$$w^\top \psi^*(X_1) = \mathbb{E}[Y|X_1].$$

Remark 1. Theorem 1 indicates two desired properties with 'good' data that (approximately) satisfy class-conditional independence. First, the data will have a good geometric property in the learned representation space, namely they become clusters that are (almost) linearly separable (by w). Second, the learned representation has *variance (top singular value of its covariance matrix)* controlled by that of $\mathbb{E}[Y|X_1]$, which is very **small** (since the label is almost determined entirely by the image itself). Such properties on the reconstruction-based SSL embedding space are not satisfied for 'bad' data. Therefore when observing data X_1 is noisy or with low quality, $\psi^*(X_1)$ tends to have higher variance.

3.2 Data Assessment on Singular Value

As shown in the Fig.1(b), to assess data from a dataset $D_S \in R^{N\times C}$, where N is the number of data and C is the dimension of data, we denote the dataset without i-th data point as $D_{S\setminus\{i\}} \in R^{(N-1)\times C}$. To begin with, we employ SSL and the unlabeled data to train an encoder that is able to extract the low-dimensional semantic information. We denote the representation of the SSL embedding space

of D_S and $D_{S\setminus\{i\}}$ as f_S and $f_{S\setminus\{i\}}$, respectively. Lastly, we perform SVD on both feature representations, and use the largest singular values (λ_S and $\lambda_{S\setminus\{i\}}$) as the assessment indicator, that is, removing a 'good' data point i results in small change in the top singular value of embedded data representation f (explained in Sec. 3.1). Thus, the EXAMINE score is defined as

$$\phi_i^{\text{EXAMINE}} = \exp\left(-(\lambda_S - \lambda_{S\setminus\{i\}})\right), \tag{3}$$

where $\phi_i^{\text{EXAMINE}} \in (0, 1)$ and a larger ϕ_i^{EXAMINE} indicates better data quality[4]. Note that EXAMINE is also a leave-one-out strategy but evalated on the change of the largest singular value.

We claim two advantages of using SVD for data assessment. First, by performing the SVD-based evaluation, we do not need any knowledge about the corresponding labels. This is not only the primary difference from previous methods, but also a perfect fit to our problem set up - finding good data to be labeled. Second, unlike previous methods (*i.e.,* LOO and Data Shapely, see Sec. 2.1) that rely on extensively training a new model for different data combinations, our proposed method is efficient by performing SVD once for each data point without additional model training after the SSL encoder has been trained offline.

3.3 Forming Embedding Space using Masked Auto-encoding

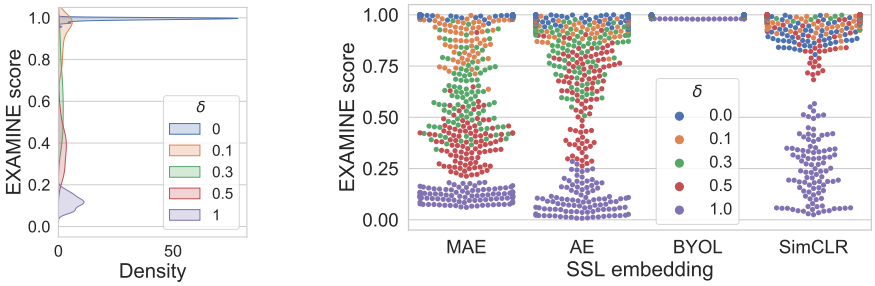
As the raw medical images are high-dimensional and have spurious features (*e.g.,* density, light, dose) that are irrelevant to their labels, directly applying SVD to them cannot capture task-related variance. Based on our theory developed on reconstruction-based SSL (Sec. 3.1), we utilize a state-of-the-art reconstructed-based strategy, Masked Auto-Encoder (MAE) [8] to learn lower-dimensional semantic feature embedding. As shown in Fig. 1(a), MAE utilizes state-of-the-art image classification framework, Vision Transformer (ViT) [3], as the encoder for semantic feature extraction, and uses a lighter version of ViT as decoder. It first divides an input image into patches, randomly blocks a certain percentage of image patches, and then feeds them into the autoencoder architecture. By blocking out a large amount of image patches, the model is forced to learn a more complete representation. With the aim of positional embedding and transformer architecture, MAE is able to generalize the relationship between each image patch and obtain the semantic information among the whole image, which achieves the state-of-the-art performance in self-supervised image representation training. *This also reduces the correlation between spurious features and labels*, compared to the traditional dimension reduction methods [2].

