An Adversarial Regularisation for Semi-Supervised Training of Structured Output Neural Networks

Mateusz Koziński[2] **Loïc Simon**[1] **Frédéric Jurie**[1]

[1]Groupe de recherche en Informatique, Image, Automatique et Instrumentation de
Caen Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
[2]CVLab, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

mateusz.kozinski@epfl.ch loic.simon@ensicaen.fr frederic.jurie@unicaen.fr

Abstract

We propose a method for semi-supervised training of structured-output neural networks. Inspired by the framework of Generative Adversarial Networks (GAN), we train a discriminator to capture the notion of a 'quality' of network output. To this end, we leverage the qualitative difference between outputs obtained on labelled training data and unannotated data. The discriminator serves as a source of error signal for unlabelled data. Initial experiments in image segmentation demonstrate that including unlabelled data with the proposed loss function into the training procedure enables attaining the same network performance as in a fully supervised scenario, while using two times less annotations.

1 Introduction

We propose an approach to semi-supervised training of structured output neural networks that enables saving significant labelling effort. We show that the performance of a network trained in a fully supervised regime on a certain amount of labelled data can be matched by training on a significantly smaller amount of labelled data, combined with a volume of unlabelled data in the proposed semi-supervised setting.

Our technical contribution consists in generating a useful error signal for unlabelled data by means of adversarial training. During training, both the labelled and the unlabelled data is forwarded through the network. The network produces qualitatively better output on the labelled images than on the unlabelled ones. Much like in GAN training, we train a discriminator network to capture this difference. The negative gradient of the discriminator with respect to its input is used as the error signal for the unlabelled data. Contrary to pre-training, our method can be applied to any structured output problem, and enables using unlabelled data to train the complete network, end-to-end, independently of its architecture.

2 Related work

The most common methods of handling limited availability of training data include training a neural network on an auxiliary task, for which large volume of data is available, and fine tuning it on a small amount of application-specific data. In self-supervised methods the auxiliary task consists in applying a perturbation to an image, like masking image regions [14], or shuffling image tiles [3, 13], and training the network to recover the original image. In case of an autoencoder [6, 11, 20, 16, 22, 21], the auxiliary task is to encode an image into a strongly regularized latent representation, and reconstruct the original image from the latter. From the perspective of structured prediction, the common

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

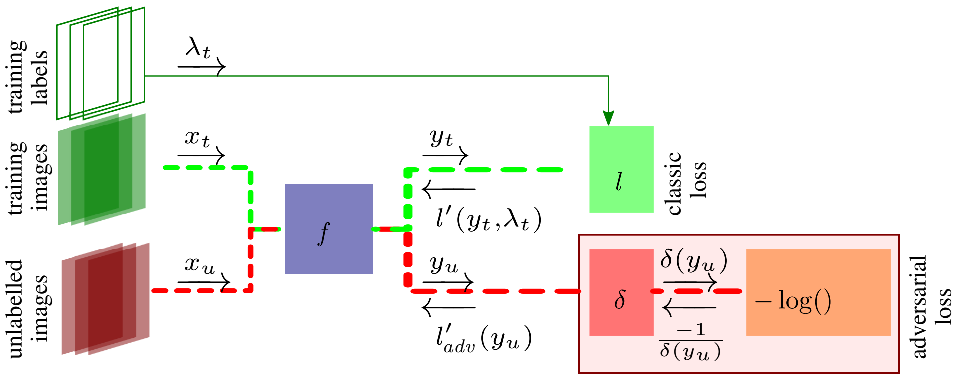

Figure 1: The flow of data and error signals when training a structured output network f with the proposed method, presented in algorithm 1. The discriminator update is not shown in the drawing. The green line denotes the flow of labelled training data and the corresponding gradients. The red line denotes the flow of unlabelled data and the corresponding gradients.

drawback of the above methods is the constraint on the network architecture. For example, only the first half of a typical segmentation network [12, 2] can be matched to an encoder of an autoencoder. In consequence, the remaining layers do not benefit from the unlabelled data. Moreover, structured output problems are characterized by correlations between output variables, and these cannot be learned by a method that is never exposed to data from the output domain. Our method is free from these limitations, enabling end-to-end training of structured-prediction networks.

