Unsupervised Feature Learning for Outlier Detection with Stacked Convolutional Autoencoders, Siamese Networks and Wasserstein Autoencoders: Application to Epilepsy Detection

Zara Alaverdyan[✉], Jiazheng Chai, and Carole Lartizien

Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne,
CNRS, Inserm, CREATIS UMR 5220, U1206, 69621 Lyon, France
zaruhi.alaverdyan@creatis.insa-lyon.fr

Abstract. In this study we tackle the problem of detecting subtle epilepsy lesions in multiparametric (T1w, FLAIR) MR images considered as normal during a visual examination by a neurologist (MRI *negative*). We cast this problem as an outlier detection problem and adapt the framework proposed in [1]. It consists in learning a oc-SVM model for each voxel in the brain volume. We generalize this approach by proposing unsupervised deep architectures as feature extracting mechanisms in order to learn representations characterizing healthy subjects. We hypothesize that such architectures may capture features that allow to distinguish pathological voxels from the normal cases used in the training. As such, we exploit and compare three architectures, a novel configuration of siamese networks, stacked convolutional autoencoders and Wasserstein autoencoders. The models are trained on 75 healthy subjects and validated on 21 patients (with 18 MRI *negatives*) with confirmed epilepsy lesions achieving the best sensitivity of 62%.

Keywords: Wasserstein autoencoders · Siamese networks · Unsupervised learning · Epilepsy detection · Anomaly detection

1 Introduction

Computer aided diagnosis (CAD) systems assist clinicians in various tasks such as organ or lesion segmentation, detection of abnormal regions in a medical image, etc. The vast majority of the existing CAD systems are built upon methods developed in supervised settings, using either manually designed features or currently ubiquitous deep learning architectures. However, when the number of labeled pathological cases in the training set is not sufficient to account for the complexity of the task, supervised learning becomes infeasible. To bypass the problem of insufficient labeled data, some authors formulate lesion detection

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 210–217, 2018.
https://doi.org/10.1007/978-3-030-00889-5_24

tasks in semi-supervised settings, by accounting for both labeled and unlabeled data in a deep architecture for MS lesion segmentation [2] or by exploiting weak labels (the number of lesions in a scan) to detect enlarged perivascular spaces in the basal ganglia [3].

Another recent tendency goes even further and casts lesion detection problem as an anomaly detection task. Anomaly detection, also referred to as outlier detection, consists in learning the boundary of the normal class in order to later identify the observations that lay outside of it. Over the recent years the challenging topic of outlier detection has been studied extensively and many algorithms have been proposed for outlier detection depending on the nature of the data and the type of anomalies [4]. In computer vision, recent works investigated approaches based on deep architectures such as autoencoders or Generative Adverserial Networks (GANs) coupled with various outlier detection algorithms [5]. In the medical imaging domain, [6,7] proposed a model defining a score function that measures how anomalous a given sample is based on the reconstruction and discrimination losses estimated by a GAN architecture trained on normal samples only. In [8,9], a latent representation of normal samples is first learned with deep unsupervised networks and then fed to a one-class support vector machine (oc-SVM) model to estimate the boundaries of the normal examples.

In this work we build on the framework proposed in [9] for the challenging application of epilepsy lesion detection in patients with *MRI negative* exams, meaning that the lesions were not visually identified by clinicians on the MR scans [10]. We propose to exploit three unsupervised deep learning architectures as feature extracting mechanisms in the outlier detection context. We consider stacked convolutional autoencoders, a novel configuration of siamese networks [9] and Wasserstein autoencoders [11] that have been shown to combine the advantages of both standard generative adversarial networks (GAN) and variational autoencoders (VAE) in generating synthetic natural images without compromising the stability of the training. We couple these architectures with voxel-level oc-SVM models and compare their performances on the epilepsy lesion detection task.

2 Method

2.1 Unsupervised Feature Extraction with Autoencoders

The first step of the proposed system is to learn patch-level representations of healthy subjects by exploiting the three types of architectures below.

