Human-Machine Collaboration for Medical Image Segmentation

F. Last[1], **T. Klein**[1], **M. Ravanbakhsh**[2], **M. Nabi**[1], **K. Batmanghelich**[3], **V. Tresp**[4]
[1] Machine Learning Research, SAP SE, Berlin, Germany
[2] University of Genoa [3] University of Pittsburgh [4] Ludwig Maximilian University

Abstract

Deep learning-based approaches achieve state-of-the-art performance in the majority of image segmentation benchmarks. However, training of such models requires sizable amount of manual annotations. In order to reduce this effort, we propose a method based on conditional Generative Adversarial Network (cGAN), which addresses segmentation in a semi-supervised setup and in a human-in-the-loop fashion. More specifically, we use the discriminator to identify unreliable slices for which expert annotation is required and use the generator in the GAN to synthesize segmentations on unlabeled data for which the model is confident. The quantitative results on a conventional standard benchmark show that our method is comparable with the state-of-the-art fully supervised methods in slice-level evaluation requiring far less annotated data.

1 Introduction

Semantic image segmentation, which aims at assigning a class label to each pixel in an image, is one of the main applications of machine learning in medical image processing. Lately, deep learning techniques have been shown to attain exceptional results in this domain, outperforming the traditional approaches. However, large amounts of manually labeled data - which are key for supervised deep learning applications - are often expensive or impractical to collect in the medical domain.
To capitalize on the effectiveness of deep learning approaches for semantic segmentation tasks, while at the same time dealing with the limited availability of labeled data in the medical field, we propose a human machine collaboration [1] framework for medical image segmentation based on the popular generative adversarial network (GAN) framework. We show that the scores produced by the adversarial discriminator, which is trained to detect out-of-distribution samples, can be interpreted as inherent uncertainty estimates for active learning. The ability to directly use the adversarial discriminator score as a measure of uncertainty results in a convenient end-to-end approach to active learning. [3] propose the combination of cross-entropy and adversarial losses for semantic segmentation. [6] perform semi-supervised image segmentation using GAN, leveraging unlabeled and generated data for estimating a proper prior. [8] employ GAN for active learning for classification problems, generating samples to query rather than selecting them from a pool. None of these approaches use the discriminator score to measure model certainty.

2 Human-Machine Collaborative Learning with GAN

The proposed approach is leveraging conditional GANs (cGAN) for facilitating the human-machine collaboration for segmentation. To that end, the generator G is trained to produce accurate label maps corresponding to the conditioned image, while the discriminator D attempts to recognize whether a given segmentation is in accordance with the input image. What is more, D can be used to estimate model uncertainty for unseen images. Specifically, we propose to use D for ranking

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

the predicted segmentations referred to as pseudo ground truth, such that annotations querying is restricted to low-confidence items. Thus expert annotations are obtained in an active learning fashion for out-of-distribution samples only, therefore incurring minimal cost.

The process of learning the model decomposes in several stages. First, a supervised base model is initialized and trained using the small set of labeled samples $I_{labeled}$. Second, the model is trained in an interactive fashion for n iterations. In each iteration, segmentation predictions are computed for the remaining unlabeled images $I_{unlabeled}$, which is followed by ranking. The top $\frac{k}{n}$ samples from the ranked pool are selected and queried for expert annotation, where k is the total annotation budget, yielding labeled set S_{expert}. All other samples from P are segmented using generator G, resulting in the labeled set S_{pseudo}.

Last step in each active learning cycle is an update of the model G, D. The full training procedure is illustrated in algorithm 1.

Algorithm 1: icGAN training

Input: $I_{labeled}$, $I_{unlabeled}$, k, n
$G, D \leftarrow initialize()$
$S \leftarrow I_{labeled}$
$P \leftarrow I_{unlabeled}$
for $i \in 1..n$ **do**
  $G, D \leftarrow train(S, G, D)$
  $P_{rank} \leftarrow rank(D(P))$
  $Q \leftarrow top(P_{rank}, \frac{k}{n})$
  $P \leftarrow P \setminus Q$
  $S_{expert} \leftarrow humanExpert(Q)$
  $S_{pseudo} \leftarrow G(P)$
  $S \leftarrow S_{true} \cup S_{pseudo}$
  $G, D \leftarrow train(S, G, D)$
end

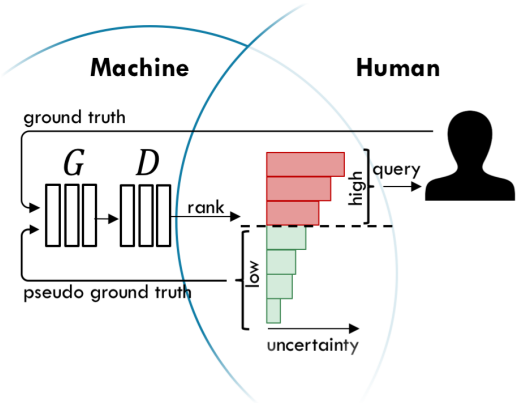

Figure 1: Algorithmic perspective (left) and concept view (right) of the proposed human-machine collaborative learning technique.

