Classification and Generation of Earth Observation Images using a Joint Energy-Based Model

Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Sébastien Lefèvre

CLASSIFICATION AND GENERATION OF EARTH OBSERVATION IMAGES USING A JOINT ENERGY-BASED MODEL

Javiera Castillo-Navarro[1,2], Bertrand Le Saux[3], Alexandre Boulch[4], Sébastien Lefèvre[2]

[1] ONERA, Université Paris-Saclay, F-91123 Palaiseau, France
[2] Université Bretagne Sud, IRISA UMR 6074, F-56000 Vannes, France
[3] European Space Agency, ESRIN Φ-lab, I-00044 Frascati (Rome), Italy
[4] valeo.ai, F-75008 Paris, France

ABSTRACT

Deep learning has changed unbelievably the processing of Earth Observation tasks such as land cover mapping or image registration. Yet, today new models are needed to push further the revolution and enable new possibilities. We propose a new framework for generative modelling of Earth Observation images. It learns an energy-based model to estimate the underlying distribution of the data while jointly training a deep neural network for classification. On the varied image types of the EuroSAT benchmark, we show this model obtains classification results on par with state-of-the-art and moreover allows us to tackle a wide range of high-potential applications: image synthesis, out-of-distribution testing for domain adaptation, and image completion or denoising.

Index Terms— Deep Learning, Energy-based Models, Generative Models, Earth Observation.

1. INTRODUCTION

The uptake of deep learning in Earth observation (EO) has been massive in the recent years and has revolutionized applications such as classification, segmentation, detection or change analysis [1, 2], enabling also for building or road extraction at global scale [3, 4]. It was made possible thanks to large datasets and well-defined tasks, i.e. settings adequate for discriminative learning of feedforward neural networks. Yet, there is now a need for addressing more complex tasks such as explaining decision-making processes or simulating complex scenarios with Earth observation data, e.g. to evaluate and mitigate the effects of climate change.

The opportunity lies in modelling the joint distribution of data and the various variables at stake rather than only a posteriori outputs. Such generative models include Generative Adversarial Networks (GANs) which have been widely used in the last years [5, 6] but have known issues such as being prone to mode collapse in the estimated distribution.

J. Castillo-Navarro's work is partially funded by a CNES grant.

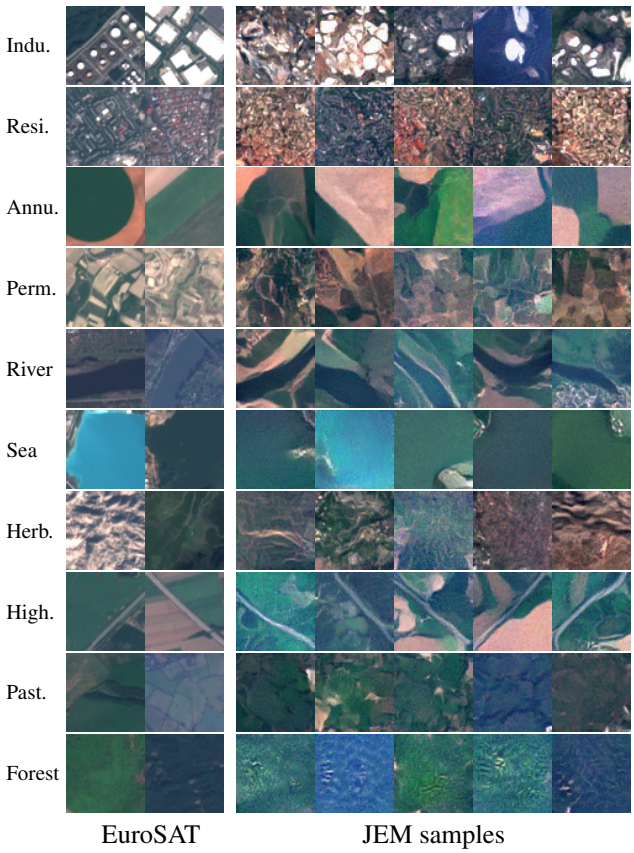

Fig. 1. Class-conditional samples generated by the model. First two columns contain real EuroSAT samples. Last five columns present JEM-generated samples.

Alternatively, we propose to use a Joint Energy-based Model (JEM) [7], which allows us to learn to classify and generate data at the same time. By plugging an energy function into a single classification neural network, we are able to generate images via Markov chain Monte Carlo sampling (as shown in Fig. 1). Furthermore, our probabilistic model can measure compatibility of new data with respect to the train data, enabling the possibility of out-of-distribution detection.

