ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Latent representations of transient candidates from an astronomical image difference pipeline using Variational Autoencoders

Pablo Huijse[1,2], Nicolas Astorga[1] and Pablo Estévez[1,2] and Giuliano Pignata[3,2]

1- Universidad de Chile - Dept. of Electrical Engineering
Av. Tupper 2007, Santiago, Chile

2- Millennium Institute of Astrophysics
Av. Vicuña Mackenna 4860, Macul, Santiago, Chile

3- Universidad Andrés Bello - Dept. of Astrophysics
Fernández Concha 700, Las Condes, Santiago, Chile

Abstract. The Chilean Automatic Supernovae SEarch (CHASE) is a survey designed to detect early Supernovae. In this paper we explore deep autoencoders to obtain a compressed latent space for a large transient candidate database from the CHASE image difference pipeline. Compared to conventional methods, the latent variables obtained with variational autoencoders preserve more information and are more discriminative towards real astronomical transients.

1 Introduction

Supernovae (SNe) are the explosive phenomenon that massive stars undergo by the end of their life time. Some types of SNe are standard candles, *i.e.* they always reach the same peak luminosity. Because of this SNe are fundamental in the measurement of distances in our Universe and have been key for the latest advances in cosmology [1]. But detecting SNe is not an easy task as large areas of the sky need to be scanned repeatedly in order to find them.

Several surveys to hunt SNe have been proposed and among them is the Chilean Automatic Supernovae SEarch (CHASE) [2], a survey with the objective of detecting SNe from its early moments that has been running since 2007. CHASE uses the six 40-cm Pancromatic Robotic Optical Monitoring and Polarimetry Telescopes (PROMPT) at Cerro Tololo, Chile. CHASE images are reduced using a custom image-difference pipeline. In summary, every pair of images are (1) flat-field corrected, (2) astrometric-solution aligned, (3) zero-point calibrated and (4) subtracted. Pixels of this subtracted image with amplitudes larger than 5σ are selected as transient candidates and saved for manual inspection. Astronomers analyze a region of the sky around these candidates and discriminate if they correspond to astrophysical transients or artifacts. Fig. 1 shows examples of real and spurious transient candidates obtained by CHASE.

In this work we analyze 19 million CHASE candidates, exploring latent representations for which discrimination of real astronomical transients can be done in a more efficient way. We use novel techniques based on the Variational Autoencoder (VAE) and compare with classical methods for latent variable extraction.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

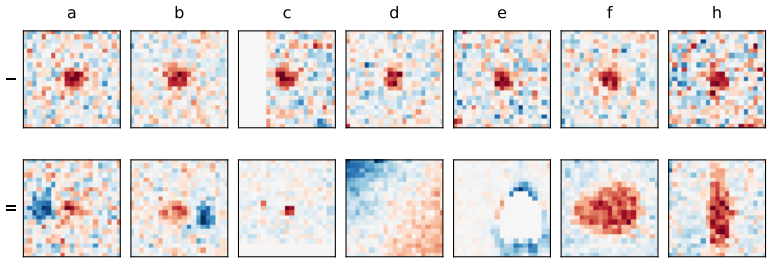

Fig. 1: Examples of CHASE candidates. The first row shows real astrophysical transients. The second row shows artifacts due to badly subtracted candidates (II.a and II.b), hot pixels (II.c), and defects in the CCD (II.e) among others.

2 Literature review

The recent advances in artificial neural networks have allowed for the development of image classifiers with super-human accuracy that can learn from millions of examples. In astronomy, convolutional neural networks (CNN) [3] have found great success in problems such as galaxy classification [4] and transient candidate discrimination [5]. On the other hand much less attention has been given to unsupervised deep architectures, e.g. autoencoders (AE) [6].

