Learning Cross-Modality Representations from Multi-Modal Images

Gijs van Tulder and Marleen de Bruijne

Abstract—**Machine learning algorithms can have difficulties adapting to data from different sources, for example from different imaging modalities. We present and analyze three techniques for unsupervised cross-modality feature learning, using a shared autoencoder-like convolutional network that learns a common representation from multi-modal data. We investigate a form of feature normalization, a learning objective that minimizes cross-modality differences, and modality dropout, in which the network is trained with varying subsets of modalities. We measure the same-modality and cross-modality classification accuracies and explore whether the models learn modality-specific or shared features. This paper presents experiments on two public datasets, with knee images from two MRI modalities, provided by the Osteoarthritis Initiative, and brain tumor segmentation on four MRI modalities from the BRATS challenge. All three approaches improved the cross-modality classification accuracy, with modality dropout and per-feature normalization giving the largest improvement. We observed that the networks tend to learn a combination of cross-modality and modality-specific features. Overall, a combination of all three methods produced the most cross-modality features and the highest cross-modality classification accuracy, while maintaining most of the same-modality accuracy.**

Index Terms—**Representation learning, Transfer learning, Autoencoders, Deep learning**

I. Introduction

Many machine learning methods that work well on data that is similar to their training data might fail on data with different characteristics. This can lead to practical problems in medical image analysis, for example when existing models need to be applied to scans acquired with different imaging protocols or with different scanners. In these cases, transfer learning approaches can help to improve results, by allowing data from different sources to be used to train a single model that works for all sources. This paper proposes one of these approaches, based on representation learning using convolutional neural networks (CNNs). We present and study several ways to encourage a CNN to learn a common feature representation from heterogeneous data, in order to obtain a source-independent representation that is similar for data from all sources. This common representation makes it possible to train a model on data from one source and apply it to data from another. We apply these methods in cross-modality experiments.

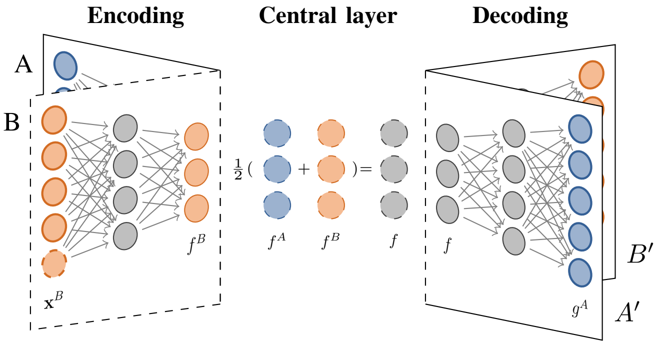


Fig. 1. Schematic overview of the axial CNN for two modalities. For each modality m, the input $\mathbf{x}^m$ is encoded into a representation $f^m(\mathbf{x}^m)$. The representations for all modalities are averaged into a mean representation $f(\mathbf{x})$, which is then used to compute reconstructions $g^m(f(\mathbf{x}))$ for all modalities. Each additional modality adds an extra input and output plane and is included in the average for the central layer.

Neural networks for cross-modality learning, such as the model presented here, have been popular in computer vision for some years (starting with [1]) and have more recently also been applied to medical images (e.g., [2]). Similar approaches to transfer knowledge between modalities have also been used to learn from incomplete datasets with missing modalities (e.g., [3] and [4]). In contrast with previous work learning a joint representation using a single transformation for all modalities (e.g., [5]), we propose cross-modality networks that learn a separate transformation for each modality. This allows the networks to model more complex transformations between modalities, such as intensity inversions, instead of merely learning modality-invariant features that are expressed in the same way in all modalities.

