Image and Vision Computing 137 (2023) 104746

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

VAESim: A probabilistic approach for self-supervised prototype discovery

Matteo Ferrante [a,*], Tommaso Boccato [a], Simeon Spasov [c], Andrea Duggento [a], Nicola Toschi [a,b]

[a] Department of Biomedicine and Prevention, University of Rome Tor Vergata, Italy
[b] Martinos Center For Biomedical Imaging, MGH and Harvard Medical School, USA
[c] Department of Computer Science and Technology, University of Cambridge, UK

ARTICLE INFO

Keywords:
Deep clustering
Medical imaging
Variational autoencoders
Prototypes discovery

ABSTRACT

In medical image datasets, discrete labels are often used to describe a continuous spectrum of conditions, making unsupervised image stratification a challenging task. In this work, we propose VAESim, an architecture for image stratification based on a conditional variational autoencoder. VAESim learns a set of prototypical vectors during training, each associated with a cluster in a continuous latent space. We perform a soft assignment of each data sample to the clusters and reconstruct the sample based on a similarity measure between the sample embedding and the prototypical vectors. To update the prototypical embeddings, we use an exponential moving average of the most similar representations between actual prototypes and samples in the batch size. We test our approach on the MNIST handwritten digit dataset and the PneumoniaMNIST medical benchmark dataset, where we show that our method outperforms baselines in terms of kNN accuracy (up to +15% improvement in performance) and performs at par with classification models trained in a fully supervised way. Our model also outperforms current end-to-end models for unsupervised stratification.

1. Introduction

1.1. Unsupervised learning in medicine

Large unlabeled image datasets in medicine greatly outnumber curated ones. When labels are available, they often describe a continuous spectrum of conditions using discrete labels. There is growing interest in developing unsupervised methods to uncover the hidden or latent structure of these disorders, aiming to identify disease subtypes or novel disease classes (see [1] for a review). However, unsupervised learning remains a challenging task, even in domains with large and curated datasets. Numerous approaches have been proposed, but they typically require vast amounts of data.

1.2. Deep learning and clustering

To make these approaches more practical, we could develop data-efficient unsupervised methods or increase reliance on self-supervised or unsupervised pretraining, followed by supervised fine-tuning, which would substantially reduce the need for labeled samples. In this context, several papers have addressed the problem of clustering using deep learning architectures that find a nonlinear mapping from the input space to the feature space where clustering is performed (see"Related Work" section).

1.3. Proposed approach

In this work, we propose an unsupervised deep learning probabilistic approach that projects inputs to a learned latent space with"prototype" points, enabling low-dimensional representations of the inputs to be useful for both reconstruction and subsequent tasks such as classification. Each"prototype" point is associated with a cluster in the latent space. During reconstruction, we condition the decoder on both a) the sample latent embedding and b) a soft cluster assignment based on the similarity of the sample embedding to all prototype vectors. We assume that downstream tasks, such as classification, are simplified by this conditioning, which provides more context on the relative position of the sample embedding in the latent space. The main challenge is defining a strategy to learn these prototype vectors, which should represent subsets that cluster together in the latent space.

1.4. Objective and contributions

The objective of this work is to augment the Variational Autoencoder

* Corresponding author.
E-mail address: matteo.ferrante@uniroma2.it (M. Ferrante).

https://doi.org/10.1016/j.imavis.2023.104746
Received 3 April 2023; Received in revised form 1 June 2023; Accepted 19 June 2023
Available online 4 July 2023

(VAE) framework to also learn prototype vectors, thus obtaining a better organization of the latent space. The prototypes are learned using a momentum update during training and are stored in the model's memory. This conditional information about similarity during decoding should carry all the information needed for reconstructions and implicitly impose a structure in the latent space. We demonstrate that this approach is more effective than two-step approaches that first train a vanilla VAE and then cluster its latent space, for example, using KMeans. Our end-to-end approach outperforms these and other deep clustering baselines on the MNIST handwritten digit dataset [9] as a benchmark, and on the Pneumonia dataset from the MedMNIST collection [21] as an example in the field of biomedical images.

1.5. Performance evaluation

We quantify the performance differences using various metrics, related to both the alignment between the label clusters and the classification obtained using kNN or linear classification, to measure the capacity of the latent space to capture the fundamental characteristics of the encoded samples.

2. Related work

2.1. Large-scale unsupervised and self-supervised approaches

General trends in unsupervised and self-supervised deep learning are dominated by large networks trained on hundreds of thousands of images. Recent examples include student-teacher approaches such as DINO [6], based on knowledge distillation without labels, and contrastive-based approaches such as [7,8,13], where models are trained through a contrastive loss and a similarity metric to position similar samples close in the latent space and different ones far apart. These approaches heavily rely on different augmentations to generate positive pairs. However, due to the intrinsic properties of medical image datasets, these methods do not translate well to the medical context.

2.2. Challenges in medical image domain

Knowledge distillation approaches like DINO depend on data-hungry architectures like Vision Transformers [12] or ResNet50 [14], whose power can only be harnessed when"big data" is available. Contrastive-based approaches can address label scarcity by utilizing augmentations. However, in medical applications, geometric augmentations like mirrors, zooms, and color distribution shifts fail to capture the full diversity of pathologies. Moreover, methods like SimCLR [7] demand large batch sizes for negative sampling, leading to high computational requirements that can only be met with nonstandard hardware. This issue can be partly mitigated using more efficient techniques such as MOCO [13], where a momentum encoder is employed for negative sampling and stores low-dimensional representations of previous batches. Although promising, these approaches face challenges when applied directly to the medical domain.

2.3. VAE-based approaches

In situations with limited data, smaller models relying on self-supervised pretraining with autoencoders or variational autoencoders may be more effective. In [18], the authors trained a variational autoencoder using a Gaussian mixture model (GMM) as an unsupervised clustering algorithm. By incorporating this information into the loss function, they facilitated the model's learning of a well-separated latent space. In [15], the authors replaced the VAE prior with a GMM-based prior, enforcing a similar separation more directly. In [19], a differentiable version of the K-Means algorithm based on the softmax function is proposed and integrated into a neural network.