3 Experiments and Results

The proposed method is evaluated based on 3D cardiovascular MR images from the HVSMR 2016 challenge [4]. The set consists of ten axial, cropped volumes from ten different patients with ground truth annotations. Patient images without ground truth annotations (test set) are not used for this work. The images are segmented according to three labels: background, ventricular myocardium, and blood pool. For the experiment, all volumes were sliced along the same axis.
The baseline method is a fully supervised cGAN employing a U-Net with skip connections as the generator network and a PatchGAN as the discriminator [2]. The model is trained and evaluated using 10-fold cross validation.

Table 1: Dice scores for different amounts of supervised data and different benchmark models

	Proposed									Benchmark		
	10%	20%	30%	40%	50%	60%	70%	80%	90%	[2]	[7]	[5]
Myocardium	0.41	0.45	0.53	0.57	0.62	0.67	0.71	0.75	0.73	0.73	0.82	0.75
Blood Pool	0.86	0.88	0.89	0.90	0.91	0.92	0.92	0.94	0.95	0.94	0.93	0.89
Average	0.64	0.66	0.71	0.74	0.77	0.80	0.82	0.85	0.84	0.84	0.88	0.82

The proposed approach is based on the same architecture as the baseline network, but trained as described in section 2. Using the slices of a single patient volume a base model is trained. In order to estimate a lower bound of accuracy attained by the proposed method - which uses a fraction of the labeled data used for the fully supervised model - an experiment consisting of a single active

learning cycle ($n = 1$) is conducted. For simplicity, active learning cycles are simulated by using different fractions of annotations. Specifically, the supervised base model was used to determine the set of queries Q, before training a new model from the joint set $I_{labeled} \cup Q$. The experiment was conducted repeatedly for different values of budget k expressed in terms of share of the total available labeled data (0%, 10%, ..., 100%).

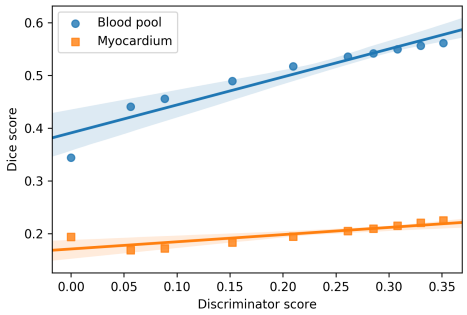

Figure 2: The average Dice score as a function of scores of D.

4 Discussion and Conclusion

First experiments suggest a strong and significant correlation between Dice score and confidence with (r = 0.98, p-value < 0.001) as shown in figure 2. As a result, the discriminator appears to be a good indicator of the quality of the label maps produced by the generator, justifying its interpretation as a measure of uncertainty. As illustrated in table 1, the performance of the proposed active learning approach increases with larger portions of data annotated interactively, reaching nearly the performance of the fully supervised benchmark methods after training with only 80% of the labels. Note that for these experiments, only one active learning cycle was conducted. Thus the obtained results constitute a lower bound on active learning performance. More active learning loops as well as incremental update of the model suggest to further improve the performance. This is because incrementally learning from new annotations is likely to change the model's ranking and selection of samples, exploring the pool of unlabeled samples more diversely.

References

[1] Azad Abad, Moin Nabi, and Alessandro Moschitti. Autonomous crowdsourcing through human-machine collaborative learning. *SIGIR*, pages 873–876, 2017.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[3] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic Segmentation using Adversarial Networks. *arXiv:1611.08408 [cs]*, November 2016. arXiv: 1611.08408.

[4] Danielle F. Pace, Adrian V. Dalca, Tal Geva, Andrew J. Powell, Mehdi H. Moghari, and Polina Golland. Interactive Whole-Heart Segmentation in Congenital Heart Disease. *MICCAI*, 2015.

[5] Rahil Shahzad, Shan Gao, Qian Tao, and Rob van der Geest. Automated cardiovascular segmentation in patients with congenital heart disease from 3d cmr scans: combining multi-atlases and level-sets. In *Reconstruction, Segmentation, and Analysis of Medical Images*. 2016.

[6] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi Supervised Semantic Segmentation Using Generative Adversarial Network. In *ICCV*, 2017.

[7] Lequan Yu, Qi Dou, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. Automatic 3d cardiovascular mr segmentation with densely-connected volumetric convnets. In *MICCAI*, 2017.

[8] Jia-Jie Zhu and Jose Bento. Generative Adversarial Active Learning. *arXiv preprint arXiv:1702.07956*, 2017.