Medical Image Analysis 33 (2016) 94–97

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Editorial

Machine learning approaches in medical image analysis: From detection to diagnosis

Marleen de Bruijne

[a] Biomedical Imaging Group Rotterdam, Departments of Medical Informatics, Radiology & Nuclear Medicine, Erasmus MC-University Medical Center Rotterdam, The Netherlands
[b] The Image Section, Department of Computer Science, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 18 April 2016
Revised 22 June 2016
Accepted 22 June 2016
Available online 23 June 2016

Keywords:
Machine learning
Classification
Computer aided diagnosis
Transfer learning

ABSTRACT

Machine learning approaches are increasingly successful in image-based diagnosis, disease prognosis, and risk assessment. This paper highlights new research directions and discusses three main challenges related to machine learning in medical imaging: coping with variation in imaging protocols, learning from weak labels, and interpretation and evaluation of results.

1. Introduction

Supervised learning techniques, which learn a mapping from input data to output (labels) from a set of training examples, have shown great promise in medical image analysis. Pattern classification has already been used for decades to detect, and later characterize, abnormalities such as masses in mammograms and nodules in chest radiographs based on features describing local image appearance (Giger et al., 2008). With improvements in computer hardware it has become feasible to train more and more complex models on more data, and in the last few years, the use of supervised learning in image segmentation, recognition, and registration has accelerated. Trained appearance models are replacing simple intensity and gradient models as a component in segmentation systems, and statistical shape models that describe the typical shape and shape variations in a set of training shapes have replaced free form deformable models in many cases. Several new methods learn to diagnose disease in a fully data driven manner, using multivariate classification or regression to directly map from imaging data to diagnosis. These techniques are not restricted by current knowledge on disease-related radiological patterns and often have higher diagnostic accuracy than more traditional quantitative analysis based on simple volume or density measures.

Supervised quantification approaches can not only assist in diagnosis, but are also increasingly used to predict future disease onset or progression. Models are then trained on data from longitudinal studies in which the disease status years after the acquisition of the baseline image is known. For example at Erasmus MC, Achterberg et al. (2014) showed that hippocampal shape classification in a healthy elderly population is predictive of onset of dementia symptoms up to ten years later. van Engelen et al. (2014) used multivariate sparse Cox regression to take time to event into account in the model and found that changes in plaque texture and volume in ultrasound images of the carotid artery could predict future vascular events better than traditional risk factors could.

Possibly the most widespread application of machine learning based diagnosis appearing in publications is in neurodegenerative diseases, where researchers aim to diagnose Alzheimer's disease or other forms of dementia, or predict conversion from mild cognitive impairment (MCI) to dementia, based on brain MR images. This is likely driven, at least in part, by the availability of large datasets with diagnostic labels, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and Open Access Series of Imaging Studies (OASIS).

Another example where availability of data has altered the course of research is the detection of diabetic retinopathy in retinal fundus photographs. Many early papers focused on optimizing detection and segmentation of retinal vessels, for which several smaller public databases with ground truth were available. A recent Kaggle competition on diabetic retinopathy detection[1] changed the field by providing 35,000 images with expert visual

E-mail address: marleen@di.ku.dk, marleen.de.bruijne@gmail.com

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection.

http://dx.doi.org/10.1016/j.media.2016.06.032
1361-8415/

scores for training. This has drawn attention from data scientists around the world with no or little prior experience in medical image analysis. Many of the 661 participating teams used no specific pre-processing or segmentation but still obtained very good results. The top performing contributions all used different layouts of convolutional networks, with extensive data augmentation to increase the amount of training data even further, and achieved performance scores surpassing those previously reported for human experts.

We need to keep in mind that this example is a specific task, performed on 2D images. Differential diagnosis or quantification based on full 3D or 4D, possibly multi-modal, imaging data would require even larger training sets to describe all biological variation adequately. Additional domain specific knowledge will therefore still be needed in many cases. Results of another big data challenge which have just become available at the time of writing this paper, the "2015 Data Science Bowl"[2], seem to point in that direction. The challenge was to automatically measure end-systolic and end-diastolic volumes from dynamic cardiac MRIs. While the list of best performing algorithms is again dominated by convolutional neural networks, all top teams also performed specific preprocessing steps to detect the relevant regions and align image sequences. Nonetheless, these two examples suggest that supplying general purpose machine learning algorithms with a large amount of training data can lead to large improvements over current state-of-the-art performance in medical image analysis and computer aided diagnosis.

