2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)
Venice, Italy, April 8-11, 2019

LEARNING AN MR ACQUISITION-INVARIANT REPRESENTATION USING SIAMESE NEURAL NETWORKS

W.M. Kouw[†⋆] *M. Loog*[⋆†] *L.W. Bartels*[‡] *A.M. Mendrik*[‡+]

[⋆]Delft University of Technology, [†]University of Copenhagen,
[‡]University Medical Center Utrecht, [+]Netherlands eScience Center

ABSTRACT

Generalization of voxelwise classifiers is hampered by differences between MRI-scanners, e.g. different acquisition protocols and field strengths. To address this limitation, we propose a Siamese neural network (MRAI-NET) that extracts acquisition-invariant feature vectors. These can consequently be used by task-specific methods, such as voxelwise classifiers for tissue segmentation. MRAI-NET is evaluated on both simulated and real patient data. Experiments show that MRAI-NET outperforms both voxelwise classifiers trained on the source data as well as classifiers trained on the limited amount of target scanner data available.

Index Terms— MRI, Acquisition-variation, Representation Learning, Siamese Neural Network.

1. INTRODUCTION

Voxelwise classifiers for brain tissue segmentation should be trained on a sufficiently large representative data set, covering all possible types of variation. However, acquiring manual labels as ground truth is both labor intensive and time consuming. Furthermore, non-standardized manual segmentation protocols and inter- and intra-observer variability add a factor of variation to an already complex problem. Instead of increasing the number of manual labels, we propose to improve generalization by teaching a neural network to minimize an undesirable form of variation, namely acquisition-based variation. The proposed network learns a representation [1], in which for example gray matter patches acquired with a 1.5T scanner and a 3T scanner are considered similar. Therefore it has the potential to fully exploit a 1.5T data set with labeled brain tissues for segmenting an unlabelled 3T data set.

Overcoming acquisition-variation is a relatively new challenge in medical imaging. Transfer classifiers have been proposed that focus on weighting classifiers, such as weighted SVM's [2] and weighted ensembles [3]. Weights are based on how well each training sample matches the test data. However, these classifiers need to be retrained for every new test data set. Furthermore, they do not remove acquisition-variation or extract acquisition-invariant feature vectors for later use by task-specific methods.

We propose to learn a task-independent representation, in which acquisition-based variation is minimized while tissue variation is maintained. Patches sampled from MRI-scans that are mapped to this new representation will become feature vectors, and can be used to train task-specific classifiers. In order to minimize one factor of variation while maintaining another, we exploit a Siamese network [4]. Our proposed network is called MRAI-NET.

2. MR ACQUISITION-INVARIANT NETWORK

Suppose that we have scans that are acquired in two different ways; S (source) and T (target). A tissue patch, e.g. gray matter, is selected from both S and T. The aim is to teach a neural network that both these patches are gray matter, regardless of their visual difference. To achieve this, we use a loss function that expresses that pairs of samples from the same tissue but different scanners should be *similar*. However, if the neural network would only receives this instruction, it would map all patches to a single point and would destroy variation between tissues. To balance out the action of making certain pairs more similar, the network is also instructed that patches from different tissues – regardless of scanner – should remain *dissimilar*.

2.1. Siamese loss

Neural networks transform data in each layer. We summarize the total transformation from input to output layer with the symbol f: patch s from S will be mapped to the new representation with $f(s)$ and patch t from T will be mapped with $f(t)$. Distance in the new representation is expressed as $d_f(s,t) = \|f(s) - f(t)\|_1$. Pairs marked as similar (y=1) should be pulled together, while those marked as dissimilar (y=0) should be pushed apart. The loss for the similar pairs consists of the squared distance, $\ell_{\text{sim}}(f \mid s,t) = d_f(s,t)^2$. The loss function for the dissimilar pairs consists of a hinge loss: $\ell_{\text{dis}}(f \mid s,t) = \max\left[0, m - d_f(s,t)\right]$ where m is the margin parameter. Pairs that are pushed past the margin, will

WMK acknowledges support from the Niels Stensen Fellowship. AMM acknowledges support from ZonMw, IMDI Grant 104002002 (Brainbox).

not suffer a loss. We can combine the similar and dissimilar losses into a single loss function:

$$\ell(f) = \sum_i y_i \, \ell_{\text{sim}}\left(f \mid s_i, t_i\right) + (1 - y_i) \, \ell_{\text{dis}}\left(f \mid s_i, t_i\right)$$
$$= \sum_i y_i \, d_f(s_i, t_i)^2 + (1 - y_i) \max\left[0, m - d_f(s_i, t_i)\right] .$$

where i iterates over pairs. This type of loss function is known as a *Siamese* loss [4].

