DEEP DIFFUSION PROCESSES FOR ACTIVE LEARNING OF HYPERSPECTRAL IMAGES

Duc Nguyen, Abiy Tasissa, James M. Murphy

Department of Mathematics, Tufts University, Medford, MA 02155, USA

ABSTRACT

A method for active learning of hyperspectral images (HSI) is proposed, which combines deep learning with diffusion processes on graphs. A deep variational autoencoder extracts smoothed, denoised features from a high-dimensional HSI, which are then used to make labeling queries based on graph diffusion processes. The proposed method combines the robust representations of deep learning with the mathematical tractability of diffusion geometry, and leads to strong performance on real HSI.

Index Terms— hyperspectral images, variational autoencoders, deep clustering, active learning, diffusion geometry

1. INTRODUCTION

Machine learning has provided revolutionary new tools for remote sensing, but state-of-the-art methods often require huge labeled training sets. In particular, supervised deep learning methods can achieve near-perfect labeling accuracy on high-dimensional hyperspectral images (HSI), provided huge libraries of labeled pixels are available [1]. This hinders the practicality of these methods, as in many settings, data is collected at a pace that far exceeds human ability to generate corresponding labeled training data.

In order to account for this, methods that require only a very small number of labels are needed. The *active learning* regime is particularly attractive for HSI labeling problems. In active learning, an algorithm is provided with an unlabeled dataset, and the algorithm iteratively queries points for labels. By choosing query points in a principled manner, the resulting parsimonious training set can provide the classification performance of a much larger training set chosen uniformly at random.

We propose an active learning method for HSI based on deep feature extraction and random walks on graphs. A variational autoencoder is used to nonlinearly denoise and compress the high-dimensional HSI. Then, the resulting features are considered as vertices of a graph, and a Markov diffusion process on the graph is used to determine label queries and label all data points. The proposed method combines the excellent feature learning of deep autoencoders with the mathematical interpretability of graph diffusion processes, and leads to strong empirical performance on real HSI.

2. BACKGROUND

Extracting essential features of data is an important part of both supervised and unsupervised machine learning problems. For example, the unsupervised technique of Gaussian mixture models (GMM) assumes that the data under consideration is an i.i.d. sample from a mixture of Gaussians. To solve the unsupervised clustering problem on these data, parameters of the GMM are then estimated using the expectation maximization algorithm [2]. In recent years, deep generative methods, such as *generative adversarial network (GANs)* and *variational autoencoder networks (VAEs)*, have been used as feature extraction tools and successfully applied to many machine learning tasks [3, 4]. In the context of the clustering problem, a set of methods, known as deep clustering, propose learning features of the data and clustering simultaneously, showing strong empirical results [5, 6, 7, 8].

2.1. Variational Autoencoder Networks (VAEs)

We now describe VAEs following the exposition in [9]. Let $x \in \mathbb{R}^N$ and consider a latent variable model $p_\theta(x) = \int p(x|z)p(z)\,dz$. The variable $z \in \mathbb{R}^k$, $k \ll N$, is a latent variable and provides a low-dimensional parametrization of the data x. The goal of VAE is to maximize the probability of data samples generated as $p(x)$. In the inference part of VAE, the training procedure maps a data point x to its latent representation z such that the z adheres to the distribution $p(z)$.

Since the posterior distribution $p(z|x)$ and $p(z)$ are unknown, VAE makes the assumption that $p(z|x) = q(z|x, \phi)$ where ϕ [where does ϕ live?] are parameters to be learned from the network. To enforce that the variational distribution $q(z|x, \phi)$ agrees with $p(z)$, the optimization

$$\min_{\phi} \; KL(q(z|x, \phi), p(z)),$$

is considered, where KL is the Kullback-Leibler divergence. With this, $q(z|x)$ can be interpreted as the encoder part of VAE providing us a latent representation z given a sample x. The generative part of VAE considers the reconstruction of a data sample given the latent representation z. In particular, the goal of the training procedure is to map a latent variable z to a data sample $\hat{x}$ such that $\hat{x}$ is similar to the true data sample x. This naturally motivates the maximization of $p(x|z, \theta)$ with θ

denoting the network parameters. Formally, we consider the following optimization program:

$$\max_{\theta} \mathbb{E}_{z\sim q(z|x,\lambda)} \log(p(x|z,\theta))$$

With this, $p(x|z)$ can be interpreted as the decoder part of VAE. The inference and generative parts of VAE i.e the encoder and decoder model can be jointly trained by maximizing the loss function

$$\mathcal{L}(x,z;\theta,\phi) = \mathbb{E}_{z\sim q(z|x,\lambda)} \log(p(x|z,\theta)] - KL(q(z|x,\phi),p(z))$$

In summary, the optimal network parameters θ^* and ϕ^* are obtained as $(\theta^*,\phi^*) = \arg\max_{\theta,\phi} \mathcal{L}(x,z;\theta,\phi)$, optimized using stochastic gradient descent [Should this be empirical, since in practice we will have a sample?]. In our proposed methodology, we first use the VAE for feature extraction of the data. Specifically, the latent space representation resulting from a trained VAE is used as input to a clustering algorithm. [Why is this a good idea?]

