How to Cure Cancer (in images) with Unpaired Image Translation

Joseph Paul Cohen, Margaux Luck, Sina Honari
Montreal Institute for Learning Algorithms, University of Montreal
{cohenjos, luckmarg, honaris}@iro.umontreal.ca

Abstract

We discuss how distribution matching losses, such as those used in CycleGAN, when used to translate images from one domain to another can lead to mis-diagnosis of medical conditions. It seems appealing to use these methods for image translation from the source domain to the target domain without requiring paired data. However, the way these models function is through matching the distribution of the translated images to the target domain. This can cause issues especially when the percentage of known and unknown labels (e.g. sick and healthy labels) differ between the source and target domains. When the output of the model is an image, current methods do not guarantee that the known and unknown labels have been preserved. Therefore until alternative solutions are proposed to maintain the accuracy of the translated features, such translated images should not be used for medical interpretation (e.g. by doctors). However, recent papers are using these models as if this is the goal.

1 Introduction

Generative Adversarial Networks (GANs) [1] have been used as an efficient and cheap method for data generation through implicit distribution matching. Recently, adversarial approaches for un-paired image translation between two domains have been proposed such as CycleGAN [2], and Adversarially Learned Inference (ALI) [3]. In medical imaging, un-paired domain translation models such as CycleGAN, have been used recently in translation tasks such as from MRI to CT. When translating images from a source to a target domain these models are trained to match a target distribution by any means necessary which includes hallucinating images by adding or removing image features. This is particularly problematic when the source and target domains have un-proportional distribution of known and unknown labels or features (e.g, being sick or healthy) which can change the image label through such image translations and implicitly change the nature of the data. Due to such a bias, we recommend until better solutions are proposed that maintain the vital information, such translated images should not be used for medical diagnosis, since they can lead to mis-diagnosis of medical conditions. This issue should be discussed because recently several papers have been published performing image translation using distribution matching. The main motivation for many of these approaches was to translate images from a source domain to a target domain such that they could be later used for interpretation (e.g. by doctors). Applications include MR to CT [4; 5], CS-MRI [6; 7], CT to PET [8], and automatic H&E staining [9].

We demonstrate the problem with a caricature example in Figure 1 where we *cure cancer* (in images) and *cause cancer* (in images) using a CycleGAN that translates between Flair and T1 MRI samples. In Figure 1(a) the model has been trained only on healthy T1 samples which learns to remove tumor from the image. This model has learned to match the target distribution regardless of maintaining features that are present in the image. We demonstrate below how these methods introduce a bias in image translation due to matching the target distribution.

Extended version of the paper online: `https://arxiv.org/abs/1805.08841`

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

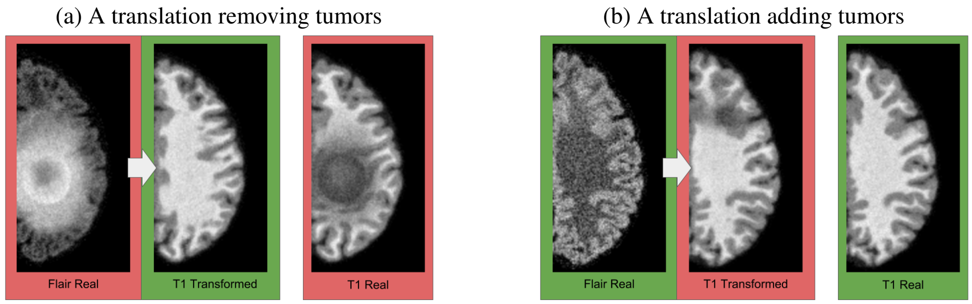

Figure 1: Examples of two CycleGANs trained to transform MRI images from Flair to T1 types. We show healthy images in green and tumor images in red. In (a) the model was trained with a bias to remove tumors. Here the target distribution did not have any tumor examples so the transformation model was forced to remove tumors in order to match the target distribution. In (b) the tumors were added to the test image to match the distribution which is composed of only tumor examples during training.

2 Bias Impact

Let's consider D_a and D_b as source and target domains. A CycleGAN learns a function $t(a)$ that maps a sample a from domain D_a to domain D_b. In this section, we construct training scenarios by setting the source domain fixed (with 50% healthy and 50% tumor samples) and change the ratio of healthy to tumor samples in the target domain D_b to observe the impact of the bias in the target domain composition on how the translation function $t(a)$ learns to match the target distribution.

We use the BRATS2013 [10] synthetic MRI dataset since we can visually observe the tumors, it is public, and has paired samples to evaluate the results. We split the dataset into 1400 training samples and 300 holdout test samples. We translate in-between Flair and T1 domains. We train 11 different CycleGAN models, as shown in Figure 2, where we keep the percentage of tumor samples in the source domain at 50% and change the percentage in the target domain from 0% to 100%. All these models are trained with 700 images in the target domain (the maximum number of images of only healthy and sick patients in the training set). In place of a doctor to classify the transformed samples we use an impartial CNN classifier which obtains 80% accuracy on the test set. The results of using this classifier on the generated T1 samples with different target domain composition is shown in Figure 2. If there was no bias in matching the target distribution due to the composition of the samples in the target domain, there would have been no difference in the percentage of the images diagnosed with tumor as we change the target domain composition in Figure 2. Moreover, at no point the translation is perfect even at the extreme of the plots. In Figure 3 we show how the translated images in the test set look as we change the composition of the target domain. The cancer tumor gradually appears and gets bigger from left to right. This is due to having the model match the target domain distribution statistics regardless of maintaining the source domain information that is vital.