2.4. Medical domain adaptations

Other unsupervised approaches have adapted these ideas for the medical field. Examples include the Mixture of Experts (MoE) [16], where a clustering network proposes a cluster and a mixture of experts reconstructs single-cell data, and [2], which combines a CNN with a KMeans approach for classifying image datasets with different modalities. Furthermore, contrastive learning translation incorporating multi-instance learning has achieved results comparable to fully supervised state-of-the-art algorithms [3]. However, these strategies may be sub-optimal in certain clinical or medical domains due to their reliance on nondifferentiable methods, the need for large data volumes, or the requirement of domain-specific knowledge for incorporating biologically or medically relevant data augmentations.

2.5. Our approach

Our work combines tempered softmax similarity-based soft labels to inform the decoder about clustering, while using nondifferentiable functions to update prototype values. This approach aims to address some of the challenges faced by previous methods in the medical domain and enhance the performance of unsupervised learning for medical image analysis.

3. Materials and methods

The primary objective of VAESim is to enhance the vanilla VAE by conditioning the decoder on not only the latent embedding of a sample but also its soft cluster assignment in the latent space. This entails two main challenges: 1) learning a descriptive latent space that contains enough information to reconstruct the input samples, and 2) learning the prototypical vectors of the latent space. The VAESim model consists of three components: the encoder e, the decoder d, and the prototype matrix $Q^{K\times q}$.

First, we encode the p-dimensional vector input x with a neural network $e : \mathbb{R}^p \to \mathbb{R}^q \times \mathbb{R}^q$ into a q-dimensional latent space z. Then, we sample a latent representation from the estimated posterior, modeled as a Gaussian, $z \sim \mathcal{N}\left(z_\mu, e^{\frac{z_{log\sigma}}{2}}\right)$.

Next, we compute the similarity between the latent embedding and the prototypical vectors using cosine similarity. We then perform soft cluster assignment by computing the tempered softmax over the similarity measures.

$$c = softmax\left(\frac{Qz}{\tau}\right) \tag{1}$$

We employ τ as the temperature parameter and c as the cluster assignment. The decoder d receives the latent representation z along with the cluster assignment c, and reconstructs the original input sample by mapping $\tilde{x} = d(z, c)$ back into the input space. To enhance the vanilla VAE's standard ELBO loss, we introduce an optional term promoting orthogonality in the prototype representations.

The updating process for the prototype vectors follows Algorithm 1, with the differentiating aspects of our method being the *cluster* equation and the *update* function. Given the latent embeddings z and the prototypes $Q^{K\times q}$, the *cluster* equation (Eq. 1) generates soft assignments. The *update* function, presented in Algorithm 2, constitutes a fundamental component of our approach. During each training iteration, we assign every sample in the training dataset to a cluster. To update each prototypical vector, we employ an exponential moving average, which is computed using the mean of all sample embeddings belonging to the corresponding cluster.

Algorithm 1. Training VAESim model.

input A set of N training images $X^{N\times p}$.
τ : 1. linearly decreases to 0.01 in half of the epochs.
Initialize prototypes matrix Q.
for x in the training set **do**
optimizer.zero_grad()
$z = e(x)$
$c = softmax(\frac{Qz}{\tau})$
$\tilde{x} = d(x, c)$
loss=$\mathcal{L}_{recon}(x, \tilde{x}) + \beta KL_{div}(q_z, p_z)$
loss.backward()
optim.step()
Update(Q,z,c)

Algorithm 2. Update function.

input Q,z,c
η : 0.95
for z_i,c_i in z,c **do**
Sample cluster label $k_i \in [0, K-1]$: $k_i \sim multinomial(c_i)$.
for $i = 0 \dots K$ **do**
Update the i-th column of Q matrix using vectors assigned to it
$\bar{z_i}$=mean(z_i) **if** $k_i == i$
$Q_i = \eta\bar{z_i} + (1-\eta)Q_i$

3.1. Variational autoencoder (VAE)

Variational autoencoders (VAEs) offer a key advantage over traditional autoencoders by imposing a well-structured, continuous latent space, which allows for the generation of meaningful samples from the latent space. Instead of merely projecting data points onto a manifold and reconstructing them, the encoder neural network in VAEs attempts to approximate the computationally intractable joint probability $p(\boldsymbol{z}|\boldsymbol{x})$ through probabilistic encoding $q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x})$. The encoder produces the mean and log-variance of a multidimensional Gaussian distribution, from which a value is sampled and passed to the decoder to map back from $\boldsymbol{z}$ to $\widetilde{\boldsymbol{x}}$.

The model is optimized by simultaneously enforcing reconstruction quality through a mean squared error (MSE) loss and a regularization term that encourages the Gaussian distribution to closely resemble a chosen posterior, typically a standard normal distribution. The latter term usually employs a weighted Kullback–Leibler (KL) divergence. Consequently, the model is optimized to reconstruct items with a latent distribution that closely approximates a standard multivariate Gaussian distribution. The loss function is thus $\mathscr{L} = |\boldsymbol{x} - \widetilde{\boldsymbol{x}}|^2 + \beta KL(q_{\boldsymbol{z}}, \mathcal{N}(0,1))$. In this work, we modify the decoder network to accept both the latent representation and a condition vector representing prototype similarity for each sample.

3.2. Prototype matrix Q

The prototype matrix Q is initialized based on a predetermined number of clusters k and the latent space dimension p. The Q matrix is of shape (k, p), with k rows and p dimensions per row. During the first forward pass, k elements from the initial batch are randomly selected and set as rows of the Q matrix.

In the early stages of training, the prototype values are not particularly meaningful, so a tempered softmax of the similarity is used with a temperature schedule. For the first quarter of training steps, the temperature is set to a high value (> 1), flattening the similarity distribution across different prototypes. As the model's reconstruction capability improves, the temperature decreases, resulting in more distinct conditioning vectors.

