Aerila image analysis with generative adversarial networks

Joar Gruneau
joar@gruneau.se

February 20, 2018

1 Introduction

2 Relevant Theory

2.1 Generative adversarial networks

Goodfellow *et al* [2] first proposed the generative adversarial network (GAN). The network consists of two parts, a generator and a discriminator. The generators task is to generate samples from some data distribution. The discriminators task is to differentiate these generated samples from the true samples. This results in a counter fitting game where the generator continuously tries to produce better generated data to fool the discriminator and the discriminator is forced to become better at differentiating these generated samples from the true samples.

A common solution to try to force the generator to generate samples from the entire distribution is to input a noise vector into the generator [10]. Since we in this work are only interested in segmentation where a deterministic mapping from the image to the segmentation map is desired we will not input any noise vector into the generator.

2.1.1 Unconditional generative adversarial networks

Unconditional GANs are the simplest form of GANs. Here the discriminator does not observe the input to the generator. This means that the discriminator will learn a loss function which does not depend on the generators input [4]. We first define the binary cross entropy loss.

$$\ell_{bce}(\hat{z}, z) = -(zln(\hat{z}) + (1-z)ln(1-\hat{z})) \tag{1}$$

Here $\hat{z}$ is the prediction and z is the ground truth. The loss function for a unconditional GAN can then be described as.

$$\mathcal{L}(G, D) = -(\ell_{bce}(D(y), 1) + \ell_{bce}(D(G(x)), 0)) \tag{2}$$

Here D stands for the discriminating network and G for the generating network. G tries to minimize this function and D tries to maximize it. Hence we get a minimax game

$$G^* = argmin_G[max_D[\mathcal{L}(G, D)]] \tag{3}$$

2.1.2 Conditional generative adversarial networks

A conditional generative adversarial network (cGAN) was proposed by [6]. By letting the discriminator observe the input to the generator we can condition the loss function the discriminator learns on this input. This is of

great importance here since we are not just trying to generate any semantic maps but semantic maps corresponding to the input image. The objective function will in this case be given by and the networks is trained with the minimax equation (3).

$$\mathcal{L}(G, D) = -(\ell_{bce}(D(x, y), 1) + \ell_{bce}(D(x, G(x)), 0)) \tag{4}$$

It has been shown that a multi term loss function can improve the quality of the generator [7, 4]. For image to image mappings a $\mathcal{L}_1$ or $\mathcal{L}_2$ loss is usually used. However for multi class image segmentation a multi class cross entropy loss is a better option to enforce the generator to assign a high probability to the correct class. The multi class cross entropy loss is given below.

$$\ell_{mce}(\hat{y}, y) = -\sum_{n=1}^{H*W}\sum_{n=1}^{C} y * log(\hat{y}) \tag{5}$$

Here y is the ground truth segmentation maps while $\hat{y}$ is the predicted maps. The discriminators objective is unchanged but the generator now has to fool the discriminator as well as minimizing the distance to the ground truth.

$$G^* = argmin_G[max_D[\mathcal{L}(G, D)] + \lambda \ell_{mce}(G)] \tag{6}$$

Here λ is just a constant which controls the importance of the second loss term.

2.2 Fully convolutional networks and U-NET

A fully convolotional network (FCN) was first proposed Shelhamer and Long *et al* [13]. A FCN is a CNN without any fully connected layers. A network with fully connected layers must have a specific input size on the image while a FCN network can take inputs of any size. The key insight is that by the authors were that fully connected layers can be viewed as convolutions with kernels that cover their entire input region. Hencea CNN with fully conected layers can be viewed as a FCN since it takes patches from a image of any size and outputs a spatial output map when the patches are aggregated. While the resulting maps are equivalent the computational cost for the FCN is greatly reduced. This is because no overlapping regions between patches has to be computed. This makes these networks ideal for generating dense output maps such as for image segmentation

Ronneberger *et al* [11] builds on the advancements of the FCN to propose a new type of segmentation network. The U-NET uses a encoder decoder structure with skip connections from bottleneck layers to upsampled layers. Thess skip connections are crucial to segmentation tasks as the initial feature maps maintain low-level features such that can be properly exploited

for accurate segmentation.The network has been shown to produce high accuracy results even on small sized datasets [14, 11, 4, 15, 16]. Ronneberger *et al* [11] attributes this to the networks structure which creates internal data augmentation.