2.2. Labeling pairs as similar or dissimilar

Assume that we have sufficient manual segmentations (voxel labels) on scans S to train a supervised classifier and a limited amount of labels from scans T. Let K be the set of tissue labels, S_k be the set of patches from scan S of tissue k, and T_k be the set of patches from scan T of tissue k.. We form similar and dissimilar pairs, designated by a similarity label y, by taking pairwise combinations of individual patches $s_k \in S_k$ and $t_k \in T_k$. The following pairs are labeled as similar ($y = 1$): source patches from the same tissue (s_k, s_k), source and target patches from the same tissue: (s_k, t_k), and target patches from the same tissue: (t_k, t_k). Conversely, the following are labeled as dissimilar ($y = 0$): source patches from different tissues (s_k, s_l), source and target patches from different tissues (s_k, t_l), and target patches from different tissues (t_k, t_l), where $k, l \in K$ but $k \neq l$.

Let N_k be the number of patches extracted from a scan S belonging to tissue k, and M_k be the number of patches extracted from scan T of tissue k. In total, the number of combinations is $\sum_{k \in K}(N_k + M_k)^2 + \sum_{(k,l) \in \binom{K}{2}}(N_k N_l + N_k M_l + M_k M_l)$, where $(k, l) \in \binom{K}{2}$ refers to all combinations of 2 that can be taken from the set of tissues K. The combinatorial explosion works in our favor, as it allows us to generate a large training data set from only a few labeled target samples. Figure 1 illustrates the process of selecting pairs of patches from different scanners.

2.3. Network architecture

The network consists of two pipelines and a Siamese loss layer that acts on the pipes' output layers. We made the following architectural choices: 15x15 input patches, 8 convolution kernels of size 3x3 with "ReLU" activation functions, a fully-connected layer of size 16, another fully-connected layer of size 8, and a final fully-connected layer of size 2. Dropout was set to 0.2 during training, and we used a standard "RMSprop" optimizer to perform backpropagation. For more implementation details, see the accompanying software repository: `github.com/wmkouw/mrai-net`. MRAI-NET is implemented in Tensorflow and Keras.

Patches represented in the final representation layer are, in fact, feature vectors. The wider the layer, the higher the feature vector dimensionality. The two pipelines share their weights, which means they are constrained to perform the same transformation. This means that *single patches* can be fed through the network. It is not necessary to form pairs at test time.

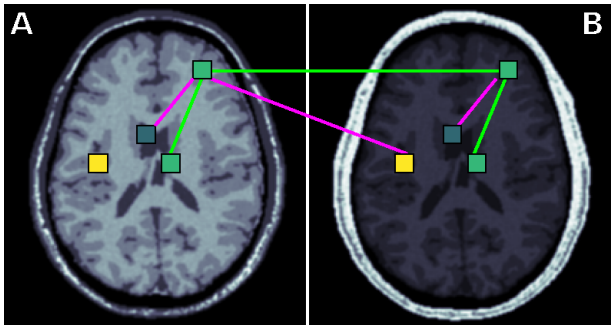


Fig. 1: Illustration of extracting pairs of patches from scans S and T. Each image shows 4 patches: 2 gray matter ones (green), 1 cerebrospinal fluid (blue) and 1 white matter (yellow). The lines mark the 6 types of combinations from Section 2.2 (green = similar, purple = dissimilar).

3. EXPERIMENT

In this experiment we test the dissimilarity between patches from the source and target scanners and we compare the performance of a linear classifier trained on MRAI-NET's feature vectors in a cross-scanner tissue segmentation task.

3.1. Data

We simulated different MR acquisitions from anatomical models of the human brain [5], using the MRI simulator SIMRI [6, 5]. The anatomical models consist of transverse slices of 20 normal brains (Brainweb). We simulated two acquisition types: (1) Brainweb1.5T, a standard gradient-echo acquisition protocol with the same parameters as the MRI-scanner in the Rotterdam Scan Study (B0 = 1.5T, $\theta = 20°$, TR=13.8 ms, TE=2.8 ms) [7], and (2) Brainweb3.0T, a standard gradient-echo protocol with the same parameters as the scanner used for MRBrainS (B0 = 3.0T, $\theta = 90°$, TR=7.9 ms, TE=4.5 ms) [8]. Magnetic field inhomogeneities and partial volume effects are not included in the simulation. There are 9 tissues, but we grouped these into "background", "cerebrospinal fluid", "gray matter", and "white matter". The simulations result in images of 256 by 256 pixels, with a 1.0x1.0mm resolution. Figure 1 shows examples of Brainweb1.5T (S) and Brainweb3.0T (T) scans. Since the same phantoms are used with both acquisition protocols, we effectively have the same patient in two different scanners. This allows us to isolate acquisition-based variation. In order to evaluate the proposed method on real data, we use the publicly available training data (5 subjects) from the MRBrainS challenge [8].