Cross-modality classification is a relatively unexplored topic in medical image analysis, but has received more attention in multimedia retrieval, most often in works on cross-modality classification of images and text (e.g., [6]–[11]). Feng et al. [9] present cross-modal retrieval experiments in cross-modal feature learning, using autoencoders and restricted Boltzmann machines to learn shared representations from images and text. They evaluate a learning objective similar to the similarity term discussed in this paper, as well as a form of modality dropout. Srivastava and Salakhutdinov [10] use deep Boltzmann machines to learn joint representations for text and images, reporting that multi-modal learning can improve results even if some modalities are not available at test time. Ngiam et al. [1] present cross-modality classification experiments with restricted Boltzmann machines and deep autoencoders, showing that speech classification can be improved

G. van Tulder and M. de Bruijne are with the Biomedical Imaging Group, Erasmus MC, Rotterdam, The Netherlands (email: g.vantulder@erasmusmc.nl or marleen.debruijne@erasmusmc.nl). M. de Bruijne is also with the Department of Computer Science, University of Copenhagen, Denmark.

by learning from video and audio. They train with a form of modality dropout to learn models that are robust to inputs with missing modalities. Vukotić et al. [11] present cross-modal deep networks based on deep autoencoders, aiming to learn a common hidden representation from text and images in a video hyperlinking task. In the medical domain, Moradi et al. [12] proposed a cross-modality neural network combining text and images for semi-automatic annotation of medical images, using a two-step approach that first extracts features from text and images and then learns a mapping between the two domains. In this paper, we propose a single-step method to learn cross-domain representations from multimodal medical images, and evaluate a number of additions to obtain representations that perform well in cross-modality classification.

Recent work using adversarial learning provides an alternative method for unsupervised domain adaptation, using an adversarial loss function. This can be done at the image level or at the feature representation level. Adversarial domain adaptation on an image level can be implemented with cycle-consistent generative adversarial networks (CycleGANs). For example, Zhang et al. [13] applied this to CT and MRI data, by training a CycleGAN to convert MRI data to CT and back. In this case, the discriminator network attempts to discriminate between CT derived from MRI data and real CT images. On a feature level, the adversarial loss can be implemented by a discriminator network that attempts to identify the source modality of a sample from its feature representation. For example, Kamnitsas et al. [14] applied this to an MRI and CT brain segmentation task, and describe how the adversarial loss helps to produce a feature representation that is more similar across modalities. Unlike the methods proposed in this paper, the adversarial methods do not use corresponding image samples from both domains, but rely solely on the adversarial loss to learn the translation.

We present results of patch-wise cross-modality classification experiments on two multi-modal datasets: a knee cartilage segmentation dataset with two different MRI sequences, and a brain tumor segmentation dataset with four MRI sequences. Voxel classification approaches such as the deep convolutional networks used in this paper have been used previously for both types of data. For example, knee cartilage segmentation has been approached with texture features (e.g., [15]) and deep neural networks [16]. Texture-based voxel classification also gave good results for the brain tumor segmentation problem (see [17] for an overview). In recent years, deep convolutional networks have also been applied to this problem (e.g., [18]).

For both datasets, we use unlabeled training data with multiple modalities per subject to train an axial CNN [2] that learns source-specific transformations that map data from each source to a single common representation. We evaluate this common representation in a transfer learning setting, training a classifier on labeled data from one source and applying it to data from another. We combine the basic cross-modality architecture with three techniques to further improve cross-modality feature learning: modality dropout [1], [4], a similarity term [2], and a normalization step. We analyze whether the models learn mostly shared features, mostly modality-specific features, or a combination of both.

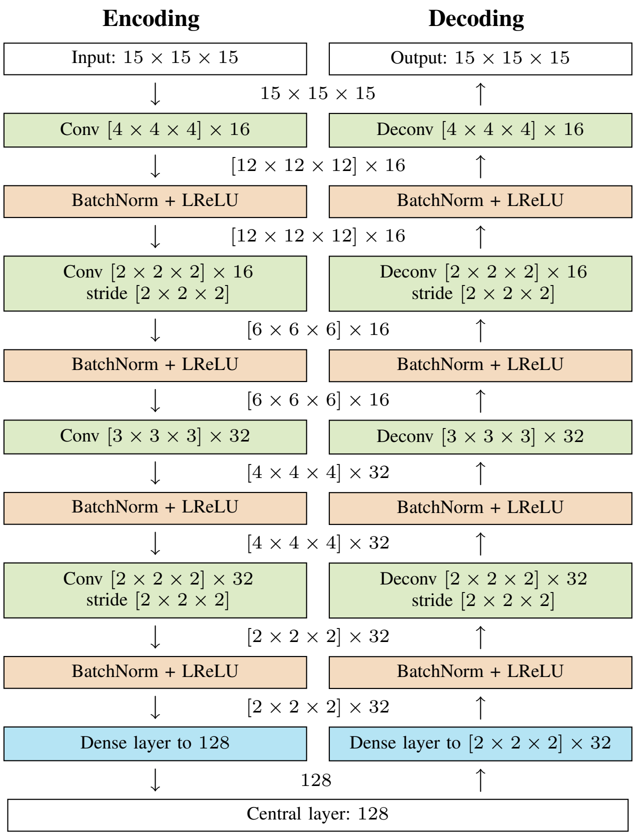


Fig. 2. The structure used for the encoding and decoding parts of the network, with the size of intermediate representations shown between the blocks.