Stacked Convolutional Autoencoders (sCAE) are a variation of autoencoders that first map the input $\mathbf{x} \in \mathcal{X}$ to a latent representation space $\mathcal{Z}$ through a series of convolutional and max-pooling operations (encoder E) and later map it back to the original input space with a series of de-convolutions and up-poolings by producing a reconstruction $\tilde{\mathbf{x}}$ of the input (decoder G). The

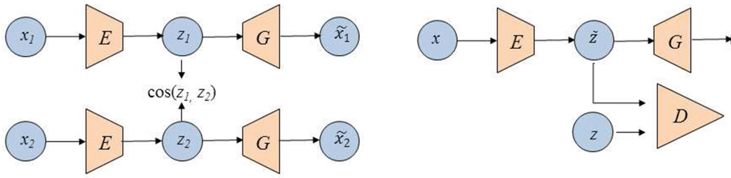

Fig. 1. Left: Siamese neural network composed of stacked convolutional autoencoders as sub-networks (sCAE). The input consists of a pair of patches ($\mathbf{x_1}$, $\mathbf{x_2}$) of 2 different subjects centered at the same voxel in the brain. The encoder E maps $\mathbf{x}$ to the latent representation $\mathbf{z}$ while the decoder G maps it back to the input space producing a reconstruction $\tilde{\mathbf{x}}$. Right: Wasserstein autoencoder (WAE) composed of an encoder E, a decoder G and an adversary discriminator D.

parameters are iteratively updated to minimize the deviation between the output $\tilde{\mathbf{x}}$ and the input $\mathbf{x}$. A sCAE is illustrated on Fig. 1 as the top sub-network of the architecture on the left.

Regularized Siamese Autoencoders (rSN), as proposed in [9], consist of two identical (same architecture, shared parameters) stacked convolutional autoencoders with K hidden layers and a cost module (shown on Fig. 1). The siamese network receives a pair of patches $(\mathbf{x_1}, \mathbf{x_2})$ at input, then each patch is propagated through the corresponding subnetwork yielding representations $(\mathbf{z_1}, \mathbf{z_2})$ respectively in the middle layer which are then passed to the loss function 1 below. The network is trained to maximize the cosine similarity of the representations of patches centered at the same voxel and belonging to different healthy subjects, at the same time imposing the subnetworks to produce reconstructions close to the original input. The loss function for a single pair hence is:

$$L_{rSN}(\mathbf{x_1}, \mathbf{x_2}; \Theta_{rSN}) = \sum_{t=1}^{2} ||\mathbf{x_t} - \tilde{\mathbf{x}}_\mathbf{t}||_2^2 - \alpha \cdot cos(\mathbf{z_1}, \mathbf{z_2}) \tag{1}$$

where $\tilde{\mathbf{x}}_\mathbf{t}$ is the reconstructed output of the patch $\mathbf{x_t}$ produced by sub-network t while $\mathbf{z_t}$ is its (vectorized) representation in the middle layer and α is a coefficient that controls the tradeoff between the two terms. Θ_{rSN} denotes the parameter set.

Wasserstein Autoencoders (WAE) have been recently introduced as generative models combining the best properties of Wasserstein GANs and Variational Autoencoders [12]. As shown on Fig. 1, a Wasserstein auto-encoder consists of three components: an encoder E mapping an input patch from the data space $\mathcal{X}$ to the latent space $\mathcal{Z}$, a decoder G mapping a latent code from the latent space $\mathcal{Z}$ to the data space $\mathcal{X}$, and an adversary network D that tries to distinguish the prior distribution of the latent code P_Z from the latent distribution Q_Z produced by the encoder. The resulting loss function can be expressed as

$$L_{WAE}(X; \Theta_{WAE}) = \frac{1}{N} \sum_{i=1}^{N} c(\mathbf{x_i}, \tilde{\mathbf{x}}_\mathbf{i}) + \lambda \cdot D_Z(P_z, Q_z) \tag{2}$$

where D_Z measures the discrepancy between a given distribution P_z and Q_z for the dataset $X = \{\mathbf{x_i}\}_{1,..,N}$ and c measures the reconstruction error. λ is a coefficient that controls the tradeoff between the two terms and Θ_{WAE} denotes the parameter set. The generic form of the WAE loss allows different reconstruction error functions and regularizers. We used the standard reconstruction error $c(\mathbf{x_i}, \tilde{\mathbf{x}}_\mathbf{i}) = ||\mathbf{x} - \tilde{\mathbf{x}}_\mathbf{i}||_2^2$ and the Jenssen-Shanon divergence as D_Z.