The prototype matrix Q update method is a critical aspect of this approach. Instead of using gradient descent, the Q matrix values are computed separately from the computational graph by applying a momentum update for all samples belonging to a specific cluster, i.e., all samples whose latent embeddings $\boldsymbol{z}$ are most similar to a specific column of the Q matrix.

At each iteration, the batch "cluster labels" are sampled from (or obtained by argmaxing) the categorical distribution c. These labels are used to compute the *update* function for the Q matrix.

3.3. Mathematical interpretation

In summary, the fundamental objective function for a typical Variational AutoEncoder (VAE) is the evidence lower bound (ELBO), formally defined as:

$$L_{ELBO}(x, \widetilde{x}) = \mathbb{E}_{z\sim q(z|x)}[log(p(x|z))] - KL(q(z|x)\|p(z)) \tag{2}$$

Here, $q(\boldsymbol{z}|\boldsymbol{x})$ signifies the approximate posterior, $p(\boldsymbol{x}|\boldsymbol{z})$ represents the likelihood, and $p(\boldsymbol{z})$ is the prior. The initial term seeks to minimize the difference between the decoded samples $\widetilde{\boldsymbol{x}}$ and the input samples $\boldsymbol{x}$, while the Kullback–Leibler (KL) divergence term ecourages the approximate posterior to mirror the prior. In our proposed VAESim model, we also utilize the ELBO loss. However, we incorporate an additional loss term to encourage orthogonality within the prototype representations, which acts as a regularizer on the prototype matrix Q, defined as:

$$L_{ortho}(Q) = \left\|QQ^T - I\right\|^2 \tag{3}$$

Here, I is the identity matrix, and $\|\cdot\|$ denotes the norm. The implementation of this loss term encourages each prototype vector to remain orthogonal to each other. Thus, the comprehensive loss function for VAESim can be formalized as:

$$L_{VAESim}(x, \tilde{x}, Q) = L_{ELBO}(x, \tilde{x}) + \lambda L_{ortho}(Q) \quad (4)$$

Here, λ serves as a hyperparameter to balance the ELBO and orthogonality losses. The entire procedure is outlined in Alg 2, with the key novelty lying in the innovative approach to updating the prototypes matrix Q. Here, c signifies all samples assigned to the k-th cluster, z_i is the mean of all sample embeddings in C_i, and η is a parameter controlling the moving average. This algorithm applies an exponential moving average which updates the prototypes, improving the model's robustness by avoiding sudden changes. Our proposed VAESim model enhances the basic VAE by supplying the decoder with both the latent representation z and the cluster assignment c. Consequently, the model learns a more descriptive latent space that is beneficial for input reconstruction and further tasks such as classification. Moreover, by imposing an orthogonality constraint on the prototypes, each prototype is encouraged to represent unique clusters within the latent space, thus potentially augmenting performance in downstream tasks.

3.4. Evaluation

The regularization constraint imposed by the KL divergence in the loss function tends to bring all inputs closer together in the latent space. As a result, using traditional metrics for unsupervised clustering could be misleading, given that they often normalize by intercluster distance. Instead, we generate a visually meaningful 2D representation using t-SNE to examine how inputs are organized in the latent space. Subsequently, we propose three performance evaluation approaches, all based on the downstream classification task. The model training is entirely self-supervised, and some available labels are utilized for comparison during the evaluation phase. We employ three standard methods for learning the downstream task: a statistical mapping approach between cluster labels and actual labels, the kNN (k-nearest neighbors) algorithm, and the training of a linear layer with categorical cross-entropy as the loss function.

In the statistical mapping approach, clusters are determined by maximizing the similarity between samples and prototypes. Each cluster label is then associated with the most frequent label present in the corresponding cluster, serving as a way to quantify how well each prototype captures the essential elements of a class. During inference, this mapping is used to predict the final class.

In the second approach, a subset of the training set is used to compute the"memory bank" for the kNN algorithm. During inference, each test set sample is compared with each element stored in the memory bank by calculating the pairwise Euclidean distance. The predicted label is defined as the mode of the label of the k closest samples.

In the third approach, a linear layer maps from z to the number of possible classes. This linear layer is trained with the Adam optimizer and a learning rate of 3×10^{-4} for 200 epochs over a subset of the training set, using categorical cross-entropy as the loss function.

3.5. Baselines

To compare our probabilistic framework with conventional two-step approaches, we trained a VAE designed to closely resemble the VAE used for image reconstruction. Subsequently, we performed a KMeans evaluation to generate cluster labels and train a kNN and linear classifier on the latent space, similar to other experiments. This allowed us to assess the VAE + KMeans method using the same metrics as our approach. We also evaluated two other approaches, namely VaDE [15] and GMVAE [11], using default parameters. Each experiment was executed 10 times to obtain mean accuracies and standard deviations. It is worth noting that the reported baseline accuracies might not match those presented in the original papers due to differences in training splits, random seeds, and particularly, distinct methods for mapping between cluster labels and ground truth. Specifically, both GMVAE and VaDE rely on a clustering accuracy measure defined by mapping N cluster labels to N real labels using the Hungarian matching algorithm [17], which requires the number of clusters to be identical to the number of unique labels. We opted for an alternative mapping approach between cluster and actual labels to allow greater flexibility in cluster numerosity selection, thus enabling the potential discovery of more refined structures while preserving the possibility of many-to-one mapping. As a comparison, we also report state-of-the-art results achievable through transfer learning from ResNet50 pretrained on ImageNet.

3.6. Experiments

We evaluated our model on two datasets: a classic handwritten digit dataset ([10]) and a medical benchmark dataset, PneumoniaMNIST, which is part of the MedMNIST dataset [21,22]. MedMNIST is a collection of medical imaging datasets curated for benchmarking purposes. We selected PneumoniaMNIST as a biomedical example in which chest X-ray images are assigned to one of two possible labels:"healthy" or"pneumonia." This dataset comprises 5856 images (4708 for training and 1148 for testing). Motivated by the advanced data preprocessing methodologies delineated in the study'Segmentation of Crop Images for Crop Yield Prediction' [1], which demonstrated considerable enhancements in predictive accuracy for crop images, we sought to extrapolate analogous techniques to our grayscale medical image dataset. Despite the fact that the aforementioned study was primarily focused on images rich in color variance, we specifically designed an experiment pertinent to our context. This involved integrating random histogram equalization with a series of geometric augmentations, including but not limited to, rotation, scaling, and translation. All code is written in PyTorch [20] and trained on a server equipped with an A6000 GPU and 512 GB RAM. The code for the VAEsim implementation and the baseline algorithms can be found at the following GitHub repository: https://github.com/matteoferrante/VAESIM.git.