Our work is inspired by the Generative Adversarial Networks (GANs) [5]. In GAN, a generator network is trained to transform a random vector drawn from a simple sampling distribution to a sample from a complicated target distribution. The flagship application is to train the generator to yield realistically looking images. The key property of GANs is that all that is required for training the generator is a collection of samples from the target distribution. The error signal is backpropagated to the generator from a discriminator network, trained to differentiate samples from the target distribution from ones output by the generator. Previous applications of GANs to semi-supervised learning focus on re-using the discriminator for feature extraction [15], generating additional training images [18], or regularizing the predictions on the generated data to have low confidence [23]. We propose to use the *gradient* of the discriminator as opposed to its outputs or weights.

GANs can also be used for mapping between two domains of interest [7]. In this case discrimination is performed between pairs of input and output items. This type of loss has been demonstrated to boost segmentation results when combined with a standard cost function [9]. In contrast to these supervised methods, we use a discriminator to generate a training signal for *unlabelled* data.

For domain adaptation [4, 8], the discriminator is trained to differentiate between features obtained for samples from two different domains, like synthetic and real images. It is then a source of an error signal, that makes the network invariant to the inter-domain shift. In contrast, in our method, the discriminator regularizes the network with use of unlabelled data from the same domain.

3 Method description

We address the problem of training a structured output network f_w, parametrised with a weight vector w, to produce predictions y on input data x. In our scenario, in addition to the training examples x_t, $t \in \mathcal{T}$, with annotations λ_t, a volume of unlabelled examples x_u, for $u \in \mathcal{U}$, is available. To handle the unlabelled data, we combine a classic supervised loss $l(y_t, \lambda_t)$, measuring the consistency of $y_t = f(x_t)$ and λ_t, with a novel and unsupervised loss $l_{adv}(y_u)$

$$C_{tot}(w) = C(w) + \alpha C_{adv}(w) = \mathbb{E}[l(f_w(x_t), \lambda_t)] + \alpha \mathbb{E}[l_{adv}(f_w(x_u))]\,, \tag{1}$$

where α is a constant. Training consists in determining the optimal network parameter by solving

$$w^* = \arg\min_w C_{tot}(w). \tag{2}$$

We describe the unsupervised cost C_{adv} in section 3.1 and the training algorithm in section 3.2.

3.1 Adversarial loss

Training a network on the labelled data (x_t, λ_t), $t \in \mathcal{T}$ results in a qualitative difference between outputs $y_t = f_w(x_t)$ and outputs produced for the unseen data $y_u = f_w(x_u)$, $u \in \mathcal{U}$. Ideally, x_t and x_u are identically distributed, so one might think the same holds for $f_w(x_t)$ and $f_w(x_u)$. In practice, the dimensionality of x is typically large and the training set is not representative of the variability of the unseen data. This biases f_w to perform better on the training examples. We leverage this difference to define the unsupervised cost C_{adv} as a regularisation term that tends to close this gap.

Inspired by GANs, we propose to train a discriminator network δ_v, parametrised by v, to capture the qualitative difference between y_t for $t \in \mathcal{T}$, and y_u, for $u \in \mathcal{U}$. We interpret the discriminator output $\delta_v(y)$ as the likelihood that y has been obtained from an element of the labelled training set, and we interpret $1 - \delta_v(y)$ as the likelihood of y originating from the unlabelled set. The optimal parameter of the discriminator $v^* = \arg\min_v CE_{disc}(v)$ is defined in terms of the cross-entropy

$$CE_{disc}(v) = -\mathbb{E}_{t\in\mathcal{T}}[\log(\delta_v(y_t))] - \mathbb{E}_{u\in\mathcal{U}}[\log(1 - \delta_v(y_u))]. \tag{3}$$

The negative logarithm of the output of the optimal discriminator can be used as a 'quality measure' for image segmentations. Indeed, $\delta_{v^*}(y)$ is the likelihood that y originates from the training set, and the outputs on the training set are qualitatively better. We therefore define the unsupervised cost as

$$C_{adv}(w) = \mathbb{E}_{u\in\mathcal{U}}[-\log(\delta_{v^*}(f_w(x_u)))]. \tag{4}$$

Minimising (4) with respect to w drives f_w towards reducing the gap between performance on labelled and unlabelled data.