2.2 Voxel-Level Outlier Detection with Oc-SVM Classifiers

A oc-SVM classifier [13] is an outlier detection method that seeks to find the optimal hyperplane that separates the given points from the origin in a dot product space defined by some kernel function ϕ. The latent representations $\mathbf{z}$ learnt by each of the networks proposed above was used to train oc-SVM classifiers at voxel level. For a given voxel v_i, the associated oc-SVM model C_i is trained on the matrix $M_i = [\mathbf{z_{i1}}, ..., \mathbf{z_{in}}]$ where $\mathbf{z_{ij}}$ is the feature vector corresponding to the patch centered at v_i of subject j and n is the number of subjects. For a new patient, each voxel v_i is matched against the corresponding model C_i and is assigned the signed score output by C_i. This yields a *distance map* D_p for the given patient. This map is later normalized by the estimated voxel-level standard deviation (computed on the healthy subjects with 1-fold evaluation). We keep the most negative scores up to the score corresponding to a pre-chosen *p-value* in the patient's distance score distribution and apply a 26-connectivity rule to identify connected components which we refer to as *clusters* (and the map - *cluster map*). The *clusters* are what we refer to as *detections* by the proposed method. The clusters are then ranked according to the size and the average score of their voxels. Such ranking favors large clusters with the most negative average score. Finally, we keep the top n detections and discard the rest. When a cluster overlaps significantly with the ground truth of a patient we consider it a *true positive* and *false positive* otherwise.

3 Experiments and Results

3.1 Dataset Description and Pre-processing

The study was approved by our institutional review board with approval numbers 2012-A00516-37 and 2014-019 B and a written consent was obtained for all participants.

Our database consists MR images (T1-weighted and FLAIR) of 75 healthy subjects and 21 patients acquired on a 1.5T Sonata scanner (Siemens Healthcare, Erlangen, Germany). All the volumes were normalized to the standard brain template of the Montreal Neurological Institute (MNI) [14] using a voxel size of $1 \times 1 \times 1$ mm with the unified segmentation algorithm [15] implemented in

SPM12 also correcting for magnetic field inhomogeneities. This spatial normalization assures a voxel-level correspondence between the subjects. We removed top 1% intensities and scaled the images between 0 and 1 at image level before feeding the patches to the networks.

The method has been validated on 21 patients admitted to our clinical center with confirmed medically intractable epileptogenic lesions: 2 of them were visually detected on the FLAIR images and only 1 lesion was identified on both T1w and FLAIR scans. The remaining 18 patients are confirmed *MRI negative* patients. The *MRI negative* patients had surgeries and have been seizure-free since. The ground truth annotations used in the performance evaluation were obtained by outlining the visible zones of the *MRI positive* patients and by combining the information of post-surgical MR images and the resected zones for *MRI negative* patients.

3.2 Feature Extraction with sCAE, rSN and WAE

As shown on Fig. 1, the three architectures consist of the same encoder E and decoder G (the stacked convolutional autoencoder is identical to the upper subnetwork of the siamese network). The architecture details are shown on Fig. 2a. The encoder E takes as input an $18 \times 18 \times 2$ patch (the third dimension corresponds to the two modalities-T1 and FLAIR) and outputs a latent representation $\mathbf{z}$ of dimension 64. LeakyReLU was used as activation in the WAE discriminator with scale 0.02 for negative input values. ReLU was used in the generator and the encoder (except for the last layer of G where sigmoid is applied). We varied the λ parameter values in loss 2 among 1, 5, 10, 20 and 100.

All the three networks were trained on the same data set of patches extracted from healthy subjects' images with a stride 8. In the case of the siamese network, each patch of a subject was randomly matched with a 'similar pair' among the remaining subjects. The α parameter in the loss 1 is set to 0 during the first 10 epochs, then grows linearly for 15 epochs until it reaches 0.5 and then plateaus for 5 more epochs. The Adam optimizer was used with the learning rate set to 0.001 with a training batch size of 128.