2.2. Learning by Active Nonlinear Diffusion

In this paper, we use an active learning algorithm for high dimensional data based on diffusion geometry [10, 11, 12]. The algorithm, learning by active nonlinear diffusion (LAND) [13], is a semisupervised algorithm that takes into account the geometry of the data to identify the most important points to be queried for labeling. A key advantage of LAND is that it provides rigorous theoretical performance guarantees. In addition, the algorithm handles general clusters that could be nonlinear, live in a high ambient dimension and may be susceptible to noise and outliers [13].

Consider an HSI $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, where each pixel is represented as a point in $\mathbb{R}^D$ where D is the number of spectral bands. Define W to be the $N \times N$ weight matrix defined as $W(x,y) = e^{-\|x-y\|_2^2/\sigma^2}, x \in NN_k(y)$ and $W(x,y) = 0$ otherwise, where $NN_k(x)$ is the set of k-nearest neighbors of y in X with respect to Euclidean distance and σ is a scale parameter. Given the weight matrix, the degree of x is $\deg(x) := \sum_{y\in X} W(x,y)$. A random walk on X can be defined using the $N \times N$ transition matrix $P(x,y) = W(x,y)/\deg(x)$. By construction, P has a spectral decomposition $\{(\lambda_n,\Psi_n)\}_{n=1}^N$, and we define the *diffusion distance* between $x, y \in X$ as $d_t^2(x,y) = \sum_{n=1}^N \lambda_n^{2t}(\Psi_n(x) - \Psi_n(y))^2$. The parameter t in diffusion distance depends on a parameter t informs how long the diffusion process runs. In this paper, we set t to be 30; diffusion learning is relatively robust to choice of t [10].

An important part of the LAND algorithm is to determine the data points which one should query for labels. This is achieved using a kernel density estimator and diffusion geometry. For instance, consider the density estimator

$$p(x) = \sum_{y\in NN_k(x)} \exp(-\|x-y\|_2^2/\sigma_0^2)$$

with σ_0 denoting a scale parameter to be set. Let

$$\begin{aligned}\rho_t(x) &= \min_{x\neq \arg\max_z p(z)} \{D_t(x,y)\ | p(y) \geq p(x), x \neq y\}, \\ &= \max_{y\in X} D_t(x,y), x = \arg\max_z p(z) \qquad (1)\end{aligned}$$

[this should be cases, right?] be the (t-dependent) diffusion distance between a point and its nearest diffusion neighbor of higher density if x is not the maximizer of $p(x)$, and the maximum diffusion distance to another point if x is the maximizer of $p(x)$. The modes of the data are determined as the maximizers of $\mathcal{D}_t(x) = p(x)\rho_t(x)$. We interpret this quantity as follows. A large $\mathcal{D}_t$ value for points indicate that they are high density and are D_t-far from other high density points. In LAND, the modes of X are characterized as maximizers of $\mathcal{D}_t$. Diffusion distances and density can also employed to label all other points relative to these modes. We refer the interested reader to [13].

3. PROPOSED ALGORITHM

We propose an active learning algorithm, VAE-LAND, which has two main stages. The first stage is feature extraction of an unlabeled high-dimensional dataset using a variational autoencoder. The second stage employs the LAND active learning diffusion based clustering algorithm to infer the true labels. Details of the method are in Algorithm 1.

The second part of the algorithm concerns deploying clustering algorithms on the latent representation of the data. Since our algorithm is in the active learning framework, we first acquire labels for small subset of the latent data. Different active learning algorithms differ in the ways in which one determines the subset of the data to query for labels. For example, in a related work on active learning using variational autoencoders [14], K-means clustering is first used to partition the space and then acquire the labels using uniform random sampling in each partition. Given the labels, a classifier is then trained in the latent space for the prediction task. The underlying assumption is that the K-means partitions represent the structure of the latent space, and empirical results in [14] show the superiority of this approach to doing active learning via K-means partitions in the original space. Since the performance of the active learning algorithm hinges on how well the geometry of the latent space is represented, the choice of the clustering method is crucial. [Some of this could/should be put into a "Section 3.1. Comparison to Related Methods" or something similar].