An enormous amount of data that could potentially be used for training exists: clinical experts assess many thousands of MRI and CT scans every day. In OECD countries alone, over 200 million CT and MRI scans are acquired per year[3] and radiographs and ultrasound images are acquired even much more frequently. Making part of this data available to train computer aided diagnosis algorithms could have tremendous impact.

In this paper I discuss three of the main challenges in approaching diagnosis with machine learning techniques and highlight several interesting research directions.

2. Varying imaging protocols

The main obstacle currently preventing wider use of machine learning in medical imaging is a lack of representative training data. While supervised learning techniques have shown much promise in relatively constrained experiments with standardized imaging protocols, their performance may quickly deteriorate on new images that are acquired under slightly different conditions. These techniques operate under the assumption that both train and test datasets are random samples drawn from the same distribution. In practice however, the available training data is often acquired earlier with a different imaging protocol, different scanner model, or from a different patient population, which would violate this assumption. An example of typical differences that can be found in multi-center MRI studies is given in Fig. 1.

One approach to cope with these issues, which is gaining increasing interest, is to apply transfer learning or domain adaptation techniques. We discern two classes of approaches that both aim to make train and test distributions more similar: weighting and feature space transformation techniques.

In weighting based transfer learning, training data with slightly different properties from the target data to analyze is used next to some labeled target data. A transfer classifier or regressor is then trained on all samples, but the additional, different-distribution samples receive a lower weight than the labeled target data. These different-distribution samples can help to regularize a classifier in a data driven manner — better than an uninformed regularizer — which makes it possible to train a reliable model with fewer labeled target samples. A similar effect can be achieved using the parameters of a classifier trained on different data to regularize a classifier on the target samples, as is done for instance in adaptive SVM. Such approaches may be easier to share between institutes as they do not require access to the original data samples that produced the classifier. Alternatively, samples, images, or image sets can be weighted in a fully unsupervised manner e.g. based on feature distribution similarity (van Opbroek et al., 2015b) or sample similarity (Heimann et al., 2014) with the target data.

In our research, we found that weighting based transfer learning approaches can significantly improve classification accuracy in MRI segmentation problems when few labeled target samples are available (van Opbroek et al., 2015a; van Engelen et al., 2015). However, the number of labeled target samples at which a classifier trained on only those samples performs as good as the transfer learning approaches was in these experiments quite low — several hundred labeled voxels distributed over all classes, up to a few well chosen, fully annotated images (Fig. 2). This depends of course on the data distribution and the model complexity. We would expect that with more complex representations, such as an increased number of image features or the representations obtained using 3D deep neural networks, the benefit of transfer learning becomes more clear. For example, in a different application using marginal space learning to localize ultrasound transducers in fluoroscopy sequences, Heimann et al. (2014) could completely eliminate localization errors by augmenting training sequences with synthetic data and subsequently downweighting less realistic synthetic images using a domain adaptation technique. Moreover, there is clearly still room for improvement in current methods; many general purpose transfer learning techniques are available but few explicitly take (medical) image properties into account.

While approaches based on sample or image weighting can compensate for some changes in distribution, they assume that the conditional distribution of the labels given the feature vectors is similar between the target data and (at least part of) the training data. This will often not be the case, for instance if intensity scale or contrast varies between images and the derived image features are not invariant to such transformations. A first step to address this will typically be image contrast normalization or standardization of image features to zero mean and unit variance, if necessary followed by a correction for intensity inhomogeneities. To compensate for further differences in distributions, a range of supervised and unsupervised techniques have been proposed in the machine learning and computer vision literature to project data into a latent space where distributions are more similar, for instance by minimizing the so-called Maximum Mean Discrepancy between distributions in a kernel space. An important remaining issue is that although transfer learning often improves results on similar tasks, without sufficient labeled target data it is not possible to detect negative transfer which undermines performance. Transfer learning techniques that could guarantee that the result of the transfer technique is never worse than the supervised solution, such as recently proposed for semi-supervised learning (Loog, 2016), are therefore of great interest.

