Title: Semantic-Fusion GANs for Semi-Supervised Satellite Image Classification
Authors: Subhankar Roy, Enver Sangineto, Begüm Demir and Nicu Sebe

SEMANTIC-FUSION GANS FOR SEMI-SUPERVISED SATELLITE IMAGE CLASSIFICATION

Subhankar Roy[1], Enver Sangineto[1], Begüm Demir[2] and Nicu Sebe[1]

[1]Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy
[2]Faculty of Electrical Engineering and Computer Science, TU Berlin, Berlin, Germany

ABSTRACT

Most of the public satellite image datasets contain only a small number of annotated images. The lack of a sufficient quantity of labeled data for training is a bottleneck for the use of modern deep-learning based classification approaches in this domain. In this paper we propose a semi-supervised approach to deal with this problem. We use the discriminator (D) of a Generative Adversarial Network (GAN) as the final classifier, and we train D using both labeled and unlabeled data. The main novelty we introduce is the representation of the visual information fed to D by means of two different channels: the original image and its "semantic" representation, the latter being obtained by means of an external network trained on ImageNet. The two channels are fused in D and jointly used to classify fake images, real labeled and real unlabeled images. We show that using only 100 labeled images, the proposed approach achieves an accuracy close to 69% and a significant improvement with respect to other GAN-based semi-supervised methods. Although we have tested our approach only on satellite images, we do not use any domain-specific knowledge. Thus, our method can be applied to other semi-supervised domains.

Index Terms— semi-supervised learning, generative adversarial networks, satellite image classification

1. INTRODUCTION

One of the reasons for which satellite image classification is challenging is due to the lack of large annotated training datasets which has prevented so far the systematic adoption of modern deep-learning based approaches in this field. Common deep-learning methods (e.g., ResNets [1]) achieve a high image classification accuracy when trained in a *supervised* regime with plenty of annotated data [2]. However, despite very recently a few satellite datasets have been publicly released which contain thousands of images, most of the current application scenarios in this field are based on training datasets of only a few hundreds of labeled images.

On the other hand, recent trends in deep learning research have shown the possibility to use a *semi-supervised* training regime for training deep networks. For instance, Salimans et al. [3] showed that Generative Adversarial Networks (GANs) [4] can be used to boost the accuracy of a classifier using semi-supervised data. The main idea is that the classifier corresponds to the discriminator D of a GAN, trained together with a generator G. However, different from a standard GAN, where D is asked to discriminate between "real" and "fake" images (the latter being produced by G), in the semi-supervised framework proposed in [3], D is also asked to predict the correct class of those subset of images which are associated with labels. Intuitively, the gain comes from the exploitation of the additional unlabeled images, from which D needs to extract dataset-specific visual information which allows it to discriminate these images from the fake ones.

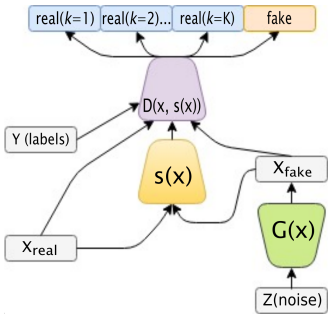

Fig. 1. SF-GAN overview: The generator G produces fake images by sampling from the noise distribution p_z. The discriminator D has access to X_{real} (containing both labeled and unlabeled images) and X_{fake}, as well as their semantic representation $s(\cdot)$, obtained using a pre-trained deep network. D outputs a probability distribution over $K+1$ classes where the first K classes are real and the final class is $fake$.

In this paper we build on this idea of adding semantics. Specifically, we exploit an external network, trained on ImageNet (which contains no satellite-image), to extract generic visual information from our domain-specific images. We feed the satellite images to the Inception Net [5] and we extract a high-level representation of these images using the activation values of its last convolutional layer. Then we *fuse* this representation with an analogous representation obtained in the last convolutional layer of D. In this way, the decision

of D depends (also) on generic visual semantics, extracted by means of the Inception Net, where the latter leverages the large dataset (ImageNet) it has been trained on (see Fig. 1). We call this approach Semantic Fusion GAN (SF-GAN) and we empirically show that SF-GAN achieves a large accuracy boost with respect to both "standard" supervised-trained deep networks and semi-supervised GANs, especially when the cardinality of the labeled training subset is very small.