In this paper, we use an axial neural network architecture that is similar to the architecture that we used in our earlier work on the similarity term [2], although that paper used a much simpler network without convolutional layers. The idea of using a separate network path for each input source also appears in work by Ngiam et al. [1] and Havaei et al. [4] on modality dropout, although the latter only applied it to the input side of a supervised classification network and not to reconstruction. To the best of our knowledge, the combination of all three methods and the extensive evaluation and analysis of the feature representations learned by the different methods is a novel contribution of this paper.

This paper is organized as follows. Section II outlines the basic model and the three techniques to improve cross-modality feature learning. Section III discusses the datasets. Section IV gives an overview of the experiments, the results of which are presented in Section V. Section VI and Section VII discuss the conclusions.

II. Methods

We investigate the axial convolutional neural network [2] (Fig. 1) for cross-modality learning. This is an autoencoder-like model that learns a common representation for data from multiple modalities, which can then be used for cross-modality classification: training a classifier on data from one modality and applying it to data from another, using the shared representation as a common feature description for samples from both modalities. In this section, we describe the model and three extensions that can further improve the cross-modality similarity of the representations.

A. Axial convolutional neural network

We construct a multi-input autoencoder network (Fig. 1) that has an input $\mathbf{x} = \left\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^M\right\}$ with corresponding input patches $\mathbf{x}^m$ for each of the M modalities. For a modality m, given an input patch $\mathbf{x}^m$, the network uses a modality-specific encoding transformation f^m to compute the representation $f^m\left(\mathbf{x}^m\right)$. Because the model should produce the same representation for each of the modalities, we compute the mean representation $f\left(\mathbf{x}\right) = \frac{1}{M}\sum_{m=1}^{M} f^m\left(\mathbf{x}^m\right)$ and use this as the input for the modality-specific decoding transformations $g^m\left(f\left(\mathbf{x}\right)\right)$. The network is trained with an autoencoder objective to minimize the sum of the reconstruction errors:

$$\mathcal{L}_{recon} = \sum_{m=1}^{M} \left|g^m\left(f\left(\mathbf{x}\right)\right) - \mathbf{x}^m\right|. \tag{1}$$

The model is trained with paired input patches $\mathbf{x} = \left\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^M\right\}$. We assume that the images are registered and that there is a voxelwise correspondence between all patches $\mathbf{x}^m$ for a given sample. Furthermore, although the network can handle incomplete training samples for which not all M modalities are available, it needs sufficient training pairs to learn the correspondences between all modalities.

The encoding and decoding transformations in our models are implemented as convolutional networks (Fig. 2) with a sequence of convolution and batch normalization layers. The encoding part of the network uses strided, valid convolutions to avoid border effects in the central layer. The decoding part is the inverse of the encoding part, using transposed convolutions to reconstruct the original input size. All inner layers use leaky rectified linear units; the reconstruction layer is linear to allow it to reproduce the full range of input values.

Taking the mean representation over all modalities encodes our goal of learning a common representation across modalities in the structure of the network: ideally, we want the representation $f^m\left(\mathbf{x}^m\right) \approx f\left(\mathbf{x}\right)$ to be the same for all modalities m. Using the average representation instead of a single shared layer makes it possible to train and test with incomplete data for which not all modalities are available: by dividing the sum by the correct number of modalities, the scale of the combined feature values becomes independent of the number of input modalities.

Averaging the representations over all modalities is not sufficient to learn cross-modality representations, because it still allows the network to learn modality-specific features. If the network is always trained with complete training samples, for which all modalities are always available, it might allocate a different part of the feature representation to each modality. This would produce a single feature vector that can be used to reconstruct all modalities, but it would not produce a true cross-modality representation, because it is still dependent on all input modalities. To obtain a true cross-modality representation, we need to change how the model is trained. The remainder of this section presents three techniques to do this.