4. Results

Our primary hypothesis is that our end-to-end deep clustering approach would yield improved performance in downstream classification tasks in comparison to a two-step approach and other baseline methods. For the VAE + KMeans baseline, we trained a VAE with an architecture closely matching the one used in our approach, specifically with a latent dimension of 32 and a batch size of 2048.

Subsequently, we employed K-Means to determine the optimal number of clusters using the elbow method and the yellow-brick library [4]. VaDE and GMVAE served as baselines for deep clustering. The results indicate that our model consistently outperforms (or performs on par across all metrics) the investigated baselines, with improvements over the two-step approach ranging from +13% up to +17% on the MNIST dataset and from +4% to +17% on the Pneumonia dataset. The most significant improvements are observed in kNN accuracy, suggesting that our modifications to the VAE framework enable the model to learn a latent space better suited for kNN estimations. In other words, our approach positions similar examples closer to each other in the latent space compared to a standard VAE + KMeans method.

Examples of the *prototypes* are visible in Fig. 3. Qualitatively, they clearly represent different ways to write digits and this pattern emerges from the update function during training. The Fig. 4 shows a tSNE representation of the latent space with both the real labels and the cluster labels. (See Figs. 1 and 2.) (See Table 1.)

Prototypes and t-SNE representations for Pneumonia datasets are visible in the Fig. 6 and Fig. 5. The prototype representations include both some artefactual structures and some realistic X-ray chest images, where in many images pneumonia-related opacities are visible.

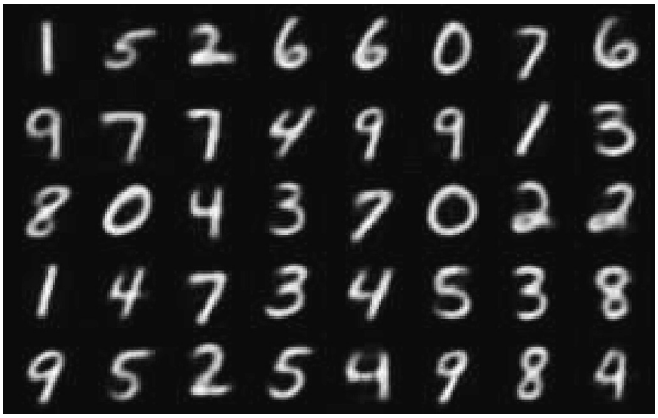

Fig. 3. Prototypes for MNIST. These are the reconstruction of the Q matrix rows.

5. Discussion

5.1. Overview of the proposed framework

In this work, we introduced a probabilistic framework that adapts the standard training of a variational autoencoder (VAE) to generate representations of prototypes or modes present in the dataset. After encoding, the representations are compared with a memory bank of prototypes, and a measure of similarity is passed to a conditional decoder, along with the latent representation, to generate reconstructions. The update method for the prototype values is detached from the computational graph, enabling the use of nondifferentiable functions for aggregation and updates. As a proof of concept, we employed an exponential moving average of the closest samples to the cluster prototype.

5.2. Training and latent space organization

During the initial phase of training, the encoder focuses on indirectly generating robust prototypes. This indirect process relies on the proximity and similarity of same-class labels in the latent space. By forcing the decoder to work with the conditioning vector and utilizing the update procedure, we establish an organization in the latent space that surpasses the results from standard VAEs. As training progresses, the temperature of the similarity conditioning vector decreases, leading to harder assignments. The encoder then concentrates on exploiting features concerning differences between samples and prototypes, improving the latent space organization for downstream classification tasks.

5.3. Comparison with baselines and potential applications

Our continuous conditional learning approach with decreasing temperature outperforms the baselines of other deep clustering VAE-based methods and demonstrates better performance compared to two-step approaches. The model retains VAE properties, allowing for sampling from the latent space and rejecting poor samples based on their similarity to a specific cluster. This capability enables various applications, such as generating similar digits from handwritten digit datasets or reconstructing specific modes of healthy or pathological medical images for further investigation. Our research presents a novel approach to image stratification with a focus on grayscale medical images. However, the concept of prototype discovery within our architecture can have significant implications for a variety of other domains, and studying these could be a promising direction for future research. For instance, agricultural imagery, as exemplified by the study 'Enhancing assessment of corn growth performance under various experimental management practices using Unmanned Aerial Vehicles (UAVs) and deep learning' [20], involves analysis of complex color patterns and spatial structures. Our prototype-based approach might help in identifying characteristic patterns associated with various crop conditions, leading to more accurate prediction and better management strategies.