The Variational AE (VAE) [7] is a recent development that extends the conventional AE to include a continuous stochastic latent variable layer. Contrary to the AE, the decoder in the VAE is a generative model, *i.e.* data x mimicking the original distribution can be sampled from the latent code z. The probabilistic decoder and encoder are modeled as neural networks $p_\theta(x|z)$ and $q_\phi(z|x)$ with parameters θ and ϕ, respectively. The VAE is trained by maximizing

$$\mathcal{L}(\theta, \phi, x^{(i)}) = -\mathrm{D}_{KL}\left[q_\phi(z|x^{(i)})||p_\theta(z)\right] + \mathbb{E}_{q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}|z)\right], \quad (1)$$

where $D_{KL}[\cdot||\cdot]$ is the Kullback Leibler divergence. Eq. (1) is the variational or evidence lower bound (ELBO) of the likelihood of sample $x^{(i)}$ [7], and it can be interpreted as a regularization term plus reconstruction error. To make computations tractable a multivariate isotropic Gaussian distribution for the variational posterior is assumed, *i.e.* $q_\phi(z|x) = \mathcal{N}(z|\mu, \sigma^2 I)$. For the latent prior a standard Gaussian is assumed, *i.e.* $p_\theta(z) = \mathcal{N}(z|0, I)$. Then, reparametrizing $z = \mu + \sigma\varepsilon$, with $p(\varepsilon) = \mathcal{N}(\varepsilon|0, I)$, yields a closed-form expression for Eq. (1)

$$\mathcal{L}(\theta, \phi, x^{(i)}) \approx \frac{1}{2}\sum_{j=1}^{J}(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2) + \frac{1}{L}\sum_{k=1}^{L}\log p_\theta(x^{(i)}|z^{(i,k)}), \quad (2)$$

which can be optimized via gradient descent after sampling L times from $p(\varepsilon)$. The user chooses the distribution of the likelihood $p_\theta(x|z)$ depending on the problem. For continuous data an isotropic Gaussian is usually chosen which after applying the log yields the mean square error (MSE) loss (assuming $\sigma^2 = I$).

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

3 Methods

We study a subset of 19.050.609 CHASE transient candidates, corresponding to 44 observation nights. The candidates are represented as 21×21 pixel stamps centered around a pixel that varied above the 5σ threshold in the difference image. In all experiments we consider a latent dimensionality of 21, *i.e.* the square root of the input dimensionality. The stamps are scaled to the range $[-1, 1]$ by subtracting their median value and dividing by their absolute maximum. The CHASE subset is divided into training, validation and test sets of equal size. The autoencoder architectures that we test are described as

1. Sparse convolutional AE (SCAE): The encoder is a CNN with three convolutional layers (CL) and one fully-connected layer (FCL). Dimensionality is reduced using strided convolutions instead of pooling layers as this yields simpler networks with equivalent performance [8]. CLs have 32 filters with size 3×3 and stride two. The latent layer is a FCL with 21 neurons. The latent layer weights are regularized by minimizing their ℓ_1 norm ($\lambda = 0.005$). In this application regularizing the latent layer is critical to obtain meaningful coding. The decoder is a mirrored version of the encoder in which upsampling is performed using nearest neighbors interpolation. All layers use Rectified Linear Units (ReLU) activations except the latent layer and the decoder output (linear activation). We found that having linear activations in the latent layer performed better than ReLU, tanh or sigmoid functions in this setting.

2. Convolutional VAE (CVAE): Uses the same architecture as SCAE with three exceptions, (1) the encoder has two FCLs with linear activation connected to the last CL to represent μ and $\log \sigma^2$, respectively; (2) it has a sampling layer before the decoder that draws $\varepsilon \sim \mathcal{N}(0, I)$ and compute $z = \mu + \sigma\varepsilon$; and (3) ℓ_1 regularization of the latent layer is not used as VAE regularizes through the KL divergence. In this application annealing the regularization term in Eq. (2) is critical to avoid the trivial solution where the latent codes are ignored. This occurs because right after initialization the variational posterior carries little information on the data (reconstruction is poor) and the optimization exploits the regularization term (codes do not depart from to the prior) [9].

In both architectures: (a) The mean square error is used for the reconstruction loss, *i.e.* Gaussian outputs with unit variance are assumed for the decoder. In this case scaling the data to $[0, 1]$ and using cross-entropy loss, *i.e.* assuming a Bernoulli output, performs worse. (b) The networks are trained for 500 epochs with mini-batches of size 128. Early-stopping in the validation set is used to prevent overfitting. (c) Learning and momentum rates are set using Adam [10] with Nesterov momentum. (d) The tensorflow library is used to train the models. The implementation can be found at `github.com/phuijse/VAE_CHASE`.