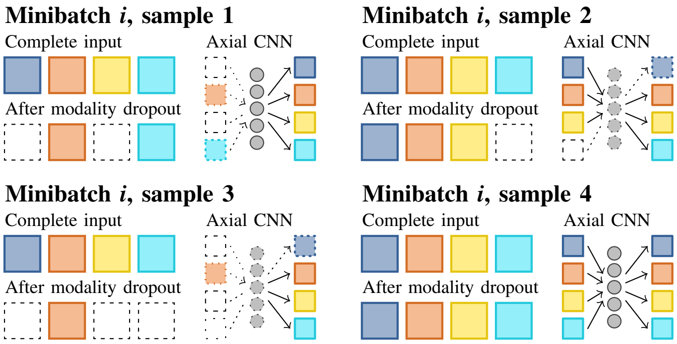


Fig. 3. Schematic illustration of modality dropout with four modalities. We select a random subset of 1 to 4 modalities for each sample in each minibatch. The network is only given the selected input modalities to compute the central representation, but we ask it to reconstruct all modalities and optimize the full reconstruction error. The subsets are generated independently for each sample, so a minibatch can contain multiple modality combinations. We generate a new random subset each time a sample is used for training.

B. Modality dropout

The first approach (used, for example, in [1] and [4]) modifies the training procedure. In the default training procedure, the network is never explicitly forced to learn to reconstruct one modality from another, because all modalities are always available for all training samples. If the representation is sufficiently large, the network might learn to use a separate part of the representation for each modality. Modality dropout prevents this by disabling modalities at random during training, computing the mean representation from a random subset of modalities while still optimizing the reconstructions for all modalities. For a model with M modalities, we select a random subset of 1 to M input modalities in each update step. We generate a random subset each time a sample is included in a minibatch: the modalities can be different each time a sample is used, and each minibatch can contain multiple modality combinations (see Fig. 3). Using incomplete inputs for training means that the network can no longer rely on the original modality for its reconstruction, but is forced to learn cross-modality reconstructions and representations.

C. Similarity term in the learning objective

The second approach explicitly adds cross-modality learning to the learning objective, similar to the approach in [2]. We compute the difference between the modality-specific representations $f^m\left(\mathbf{x}^m\right)$ and the mean representation $f\left(\mathbf{x}\right)$. We add this to the original learning objective (1) with a tunable weight $\alpha \in [0, 1]$:

$$\mathcal{L}_{sim} = \sum_{m=1}^{M} \left|f^m\left(\mathbf{x}^m\right) - f\left(\mathbf{x}\right)\right|, \tag{2}$$

$$\mathcal{L}_{combined} = \left(1 - \alpha\right)\mathcal{L}_{recon} + \alpha\mathcal{L}_{sim}. \tag{3}$$

Choosing α large enough will cause the network to reduce the differences between the representations for each modality. However, it is equally important not to set α too high: choosing a value very close to 1 will disregard the reconstruction error and can produce representations that may be very similar, but are also very uninformative.

The similarity term as defined in (2) can have another undesired effect: it can be trivally minimized by reducing the absolute feature values, so it might lead to very small or completely disabled feature values. This reduces the loss but does nothing to improve the cross-modality similarity. To prevent this trivial optimization, we normalize all feature vectors to zero mean and unit standard deviation.

D. Per-feature normalization

Global normalization across all features still allows cross-modality differences between individual features: they can be active for one modality and disabled in another. Our third approach is therefore to normalize each individual feature to zero mean and unit standard deviation, before averaging the modality-specific representations to get the mean representation. This per-feature normalization helps to remove a large part of the differences between modalities, and allows the network to focus on more meaningful ways to improve the representation similarity. We implement this normalization using a standard batch normalization procedure [19] to learn estimates of standard deviation and mean for each feature, per-modality, and to normalize the feature to zero-mean and unit standard deviation. The batch normalization formula provides scaling and shift parameters (β and γ in [19]), which allow the model to scale and shift the features away from a zero mean and unit standard deviation. In our case, doing so could reintroduce differences between modalities. We fix the parameters to $\beta = 1$ and $\gamma = 0$ to prevent this. (Note that we only make this change for this specific per-feature normalization step. We use the standard batch normalization formula for the batch normalization layers in the network, as shown in Fig. 1.)