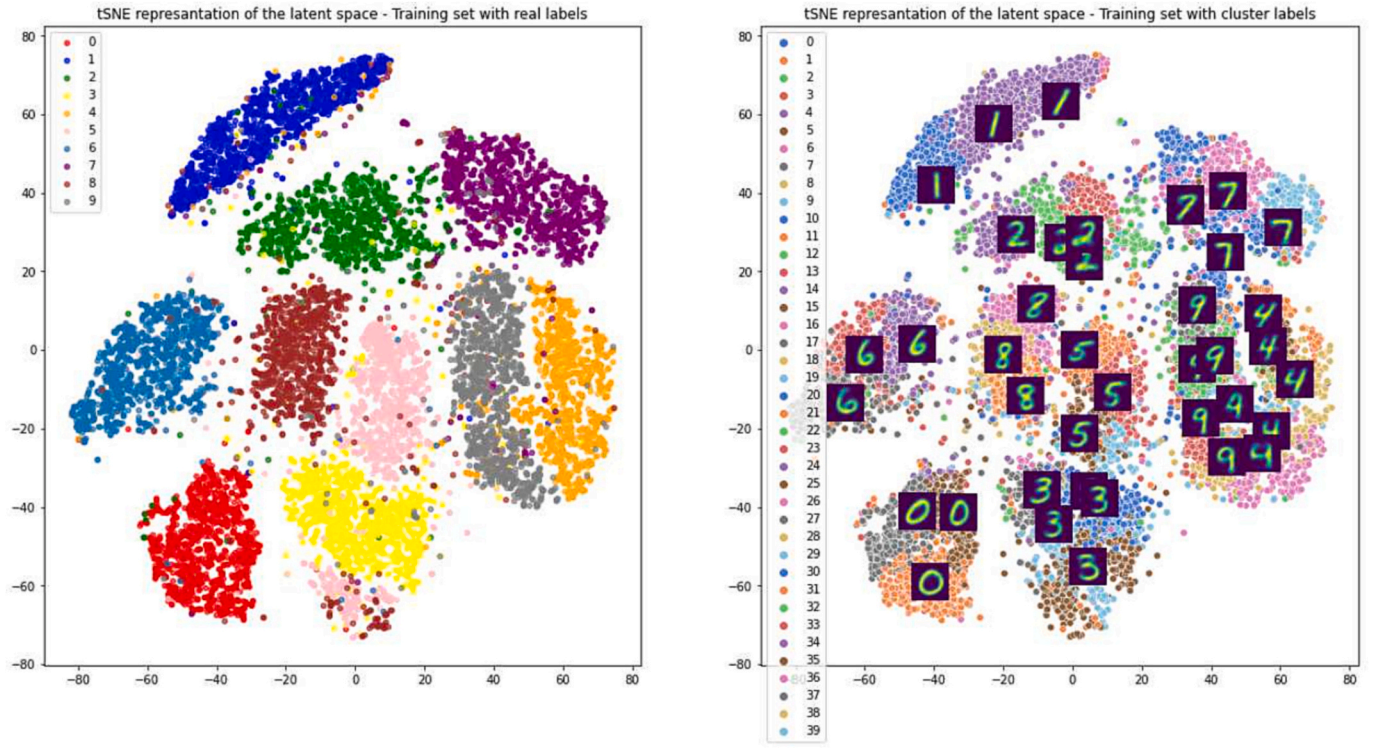

Fig. 4. tSNE visualization of the latent space. Left: colors are real labels. Right: colors are cluster labels, with superimposed cluster prototypes.

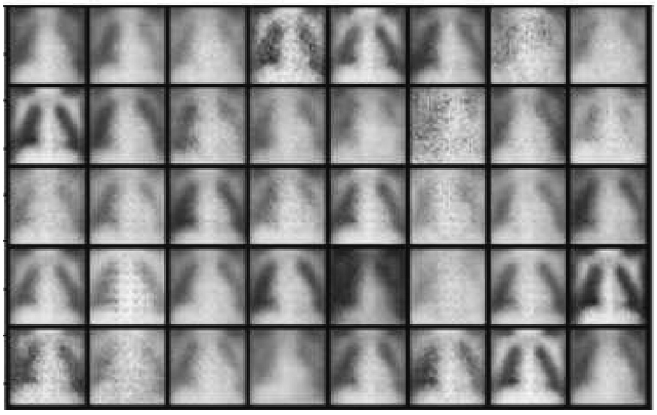

Fig. 5. Prototypes for Pneumonia. These are the reconstruction of the Q matrix rows.

Similarly, in the field of sports analytics, which often involves the study of dynamic video data, the feature extraction techniques illustrated in'Camera Position Estimation using 2D Image Dataset' [?] could potentially be adapted in conjunction with our model. By identifying prototypical movement patterns or game situations, coaches and analysts could gain deeper insights into player performance and game strategy. Moreover, the exploration of our method can be extended to fields such as geospatial analysis, urban planning, and even astrophysics. In these domains, the discovery of prototypes could assist in understanding common patterns or anomalies, which could lead to more effective decision-making and forecasting. While the application of our modified VAE architecture in these domains would require appropriate adaptation and validation, we believe that the fundamental principle of prototype discovery could offer significant benefits across a wide range of fields.

5.4. Challenges and considerations in applying self supervised deep learning to medical imaging

A critical aspect in the application of deep learning methods to medical imaging is understanding and addressing the inherent challenges associated with these images. Medical imaging modalities often produce images characterized by low resolution, high noise levels, low contrast, and potential geometric deformations. Such issues can impose significant difficulties for machine learning models, potentially impacting the quality of feature extraction, model generalization, and ultimately, performance in medical imaging tasks. High noise levels, often resulting from the imaging equipment, patient movement, or other environmental factors, may obscure important features in the images, thereby affecting the precision of prototype discovery. Similarly, low resolution and contrast can result in the loss of critical information, compromising the quality of the latent representations our model learns. Furthermore, geometric deformations, whether introduced during the imaging process or resulting from the patient's unique anatomy, can present challenges in generalizing learned patterns to unseen data. However, our self-supervised approach and the inherent properties of the Variational Autoencoder (VAE) architecture may provide a degree of robustness against these issues. By learning to encode salient information in a latent space, the model might counterbalance the effects of noise and resolution issues to some extent. Still, integrating specialized pre-processing steps such as noise reduction, contrast enhancement, and geometric correction could further enhance the performance of our model. Future work should focus on the development and integration of such tailored pre-processing techniques, as well as exploring architectures that are more resilient to these challenges in medical imaging. Acknowledging and addressing these inherent difficulties will be critical for the continued advancement and successful application of deep learning methods in the field of medical image analysis.

