Energy-based Models for Earth Observation Applications

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

Energy-based Models for Earth Observation Applications

J. Castillo Navarro[1,2] B. Le Saux[3] A. Boulch[4] S. Lefèvre[2]

[1] ONERA, *Université Paris Saclay*, F-91761 Palaiseau, France.
javiera.castillo_navarro@onera.fr
[2] Université Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France.
[3] European Space Agency, ESRIN Φ-lab, I-00044 Frascati (Rome), Italy.
[4] valeo.ai, F-75008 Paris, France.

Résumé

Aujourd'hui, la disponibilité de larges bases de données de télédétection a rendu possible l'utilisation des techniques d'apprentissage profond pour l'observation de la Terre. Néanmoins, les modèles actuels sont conçus pour résoudre un problème à la fois: la classification ou la génération d'images. Nous présentons un nouveau modèle d'énergie qui estime la distribution jointe entre les images et leur annotation. Nous montrons que ce modèle a des performances de classification comparables à celles de l'état de l'art. De plus, il permet de s'attaquer à un large éventail d'applications: de la synthèse d'images à l'apprentissage semi-supervisé.

Mots Clef

Apprentissage profond, modèles d'énergie, modèles génératifs, observation de la Terre.

Abstract

The large amount of data, available thanks to the recent sensors, have made possible the use of deep learning for Earth Observation. Yet, actual approaches tend to tackle one problem at a time, e.g. classification or image generation. We propose a new framework for Earth Observation images processing which learns an energy-based model to estimate the underlying distribution, possibly estimated using non-annotated images. 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 to tackle a high range of high-potential applications, from image synthesis to high performance semi-supervised learning.

Keywords

Deep Learning, Energy-based Models, Generative Models, Earth Observation.

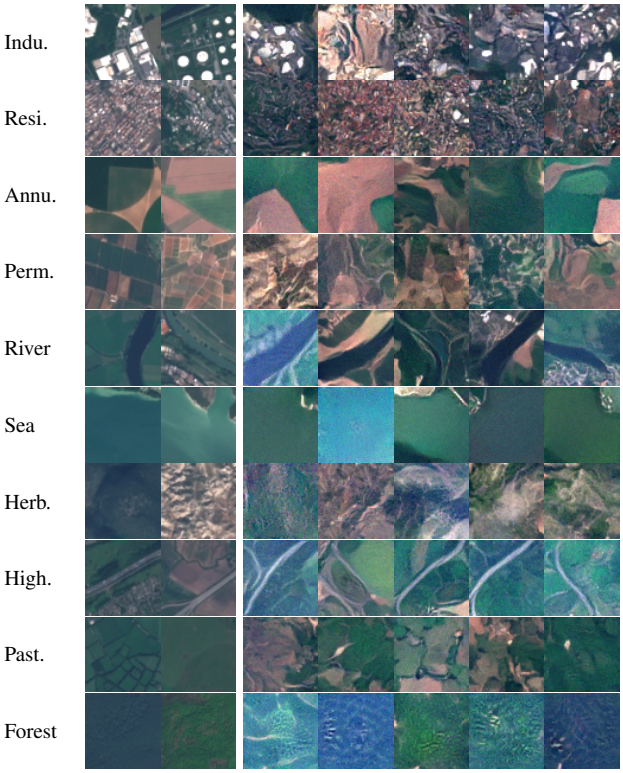

Figure 1: Class-conditional samples generated by the model. First two columns contain real EuroSAT samples. Last five columns present JEM-generated samples.

1 Introduction[1]

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 [3], [18], enabling also for building or road extraction at global scale [10], [16]. It was

[1] *Previous submission note*: This work has been presented at the International Conference on Learning Representations (ICLR), workshop on Energy-based Models, 2021.

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 developing models able to generalize well from few and scarce labeled data for global mapping of the Earth or explaining decision-making processes and 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 [1], [9] but have known issues such as being prone to mode collapse in the estimated distribution.

Alternatively, we propose to use a Joint Energy-based Model (JEM) [5], 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). Additionally, our probabilistic model can measure compatibility of new data with respect to the train data, enabling the possibility of out-of-distribution detection. Furthermore, this hybrid generative-discriminative model is particularly well-suited for semi-supervised learning.

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:

- High quality image generation following the global distribution of the training data;
- Classification performances comparable with the state-of-the-art approaches;
- Semi-supervised classification, even with very few labels;
- Domain comparison using the energy function for reliable applicability on new data;
- EO image inpainting for incomplete data.

The paper is organized as follows: Sec. 2 presents joint classification-generation models and we report experimental results for several applications in Sec. 3. Finally, we conclude 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 be interpreted as probabilistic models, expressing the density $p(\mathbf{x})$ as:

$$p(\mathbf{x}) = \frac{\exp(-E(\mathbf{x}))}{Z}, \text{with } Z = \int_{\mathcal{X}} e^{-E(\mathbf{x})}. \quad (1)$$

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 [4], [5], [15]. However, applications to remote sensing are scarce [11], and have never been coupled with image generation in this context.

