Latent representations learned by autoencoders are well transferred to pattern recognition tasks, aren't they?

Mikhail Goncharov [1] **Evgenia Soboleva** [1]

1. Introduction

Machine learning can be supervised or unsupervised. *Supervised learning* aims to build a model for solving a particular task: given an input x, predict the target y. From the probabilistic prospective, (x, y) comes from some unknown distribution $p(x, y)$. To make predictions, supervised models learn $p(y \mid x)$ from a finite training set $\{(x_1, y_1), \ldots, (x_n, y_n)\} \sim p(x, y)$. Depending on the type of the target, there are two types of tasks:

- *classification* (also called *pattern recognition*), if the target is a class from a set of classes: $y \in \{1, \ldots, K\}$;
- *regression*, if the target is a real number: $y \in \mathbb{R}$.

Unsupervised learning aims to build a model of input data itself. Probabilistically speaking, given a set of data $\{x_1, ..., x_n\} \sim p(x)$, unsupervised models learn the unknown data distribution $p(x)$. Unsupervised learning involves many tasks:

- *generative modeling* aims to sample new data from the distribution $p(x)$;
- *density estimation* aims to estimate density $p(x)$ for a given x;
- *dimensionality reduction* aims to find a low-dimensional representations of data points which preserve the most part of data variability;
- *metric learning* aims to learn a metric between data points, which can in turn be useful for such tasks as *clustering*, *retrieval* or *anomaly detection*.

Another possible goal of unsupervised learning is to learn a feature extractor, that eventually can be used for *transfer learning*, i.e. can be incorporated into other architectures which are then fine-tuned in a supervised fashion for solving different downstream tasks. This type of unsupervised learning is traditionally refered to as *feature learning* or *representation learning*. Recently, it also has become known as *self-supervised learning*. The problem of self-supervised representation learning is the key interest of this work.

Self-supervised representation learning is important because it suggests a way to overcome the limitations of the supervised learning. Training of large supervised models from scratch requires enormous amount of human-labeled data and energy costs. Also supervised models suffer from overfitting and domain shift problems. Fine-tuning of small models on top of large frozen pre-trained feature extractors trained in self-supervised fashion potentially allows to overcome all these challenges.

1.1. Related work

Autoencoder (AE) (Baldi, 2012) is a classical model for unsupervised learning of low-dimensional representations of complex high-dimensional data. Today, many practitioners still consider it as a default model for representation learning.

Variational autoencoder (VAE) (Kingma & Welling, 2013) is a latent variable probabilistic model which is primarily intended for generative modeling. However it also can be used for representation learning. Compared to its non-probabilistic counterpart – standard AE, the VAE learns more regular and structured, i.e. disentangled (Higgins et al., 2022), latent space, which may be beneficial for transfer learning.

However, current state-of-the-art representation learning methods in domains of computer vision (CV), natural language processing (NLP), times series modeling, etc. are different from AE or VAE. They can be split in two groups:

- *contrastive methods*, that directly penalize the learned representations to be closer for pairs of ”close” objects and futher for pairs of random objects.
- *masked autoencoders* (MAE), which are trained to reconstruct missing parts of input data, that are randomly masked during training (He et al., 2021).

[*] Equal contribution [1] Skoltech. Correspondence to: Mikhail Goncharov <Mikhail.Goncharov2@skoltech.ru>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

The ideas underlying these methods have historically originated in the field of NLP. The idea of contrastive methods comes from `word2vec` (Mikolov et al., 2013a;b) model. MAE is inspired by BERT (Devlin et al., 2018).

1.2. Contribution

In this work we study the following question: are contrastive methods really better representation learners than standard autoencoders? We focus on CV domain and provide comprehensive experiments comparing representations learned by 1) standard AE, 2) VAE and 3) SimCLR (Chen et al., 2020) – most popular contrastive method in CV. We evaluate how well the representations are transferred to classification task on MNIST and CIFAR10 datasets. We show that all methods work on par with each other on MNIST dataset. However, when it comes to natural images, contrastive representations are better aligned with pattern recognition tasks than representations learned by AE or VAE.

2. Experiments

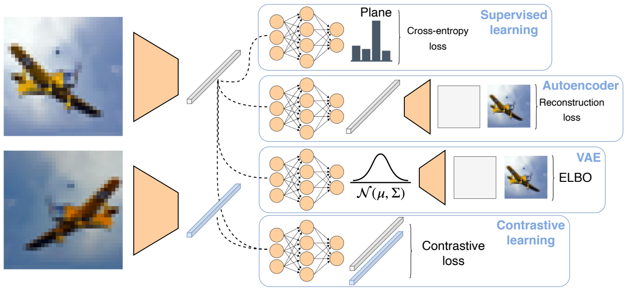

Figure 1. Schematic illustration of our framework for evaulation of different representation learning methods.