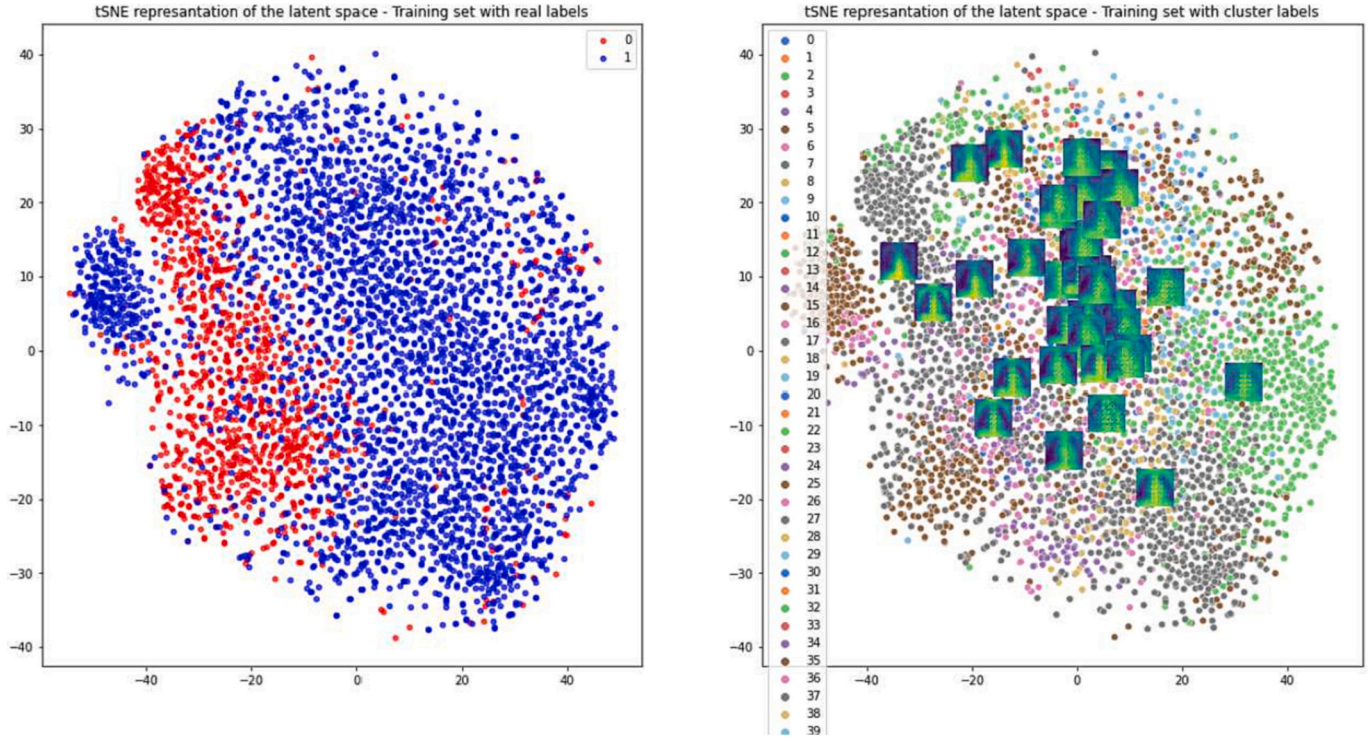

Fig. 6. tSNE visualization of the latent space. Left: colors are real labels. Right: colors are cluster labels, with superimposed cluster prototypes.

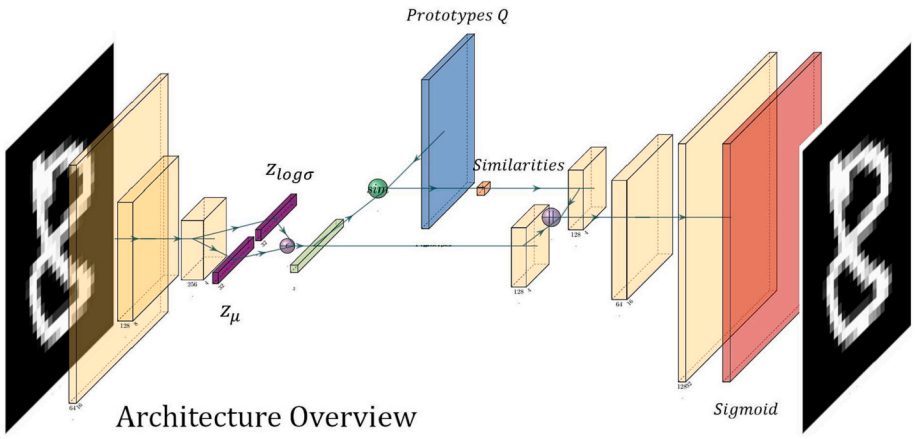

Fig. 1. Overview of VAESim architecture. We use ε to denote sampling with the reparametrization trick, *sim* denotes the similarity measure, and $\|$ represents the concatenation operation.

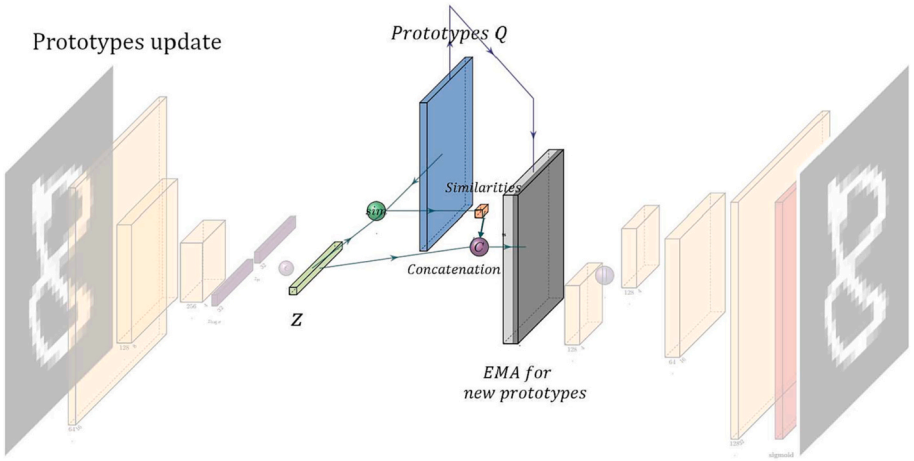

Fig. 2. Graphical description of the Q matrix update mechanism. See Algorithm 2 for details. We use the similarity between prototypes and sample embeddings to assign each sample embedding to the cluster with the most similar prototype vector. Then, we compute the mean of all z embeddings in the same cluster and use the exponential moving average between them to update the prototypes matrix.