3.3 oc-SVM Classifier Design

We used oc-SVM classifiers with RBF kernel by setting the kernel width γ for each voxel v_i individually to the estimated median of the standardized euclidean pairwise distances of the corresponding matrix M_i (see Sect. 2.2) as in [16]. The allowed fraction of outliers for all models was set to 0.03 (this parameter does not impact the results).

3.4 Results and Discussion

Below we evaluate the performance of the system on 21 patients with confirmed epilepsy lesions. Figure 2b shows the performance obtained with each of the architectures: the y-axis shows the detection rate among the top n clusters, ranked according to their average score and size. The rSN features seem to outperform the features learnt with WAE and sCAE, WAE performing better than sCAE for certain values of λ ($\lambda = 1$ and $\lambda = 100$ did not yield a good performance). The latter confirms our hypothesis that the reconstruction error, when enhanced with a regularization, fits better to the anomaly detection context. The WAE performance is still inferior to that of rSN which might be due to a limitation of the model itself or the experimental choice of the hyper-parameters (we can see how the performance is affected by the choice of λ; the value 20 is less successful, probably since it prioritizes too much the adversarial term; the value 100 entirely degraded the results and, hence, is not shown). Figure 3 shows the output of the system with the considered architectures. The patient has a visible lesion outlined in green. The detection quality varies, especially WAE with $\lambda = 20$ almost misses the lesion.

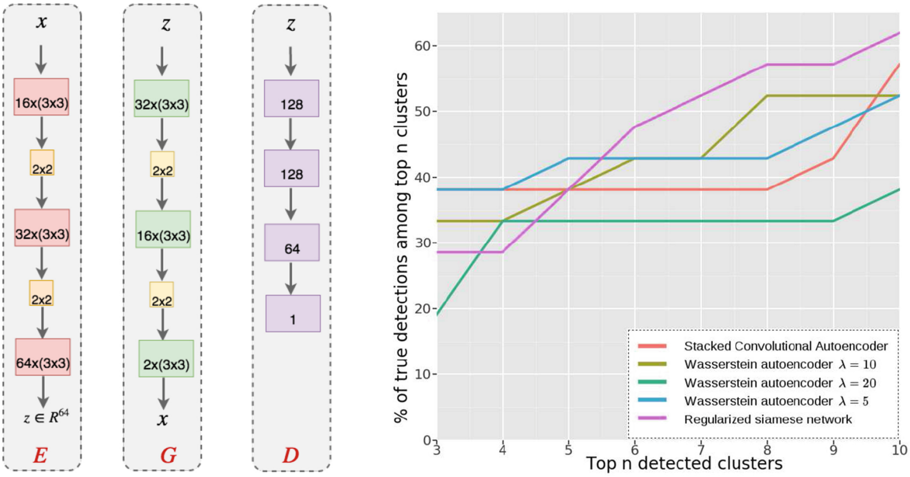

Fig. 2. (a) The encoder, decoder and discriminator architectures respectively. Red/green/violet boxes denote convolutional/deconvolutional/fully connected layers respectively. Orange/yellow boxes stand for maxpooling/uppooling. (b) The performance of the CAD system with sCAE, WAE and rSN features. x-axis: Top n clusters, y-axis: Detection rate among the top n clusters. Ranking based on average score and size. (Color figure online)

Unlike most recent studies that focus on a single epilepsy type (FCD) and use handcrafted features characterizing it [1,17–20], our method seeks to find more complex features in an unsupervised manner in order to identify lesions with rather unknown signatures. Naturally, such an approach, when applied to a specific pathology, is likely to produce more false positive detections. Although a fair comparison with the published results is difficult because of the differences in the patient groups, the obtained results (62% sensitivity for 9 false positives