In this work, we establish the potential of joint energy-based models for classification and image generation in Earth observation. The key features of our approach are:

- Classification performances comparable with the state-of-the-art approaches;
- High quality image generation following the global distribution of the training data;
- Domain comparison using the energy function for reliable applicability on new data;
- Energy-based insight on model confidence when applied to unseen locations;
- EO image inpainting for incomplete data.

We present in Sec. 2 energy-based models and the procedure to train a joint classification-generation model. We then report experimental results for several applications in Sec. 3. Finally, conclusion and future works are discussed in Sec. 4.

2. ENERGY-BASED MODELS AND JEM

Energy-based models. Inspired from statistical physics, energy-based models [8] (EBMs) aim to capture dependencies between variables, $\mathbf{x} \in \mathcal{X}$, through a scalar function $E : \mathcal{X} \to \mathbb{R}$, referred as the *energy function*. Learning an EBM consists in finding an energy function that associates low energy values to correct configurations of variables, and higher energy values to incorrect configurations. Then, the energy can be considered as a measure of compatibility.

EBMs can easily be interpreted as probabilistic models using the Gibbs distribution, expressing the density $p(\mathbf{x})$ as:

$$p(\mathbf{x}) = \frac{\exp(-E(\mathbf{x}))}{Z}, \tag{1}$$

where $Z = \int_{\mathcal{X}} e^{-E(\mathbf{x})}$ is a normalization constant.

The advantage of training EBMs is that the energy value parameterizes all the information about inputs. This alleviates the burden of computing the normalization constant Z, which is often intractable. Moreover, this provides much more flexibility in the design of learning models.

Recently, EBMs have benefited from the expressive power of deep neural networks to model complex energy functions, with impressive results in generation, hybrid generation-classification and other applications [7, 9]. However, EBMs have been scarcely used in remote sensing [10], and have never been coupled with image generation in this context.

Joint energy-based models [7] can be used to extend a classic classifier architecture into an hybrid discriminative-generative model, by simply re-interpreting the outputs of the classification network. Let $f_\theta : \mathbb{R}^D \to \mathbb{R}^K$ be a classification neural network, with K the number of classes. The key idea of JEM is to express the joint distribution of images and labels as a joint energy-based model:

$$p_\theta(\mathbf{x}, y) = \frac{\exp(f_\theta(\mathbf{x})[y])}{Z_\theta} \tag{2}$$

The marginal distribution $p_\theta(\mathbf{x})$ can be obtained by:

$$p_\theta(\mathbf{x}) = \sum_{y=1}^{K} p_\theta(\mathbf{x}, y) = \frac{\sum_{y=1}^{K} \exp(f_\theta(\mathbf{x})[y])}{Z_\theta} \tag{3}$$

where $f_\theta(\mathbf{x})[y]$ is the y-th entry of $f_\theta(\mathbf{x})$.

From (3), one may observe that the distribution $p_\theta(\mathbf{x})$ is also an energy-based model, with the energy given by:

$$E_\theta(\mathbf{x}) = -\log\left(\sum_{y=1}^{K} \exp(f_\theta(\mathbf{x})[y])\right) \tag{4}$$

The model is then trained to maximize the joint log-likelihood, $\log p_\theta(\mathbf{x}, y)$, factorized as:

$$\log p_\theta(\mathbf{x}, y) = \log p_\theta(\mathbf{x}) + \log p_\theta(y|\mathbf{x}) \tag{5}$$

As shown below, (5) is the key to obtain an hybrid model.

Classification. The second term is related to $p_\theta(y|\mathbf{x})$, which written as $p_\theta(y|\mathbf{x}) = p_\theta(\mathbf{x}, y) \;/\; p_\theta(\mathbf{x})$ corresponds to the softmax output of a usual classifier. Thus it can be optimized using the cross-entropy loss, as a standard neural network.

Generation. The first term $\log p_\theta(\mathbf{x})$ corresponds to the generative part. It is trained as an energy-based model by approximating the gradient $\nabla_{\mathbf{x}} p_\theta(\mathbf{x})$ using a sampler based on Stochastic Gradient Langevin Dynamics (SGLD) [9] and thus, generates samples following:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \tfrac{\alpha}{2}\nabla_{\mathbf{x}} E_\theta(\mathbf{x}_i) + \varepsilon, \quad \mathbf{x}_0 \sim p_0(\mathbf{x}), \tag{6}$$

with $\varepsilon \sim \mathcal{N}(0, \alpha)$ and $p_0(\mathbf{x})$ usually a Uniform distribution.