We propose a simple framework to evaluate AE, VAE and SimCLR representations. We build the following four models on top of the standard ResNet18 encoder (see. Figure 1):

- We attach a two-layer MLP which classifies the representations returned by the encoder. It is trained in a supervised fashion using standard cross-entropy loss. We refere to it as Supervised model.
- We build the AE model, by attaching a two-layer MLP, which can optionally reduce the dimension of representations, followed by a decoder network. This model is trained using a standard reconstruction mean-squared error loss.
- We build the VAE model in the same way as AE model, with only difference that two-layer MLP predicts the parameters of the posterior distribution in the latent space. This model is trained to maximize ELBO (Kingma & Welling, 2013).
- We also attach a two-layer MLP projector which maps the representations to embeddings on which the SimCLR loss (Chen et al., 2020) is applied.

To evaluate representations learned in a self-supervised fashion we train the encoder as a part of AE, VAE or SimCLR model, and simultaneously train a supervised two-layer MLP classifier on top of the encoder. Note that we apply `stopgrad` operator to the representations before fedding them into the classifier. Therefore, encoder is trained in a fully unsupervised fashion.

3. Results

Classification accuracies of the two-layer MLP classifier trained on frozen representations learned by AE, VAE and SimCLR models, as well as accuracy of the Supervised model are given in Table 1. We also visualize the representations learned by all the models using t-SNE (Van Der Maaten & Hinton, 2008) on MNIST and CIFAR datasets in Figures 2, 3.

Table 1. Accuracy of the two-layer MLP classifier trained on frozen representations learned by AE, VAE and SimCLR models. Classification accuracy of Supervised model is also reported as a strong baseline.

Dataset	AE	VAE	SimCLR	Supervised
MNIST	98.3	97.8	98.5	99.5
CIFAR10	55.6	54.5	**81.9**	**83.2**

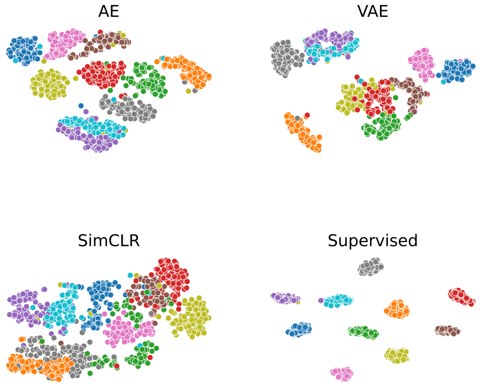

Figure 2. t-SNE projection of representations learned by 1) AE, 2) VAE and 3) SimCLR and 4) Supervised models on MNIST dataset. Colors show the MNIST ground truth classes.

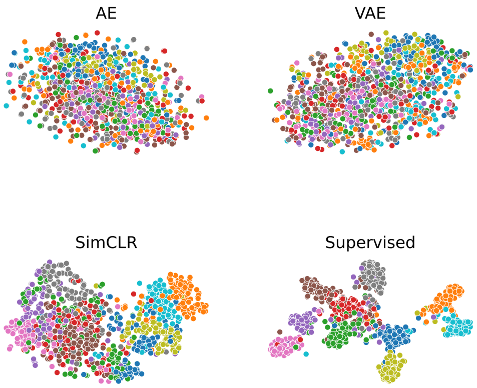

Figure 3. t-SNE projection of representations learned by 1) AE, 2) VAE and 3) SimCLR and 4) Supervised models on CIFAR10 dataset. Colors show the CIFAR10 ground truth classes. Note that, SimCLR maps objects of the same class to the same cluster of representations, while semantics of the latent space learned by AE and VAE is much worse aligned with the semantics of the CIFAR10 classes.

4. Conclusion

From our results we conlcude that

- Representations learned by autoencoders can be easily transferred to pattern recognition tasks only on very simple data, like MNIST.
- Modern contrastive learning methods, e.g. SimCLR, learn representations which contain information about semantics of the images in a more explicit form, i.e. can be classified by a simple models, like two-layer MLP.

We believe that the results of this work motivate readers to contribute to the development of new self-supervised representation learning methods.

References

Baldi, P. Autoencoders, Unsupervised Learning, and Deep Architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pp. 37–49. JMLR Workshop and Conference Proceedings, June 2012. URL https://proceedings.mlr.press/v27/baldi12a.html.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv*, February 2020. doi: 10.48550/arXiv.2002.05709.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, October 2018. doi: 10.48550/arXiv.1810.04805.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked Autoencoders Are Scalable Vision Learners. *arXiv*, November 2021. doi: 10.48550/arXiv.2111.06377.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *OpenReview*, July 2022. URL https://openreview.net/forum?id=Sy2fzU9gl.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *arXiv*, December 2013. doi: 10.48550/arXiv.1312.6114.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv*, January 2013a. doi: 10.48550/arXiv.1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *arXiv*, October 2013b. doi: 10.48550/arXiv.1310.4546.

Van Der Maaten, L. and Hinton, G. Viualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2605):2579–2605, November 2008. ISSN 1533-7928. URL https://www.researchgate.net/publication/228339739_Viualizing_data_using_t-SNE.