3.2. Measuring acquisition variation

The proxy $\mathcal{A}$-distance is a measure of discrepancy between two data sets [9]. Denoted by $d_{\mathcal{A}}$, it is defined as: $d_{\mathcal{A}}(s,t) = 2(1-2e(s,t))$, where e represents the test error of a classifier trained to discriminate patches s from scans S and patches t from scans T. For computing the proxy $\mathcal{A}$-distance, we draw 1500 patches from all source and 1500 from all target scans. A linear support vector machine is trained to discriminate between them, and the cross-validation error is used to produce $e(s,t)$.

3.3. Measuring tissue variation

Ultimately, we know that tissue variation is preserved if the extracted feature vectors can be used for tissue segmentation. A tissue classifier is used to measure how much variation between tissues is preserved in MRAI-NET's representation, specifically gray matter, white matter and cerebrospinal fluid. For evaluation, we use scans from target subjects that have been held back (10 subjects from Brainweb and 1 subject from MRBrainS). From these scans, we draw 50 patches per tissue at random, for a total of 1500 patches. We apply the tissue classifier to these test samples and compute the classification error rate.

3.4. Experimental setup

We compare a linear support vector machine trained on MRAI-NET's extracted feature vectors (also referred to as MRAI-NET) to two other supervised classifiers. First, the SOURCE classifier, which constitutes a convolutional neural network (CNN) trained on scans from the source (4 subjects for both Brainweb and MRBrainS) and target data (1 subject for both Brainweb and MRBrainS). This classifier represents the scenario where you would use a state-of-the-art method but would not account for acquisition-variation. Secondly, the TARGET classifier, a CNN trained on the few available patches from the target scan (1 subject for both Brainweb and MRBrainS). This classifier represents the scenario where you would disregard the source domain and work with what little labeled data is available.

SOURCE and TARGET's network architecture is the same as that of each pipeline in MRAI-NET. This rules out that differences in behavior between SOURCE, TARGET and MRAI-NET are due to choices for specific architectures. We construct learning curves by varying the number of labeled target patches available, from 1 to 1000 labeled patches per tissue.

We first performed this experiment using Brainweb1.5T as the source scanner and Brainweb3T as the target scanner. Since these are scans of the same subjects with different acquisition protocols, all variation between two scans is acquisition-based. Secondly, we performed the same experiment using Brainweb1.5T as the source scanner and MRBrainS as the target scanner. Now variation is both acquisition-based and patient-based.

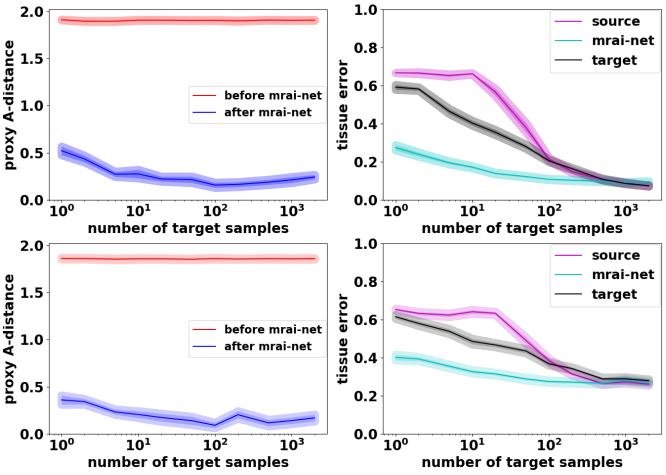
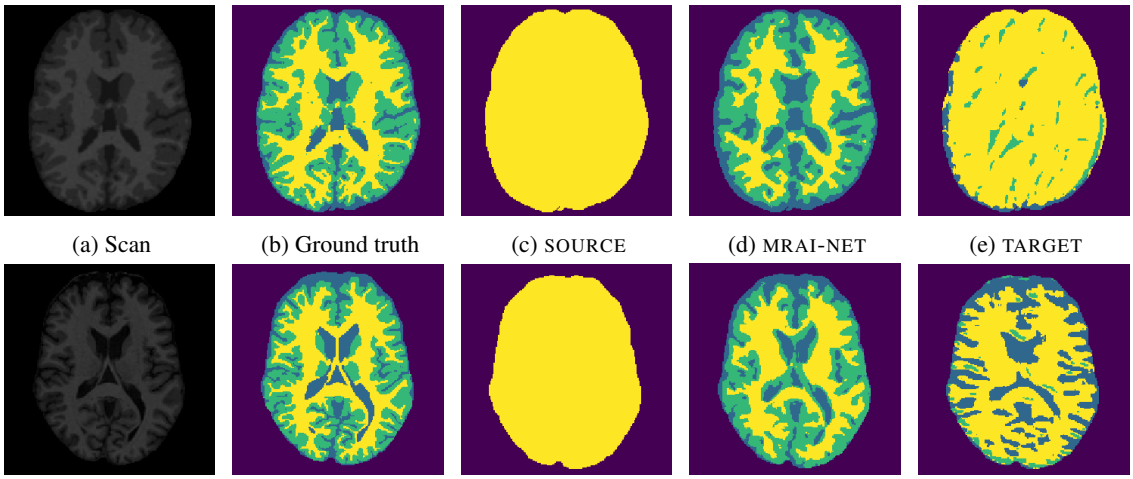