3 Related work

In most cases the segmentation networks needs some post processing to improve the accuracy of the segmentation maps. Conditional random markov fields CRF have been very successfully to enforce spatial contiguity in the output maps [1, 5]. There have been wor that used mean field inference expressed as a recurrent convolutional networks to do CRF like post processing [12, 17]. Luc *et al* [5] proposed a adveserial segmentation network to enforce higher order potentials without being limited to a single class. Instead on directly enforcing these higher order potentials in a CRF model as post processing the goal was to enforce them in the generator directly with adversarial training. This technique also has the benefit of lower complexity since at test time only the generator will be used.

The generators task was to produce segmentation maps for the C classes. One initial concern was that the discriminator would trivially be able to differentiate the generated segmentation maps from the ground truth by only examining if they were continuous or discrete. To combat this a scaling method was proposed where the ground truth segmentation maps were processed so that a mass of τ where placed on the correct label but were otherwise made as similar as possible to the generated maps (in regard to KL divergence). The scaling method showed no improvement over the basic method with no pre processing.

Son and Jung *et al* [14] showed that a U-NET combined with an adversarial loss could achieve state of the art performance for retinal vessel segmentation in fundoscopic images. The team investigated several types on adversarial networks proposed in [4] such as image-GAN, patch-GAN and pixel-GAN. For Image-GAN the discriminator make a decision on a image level if the image is generated or not. For patch-GAN the images are split into patches and the discriminator analyses each individually. The result is the aggregated result from all patches. For pixel-GAN the discriminator makes it decision on pixel per pixel level. The team found that a image-GAN togheter with a cross entropy term preformed the best and outperformed the non adversarial segmentor trained only with the cross entropy loss by a significant margin.

4 Network Architecture

For the generator network a U-NET will be used. The objective function for the GAN will be. Using the definition of the binary cross entropy loss (1) and the multi class cross entropy loss (5) we can now define our loss function as.

$$\mathcal{L}(G, D) = \ell_{mce}(G(x), y) - \lambda(\ell_{bce}(D(x, y), 1) + \ell_{bce}(D(x, G(x)), 0)) \quad (7)$$

The generator will try to minimize this loss while the descriminator will try to maximize it. Following the example of [3, 5] and replace the term $-\lambda\ell_{bce}(D(G(x), y), 0)$ with $+\lambda\ell_{bce}(D(G(x), y), 1)$. Hence instead of minimizing the probability of the descriminative network to predict the generated map to be synthetic we maximize the probability of predicting the generated map as ground truth. The reason for this is that it leads a stronger gradient for the discriminator when making predictions on ground truth and generated maps. The binary cross entropy loss then becomes,

$$\mathcal{L}_{bce}(G, D) = \lambda(\ell_{bce}(D(x, G(x)), 1) - \ell_{bce}(D(x, y), 1)) \quad (8)$$

The objective for the network hence becomes.

$$G^* = argmin_G[max_D[\mathcal{L}_{bce}(G, D)] + \ell_{mce}(G(x), y)] \quad (9)$$

The generated segmentation maps will be concatinated with the image the basic manner since [5] did not gain any improvements with a product or scaling approach above the basic. Pinheiro *et al* [8] showed that it is preferable to have the same number of channels for each input to avoid that one input becomes dominating. Therefore the individual kernels will be applied to the image before concatenating them with the semantic maps. This also allows for a different low level representation of the image.

5 The dataset

5.1 The VEDAI dataset

The VEDAI dataset [9] consists of 9 different classes, these classes and the number of objects are given in the table below.