4 Experiment

4.1 Experiment setup and dataset

We evaluate EXAMINE on a binary classification task for PCam [18], a microscopic dataset (image size 96×96) for identifying metastatic tissue in histopatho-

[4] $\lambda_S > \lambda_{S\setminus\{i\}}$ is for sure given the properties of singular value.



(a) EXAMINE with MAE (b) Comparison with different embedding

Fig. 2: Proof of concept and comparison with baseline embedding methods. Observe that EXAMINE scores (ϕ_i^{EXAMINE}) get lower when noise level increases (a), and the EXAMINE scores (ϕ_i^{EXAMINE}) in (b) shows that reconstrucion-based SSL methods perform better in separating different noise levels.

logic scans of lymph node sections. Since noise is usually the main corruption in medical images, we add non-zero mean Gaussian noise to a portion of the data to simulate real world scenario.

We split the dataset into four disjoint sets following the scale of [6]:
- **SSL Pre-Training Set** and **Assessed Set**: 160,000 and 500 unlabeled data points randomly sampled from PCam, respectively. We add 4 different level ($\mathcal{N}(\delta, \delta \times m)$, where m is the mean of the dataset and $\delta = \{0.1, 0.3, 0.5, 1\}$) of noise to 60000 data points of **SSL Pre-Training Set** and 400 data points in **Assessed Set**.

- **Clean Train Set** and **Validation Set**: 100 and 20,000 labeled data points randomly sampled from PCam. They are used to validate the data selection in an example downstream task after obtaining EXAMINE scores (ϕ_i^{EXAMINE}).

The experiments are run on NVIDIA GeForce RTX 3090 Graphics card with PyTorch. For MAE training, we select Cosine Annealing LR scheduler [14] and AdamW LR optimizer [15] with weight decay 0.05 and momentum $\{0.9, 0.95\}$. We train MAE for 200 epochs using batch size 256 ,and set image size 72, patch size 8, masking ratio 40%. As indicated in [1, 7, 8] that SSL training requires a large amount of data, we begin with training a MAE on **SSL Pre-Training Set**. To ensure the training stability, we first train MAE without any noisy data, and finetune it afterward. This step is to distinguish our setting from detection out of distribution samples. After pretraining the MAE encoder, we use the frozen encoder layers as our backbone to extract the low dimensional representations. All of our experiments are repeated five times with different random seeds and we report the mean value of the five trials.

4.2 Proof of Concept with 'Ground-truth'

To validate the correctness of ranking the samples in **Assessed Set**, we plot the distributions of EXAMINE scores (ϕ_i^{EXAMINE}) at different data corruption levels of *noisy* setting in Fig 2a Note that $\phi_i^{\text{EXAMINE}} \in (0, 1)$ and data point i with larger ϕ_i^{EXAMINE} indicates that it is considered a good data. Specifically, EXAMINE score (ϕ_i^{EXAMINE}) approaching 1 indicates the difference between the top singular values are small, thus it will be considered good data. The EXAMINE scores (ϕ_i^{EXAMINE}) for the high-quality data ($\delta = 0$) are close to 1 and significantly higher than the corrupted data. The EXAMINE scores (ϕ_i^{EXAMINE}) of data with low-level corruption ($\delta = 0.5$) are also separable from those with high-level corruption($\delta = 1$).