5.5. Hyperparameters and future work

One challenge of our model is the need to initialize several hyperparameters, particularly the number of prototypes. Selecting the optimal number of prototypes and determining the best way to initialize the Q matrix in terms of values and dimensions are subjects for future investigations. While increasing the number of prototypes may improve performance, it also requires a larger batch size due to the greater number of samples per cluster. Identifying the optimal balance between these factors will be crucial for the success of our approach in various applications, including guided data augmentation and patient stratification.

6. Conclusion

We proposed a probabilistic framework that enhances the latent space organization for both reconstruction and downstream tasks by learning to reconstruct images through a latent representation conditioned on similarity measures to learned prototypes that represent modes present in the dataset. Our dynamic representation updating procedure implicitly incorporates features that enhance similarities and differences between samples and prototypes, leading to a well-organized latent space. We demonstrated the superiority of our approach over a two-step approach based on the combination of a VAE and KMeans, as well as other baselines based on different modifications of the VAE framework for deep clustering. Our end-to-end approach consistently outperformed these baselines on multiple metrics, resulting in up to + 10% improvement on average. The latent space produced by our model also generalizes well when used for downstream tasks, as shown by our evaluation of cluster label recall, kNN classification, and separability using a linear classifier. Finally, we showed that our approach achieves performances similar to fine-tuning very large models in a supervised way by starting with a self-supervised pretraining of our modified VAE.

Future work could investigate different strategies for prototype initialization and update. Further improvements could be obtained by exploring the use of different nondifferentiable functions for updating prototypes, as well as evaluating our approach on more challenging and diverse datasets in the medical domain. We believe that our model can be extended to address other problems in medical imaging beyond classification, such as clustering based on disease subtype, disease severity, or genetic information. Our work represents a step toward developing data-efficient unsupervised methods that can augment the interpretability and generalizability of medical imaging datasets.

Credit author statement

MF conceptualized the study, developed the idea, code, and conducted the experiments, and took the lead in writing the paper. TB, AD,

Table 1
Results on MNIST and PneumoniaMNIST datasets. The results are compared to different baselines and evaluated with accuracy measured with the statistical mapping approach to evaluate overlap between cluster labels and real labels, kNN to measure the closeness between similar samples in the latent space, and the classification done with a linear layer. The last column refers to the accuracy of a fine-tuned model from ResNet50 pre-trained on ImageNet.

Approach	dataset	statistical acc	knn acc	linear acc	resnet50 acc (supervised)
VAE + KMeans	MNIST	0.68 ± 0.01	0.80 ± 0.004	0.68 ± 0.04	0.99
GMVAE	MNIST	0.76 ± 0.01	0.91 ± 0.001	0.76 ± 0.01	0.99
VaDE	MNIST	0.81 ± 0.05	0.964 ± 0.001	**0.83 ± 0.008**	0.99
VAEsim	MNIST	**0.83 ± 0.01**	**0.970 ± 0.001**	0.81 ± 0.003	0.99
VAEsim (aug)	MNIST	0.40	0.90	0.72	0.99
VAE + KMeans	Pneumonia	0.631 ± 0.01	0.613 ± 0.014	0.451 ± 0.033	0.85
GMVAE	Pneumonia	0.625 ± 0.001	0.627 ± 0.041	0.429 ± 0.061	0.85
VaDE	Pneumonia	0.625 ± 0.001	0.747 ± 0.012	0.590 ± 0.116	0.85
VAEsim	Pneumonia	**0.677 ± 0.001**	**0.778 ± 0.024**	**0.671 ± 0.010**	0.85

and SS contributed substantially to the discussions, providing valuable insights and ideas. SS also played a key role in defining the methodology. NT supervised the project. All authors critically reviewed and revised the manuscript. The final version of the paper has been approved by all authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Part of this work is supported by the EXPERIENCE project (European Union's Horizon 2020 research and innovation program under grant agreement No. 101017727).

Matteo Ferrante is a Ph.D. student enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma.

Appendix A. Impact of hyperparameters

This section explores the impact of different hyperparameters on the performance of our proposed model. Specifically, we investigate the effects of varying the latent dimension, number of prototypes, and batch size, which are key factors in determining the model's performance.

To evaluate the impact of these hyperparameters, we performed a sweep over different values while keeping all other configurations fixed. We used the Weights & Biases library for this purpose [5]. We employed a fixed architecture consisting of an encoder with three convolutional layers using ReLU activation and batch normalization, followed by flattening and dense linear layers to output z_μ and $z_{log\sigma}$. The convolutional layers have a stride of 2, kernel size of 4, and padding of 1. The Q matrix of prototypes is a PyTorch matrix, and its values are updated using different functions that are detached from the computational graph. The decoder is almost symmetrical to the encoder, using transposed convolutions instead of standard convolutions. A linear layer maps the conditioning vector and the latent vector to the features of the first convolutional layer by concatenating them. The output convolutional layer has as many channels as the input image (1 for grayscale images and 3 for RGB images) with sigmoid activation.

Fig. 7 shows the impact of varying the latent dimension, batch size, and number of prototypes on classification accuracy, evaluated using the MNIST dataset. The statistical accuracy, kNN accuracy, and linear accuracy are reported as functions of the investigated parameter.