For reference we perform latent variable extraction using incremental PCA [11] and online sparse dictionary learning (SDL) [12]. Note that due to the size of this dataset we are limited to methods that perform updates on mini-batches.

To evaluate the latent spaces we first build a training set of 315 asteroids by crossmatching the IAU's minor planet center catalog with the CHASE can-

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

didates. Given the characteristics of the CHASE survey the asteroids in the difference image are equivalent to the stellar transients we search. We find the candidates that are closest to the training set by measuring Euclidean distances in the latent space. Then for any given candidate we measure the discrimination normalized MSE (DNMSE), which is defined as the normalized mean squared error between the candidate and the asteroids in the input space.

We also evaluate the unsupervised-trained latent codes in a supervised setting using a subset of 20,000 labeled transients. The positive examples were obtained by augmenting the asteroids through rotations and simulating transients by inputting point-sources before the difference image procedure. The negative half was randomly sampled from the CHASE dataset. We train a MLP with one hidden layer over the latent codes using 15,000 objects and evaluate the classification performance on the remaining 5,000 objects.

4 Results

Fig. 2a shows the DNMSE of the test-set data. The lower the DNMSE the more similar the candidates are to the real transients (asteroids set). The DNMSE is computed for the top candidates in terms of their latent-space distance to the training set. This experiment shows that the candidates recovered by measuring distances in SCAE and CVAE latent codes are more similar to real transients than those recovered by linear methods. Ideally, if two candidates are similar in latent space they should also be similar in input space, *i.e.* neighbourhoods should be preserved, as this facilitates subsequent classification of the samples.

Fig. 2b shows four examples of transient candidate reconstructions using PCA and CVAE. First and second columns correspond to real transients. Third and fourth columns correspond to artifacts. This example shows that the artifacts reconstructed by PCA look similar to actual real transients, *i.e.* PCA misses the finer details that characterize them. A large fraction of the candidates recovered from the PCA/SDL latent codes correspond to poorly-reconstructed artifacts, which helps to explain the DNMSE difference observed in Fig. 2a.

Fig. 2c shows the ROC curves for the classification of the 5,000 labeled transient test subset. The best performance is obtained by the classifier trained over the CVAE, showing the potential of this latent space for transient classification.

Fig. 2d shows the evolution of the cost function of CVAE. The negative log-likelihood (NLL) and KL divergence correspond to the right-hand and left-hand side terms of Eq. (2), respectively. If annealing is not used the codes do not depart from the prior (zero divergence) and a trivial solution is obtained. Annealing prioritizes the NLL term on early training, allowing the decoder to learn good reconstructions and avoiding the collapse of the latent codes.

5 Conclusion and Future work

A procedure to obtain compressed latent representations for large databases of astronomical transient candidates using variational autoencoders has been pre-