2. RELATED WORK

Semi-supervised learning has been largely addressed in the past years using kernel-based methods. For instance, Chang et. al. [6] extend Locality-Constrained Linear Coding (LLC) [7] to a semi-supervised scenario where a kernelized LLC is used to learn the underlying data manifold, given only a subset of labeled images. Blanchart et al. [8] use SVMs in a semi-supervised setting for satellite image classification.

More recently, Salimans et al. [3] showed that the combination of a supervised and a semi-supervised loss in a GAN framework helps in boosting the target classification problem (more details in Sec. 3). Springenberg et al. [9] extend this idea combining the optimization of the Shannon entropy as the adversarial objective with minimizing the cross-entropy loss for the labeled samples. The *feature matching loss*, introduced in [3], which compares real and fake images using an intermediate layer of the discriminator, is extended in [10] (*perceptual loss*) using the feature space of a layer of an externally-trained network. We also use an externally-trained network to inject "semantics" in our framework. However, while the perceptual loss in [10] can be used only for *conditional* GANs, in which the generator's outcome depends on a real input image, our SF-GAN operates in an *unconditional* regime. Moreover, differently from [10], the external network in our case is not used as an auxiliary loss function but for providing semantic information to aid the discriminator decision.

Semi-supervised classification using GANs is also proposed in [11] where the discriminator outputs a multi-class probability distribution. Unsupervised and fully-supervised learning are combined in [12] in a two-stage approach. In the first stage, unlabeled data are used in the GAN setting to train the discriminatror D. Once fully trained, D is used as a feature extractor to obtain a representation of the labeled samples. In the second stage these representations are used to train an SVM classifier in a standard supervised framework.

3. PROBLEM SETTING

In this section we review the standard GAN [4] and the semi-supervised GAN approach [3] and we introduce our notation. Our proposed SF-GAN is presented in the next section.

Let $X = \{x_1, ..., x_N\}$ be the set of training images which are partly associated with class labels. Specifically, $X_l = \{x_1, ..., x_M\}$ is the subset of images associated with labels, respectively collected in $Y = \{y_1, ..., y_M\}$, $y_i \in \{1, ..., K\}$. On the other hand, $X_u = \{x_{M+1}, ..., x_N\}$ is the subset of unlabeled images, where typically $M << N$. The goal of a semi-supervised approach is to train a classifier simultaneously exploiting both (X_l, Y) and X_u.

The standard GAN framework consists of two antagonistic networks: a generator G and a discriminator D. G takes as input a noise vector, randomly generated using an a-priori distribution ($z \sim p_z$) and deterministically generates a fake image $\hat{x} = G(z; \theta_G)$, typically using an up-convolutional neural network [12], where θ_G are the parameters of G. On the other hand, D takes as input an image, which is either real, x, or fake, $\hat{x}$. The outcome of D is a binary classification probability of the input image being extracted from the real dataset or produced by G, which can be denoted as $p_D(x) = D(x; \theta_D)$, θ_D being the parameters of D. The goal of D is to assign a high probability to $x \sim p_{data}$ and a low probability to $\hat{x} = G(z)$, $z \sim p_z$. On the other hand, G aims to maximize the probability of the fake images being classified as real without having access to the real data. The overall GAN objective function can be written as follows:

$$\min_G \max_D \begin{array}{c} \mathbb{E}_{x \sim p_{data}(x)} [\log(D(x))] \\ + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \end{array} \quad (1)$$