The approaches discussed so far use training data from different sources more wisely and can compensate for possible differences between distributions. An alternative strategy would be to collect a very large and heterogeneous database for each task that contains all possible variations in imaging protocols, similar to the approach taken in the diabetic retinopathy competition described earlier. Combined with a sufficiently rich feature representation

[2] https://www.kaggle.com/c/second-annual-data-science-bowl.

[3] Health at a glance 2015.

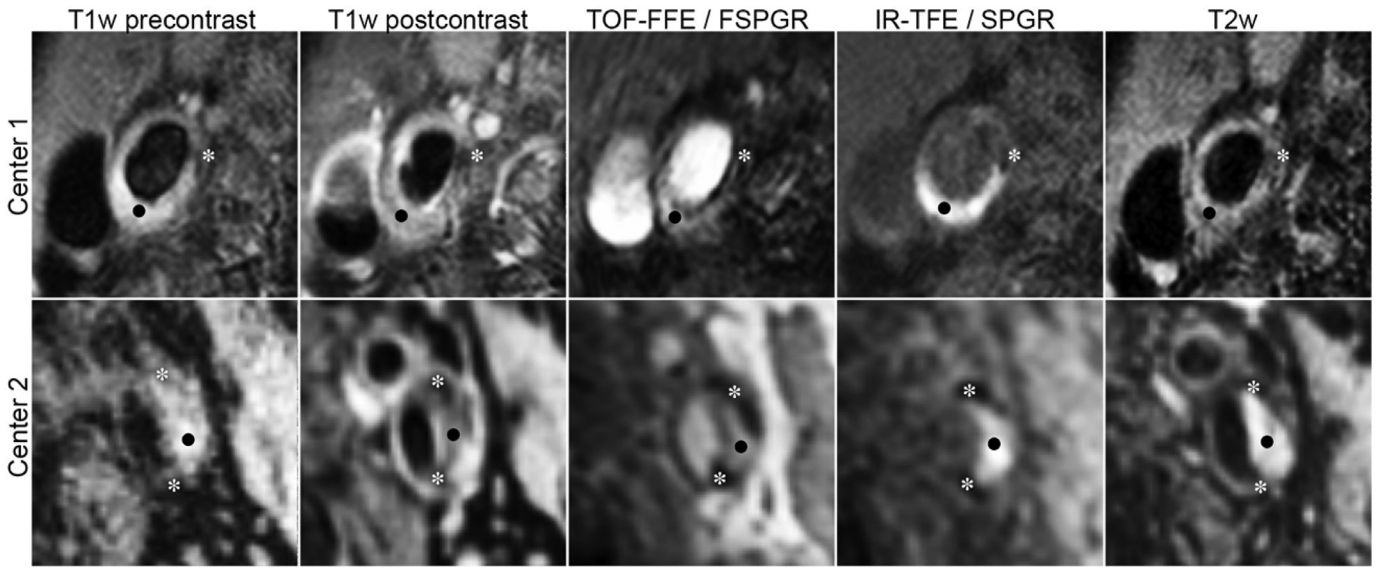

Fig. 1. MRI of the carotid artery obtained at two different sites in a multi-center study to improve diagnosis of high-risk carotid plaques. The imaging protocols in this study were carefully aligned, but due to different scanning equipment and different practices in different centers, some changes are unavoidable. Lumen, plaque, calcium spots (∗) and intraplaque hemorrhage (black dot) can clearly be distinguished in both protocols, but visual appearance differs. Reproduced with permission from van Engelen et al. (2015).