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

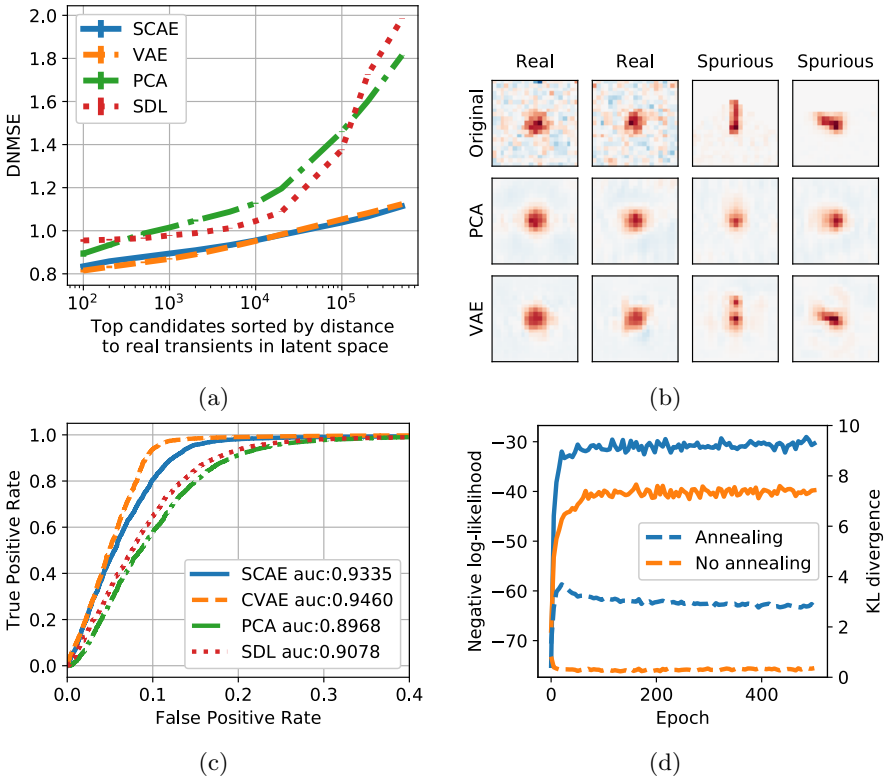

Fig. 2: (a) Normalized MSE between candidates and real transients in input space as a function of how similar they are in latent space. (b) Two real and two spurious transient candidates reconstructed with PCA and CVAE. (c) ROC curves of CVAE and other methods for the labeled transient test subset. (d) Evolution of the cost function of CVAE. When annealing is not used the KL divergence (dotted) collapses and the likelihood (solid) sets to a local optimum.

sented. The nonlinear latent features (a) retain more information from the data (better reconstruction), (b) preserve distance relations more faithfully between latent and input space and (c) allow for the training of more accurate transient classifiers, with respect to online linear methods. In the near-future we plan to (1) exploit the latent code uncertainties to improve the classification and (2) use the VAE generative model to characterize the transient behavior in the data.

In our experiments we noted that annealing the KL divergence was required in order to learn meaningful latent codes. This ad-hoc solution could be replaced with a more principled way to regularize the VAE. In the future we will test tighter bounds for the ELBO and also more flexible posteriors and priors. We also plan to test if semi-supervision can help in better guiding the optimization.

Previous results have shown that the images from which the difference was

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

obtained provide additional information that boost the transient discrimination capability significantly. We plan to extend the latent variable extraction procedure to tensors containing the difference and its two original components.

Acknowledgment

Pablo Huijse and Pablo A. Estévez acknowledge support from FONDECYT through grants 1170305 and 1171678, respectively. The authors acknowledge support from the Chilean Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC12009, awarded to The Millennium Institute of Astrophysics, MAS.

References

[1] F. Olivares E., M. Hamuy, G. Pignata, J. Maza, M. Bersten, M. M. Phillips, N. B. Suntzeff, A. V. Filippenko, N. I. Morrel, R. P. Kirshner, and T. Matheson. The Standardized Candle Method for Type II Plateau Supernovae. *The Astrophysical Journal*, 715(2):833–853, 2010.

[2] G. Pignata, J. Maza, R. Antezana, R. Cartier, G. Folatelli, F. Förster, L. Gonzalez, P. Gonzalez, M. Hamuy, D. Iturra, P. Lopez, S. Silva, B. Conuel, A. Crain, D. Foster, K. Ivarsen, A. Lacluyze, M. Nysewander, and D. Reichart. The CHilean Automatic Supernova sEarch (CHASE). In G. Giobbi, A. Tornambe, G. Raimondo, M. Limongi, L. A. Antonelli, N. Menci, and E. Brocato, editors, *American Institute of Physics Conference Series*, volume 1111 of *American Institute of Physics Conference Series*, pages 551–554, 2009.

[3] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–104. IEEE, 2004.

[4] S. Dieleman, K. W. Willett, and J. Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015.

[5] G. Cabrera-Vives, I. Reyes, F. Förster, P. A. Estévez, and J.C. Maureira. Deep-HiTS: Rotation Invariant Convolutional Neural Network for Transient Detection. *The Astrophysical Journal*, 836(1):97–104, 2017.

[6] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[7] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014.

[8] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *2nd International Conference on Learning Representations (workshop track)*, 2014.

[9] X. Chen, D. P Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

[10] D. P. Kingma and J. L. Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*, 2014.

[11] D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1):125–141, 2008.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. ACM, 2009.