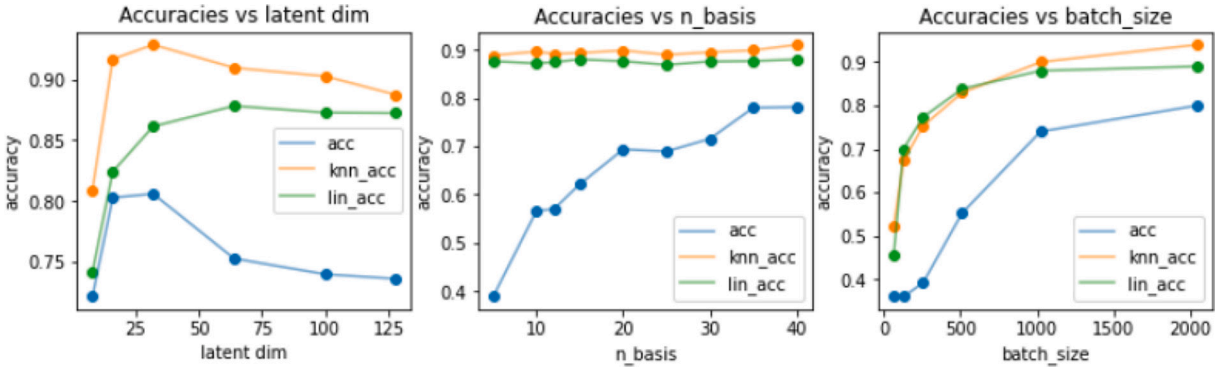

Fig. 7. Comparison between different metrics across values of *latent dim* (Left), *number of prototypes* (Center) and *batch size* (Right) for the MNIST dataset. Blue line report the accuracy computed with the statistical classification, the orange line the accuracy obtained with the kNN approach while the green line is related to the classification obtained with the linear classification. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Our results suggest that a latent dimension of around 32 is optimal, as it produces a peak in both the statistical and kNN accuracy, and the linear accuracy begins to converge to the optimal value. Increasing the number of prototypes improves the statistical accuracy, as more clusters reduce the variance within each assignment, but has little influence on kNN and linear accuracy. Increasing the batch size has a beneficial effect on all metrics.

These findings indicate that careful selection of hyperparameters is crucial for obtaining optimal performance of our proposed model. Further investigations are necessary to determine the optimal hyperparameter values for different datasets and applications.

References

[1] A.K. Aggarwal, P. Jaidka, Segmentation of crop images for crop yield prediction, Int. J. Biol. Biomed. 7 (2022) (2022) 40–44.

[2] E. Ahn, A. Kumar, D. Feng, M.J. Fulham, J. Kim, Unsupervised feature learning with k-means and an ensemble of deep convolutional neural networks for medical image classification, CoRR abs/1906.03359 (2019). http://arxiv.org/abs/1906.03359.

[3] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, M. Norouzi, Big self-supervised models advance medical image classification, 2021, https://doi.org/10.48550/ARXIV.2101.05224. https://arxiv.org/abs/2101.05224.

[4] B. Bengfort, R. Bilbro, N. Danielsen, L. Gray, K. McIntyre, P. Roman, Z. Poh, et al., Yellowbrick (2018), https://doi.org/10.5281/zenodo.1206264. http://www.scikit-yb.org/en/latest/.

[5] L. Biewald, Experiment Tracking with Weights and Biases, 2020 https://www.wandb.com/, software available from wandb.com.

[6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging Properties in Self-Supervised Vision Transformers, 2021, https://doi.org/10.48550/ARXIV.2104.14294. https://arxiv.org/abs/2104.14294.

[7] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, 2020, https://doi.org/10.48550/ARXIV.2002.05709. https://arxiv.org/abs/2002.05709.

[8] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big Self-Supervised Models are Strong Semi-Supervised Learners. arXiv preprint. arXiv:2006.10029, 2020.

[9] L. Deng, The mnist database of handwritten digit images for machine learning research, IEEE Signal Process. Mag. 29 (6) (2012) 141–142.

[10] L. Deng, The mnist database of handwritten digit images for machine learning research, IEEE Signal Process. Mag. 29 (6) (2012) 141–142.

[11] N. Dilokthanakul, P.A.M. Mediano, M. Garnelo, M.C. Lee, H. Salimbeni, K. Arulkumaran, M. Shanahan, Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders. https://openreview.net/forum?id=SJx7Jrtgl, 2017.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020, https://doi.org/10.48550/ARXIV.2010.11929. https://arxiv.org/abs/2010.11929.

[13] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, 2019, https://doi.org/10.48550/ARXIV.1911.05722. https://arxiv.org/abs/1911.05722.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2015, https://doi.org/10.48550/ARXIV.1512.03385. https://arxiv.org/abs/1512.03385.

[15] Z. Jiang, Y. Zheng, H. Tan, B. Tang, H. Zhou, Variational deep embedding: An unsupervised and generative approach to clustering, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 1965–1972, https://doi.org/10.24963/ijcai.2017/273.

[16] Kopf, A., Fortuin, V., Somnath, V.R., Claassen, M.: Mixture-of-experts variational autoencoder for clustering and generating from similarity-based representations on single cell data. PLoS Comput. Biol. 17(6), 1–17 (06 2021). doi:https://doi.org/10.1371/journal.pcbi.1009086.

[17] H.W. Kuhn, B. Yaw, The hungarian method for the assignment problem, Naval. Res. Logist. Quart (1955) 83–97.

[18] K.L. Lim, X. Jiang, C. Yi, Deep clustering with variational autoencoder, IEEE Signal Proc. Lett. 27 (2020) 231–235, https://doi.org/10.1109/LSP.2020.2965328.

[19] M. Moradi Fard, T. Thonet, E. Gaussier, Deep k-means: jointly clustering with k-means and learning representations, Pattern Recogn. Lett. 138 (2020) 185–192, https://doi.org/10.1016/j.patrec.2020.07.028. https://www.sciencedirect.com/science/article/pii/S0167865520302749.

[20] J. Xiao, S.A. Suab, X. Chen, C.K. Singh, D. Singh, A.K. Aggarwal, A. Korom, W. Widyatmanti, T.H. Mollah, H.V.T. Minh, K.M. Khedher, R. Avtar, Enhancing assessment of corn growth performance using unmanned aerial vehicles (uavs) and deep learning, Measurement 214 (2023) 112764, https://doi.org/10.1016/j.measurement.2023.112764. https://www.sciencedirect.com/science/article/pii/S0263224123003287.

[21] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. arXiv preprint. arXiv:2110.14795, 2021.

[22] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification, 2021, https://doi.org/10.48550/ARXIV.2110.14795. https://arxiv.org/abs/2110.14795.