3.2 Algorithm

The minimization can be performed with a gradient-based optimization routine, for example SGD. The gradient of the objective consists of two components and its estimate on a training batch T and unlabelled batch U can be denoted as

$$\nabla_w C_{tot}^{TU}(w) = \nabla_w C^T(w) + \alpha \nabla_w C_{adv}^U(w). \tag{5}$$

Likewise, we denote a batch estimate of the cross entropy gradient (3) by $\nabla_v CE_{disc}^{TU}(v)$. The component gradients can be computed by backpropagation. The flow of data and gradients forward and back through the networks is depicted in Figure 1. In practice, we train the network using algorithm 1. The $\mathrm{update}(w, g)$ procedure accepts the network weights w and a gradient of the cost function g and performs an update on w. While we used SGD with momentum, any update rule used for training neural networks is applicable. Instead of training the discriminator to optimality at each iteration, we perform k updates of the discriminator for a single update of the network f_w itself. There is no guarantee of convergence of the algorithm. However, our experiments demonstrate its practical utility.

Algorithm 1 Training a structured output network with adversarial cost for unlabelled data

```
$v, w \leftarrow \mathrm{randInit}()$
while not converged do
    for $IterNum = 1$ to $k$ do
        $T \leftarrow \mathrm{pickBatch}(\mathcal{T}, batchSize)$
        $U \leftarrow \mathrm{pickBatch}(\mathcal{U}, batchSize)$
        $g \leftarrow \nabla_v CE_{disc}^{TU}(v)$    ▷ backpropagation of batches $T$ and $U$ through the discriminator
        $v \leftarrow \mathrm{update}(v, g)$
    end for
    $T \leftarrow \mathrm{pickBatch}(\mathcal{T}, batchSize)$
    $g_T \leftarrow \nabla_w C^T(w)$    ▷ standard backpropagation
    $U \leftarrow \mathrm{pickBatch}(\mathcal{U}, batchSize)$
    $g_U \leftarrow \alpha \nabla_w C_{adv}^U(w)$    ▷ backpropagation through the discriminator and the network
    $w \leftarrow \mathrm{update}(w, g_T + \alpha g_U)$
end while
```

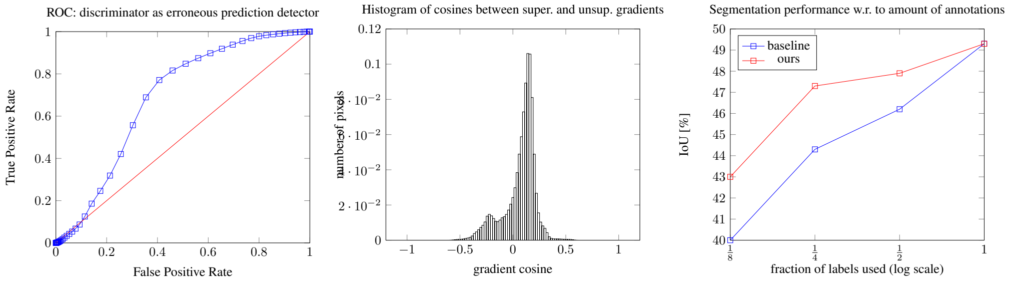

Figure 2: Left: The ROC of discriminator as a detector of erroneous predictions. Middle: The histogram of cosines between error signals originating from a discriminator and ones originating from a cross-entropy loss function. Right: The IoU attained by segnet-basic on the CamVid dataset with respect to the number of annotations used for training. See section 4 for details.

4 Experimental evaluation

Due to space limitations, we only report the key aspects of the experiments. We refer the reader to our online report for technical details.

To evaluate the adversarial regularization we reproduce the road scene segmentation setup of Badrinarayanan, Kendall and Cipolla [2]. It consists of the CamVid dataset and the segnet-basic neural network. It has 367 training images captured by a forward-looking vehicle-mounted camera, annotated in terms of 11 semantic classes. The segnet-basic architecture has a symmetric encoder-decoder architecture, with max pooling-unpooling coupling between the corresponding layers of the encoder and decoder. In total, the network has eight convolutional layers, each with 64 filters.

Our discriminator consists of three blocks of convolution with 64 filters and stride 2, batch normalization and leaky ReLU. It outputs a single variable *per patch* of the input image.

Correlation of the discriminator to prediction error. First, we verify that the discriminator is capable of capturing the quality of the structured output. We train segnet-basic on $\frac{1}{8}$-th of its original training set. Then, we train the discriminator to differentiate the outputs on the training images from ones obtained on the remaining $\frac{7}{8}$ of the original data set. We then plot the ROC curve of the discriminator as a predictor of correct pixel classification on the held out data in figure 2. We conclude that the discriminator output is informative of whether the prediction is correct.