The standard way of learning EBMs with deep learning today is by maximum likelihood training. Let p_θ be the probability density of an EBM, whose energy function, E_θ, is parameterized by a neural network of parameters θ. The density of the model, $p_\theta(\mathbf{x})$, can be fit to the distribution of data, $p_{\text{data}}(\mathbf{x})$, by maximizing the expected log-likelihood function over the data distribution:

$$\begin{aligned} \mathcal{L}_{\text{ML}}(\theta) :=& \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log p_\theta(\mathbf{x})] \\ =& \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[-E_\theta(\mathbf{x})] - \log Z_\theta \end{aligned} \quad (2)$$

The gradient of the log-likelihood can be expressed as:

$$\nabla_\theta \mathcal{L}_{\text{ML}}(\theta) = \mathbb{E}_{p_\theta(\tilde{\mathbf{x}})}[\nabla_\theta E_\theta(\tilde{\mathbf{x}})] - \mathbb{E}_{p_{\text{data}}(\mathbf{x})}[\nabla_\theta E_\theta(\mathbf{x})] \quad (3)$$

To compute the gradient expressed in Eq. (3) one needs to be able to sample from the model distribution p_θ, which is not possible. Current approaches approximate p_θ using MCMC methods, like Langevin dynamics [14]. This allows to approximately optimize the log-likelihood objective and generate samples from the model.

Joint energy-based models [5] 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 idea of JEM is to express the joint distribution of images and labels as a joint energy-based model:

$$p_\theta(\mathbf{x}, y) = \frac{\exp(f_\theta(\mathbf{x})[y])}{Z_\theta} \quad (4)$$

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

$$p_\theta(\mathbf{x}) = \sum_{y=1}^{K} p_\theta(\mathbf{x}, y) = \frac{\sum_{y=1}^{K} \exp(f_\theta(\mathbf{x})[y])}{Z_\theta} \quad (5)$$

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

From (5), one may observe that the distribution $p_\theta(\mathbf{x})$ is also an energy-based model, with the energy given by $E_\theta(\mathbf{x}) = -\log(\sum_{y=1}^{K} \exp(f_\theta(\mathbf{x})[y]))$. The model is then

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

$$\log p_\theta(\mathbf{x}, y) = \log p_\theta(\mathbf{x}) + \log p_\theta(y|\mathbf{x}) \tag{6}$$

As shown below, (6) 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) [4] and thus, generates samples following:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \tfrac{\alpha}{2}\nabla_{\mathbf{x}} E_\theta(\mathbf{x}_i) + \varepsilon, \quad \mathbf{x}_0 \sim p_0(\mathbf{x}), \tag{7}$$

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

Pipeline. In practice, given the usual classification neural network (before softmax) f_θ, the input image $\mathbf{x} \in \mathbb{R}^D$ passes through the network obtaining $f_\theta(\mathbf{x}) \in \mathbb{R}^K$. Then two branches are applied:

- Classification: the softmax function is applied to $f_\theta(\mathbf{x})$ to perform classification and compute the corresponding loss (usually, cross-entropy).
- Generation: we compute the energy $E_\theta(\mathbf{x}) = -\log(\sum_{y=1}^{K} \exp(f_\theta(\mathbf{x})[y]))$, generate samples using Eq. (7), and compute the loss corresponding to the maximum likelihood objective of EBMs in Eq. (2).

These two losses are combined to optimize the final objective in Eq. (6).

3 Experiments

We perform experiments using the EuroSAT Dataset [6] which comprises 64×64 patches from Sentinel-2 images over 34 countries in Europe. Each 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, 80% of which are used for training. We use the EuroSAT RGB version.

Implementation details Following [5], we perform our experiments using a WideResNet-28-10 architecture [17], with no batch normalization. We train our networks with the Adam optimizer [7], during 200 epochs, following the JEM training scheme. Pytorch [12] is used for all implementations.

3.1 Generation 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 trained on the EuroSAT dataset. First two columns present real samples from the dataset, while 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 [2] 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 appealing result means the model is able to learn the true distribution behind the dataset and leads to compelling applications. Generated examples may be used for simulation or even for training new models.

3.2 Classification results

Labeled samples/class	% of labels	Wide-ResNet	JEM
2000 on avg.	100%	**98.3%**	97.6%
100	$\sim$ 5%	**86.6%**	85.1%
20	$\sim$ 1%	61.5%	**67.7%**
10	$\sim$ 0.5%	50.3%	**59.5%**
5	$\sim$ 0.25%	40.0%	**49.8%**
1	$\sim$ 0.05%	26.8%	**37.5%**

Table 1: Classification results (Accuracy %) for JEM applied on EuroSAT. First row: classic JEM setting trained on the entire dataset. Following rows: models trained using a fraction of labeled samples. Gray cells indicate semi-supervised training, using the rest of the dataset as unlabeled samples.

We report in Table 1 classification results over EuroSAT for both fully supervised and semi-supervised settings. We also compare each result to the Wide-ResNet baseline, i.e., our backbone trained without energy modeling.