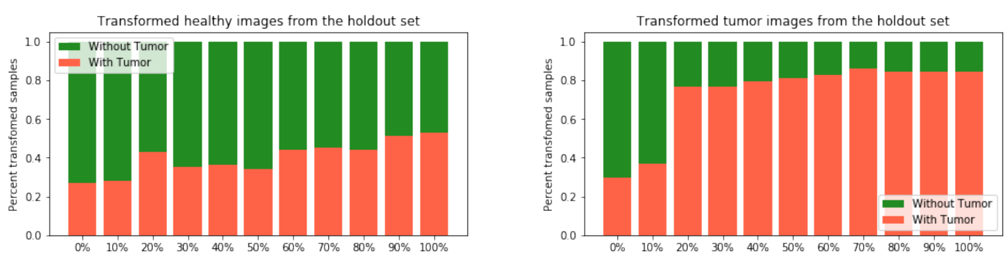

Figure 2: We plot the classifier prediction on 300 (53% tumor) unseen samples (holdout test set) as we vary the distribution of tumor samples in the target domain from 0% to 100% of cycleGAN models. This correspond to training 11 different models. We split the source domain of the holdout test set into healthy (left) and tumor (right) and apply a classifier on the translated images. Green represents samples predicted as healthy and red represents samples predicted with tumors. We observe that the percentage of the images diagnosed with tumors increases as the percentage of tumor images in the target distribution increases.

3 Conclusion

In this work we discussed concerns about how distribution matching losses, such as those used in CycleGAN, can lead to mis-diagnosis of medical conditions. We have presented evidence that when an algorithm uses distribution matching for unpaired data translation, all known and unknown class

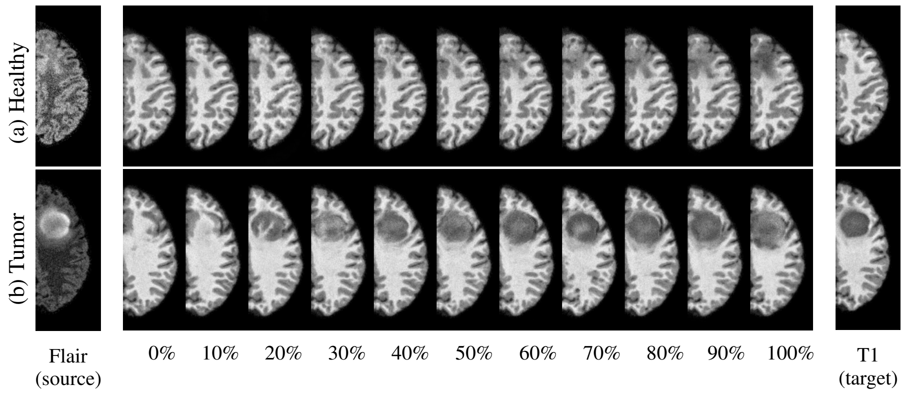

Figure 3: Illustration of healthy (a) and tumor (b) class change through domain translation while changing the ratio of the tumor samples in the target domain D_b from 0% to 100%. We show images of the source domain (Flair) on the left and the corresponding ground truth image in the target domain (T1) on the right.

labels may not be preserved. Therefore, these translated images should not be used for interpretation (e.g. by doctors) without proper tools to verify the translation process. We illustrate this problem using dramatic examples of tumors being added and removed from MRI images. We hope that future methods will take steps to ensure that this bias does not influence the outcome of a medical diagnosis.

Acknowledgements

We thank Adriana Romero Soriano, Michal Drozdzal, and Mohammad Havaei for their input and assistance on the project. This work is partially funded by a grant from the U.S. National Science Foundation Graduate Research Fellowship Program (grant number: DGE-1356104) and the Institut de valorisation des donnees (IVADO). This work utilized the supercomputing facilities managed by the Montreal Institute for Learning Algorithms, NSERC, Compute Canada, and Calcul Quebec.

References

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. 2014.

[2] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. 2017.

[3] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially Learned Inference. In *International Conference on Learning Representations*, 2017.

[4] Jelmer M. Wolterink, Anna M. Dinkla, Mark H.F. Savenije, Peter R. Seevinck, Cornelis A.T. van den Berg, and Ivana Išgum. Deep MR to CT synthesis using unpaired data. In *Workshop on Simulation and Synthesis in Medical Imaging*, 2017.

[5] Dong Nie, Roger Trullo, Caroline Petitjean, Su Ruan, and Dinggang Shen. Medical Image Synthesis with Context-Aware Generative Adversarial Networks. In *Medical Image Computing and Computer-Assisted Intervention*, 2016.

[6] Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. Compressed Sensing MRI Reconstruction using a Generative Adversarial Network with a Cyclic Loss. *IEEE Transactions on Medical Imaging*, 2018.

[7] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, and David Firmin. DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction. *IEEE Transactions on Medical Imaging*, 2018.

[8] Avi Ben-Cohen, Eyal Klang, Stephen P. Raskin, Michal Marianne Amitai, and Hayit Greenspan. Virtual PET Images from CT Data Using Deep Convolutional Networks: Initial Results. 2017.

[9] Neslihan Bayramolu, Mika Kaakinen, and Lauri Eklund. Towards Virtual H&E Staining of Hyperspectral Lung Histology Images Using Conditional Generative Adversarial Networks. In *International Conference on Computer Vision*, 2017.

[10] Bjoern H. Menze, Andras Jakab, Stefan Bauer, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 2015.