Fig. 2: Learning curves for Brainweb1.5T → Brainweb3T (Top row) and Brainweb1.5T → MRBrainS (Bottom row). (Left column) Proxy A-distance between source and target patches before (red) and after (blue) learning the new representation (smaller is better). (Right column) Tissue classification error for SOURCE, MRAI-NET and TARGET.

3.5. Results

Figure 2 shows the proxy $\mathcal{A}$-distance and the tissue classification error, for an increasing number of labeled target patches available for training. In general, the experiment with real patient data follows the same pattern as the simulated data. By using MRAI-NET, the distance between the source and target scanner data sets (proxy $\mathcal{A}$-distance) drops substantially, even with only one labeled target sample per class. For ten labeled target samples per tissue, MRAI-NET's error is 0.17 (Brainweb3T) and 0.33 (MRBrainS data), while SOURCE still performs at a 0.66/0.64 error (Brainweb3T/MRBrainS) and TARGET performs at 0.40/0.49. With one hundred target training samples the $\mathcal{A}$-distance approaches 0 (small acquisition variation means the data sets overlap), while tissue variation is preserved (tissue classification error 0.11 for Brainweb3T simulated data and 0.27 for MRBrainS real patient data). For Brainweb3T, the tissue classification error for the SOURCE and TARGET classifiers is 0.21 and 0.37, respectively. For MRBrainS, the error of SOURCE is 0.47 and the error of TARGET is 0.44. Given sufficient samples, all three classifiers reach similar performances. Figure 3 illustrates the difference in tissue classification performance when only one labeled target sample per tissue is used for training.

Note furthermore that SOURCE shows worse performance than TARGET for less than 50 samples. Apparently, the scans are so different that including the SOURCE samples in the training set actually *interferes* with learning. Given enough



(a) Scan (b) Ground truth (c) SOURCE (d) MRAI-NET (e) TARGET

(f) Scan (g) Ground truth (h) SOURCE (i) MRAI-NET (j) TARGET

Fig. 3: Example segmentations into white matter (yellow), gray matter (green) and cerebrospinal fluid (blue) using only one labeled target patch per class, for Brainweb1.5T → Brainweb3T (top row) and Brainweb1.5T → MRBrainS (bottom row).

target samples, however, SOURCE finds a good balance between source and target samples and matches the performance of TARGET.

4. CONCLUSION

We proposed to learn a representation of the data where acquisition-based variation is minimal and tissue variation is maintained. A linear classifier trained on feature vectors extracted by MRAI-NET outperforms conventional CNN classifiers trained on the source and target data sets in a cross-scanner tissue segmentation task, when few labeled target samples are available.

5. REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[2] A. van Opbroek, M.A. Ikram, M.W. Vernooij, and M. De Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," *IEEE Transactions on Medical Imaging*, vol. 34, no. 5, pp. 1018–1030, 2015.

[3] V. Cheplygina, A. van Opbroek, M.A. Ikram, M.W. Vernooij, and M. de Bruijne, "Asymmetric similarity-weighted ensembles for image segmentation," in *International Symposium on Biomedical Imaging*, 2016, pp. 273–277.

[4] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1735–1742.

[5] B. Aubert-Broche, M. Griffin, G. B. Pike, A. C. Evans, and D. L. Collins, "Twenty new digital brain phantoms for creation of validation image data bases," *IEEE Transactions on Medical Imaging*, vol. 25, no. 11, pp. 1410–1416, 2006.

[6] H. Benoit-Cattin, G. Collewet, B. Belaroussi, H. Saint-Jalmes, and C. Odet, "The SIMRI project: a versatile and interactive MRI simulator," *Journal of Magnetic Resonance*, vol. 173, no. 1, pp. 97–115, 2005.

[7] M. A. Ikram, A. van der Lugt, W. J. Niessen, P. J. Koudstaal, G. P. Krestin, A. Hofman, D. Bos, and M. W. Vernooij, "The Rotterdam Scan Study: design update 2016 and main findings," *European Journal of Epidemiology*, vol. 30, no. 12, pp. 1299–1315, 2015.

[8] A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, W. H. Bouvy, J. De Bresser, A. Alansary, M. De Bruijne, A. Carass, A. El-Baz, et al., "MRBrainS challenge: Online evaluation framework for brain image segmentation in 3T MRI scans," *Computational Intelligence and Neuroscience*, 2015.

[9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, 2010.