3. EXPERIMENTS

We perform experiments using the EuroSAT Dataset [11] which comprises 64×64 patches from Sentinel-2 images, including scenes from 34 countries in Europe. Each image patch is labeled with one of 10 land cover/land use classes (e.g. industrial, residential, highway, pasture, forest, etc.). Classes are well-balanced, with 2,000 to 3,000 examples per class. In our experiments, we use the EuroSAT RGB version.

Implementation details. Following [7], we perform our experiments using a WideResNet-28-10 architecture [12], with no batch normalization. We train our networks with the Adam optimizer [13], during 200 epochs, following the JEM training scheme. Pytorch [14] is used for all implementations.

3.1. Hybrid Generative-Discriminative Results

As stated before, JEM, as a new training paradigm, allows us to train a standard classifier not only to classify images, but also to generate new ones.

Fig. 1 shows some class-conditional examples generated by the network after being trained on the EuroSAT dataset. First two columns present real samples from the dataset, while

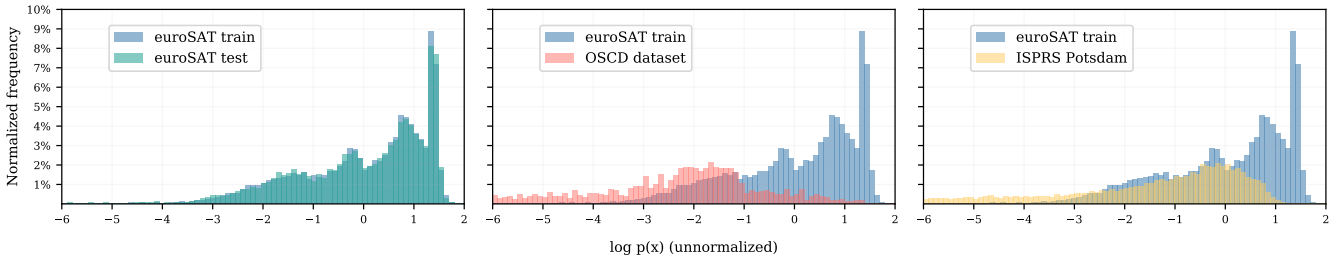

Fig. 2. Out-of-Distribution Detection using JEM. Out-of-distribution samples are assigned lower $\log p(\mathbf{x})$ values. Comparison between EuroSAT, OSCD and ISPRS Potsdam.

the five last columns show images generated by the model. Each row represents a class in the dataset. We observe that JEM-generated samples are akin to real EuroSAT samples, which is quantitatively supported by a KID score [15] of 0.06. Moreover, the model is capable to produce samples for every class on the dataset, with a large variety of images per class.

However, some classes remain challenging. For instance, forests (last row in Fig. 1) seem to be difficult to generate, maybe due to the lack of texture on forests patches. As a result, only 0.5% of generated samples correspond to forests, even though the training set is well-balanced. Industrial buildings (first row in Fig. 1) would require finer and more rectangular outlines to correctly match industrial buildings in the EuroSAT dataset. Conversely, generated samples for highways, rivers and various types of fields are remarkably similar to real images. This is a very good result because it means that the model is able to learn the true distribution behind the dataset and leads to compelling applications. Examples generated from the learnt distribution may be used for simulation or even for training new models. One could also use the learnt distribution for semi-supervised learning algorithms or in continuous learning applications.

We report in Table 1 classification results over EuroSAT. JEM results reach the same level of performances as previous methods. The slight discrepancy of the multi-task JEM with classification-only Wide-ResNet might be explained by the intrinsic regularization of the JEM model.

Table 1. Classification results comparison of standard discriminative networks vs. JEM on EuroSAT.

	Wide-ResNet	JEM	ResNet-50 [11]	GoogLeNet [11]
Acc.	98.3%	97.6%	98.6%	98.2%

3.2. Out-of-Distribution Testing

Out-of-distribution (OOD) detection is the task of identifying anomalous or significantly different examples from the training ones. This is an essential capacity to assert if the model is able to correctly classify new samples, especially in applications involving real-world decisions.

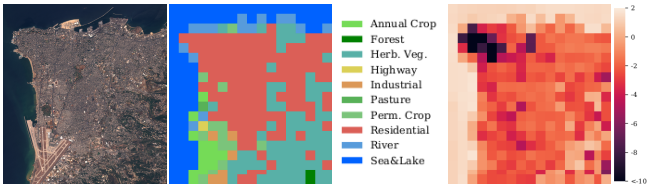

Fig. 3. Classification on a never-seen OSCD city (Beirut). From left to right: Image, classification map and confidence map (unnormalized $\log p(\mathbf{x})$).