The proposed algorithm combines the power of variational autoencoders to extract features and uses diffusion geometry on graphs to find impactful labels to query, which then propogate to other points. To summarize, our proposed algorithm has two main steps. We first use a variational autoencoder to train the unlabeled data and then cluster the latent representation of the data using LAND. [This paragraph seems a bit redundant.]

Algorithm 1: Variational Autoencoder Learning by Active Nonlinear Diffusion (VALAND)

```
Train VAE on $\{x_i\}_{i=1}^n$, the unlabeled data, to obtain
the latent representation $\{\hat{x}_i\}_{i=1}^n$. With abuse of
notation, we drop the the hats here on.

Input:[the spacing seems a bit much here]

    • $\{x_i\}_{i=1}^n$ (Unlabeled Data)
    • $\{(\lambda_\ell, \psi_\ell)\}_{\ell=1}^M$ (Spectral Decomposition of $P$)
    • $\{p(x_i)\}_{i=1}^n$ (Kernel Density Estimate)
    • $\{\rho_t(x_i)\}_{i=1}^n$ (1)
    • $t$ (Time Parameter)
    • $B$ (Budget)
    • $\mathcal{O}$ (Labeling Oracle)

Output:

    • $Y$ (Labels)

Compute $\mathcal{D}_t(x_i) = p(x_i)\rho_t(x_i)$.
Sort the data in decreasing $\mathcal{D}_t$ value to acquire the
    ordering $\{x_{m_i}\}_{i=1}^n$.
for $i = 1 : B$ do
    Query $\mathcal{O}$ for the label $L(x_{m_i})$ of $x_{m_i}$.
    Set $Y(x_{m_i}) = L(x_{m_i})$.
end for
Sort $X$ according to $p(x)$ in decreasing order as
    $\{x_{\ell_i}\}_{i=1}^n$.
for $i = 1 : n$ do
    if $Y(x_{\ell_i}) = 0$ then
        $Y(x_{\ell_i}) = Y(z_i)$, $z_i = \arg\min_z\{D_t(z, x_{\ell_i})$ —
        $p(z) > p(x_{\ell_i})$ and $Y(z) > 0\}$.
    end if
end for
```

4. EXPERIMENTAL RESULTS

We demonstrate the accuracy of the proposed algorithm by running numerical experiments. The training of the VAE is done using Tensorflow in Python. For doing the clustering via LAND, we use the publicly available MATLAB code at `https://jmurphy.math.tufts.edu/Code/`. Our test HSI dataset is the Salinas A hyperspectral dataset. The Salinas scene was captured over Salinas Valley, California. The image has a spatial resolution of 3.7-meter pixels and contains 224 spectral bands. The ground truth consists of 16 classes. We consider the Salinas A dataset, which is a subset of the Salinas dataset, and contains 6 classes. The Salinas A dataset and the ground truth data are publicly available at `http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Salinas-A_scene`. Figure 1 shows a visual of the high dimensional data and the ground truth labels. The performance of the algorithm is assessed using overall accuracy. This is defined as the ratio of correctly estimated labels to total number of labels.

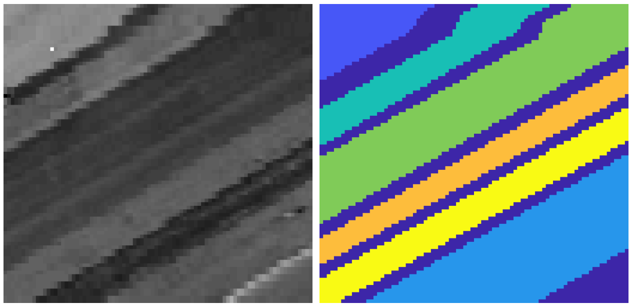

Fig. 1: The 86 × 83 Salinas A HSI data consists of 6 classes. Left: the sum of all spectral bands. Right: the ground truth.