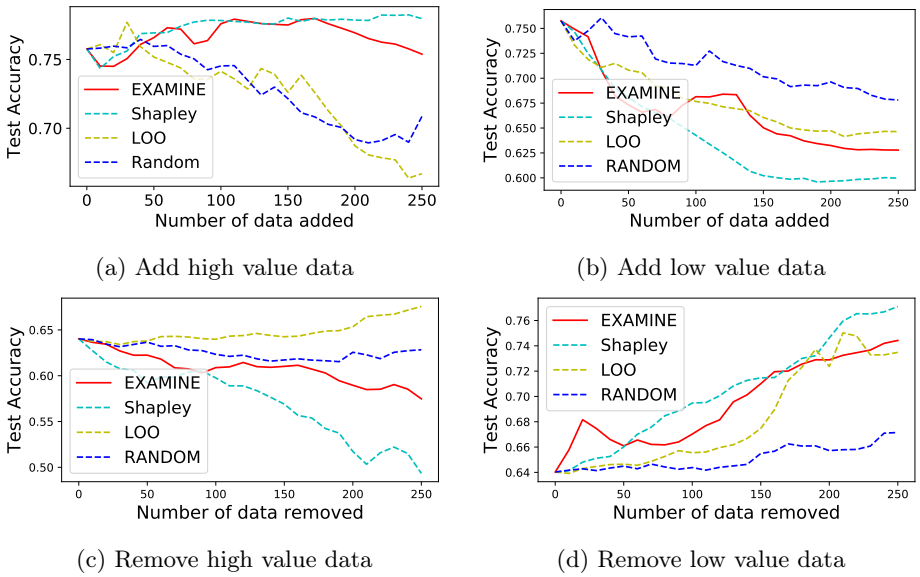
4.3 Comparison with Alternative Embedding Methods

We investigate the alternative feature encoders and compare their performance with MAE. Specifically, we replace MAE with SimCLR [1] and BYOL [7], two alternative SSL algorithms, and Autoencoder (AE) [9], a naïve reconstruction-based embedding strategy. SimCLR learns embedding by enforcing the closeness of an image and its augmented views while enlarging the distance from other images in the dataset (or batch). BYOL regularizes the multi-views of an image without sampling negative samples by training two similar networks (the online network and the target network) simultaneously. We use the same strategy as training MAE for these alternative encoders. Fig. 2b shows that using MAE embedding to calculate EXAMINE scores (ϕ_i^{EXAMINE}) provides the best separability for the clean data from the corrupted data. The reason is that MAE best satisfies the theoretical conditions that support our proposal (Theorem 1). AE is second to MAE, but separation boundaries are less clear.

4.4 Comparison with Baseline Data Valuation Methods

We compare EXAMINE with supervised data valuation methods, LOO (Eq. 1 and Truncated Monte Carlo (TMC) version of Data Shapley (Eq. 2) [6], as well as a baseline method that randomly assigns data values. All these methods are applied on **Assessed Set**'s features extracted by the pre-trained MAE. We design four experiments to evaluate how selecting data using the different data assessment methods can affect the classification accuracy. We report the averaged test accuracy on **Validation Set** using logistic regression models (LRM).

Fig. 3a and Fig. 3b show the results of adding data for training. We start with a LRM trained on the small **Clean Training Set**, and then add good/bad data from **Assessed Set** following the descending/ascending orders of their data values. Our results show that adding data with high EXAMINE score (ϕ_i^{EXAMINE}) achieves comparable accuracy curve as Data Shapley, while adding data with low EXAMINE scores (ϕ_i^{EXAMINE}) results in similar curve as Data Shapley in the beginning and overall lies in between Data Shapley and LOO. This indicates that EXAMINE score (ϕ_i^{EXAMINE}) is able to identify what data to be labeled and



(a) Add high value data

(b) Add low value data

(c) Remove high value data

(d) Remove low value data

Fig. 3: Comparison with baseline data valuation methods. Adding good data(a) and removing bad data(d) should increase accuracy. Adding bad data(b) and removing good data(c) should result in accuracy drop. We conclude that EXAMINE can help model training by identifying good and bad data.