Salimans et al. [3] extend the above framework to deal with semi-supervised learning by adding K final neurons to D, one per target class. The outcome of D is now a multi-class prediction represented by a $K + 1$ dimensional logit output which comprises of K real classes and a $(K + 1)$-th class representing the fake images. The loss function of D is consequently split into a supervised and an unsupervised loss: $\mathcal{L}_D = \mathcal{L}_{sup} + \mathcal{L}_{unsup}$, where:

$$\mathcal{L}_{sup} = - \mathbb{E}_{x,y \sim p_{data}(x,y)} [\log(p_D(y|x, y < K + 1))] \quad (2)$$

and

$$\mathcal{L}_{unsup} = \begin{array}{l} - \mathbb{E}_{x \sim p_{data}(x)} [\log(1 - p_D(y = K + 1|x))] \\ - \mathbb{E}_{z \sim p_z(z)} [\log(p_D(y = K + 1|G(z)))] \end{array} \quad (3)$$

The loss function of G remains unchanged. In the next section we show how to modify the posterior probabilities computed by D (i.e., $p_D(x)$) in order to embed visual semantics extracted from a generic, external network.

4. PROPOSED SF-GAN

The main idea behind SF-GAN is to enrich the image representation fed to D using generic visual semantics extracted by means of an external network, trained on a generic, large and fully-supervised dataset (ImageNet). Specifically, let $s(x)$ be the vector of the activation values of the last convolutional layer ($Mixed\_7c$) of the Inception Net [5] when input with image x. We write $D(x, s(x))$ to highlight the dependence of

D from both the original image x and its semantic representation $s(x)$ (see below for details). The posterior probability of class k is computed using:

$$p_D(y = k|x, s(x)) = \frac{e^{D_k(x,s(x))}}{\sum_{k'=1}^{K} e^{D_{k'}(x,s(x))}}, \quad (4)$$

where $D_k(\cdot, \cdot)$ is the score assigned to class k by D. Using Eq. 4 to compute $p_D()$ in Eq. 2-3 we obtain our discriminator loss. For training G, we use a standard generator loss with the addition of the feature matching loss (see Sec. 2).

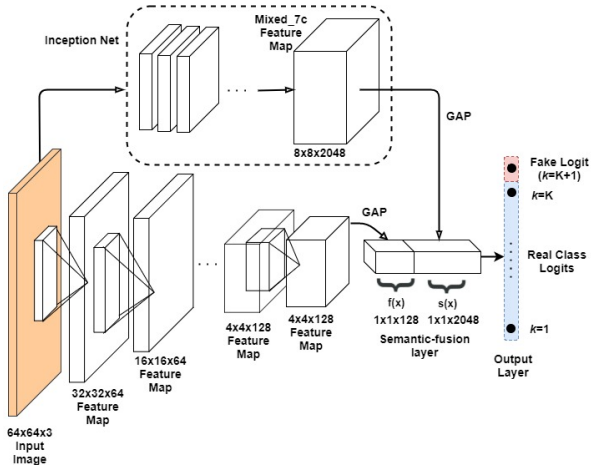

Fig. 2. The proposed SF-GAN discriminator D takes as input both a $64 \times 64 \times 3$ RGB image x and its semantic representation $s(x)$ and outputs a *K*+1 logit. The vector $s(x)$ is fused with $f(x)$, the internal representation of x, in the penultimate layer of D.

4.1. THE DISCRIMINATOR ARCHITECTURE

As shown in Fig. 2, D takes as input an RGB image (either real or fake), of spatial dimension 64×64. This input is passed through a sequence of convolutional layers, batch normalizations and Leaky ReLU non-linearities, finally producing a $4 \times 4 \times 128$ tensor, where 4×4 is the spatial resolution and 128 is the number of feature maps. We extract a representation $f(x)$ from this tensor using *Global Average Pooling* (GAP) [13]. GAP averages the information content of the feature maps spatially, each map being averaged independently of the others. In our case, the content of each feature map is averaged over the 4×4 spatial grid to produce a single scalar value. $f(x)$ is the concatenation of all the 128 average values and is further concatenated with $s(x)$. The latter is obtained by feeding a pre-trained Inception Net with x. Using the last convolutional layer of the Inception Net we obtain a representation of x as a tensor of dimension $8 \times 8 \times 2048$. Similarly to D, we apply GAP to this second tensor to get a 2048-dimensional feature vector $s(x)$. After fusion, $[f(x), s(x)]$ is processed by a final fully-connected layer which outputs the $(K + 1)$-dimensional logit.