The Salinas A HSI dataset of size 83 × 86 with 224 spectral bands is represented as image of size 7138 × 224. We use the unlabeled data to learn a latent space representation of Salinas A in $\mathbb{R}^{40}$ dimensions. The encoder network consists of three dense layers with 128 units. The activation function is the rectified linear unit (RELU). The loss function is optimized using the Adam algorithm with learning rate set to 0.0001. After training the VAE, we input the latent space representation of the Salinas A dataset to the LAND algorithm for the task of inferring the ground truth labels of the HSI data. We compare our result to the “plain” LAND algorithm that simply clusters the Salinas A dataset in its original representation. Since LAND is an active learning framework, we consider varying number of labeled data points ranging from 10 to 2000. Figure 1 compares the performance of LAND and performance of VAE-LAND. First, for both VAE-LAND and standard LAND, LAND queries lead to significantly better accuracy than random queries. The proposed algorithm, VAE-LAND, attains an accuracy of 96.97% with just 10 labeled points. This is a 12.5% improvement to the accuracy of competitive LAND algorithm for the same number of labeled points. The standard LAND algorithm requires 400 labeled points to reach accuracy of 90% while for the same number of labeled points, VAE-LAND has an accuracy of 98.35%. [Let’s break down the model parameter description from the data description.]

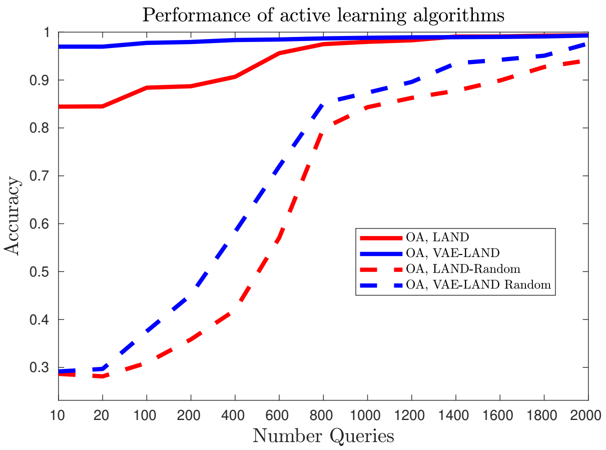

Fig. 2: For the Salinas A dataset, the performance of variational autoencoder LAND learning achieves a higher accuracy than the standard LAND algorithm. With just 10 points, the overall accuracy of VAE-LAND is 96.97%, a 12.5% improvement to the competitive LAND algorithm. [Asymptotics?]

4.1. Discussion of Experimental Results

4.2. Computational Complexity

5. CONCLUSIONS AND FUTURE DIRECTIONS

The proposed active learning algortihm, VAE-LAND, improves over the standard LAND and gives accurate results even when the number of queries are limited. The method uses VAE to generate good features, and uses the diffusion geometry based LAND algorithm to determine query points. The LAND algorithm then uses this queried labels to predict the labels of the unlabeled data samples. In future work, we would like to explore the kind of data models for which the algorithm has theoretical performance guarantees.

Acknowledgements

This research is partially supported by the US National Science Foundation grants NSF-DMS 1912737, NSF-DMS 1924513, and NSF-CCF 1934553.

6. REFERENCES

[1] X.X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[2] K.P. Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012.

[3] M.E. Abbasnejad, A. Dick, and A. van den Hengel, "Infinite variational autoencoder for semi-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5888–5897.

[4] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[5] Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu, "Learning deep representations for graph clustering," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[6] Chunfeng Song, Feng Liu, Yongzhen Huang, Liang Wang, and Tieniu Tan, "Auto-encoder based data clustering," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2013, pp. 117–124.

[7] Junyuan Xie, Ross Girshick, and Ali Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, 2016, pp. 478–487.

[8] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou, "Variational deep embedding: an unsupervised and generative approach to clustering," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 1965–1972.

[9] T. Yang, G. Arvanitidis, D. Fu, X. Li, and S. Hauberg, "Geodesic clustering in deep generative models," *arXiv preprint arXiv:1809.04747*, 2018.

[10] J.M. Murphy and M. Maggioni, "Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1829–1845, 2019.

[11] M. Maggioni and J.M. Murphy, "Learning by unsupervised nonlinear diffusion," *Journal of Machine Learning Research*, vol. 20, no. 160, pp. 1–56, 2019.

[12] J.M. Murphy and M. Maggioni, "Spectral-spatial diffusion geometry for hyperspectral image clustering," *IEEE Geoscience and Remote Sensing Letters*, 2019.

[13] M. Maggioni and J.M. Murphy, "Learning by active nonlinear diffusion," *Foundations of Data Science*, vol. 1, no. 3, pp. 271, 2019.

[14] F. Pourkamali-Anaraki and M.B. Wakin, "The effectiveness of variational autoencoders for active learning," *arXiv preprint arXiv:1911.07716*, 2019.