added to training set. Fig. 3c and Fig. 3d show the results of removing data for training. We first train LRM on **Assessed Set**, and then remove good/bad data following the descending/ascending orders of their data values. Our result shows that removing high/low EXAMINE score (ϕ_i^{EXAMINE}) data results in accuracy curve that is slightly worse than Data Shapley. We would like to emphasize Data Shapley uses utility function which requires labels to determine the data value, while EXAMINE score (ϕ_i^{EXAMINE}) is calculated only on data itself, which is more efficient in real-world scenario. Overall, EXAMINE shows comparable(or at best slightly worser) data assessment performance to Data Shapley *without* knowing the labels of the data.

In addition to successfully providing correct inspection on data quality, our method significantly reduces the computational cost without requiring training utility functions. The running time to obtain the data values using EXAMINE, LOO, and TMC Data Shapley for the whole **Assessed Set** under our experiment setting are 23 seconds, 10 minutes, and 460 minutes, respectively[5].

[5] The running time for LOO and Data Shapley can significantly increase if we use a deep neural network as the utility model.

5 Discussion and Conclusion

We present a new and efficient unsupervised data evaluation method, EXAMINE scores ϕ_i^{EXAMINE}, to assess data quality. With the help of MAE encoder, we can map data to the provable low-dimensional embedding space. The marginal differences on the largest singular value of data representation matrices can effectively separate data at different quality levels and achieve comparable performance with supervised data valuation methods when considering a specific task. This work takes a novel approach to promote AI in healthcare by identifying low quality data. We plan to test on larger scale medical datasets and collect domain experts' evaluations in the future.

Acknowledgement

This work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and NVIDIA Hardware Award.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
2. Chen, Y., Wei, C., Kumar, A., Ma, T.: Self-training avoids using spurious features under domain shift. Advances in Neural Information Processing Systems **33**, 21061–21071 (2020)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Fadahunsi, K.P., Akinlua, J.T., O'Connor, S., Wark, P.A., Gallagher, J., Carroll, C., Majeed, A., O'Donoghue, J.: Protocol for a systematic review and qualitative synthesis of information quality frameworks in ehealth. BMJ open **9**(3), e024722 (2019)
5. Fadahunsi, K.P., O'Connor, S., Akinlua, J.T., Wark, P.A., Gallagher, J., Carroll, C., Car, J., Majeed, A., O'Donoghue, J.: Information quality frameworks for digital health technologies: systematic review. Journal of medical Internet research **23**(5), e23479 (2021)
6. Ghorbani, A., Zou, J.: Data shapley: Equitable valuation of data for machine learning. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2242–2251. PMLR (09–15 Jun 2019)
7. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)

9. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. science **313**(5786), 504–507 (2006)
10. Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Gürel, N.M., Li, B., Zhang, C., Song, D., Spanos, C.J.: Towards efficient data valuation based on the shapley value. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 1167–1176. PMLR (2019)
11. Jia, R., Sun, X., Xu, J., Zhang, C., Li, B., Song, D.: An empirical and comparative analysis of data valuation with scalable algorithms (2019)
12. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. pp. 4171–4186 (2019)
13. Lee, J.D., Lei, Q., Saunshi, N., Zhuo, J.: Predicting what you already know helps: Provable self-supervised learning. Advances in Neural Information Processing Systems **34** (2021)
14. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Redman, T.C.: Data driven: profiting from your most important business asset. Harvard Business Press (2008)
17. Tosh, C., Krishnamurthy, A., Hsu, D.: Contrastive learning, multi-view redundancy, and linear models. In: Algorithmic Learning Theory. pp. 1179–1206. PMLR (2021)
18. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: International Conference on Medical image computing and computer-assisted intervention. pp. 210–218. Springer (2018)
19. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016)