Classes	Number
Car	1340
Pick-up	950
Truck	300
Plane	47
Boat	170
Camping car	390
Tractor	190
Vans	100
Other	100

To make the results comparable with the extensive research of different methods done by [18] the classes plane, boat tractor van and other were removed due to scarcity of data. The annotations for each target consists on the centre coordinates of the target the angle of the center line of the bounding box as well as the corners of the bounding box. The bounding box fits the target closely so no extra information is given on the sides. The evaluation metrics on the VEDAI dataset are:

$$Precision = \frac{True\ positive}{True\ Positive + False\ poitive} \tag{10}$$

$$Recall = \frac{True\ positive}{True\ Positive + False\ negative} \tag{11}$$

$$F1\ score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{12}$$

6 Related Work

References

[1] A. Arnab, S. Jayasumana, S. Zheng, and P. Torr. Higher Order Conditional Random Fields in Deep Neural Networks. *arXiv:1511.08119 [cs]*, Nov. 2015. URL `http://arxiv.org/abs/1511.08119`. arXiv: 1511.08119.

[2] I. Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv:1701.00160 [cs]*, Dec. 2016. URL `http://arxiv.org/abs/1701.00160`. arXiv: 1701.00160.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. URL `http://arxiv.org/abs/1406.2661`. arXiv: 1406.2661.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv:1611.07004 [cs]*, Nov. 2016. URL http://arxiv.org/abs/1611.07004. arXiv: 1611.07004.

[5] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic Segmentation using Adversarial Networks. *arXiv:1611.08408 [cs]*, Nov. 2016. URL http://arxiv.org/abs/1611.08408. arXiv: 1611.08408.

[6] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv:1411.1784 [cs, stat]*, Nov. 2014. URL http://arxiv.org/abs/1411.1784. arXiv: 1411.1784.

[7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context Encoders: Feature Learning by Inpainting. *arXiv:1604.07379 [cs]*, Apr. 2016. URL http://arxiv.org/abs/1604.07379. arXiv: 1604.07379.

[8] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollàr. Learning to Refine Object Segments. *arXiv:1603.08695 [cs]*, Mar. 2016. URL http://arxiv.org/abs/1603.08695. arXiv: 1603.08695.

[9] S. Razakarivony and F. Jurie. Vehicle Detection in Aerial Imagery : A small target detection benchmark. *Journal of Visual Communication and Image Representation, Elsevier*, Mar. 2015. URL https://hal.archives-ouvertes.fr/hal-01122605.

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. May 2016. URL https://arxiv.org/abs/1605.05396.

[11] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015. URL http://arxiv.org/abs/1505.04597. arXiv: 1505.04597.

[12] A. G. Schwing and R. Urtasun. Fully Connected Deep Structured Networks. *arXiv:1503.02351 [cs]*, Mar. 2015. URL http://arxiv.org/abs/1503.02351. arXiv: 1503.02351.

[13] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *arXiv:1605.06211 [cs]*, May 2016. URL http://arxiv.org/abs/1605.06211. arXiv: 1605.06211.

[14] J. Son, S. J. Park, and K.-H. Jung. Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks. *arXiv:1706.09318 [cs]*, June 2017. URL http://arxiv.org/abs/1706.09318. arXiv: 1706.09318.

[15] Y. Xue, T. Xu, H. Zhang, R. Long, and X. Huang. SegAN: Adversarial Network with Multi-scale L_1 Loss for Medical Image Segmentation. *arXiv:1706.01805 [cs]*, June 2017. URL http://arxiv.org/abs/1706.01805. arXiv: 1706.01805.

[16] D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas, and D. Comaniciu. Automatic Liver Segmentation Using an Adversarial Image-to-Image Network. *arXiv:1707.08037 [cs]*, July 2017. URL http://arxiv.org/abs/1707.08037. arXiv: 1707.08037.

[17] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. *arXiv:1502.03240 [cs]*, pages 1529–1537, Dec. 2015. doi: 10.1109/ICCV.2015.179. URL http://arxiv.org/abs/1502.03240. arXiv: 1502.03240.

[18] J. Zhong, T. Lei, and G. Yao. Robust Vehicle Detection in Aerial Images Based on Cascaded Convolutional Neural Networks. *Sensors (Basel, Switzerland)*, 17(12), Nov. 2017. ISSN 1424-8220. doi: 10.3390/s17122720. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5751529/.