Full supervision. Presented on the first row of Table 1, JEM results reach the same level of performances as clasification-only Wide-ResNet. The slight discrepancy of the multi-task JEM might be explained by the intrinsic regularization of the JEM model.

Semi-supervision. The following rows on Table 1 are dedicated to semi-supervised learning. Here, we leverage the use of unlabeled data with (extremely) few labeled samples (from 5% to 0.05% of the EuroSAT training set). If labels are available, we optimize $\log p_\theta(\mathbf{x}, y)$ as in eq. (6), otherwise we marginalize it out and optimize $\log p_\theta(\mathbf{x})$ only. We observe that with only 5% of labeled data, the

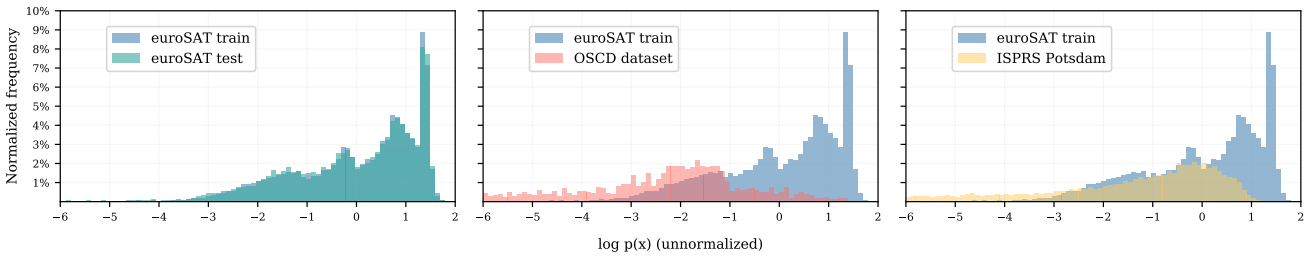

Figure 2: Out-of-Distribution Detection using JEM. Out-of-distribution samples are assigned lower $\log p(\mathbf{x})$ values. Comparison between EuroSAT, OSCD and ISPRS Potsdam.

semi-supervised JEM and supervised Wide-ResNet still reach the same level of performances. However, trained with extremely few labeled examples (1% of the original training set or less), the semi-supervised JEM model shows its potential. The gap of performance between JEM and Wide-ResNet gets bigger as labeled samples decrease in number, from 6.2% when trained on 1% of labeled samples to 10.8% of accuracy gap when trained on only 0.05% of labeled samples.

This result shows that: first, the energy function can be learned from very few annotated data, and, second, that the image distribution is well estimated such that conditional distribution $p_\theta(y|\mathbf{x})$ is easily estimated from a small set of annotated training samples.

3.3 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.

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 [3] and ISPRS Potsdam [13]. 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.

3.4 Measuring the Confidence of the Classifier

Since our model is able to perform OOD detection, we can use the unnormalized $\log p(\mathbf{x})$ 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.

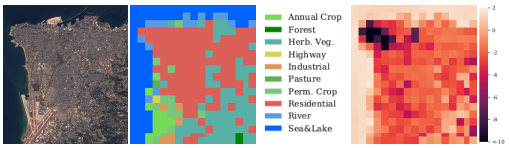

Figure 3: Classification on a never-seen OSCD city (Beirut). From left to right: Image, classification map and confidence map (unnormalized $\log p(\mathbf{x})$).

We 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.

3.5 Image Completion

The generative power of JEM can also be exploited to perform image completion. By using the iterative process described by Eq. (7), and the incomplete image as starting point $\mathbf{x_0}$, we can restore the missing pixels of an image. Fig. 4 shows some examples where the model is used 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 ability to learn the energy function from unlabeled data and thus boost classification results with respect to a model

Figure 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.

trained only with labeled data;the capacity of detecting out-of-distribution samples to decide if the model can be reliably used in a new domain or use-case; 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.

References

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

[2] M. Bińkowski *et al.*, "Demystifying MMD GANs," in *ICLR*, 2018.

[3] R. Caye Daudt *et al.*, "Urban change detection for multispectral Earth observation Using convolutional neural networks," in *IEEE IGARSS*, (Valencia, Spain), 2018.

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

[5] W. Grathwohl *et al.*, "Your classifier is secretly an energy based model and you should treat it like one," *ICLR*, 2020.

[6] P. Helber *et al.*, "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.

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

[8] Y. LeCun *et al.*, "A tutorial on energy-based learning," *Predicting structured data*, 2006.

[9] N. Merkle *et al.*, "Exploring the potential of conditional adversarial networks for optical and SAR image matching," *IEEE JSTARS*, vol. 11, no. 6, pp. 1811–1820, 2018.

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

[11] 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.

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

[13] 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.

[14] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *ICML*, 2011.

[15] J. Xie *et al.*, "A theory of generative convnet," in *ICML*, 2016.

[16] H. L. Yang *et al.*, "Building extraction at scale using convolutional neural network: Mapping of the United States," *IEEE JSTARS*, vol. 11, no. 8, pp. 2600–2614, 2018.

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

[18] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE GRSM*, vol. 5, no. 4, pp. 8–36, 2017.