Comparison of discriminator gradient to cross-entropy. To verify the utility of the discriminator gradient, we compare it to the gradient of a standard, fully supervised cross entropy loss function. In the same setup as for the previous experiment, we compute the cosine between these gradients for each pixel of every image held out for training. As can be seen in figure 2, most of the histogram mass accumulates on the positive side of the plot. This suggest that discriminator gradients point to the direction of more accurate predictions more often than not.

Segmentation performance. To check the performance of the adversarial regularization in a realistic application scenario, we compare semi-supervised training with the adversarial loss to fully supervised training, while decreasing the number of annotated training data in the original setup by a factor of 2, 4 and 8. The images excluded from the annotated training set are used as unannotated data for the adversarial regularization.

We train the network by SGD with momentum for 50000 iterations with a decaying learning rate. We jitter the training images. We perform the accuracy tests using the weight vectors after the last update in this procedure, instead of cherry-picking the best models.

We present numerical results in figure 2. When trained on the whole dataset, the baseline attains an accuracy of 49.3% Intersection-over-Union (IoU), exceeding the performance of 47.7% reported in the original paper [2]. Our method consistently outperforms the baseline when trained on a fraction of the dataset. Besides, when labelled data constitutes $\frac{1}{8}$ and $\frac{1}{4}$ of the training set, the regularized network is nearly as good as the baseline trained with twice as many labels.

Table 1: Left: Impact of weight decay and adversarial regularisation on the accuracy of segnet-basic trained on $\frac{1}{8}$-th of the CamVid training set. Right: The accuracy of segnet-basic trained on $\frac{1}{8}$-th of the data set with adversarial regularisation for different discriminator losses.

	weight decay alone						ours
decay factor	0	5e-4	1e-3	5e-3	1e-2	5e-2	1e-3
IoU	38.5	38.5	40.0	40.0	39.5	29.8	43.0

	IoU
baseline - $\frac{1}{8}$	42.3
ours - Cross Entropy	45.1
ours - Wasserstein	47.0

Comparison to weight decay To compare the performance gain due to the adversarial regularisation and weight decay, we train the network according to the protocol used in the previous experiment, on $\frac{1}{8}$ of the original training set, for several values of the weight decay coefficient. We present the results in table 1. Although weight decay leads to improved test performance, the improvement is limited. The adversarial regularisation enables breaking this limit.

Comparison of different discriminator variants We compared different variants of discriminator loss: the standard binary cross entropy (3) and the Wasserstein-GAN-like criterion [1]. We modified the Wasserstein criterion: instead of clipping the discriminator weights, we use weight decay and a hard hyperbolic tangent layer on top of the discriminator. These modifications consistently lead to improved results of our experiments. We changed the experimental setup by switching from SGD to the ADAM optimizer, leading to increased baseline performance. We present the results in table 1. The proposed regularization still improves performance, with the Wasserstein-like discriminator loss yielding significantly better results than the classical cross-entropy-based formulation.

Retinal vein segmentation We test the performance of adversarial regularisation in retinal vein segmentation. We use the DRIVE dataset [19] with 20 training eye fundus images, with binary annotations of retinal blood vessels. We adopt the U-Net symmetric encoder-decoder architecture [17], with skip connections copying feature maps of the encoder and concatenating them to the corresponding decoder feature maps. We modify it by adding batch normalization after each convolution, inserting residual skip connections over each pair of convolutions, and adding dropout after each such block. The network is trained for 20 thousands iterations using ADAM, on randomly cropped and rotated data. We vary the number of annotated training images and compare the results of supervised training to semi-supervised training, where the adversarial loss function is used for unlabelled images. We present the results in table 2. With the complete set of 20 labelled training images our baseline attains the F1 score of 0.819, close to the current state of the art of 0.821 attained by a network pre-trained on ImageNet [10]. Interestingly, with the adversarial regularisation we obtain competitive performance for just 3 training images.

Table 2: The performance of U-Net in retinal vessel segmentation vs. number of labelled training images (F1-score).