Generator		Discriminator	
Layer	Configuration	Layer	Configuration
FC_1	2048	Conv_1	filter: 64x[3,3,3]; stride: 2
UpConv_1	filter: 64x[5,5,128]; stride: 0.5	Conv_2	filter: 64x[3,3,64]; stride: 2
UpConv_2	filter: 32x[5,5,64]; stride: 0.5	Conv_3	filter: 64x[3,3,64]; stride: 2
UpConv_3	filter: 32x[5,5,32]; stride: 0.5	Conv_4	filter: 64x[3,3,64]; stride: 2
UpConv_4	filter: 3x[5,5,32]; stride: 0.5	Conv_5	filter: 128x[3,3,64]; stride: 1
-	-	Conv_6	filter: 128x[3,3,128]; stride: 1
-	-	Conv_7	filter: 128x[3,3,128]; stride: 1
-	-	Avg_pool_7	pool: 4x4
-	-	FC_8	2176 (=128+2048)

Table 1. Details of G and D. The filter configuration is described as: number of filters x [*height, width, input channels*].

4.2. IMPLEMENTATION DETAILS

Since the number of labeled images is usually small, we use *dropout* [14] in the discriminator network to help regularizing the learning process. We do not use batch normalization in the intermediate layer (Conv_7) utilized for computing the feature matching loss. This is done in order to make the mean of the intermediate features of the real data different from the generated samples.

The generator G is a standard DCGAN [12] network composed of a sequence of up-convolutional layers with fractional stride, each layer except the last being followed by a batch normalization layer and a Leaky ReLU non-linearity. Table 1 shows the architectural details of both G and D.

5. EXPERIMENTAL RESULTS

In our experiments we use the recently published EuroSAT dataset [15], composed of 27,000 annotated satellite images acquired by the Sentinel-2 satellite and grouped into 10 different land-use categories where each image belongs to a single category (e.g., "Industrial", "Residential", etc.). Each image consists of 13 bands, however, in our experiments we have considered RGB bands only as in [15]. The image spatial resolution is 64×64. Following the protocol in [15], we use 21,600 images for training. Moreover, we further split the remaining 5,400 images in 4,860 samples used for testing and 540 images used for validation.

Note that this dataset is much larger than common public satellite image datasets, and we chose EuroSAT in order to show results obtained varying the amount of labels accessible during the training process. Specifically, we simulate a scenario in which *we have access to only a limited amount of labeled data* M ($M = |X_l|$, see Sec. 3), varying M between 100 and 21,600. For a fixed value of M, the remaining train-

Method	Training regime	# of labels M (% over the full training set)			
		100 (0.46)	1000 (4.6)	2000 (9.25)	21,600 (100)
CNN (from scratch)	Supervised	29.3	46.1	59.0	83.2
Inception Net [5] (fine tuned)	Supervised	63.9	84.6	87.9	91.5
SS-GAN [3]	Semi-supervised	63.0	75.8	78.3	86.9
Proposed SF-GAN	Semi-Supervised	**68.6**	**86.1**	**89.0**	**93.2**

Table 2. Classification accuracy (%) on the EuroSAT test set.

ing data are used without labels (X_u). We train our SF-GAN[1] using Adam with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ and a batch size of 128. G and D are trained for 30 epochs and in every epoch the learning rate is shrank by a factor of 0.9 starting from an initial value of $3 * 10^{-4}$.