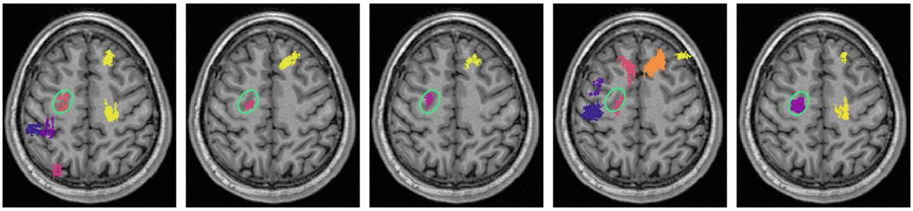

Fig. 3. CAD output for a MRI positive patient with sCAE, WAE $\lambda = 5$, WAE $\lambda = 10$, WAE $\lambda = 20$ and rSN features respectively. The images show the maximum intensity projections of the cluster maps onto an MRI transverse slice (ground truth is outlined in green circles). The maps show the top 6, 2, 2, 6 and 3 clusters, respectively. (Color figure online)

per scan for rSN features and between 52–58% for WAE and sCAE) are of the same order as those reported in recent studies for the difficult task of automated epilepsy detection in MRI negative patients ([17] reports a detection rate of 70% when individual SBM-based features are used; the results vary between 60 and 70% when considering combinations of some of these SBM features). *MRI positive* lesions are detected quite soon (usually among top 2–4 clusters) which is due to the fact that such lesions have visible markers that allow to distinguish them easily unlike the *MRI negative* patients whose lesions may be detected along with other outliers of similar 'suspiciousness'. Finally, the method with all the networks is quite straightforward to implement and to apply in daily practice as the output of the system can be obtained under a couple of minutes.

Acknowledgements. This work received funding from the French Foundation for Research in Epilepsy (FFRE). It was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

1. El Azami, M., Hammers, A., Jung, J., Costes, N., Bouet, R., Lartizien, C.: Detection of lesions underlying intractable epilepsy on t1-weighted MRI as an outlier detection problem. PloS ONE **11**(9), e0161498 (2016)
2. Baur, C., Albarqouni, S., Navab, N.: Semi-supervised deep learning for fully convolutional networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 311–319. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_36
3. Dubost, F., et al.: GP-Unet: lesion detection from weak labels with a 3D regression network. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 214–221. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_25
4. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 15:1–15:58 (2009)

5. Kiran, B.R., Thomas, D.M., Parakkal, R.: An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. J. Imaging **4**(2), 36 (2018)
6. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised Anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 146–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_12
7. Zenati, H., Foo, C.S., Lecouat, B., Manek, G., Chandrasekhar, V.R.: Efficient GAN-based anomaly detection
8. Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C.: High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. Pattern Recogn. **58**, 121–134 (2016)
9. Alaverdyan, Z., Jung, J., Bouet, R., Lartizien, C.: Regularized Siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. In: First Conference on Medical Imaging with Deep Learning (MIDL 2018)
10. Kini, L.G., Gee, J.C., Litt, B.: Computational analysis in epilepsy neuroimaging: a survey of features and methods. Neuroimage **11**, 515–529 (2016)
11. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. arXiv preprint arXiv:1711.01558 (2017)
12. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. ArXiv e-prints, November 2017
13. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)
14. Mazziotta, J., et al.: A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos. Trans. Roy. Soc. Lond. B Biol. Sci. **356**(1412), 1293–1322 (2001)
15. Ashburner, J., Friston, K.: Unified segmentation. Neuroimage **26**, 839–851 (2005)
16. Caputo, B., Sim, K., Furesjo, F., Smola, A.: Appearance-based object recognition using SVMs: which kernel should I use? In: Proceedings of NIPS Workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision, vol. 2002. Whistler (2002)
17. Ahmed, B., Thesen, T., Blackmon, K.E., Kuzniekcy, R., Devinsky, O., Brodley, C.E.: Decrypting "cryptogenic" epilepsy: semi-supervised hierarchical conditional random fields for detecting cortical lesions in MRI-negative patients. J. Mach. Learn. Res. **17**(112), 1–30 (2016)
18. Thesen, T.: Detection of epileptogenic cortical malformations with surface-based MRI morphometry. PloS ONE **6**(2), 16430 (2011)
19. Hong, S.-J., Kim, H., Schrader, D., Bernasconi, N., Bernhardt, B.C., Bernasconi, A.: Automated detection of cortical dysplasia type II in MRI-negative epilepsy. Neurology **83**(1), 48–55 (2014)
20. Gill, R.S., et al.: Automated detection of epileptogenic cortical malformations using multimodal MRI. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 349–356. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_40