# images	baseline	ours
20	0.819	
3	0.785	0.813
1	0.725	0.769

5 Discussion

We proposed a loss function for semi-supervised learning, capable of generating useful error signals based exclusively on predictions. Contrary to the pre-training and co-training approaches, it enables end to end training on unlabelled data, irrespective of the task or network architecture. We have demonstrated that it allows to capitalize on unannotated data, narrowing the performance gap between predictors trained on the fully- and partly-labelled training sets, and enabling training useful predictors on just one annotated image. These advantages come at a cost of computation time and memory needed to train the discriminator network. Moreover, the regularised networks typically took more training iterations to attain their maximum performance. Finally, the presented experiments are performed on small data sets. It remains to be seen if the proposed method yields considerable improvements when hundreds, as opposed to tens, annotated images are available.

Acknowledgement This work was partially funded by the French National Research Agency, grant number ANR-13-CORD-003 (project SEMAPOLIS).

References

[1] ARJOVSKY, M., CHINTALA, S., AND BOTTOU, L. Wasserstein GAN. *CoRR abs/1701.07875* (2017).

[2] BADRINARAYANAN, V., KENDALL, A., AND CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561* (2015).

[3] DOERSCH, C., GUPTA, A., AND EFROS, A. A. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015* (2015), pp. 1422–1430.

[4] GANIN, Y., USTINOVA, E., AJAKAN, H., GERMAIN, P., LAROCHELLE, H., LAVIOLETTE, F., MARCHAND, M., AND LEMPITSKY, V. S. Domain-adversarial training of neural networks. *CoRR abs/1505.07818* (2015).

[5] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[6] HINTON, G. E., AND SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. *Science 313*, 5786 (July 2006), 504–507.

[7] ISOLA, P., ZHU, J.-Y., ZHOU, T., AND EFROS, A. A. Image-to-image translation with conditional adversarial networks. *arxiv* (2016).

[8] JUDY HOFFMAN, DEQUAN WANG, F. Y., AND DARRELL, T. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR abs/1612.02649* (2016).

[9] LUC, P., COUPRIE, C., CHINTALA, S., AND VERBEEK, J. Semantic segmentation using adversarial networks. *CoRR abs/1611.08408* (2016).

[10] MANINIS, K., PONT-TUSET, J., ARBELÁEZ, P. A., AND GOOL, L. J. V. Deep retinal image understanding. *CoRR abs/1609.01103* (2016).

[11] MASCI, J., MEIER, U., CIREŞAN, D., AND SCHMIDHUBER, J. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks* (2011), Springer, pp. 52–59.

[12] NOH, H., HONG, S., AND HAN, B. Learning deconvolution network for semantic segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015* (2015), pp. 1520–1528.

[13] NOROOZI, M., AND FAVARO, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI* (2016), pp. 69–84.

[14] PATHAK, D., KRÄHENBÜHL, P., DONAHUE, J., DARRELL, T., AND EFROS, A. A. Context encoders: Feature learning by inpainting. *CoRR abs/1604.07379* (2016).

[15] RADFORD, A., METZ, L., AND CHINTALA, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR abs/1511.06434* (2015).

[16] RANZATO, M., HUANG, F., BOUREAU, Y., AND LECUN, Y. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proc. Computer Vision and Pattern Recognition Conference (CVPR'07)* (2007), IEEE Press.

[17] RONNEBERGER, O., FISCHER, P., AND BROX, T. U-net: Convolutional networks for biomedical image segmentation. *CoRR abs/1505.04597* (2015).

[18] SALIMANS, T., GOODFELLOW, I. J., ZAREMBA, W., CHEUNG, V., RADFORD, A., CHEN, X., AND CHEN, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain* (2016), pp. 2226–2234.

[19] STAAL, J., ABRAMOFF, M., NIEMEIJER, M., VIERGEVER, M., AND VAN GINNEKEN, B. Ridge based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging 23*, 4 (2004), 501–509.

[20] ZEILER, M. D., TAYLOR, G. W., AND FERGUS, R. Adaptive deconvolutional networks for mid and high level feature learning. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011* (2011), pp. 2018–2025.

[21] ZHANG, Y., LEE, K., AND LEE, H. Augmenting supervised neural networks with unsupervised objectives for large-scale image classification. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016* (2016), pp. 612–621.

[22] ZHAO, J., MATHIEU, M., GOROSHIN, R., AND LECUN, Y. Stacked what-where auto-encoders. *CoRR abs/1506.02351* (2015).

[23] ZHENG, Z., ZHENG, L., AND YANG, Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. *CoRR abs/1701.07717* (2017).