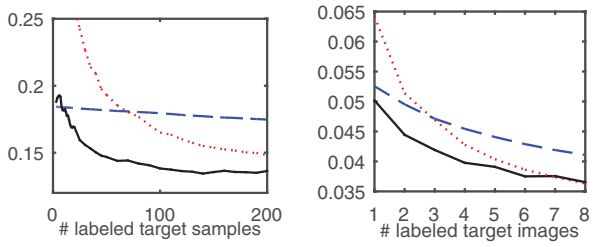

Fig. 2. Performance improvement of a weighting-based transfer learning approach compared to regular supervised classification using the same base classifier (SVM with a Gaussian kernel). Classification errors are shown as a function of the amount of labeled target data used, for the classifier trained on all data (blue, dashed line); the classifier trained on labeled target data only (red, dotted line); and the best transfer classifier of van Opbroek et al. (2015a) (black solid line), for brain tissue segmentation (left) and for white matter lesion segmentation (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and a sufficiently flexible learning model, such a simple approach could work well in practice.

3. Weak labels

Related to the lack of representative training data is a general lack of annotated data that could be used for training. Most current methods for segmentation or abnormality detection need manually segmented images to train on. This requires that a) humans are able to not only visually assess the images, but indicate boundaries reliably, which may be problematic for example for diffuse abnormalities; and b) resources are available to perform segmentation for the sole purpose of developing image analysis systems. Much more training data would be readily available if weaker labels that indicate for instance the presence, but not the location, of an abnormality could be exploited as well.

Learning with such weak, image-level labels can be addressed using multiple instance learning techniques. An image is then represented as a collection of instances (e.g., image patches) and the relation between the image label and the collection, rather than the individual instances, is learned. An example is the work of Sørensen et al. (2012), which uses multiple instance classification based on feature histograms of 3D patches randomly sampled within the lungs to discriminate between participants in a lung cancer CT screening study who had COPD and those who had normal lung function. Discrimination using this classification approach was found to be more accurate than using classic density measures and less sensitive to confounding variables such as gender and changes in inspiration level. The main advantage of this approach over other supervised texture analysis methods is that it does not require local, manual annotation by experts. In a similar approach applied to a dataset where also local annotations were available, Melendez et al. (2015) found that performance of multiple instance learning was not significantly different from its fully supervised counterpart in diagnosis of tuberculosis on chest X-rays.

These approaches have gone a long way in reducing the labeling effort required, however, they still rely on standardized diagnostic labels that may not always be available. Of special interest are therefore recent efforts to link supervised learning with semantic representations derived from free-text radiology reports such as presented by Schlegl et al. (2015). This work showed that learning based on a more complete, semantic representation outperformed multiple instance learning based on image-level labels alone in the interpretation of retinal OCT images. Such techniques need to be developed further to optimally use information from clinical reports and will eventually allow imaging biomarker discovery using large scale machine learning in routine clinical data.

4. Interpretation and evaluation

There are risks associated with applying learning techniques as a “black box” to perform diagnosis and risk assessment. A flexible learning system in a high-dimensional feature space can behave unexpectedly and this can be difficult to detect. When derived biomarkers show good separability between disease groups, it is tempting to assume that they must therefore be good at detecting the underlying signs of disease. However, depending on the training data, diagnosis decisions could well be driven not by signs of disease, but by signs of a confounding factor that is correlated with disease status in the training set. For instance, if a disease has

higher prevalence in men than in women, a complex learning algorithm might decide that the size of certain structures is a good indicator for the risk of disease, while in a study covering a large age range, signs of normal aging might be highlighted as strongly suspicious of dementia. Remedies are to collect a training set that is carefully balanced for confounding factors by e.g., age and gender matching between case and control groups, or — probably better — to incorporate possible other predictors in the learning and thus learn the joint relation between confounders and image appearance.

Due to these issues, it is not sufficient to know that a given learning approach has excellent performance on a given dataset. We should aim to understand what features drive the decisions and what are the corresponding pitfalls. While our field has become better at comparing results on common, public databases, among others in the many challenge workshops of the past few years, it is still difficult to predict the value of new algorithms outside the evaluated dataset. Generalization performance depends strongly on the size of the training set in relation to the complexity of the representation and learning model, and it is therefore surprising that most papers in the field only show a single point on a learning curve when comparing techniques. It would be more informative to show not only average performance on a specific training set, but also learning curves visualizing performance as a function of amount of training data as well as examples of cases in which the algorithm fails.