We measure the capacity of the model to detect OOD samples by comparing in Fig. 2 the histograms of unnormalized log-likelihood values of the EuroSAT training set with different public EO datasets: OSCD [2] and ISPRS Potsdam [16]. Samples which match EuroSAT distribution should get higher values of $\log p(\mathbf{x})$. On the leftmost histogram, we observe no difference between EuroSAT training and test sets, while for OSCD and Potsdam datasets, the $\log p(\mathbf{x})$ can be extremely small compared to the EuroSAT train set. This is quantitatively confirmed by computing the Kullback-Leibler (KL) divergence with respect to the model trained on EuroSAT. Indeed, KL is only 0.2 for EuroSAT test data, while for OSCD and Potsdam values are 28.2 and 25.6, respectively: more information would be needed to represent these datasets which differ in terms of location or appearance.

3.3. Measuring the Confidence of the Classifier

Since our model is able to perform OOD detection, we can use the unnoormalized $\log p(\mathbf{x})$ value as a proxy for the confidence of its prediction. To illustrate, we apply EuroSAT-trained JEM to OSCD tiles. The tiles are split into 64×64 patches which go through the network to obtain the corresponding class and the estimated log-likelihood value per patch, leading to both classification and confidence maps.

We observe in Fig. 3 the results on a never-seen location from OSCD: Beirut. The segmentation map produced by the classifier is globally correct, however the model confidence, expressed as the model log-likelihood, varies. Indeed, low confidence happens on the most peculiar downtown districts, near the harbor and in Ras Beirut, which are areas the more likely to be different from training European cities.

Fig. 4. Image completion on EuroSAT dataset. Two up rows: inpainting, 12.5% information missing at the center. Two bottom rows: pixel defect correction, 10% salt and pepper noise.

3.4. Image Completion

The generative power of JEM can also be exploited to perform image completion. Fig. 4 shows some examples where the model is used to reconstruct missing pixels of an image, for tasks such as inpainting (missing regions) or restoration (missing pixels due e.g. to sensor defects).

4. DISCUSSION

We have introduced a new hybrid discriminative-generative framework applied to Earth observation data. The joint energy-based model leads to simultaneous classification and generation of images. Classification results are on par with state-of-the-art discriminative methods, while generated samples are, in general, of good quality and remarkably similar to real examples. We have also shown appealing remote sensing applications for this model: the capacity of detecting out-of-distribution samples to decide if the model can be reliably used in a new domain or use-case; the ability to classify unseen zones with a confidence map based on the OOD metric; and image completion or restoration of corrupted images.

However, large-scale deployment of JEM remains an open issue, mostly due to computation time of the Monte Carlo sampling. Yet, our promising results show how interesting JEM can be to benefit a wide range of high potential EO applications: simulation, domain adaptation, interpretability.

5. REFERENCES

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

[2] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation Using convolutional neural networks," in *IEEE IGARSS*, 2018.

[3] V. Mnih and G. Hinton, "Learning to Detect Roads in High-Resolution Aerial Images," in *ECCV*, 2010.

[4] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the United States," *IEEE JSTARS*, vol. 11, no. 8, pp. 2600–2614, 2018.

[5] N. Merkle, S. Auer, R. Müller, and P. Reinartz, "Exploring the potential of conditional adversarial networks for optical and SAR image matching," *IEEE JSTARS*, vol. 11, no. 6, pp. 1811–1820, 2018.

[6] N. Audebert, B. Le Saux, and S. Lefèvre, "Generative adversarial networks for realistic synthesis of hyperspectral samples," in *IEEE IGARSS*. IEEE, 2018.

[7] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," *ICLR*, 2020.

[8] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting structured data*, 2006.

[9] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," in *NeurIPS*, 2019.

[10] L. Mou, X. Zhu, M. Vakalopoulou, et al., "Multitemporal Very High Resolution from space: Outcome of the 2016 IEEE GRSS Data Fusion Contest," *IEEE JSTARS*, vol. 10, no. 8, pp. 3435–3447, 2017.

[11] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE JSTARS*, vol. 12, no. 7, pp. 2217–2226, 2019.

[12] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *BMVC*, 2016.

[13] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[14] A. Paszke, S. Gross, F. Massa, et al., "PyTorch: An imperative style, high-performance deep learning library," in *NeurIPS*. 2019.

[15] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton, "Demystifying MMD GANs," in *ICLR*, 2018.

[16] F. Rottensteiner, G. Sohn, et al., "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Annals*, vol. 1, pp. 293–298, 2012.