We compare the classification accuracy of SF-GAN with: 1) A Convolutional Neural Network (CNN) trained from scratch, with a network capacity similar to the SF-GAN's discriminator network capacity; 2) Inception Net [5] with its final layer fine tuned on EuroSAT; and 3) The Semi-Supervised GAN (SS-GAN) approach proposed in [3]. Note that, to the best of our knowledge, no other semi-supervised method has been tested on EuroSAT yet. The results reported in Table 2 show that, as expected, when $M = 100$, the CNN trained from scratch performs very poorly. Note that, being the CNN trained in a fully-supervised fashion, it cannot use X_u. The same situation applies to the fine-tuning of the Inception Net. Conversely, with the same number of labeled images, $M = 100$, the proposed SF-GAN surpasses all the other classification methods including SS-GAN [3]. As we increase the number of labels M, the accuracy increases monotonically for every method. For instance, at $M = 2,000$, the accuracy of the fine-tuned Inception Net comes pretty close to our method. However, when compared to [3], our method is still 10.7% better. Interestingly, SF-GAN achieves a higher accuracy with respect to Inception Net even when all the training data are associated with their corresponding labels. This is likely due to the fact that the discriminator D in SF-GAN has access to X_{fake} (see Fig. 1), an additional source of information which is not available to the Inception Net, and needs to additionally discriminate fake images from real ones.

Finally, in our experiments we observed that SF-GAN reaches a faster convergence with less number of epochs when compared with SS-GAN [3]. As shown in Fig. 3, the accuracy on the validation set of our SF-GAN converges after epoch 9, whereas SS-GAN is still rising even after the 15-th epoch. Note that the Inception Net needs 200 epochs to converge; however, being only the last layer involved in the fine-tuning process, its overall training time is shorter. Both the faster convergence and the higher final accuracy results of SF-GAN with respect to SS-GAN show that the injection of semantic information into D helps the discriminator (and, consequently, also the generator) to quickly learn the underlying real data distribution.

[1]Code is available at `https://github.com/MLEnthusiast/SFGAN`

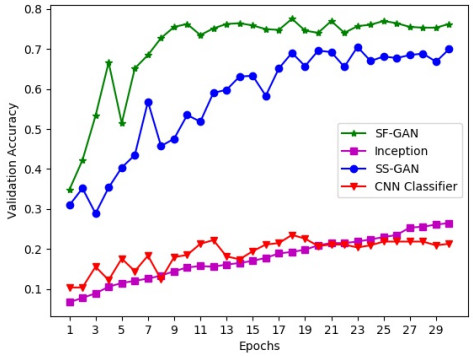

Fig. 3. Accuracy on the validation set over different training epochs of the tested methods when $M = 100$.

6. CONCLUSIONS

In this paper we proposed SF-GAN, a semi-supervised classification approach based on a GAN framework, for satellite image classification with scarcity of annotated data. The SF-GAN discriminator fuses the high-level representation of an image, obtained using a pre-trained, external deep network, with the image representation of the standard DCGAN discriminator. Experimental results show that the proposed architecture: 1) achieves a significantly higher overall accuracy when compared with other semi-supervised and fully-supervised classification methods, especially in a scenario in which only a few images are annotated; 2) achieves a faster convergence while training.

Even if the proposed method has been tested with satellite images, no domain-specific constraint or a-priori knowledge is used in our approach. Consequently, we believe that SF-GANs can be easily adopted in other semi-supervised image classification tasks.

Acknowledgements: This work was supported by the European Research Council under the ERC Starting Grant BigEarth-759764. We also want to thank the NVIDIA Corporation for the donation of the GPUs used in this project.

7. REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[6] Y.-J. Chang and T. Chen, "Semi-supervised learning with kernel locality-constrained linear coding," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 2977–2980.

[7] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[8] P. Blanchart and M. Datcu, "A semi-supervised algorithm for auto-annotation and unknown structures discovery in satellite image databases," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 3, no. 4, pp. 698–717, 2010.

[9] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.

[10] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[11] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[13] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[14] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *arXiv preprint arXiv:1709.00029*, 2017.