5. Conclusion

Machine learning approaches appear to be taking over the field and are increasingly successful in image-based diagnosis, disease prognosis, and risk assessment. Many scientific and practical challenges still need to be addressed to unlock their full potential, including how to train strong models on little data, how to improve access to data, how to best make use of the image structure and specific properties of medical imaging data in designing our models, how to interpret results, and how to apply these results in clinical practice.

Acknowledgments

I would like to thank the current and former group members of the Model-based Medical Image Analysis research group of the Departments of Radiology and Medical Informatics of Erasmus MC, the Netherlands, and members of the Image Section of Department of Computer Science, University of Copenhagen, for many fruitful discussions on machine learning and medical imaging. Thanks also to Gijs van Tulder, Annegreet van Opbroek, and Marco Loog for critically reading the manuscript. This research was partly funded by the Netherlands Organisation for Scientific Research (NWO).

References

Achterberg, H.C., van der Lijn, F., den Heijer, T., Vernooij, M.W., Ikram, M.A., Niessen, W.J., de Bruijne, M., 2014. Hippocampal shape is predictive for the development of dementia in a normal, elderly population.. Hum. Brain. Mapp. 35 (5), 2359–2371. doi:10.1002/hbm.22333.

Giger, M.L., Chan, H.-P., Boone, J., 2008. Anniversary paper: history and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM.. Med. Phys. 35 (12), 5799–5820. doi:10.1118/1.3013555.

Heimann, T., Mountney, P., John, M., Ionasec, R., 2014. Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data.. Med. Image Anal. 18 (8), 1320–1328. doi:10.1016/j.media.2014.04.007.

Loog, M., 2016. Contrastive pessimistic likelihood estimation for semi-supervised classification.. IEEE Trans. Pattern Anal. Mach. Intell. 38 (3), 462–475. doi:10.1109/TPAMI.2015.2452921.

Melendez, J., van Ginneken, B., Maduskar, P., Philipsen, R.H.H.M., Reither, K., Breuninger, M., Adetifa, I.M.O., Maane, R., Ayles, H., Sánchez, C.I., 2015. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays.. IEEE Trans. Med. Imaging 34 (1), 179–192. doi:10.1109/TMI.2014.2350539.

Schlegl, T., Waldstein, S.M., Vogl, W.-D., Schmidt-Erfurth, U., Langs, G., 2015. Predicting semantic descriptions from medical images with convolutional neural networks.. Inf. Process. Med. Imaging 24, 437–448.

Sørensen, L., Nielsen, M., Lo, P., Ashraf, H., Pedersen, J.H., de Bruijne, M., 2012. Texture-based analysis of COPD: a data-driven approach.. IEEE Trans. Med. Imaging 31 (1), 70–78. doi:10.1109/TMI.2011.2164931.

van Engelen, A., van Dijk, A.C., Truijman, M.T.B., Van't Klooster, R., van Opbroek, A., van der Lugt, A., Niessen, W.J., Kooi, M.E., de Bruijne, M., 2015. Multi-center MRI carotid plaque component segmentation using feature normalization and transfer learning.. IEEE Trans. Med. Imaging 34 (6), 1294–1305. doi:10.1109/TMI.2014.2384733.

van Engelen, A., Wannarong, T., Parraga, G., Niessen, W.J., Fenster, A., Spence, J.D., de Bruijne, M., 2014. Three-dimensional carotid ultrasound plaque texture predicts vascular events.. Stroke 45 (9), 2695–2701. doi:10.1161/STROKEAHA.114.005752.

van Opbroek, A., Ikram, M.A., Vernooij, M.W., de Bruijne, M., 2015a. Transfer learning improves supervised image segmentation across imaging protocols.. IEEE Trans. Med. Imaging 34 (5), 1018–1030. doi:10.1109/TMI.2014.2366792.

van Opbroek, A., Vernooij, M.W., Ikram, M.A., de Bruijne, M., 2015b. Weighting training images by maximizing distribution similarity for supervised segmentation across scanners.. Med. Image Anal. 24 (1), 245–254. doi:10.1016/j.media.2015.06.010.