III. Data

We performed experiments for two tasks: knee cartilage segmentation and brain tumor segmentation. In both cases, we evaluate our methods on a patch-based classification task in which we train classifiers to label the center voxel of a $15 \times 15 \times 15$ voxel neighborhood. We take paired patches from all modalities of a subject, such that the patch in each modality represents the same physical location.

For the experiments on knee segmentation, we used knee MRI images from the Osteoarthritis Initiative (OAI) [20], with the manual cartilage and meniscus segmentations from the iMorphics subset. For each subject, the dataset provides normal (N) and fat-suppressed (FS) MRI scans (Fig. 4a), made shortly after each other, which disagree on the intensity of some tissue types. The normal scans also have a somewhat better resolution. The dataset provides registered and resampled scans for each subject, to a common voxel spacing of $0.36 \times 0.36 \times 0.7$mm. We extracted paired patches of $15 \times 15 \times 15$ voxels, using the annotation of the center voxel in the normal scan as the patch label to define a three-class classification problem (cartilage, meniscus and background). The background voxels were sampled from a background mask, which we constructed by dilating the cartilage and meniscus segmentations with 10 voxels. We used N–FS pairs from baseline and 12-month follow-up sessions from 88 subjects, excluding two pairs that were not properly aligned. For each of the 172 pairs we extracted a randomly sampled, balanced set of 5000 cartilage, 5000 meniscus and 5000 background patches. Before extracting the patches, we normalized each scan to have a zero mean and unit standard deviation in the background and foreground voxels.

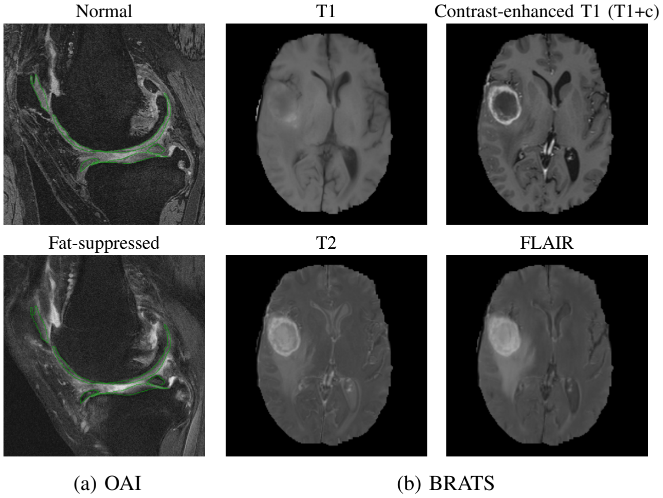


Fig. 4. Example scans from OAI and BRATS, showing the two knee MRI modalities and four brain MRI modalities.

Our second dataset uses data from the BRATS brain tumor segmentation challenge [17], which provides T1, contrast-enhanced T1 (T1+c), T2 and FLAIR scans for each subject (Fig. 4b). The challenge dataset (BRATS 2015) provides manual segmentations of four tumor components and a brain mask for each subject. The images and segmentations for each subject have been registered to the contrast-enhanced T1 scan and resampled to a $1 \times 1 \times 1$mm voxel size. For each subject, we extract patches from $15 \times 15 \times 15$ at the same position in each modality and use the label of the center voxel as the label of this sample. Because some of the tumor components are only visible on some of the modalities and we evaluate single-modality cross-modality classification, we merged the four tumor components into a single class to formulate a two-class classification problem (tumor vs. non-tumor brain tissue). The dataset contains scans of 220 subjects, for each of which we extracted a balanced set of 5000 foreground and 5000 background patches. Before extracting the patches, we normalized each scan to have a zero mean and unit standard deviation in the brain mask.

IV. Experiments

We present a comparison of all combinations of the three techniques: modality dropout, per-feature normalization and a range of weights α for the similarity term. For each combination, we trained axial neural networks to learn a common feature representation. We then used the resulting networks to compute a feature vector for each modality.

To evaluate the suitability of the common representation for classification, we trained random forest classifiers on the features extracted by each axial neural network. We distinguish two scenarios: same-modality and cross-modality classification. For same-modality classification, we trained the classifier on the features obtained from one modality and evaluate it on a feature vector obtained from the same modality. For cross-modality classification, we trained the classifier using the features derived from a different modality than the one used in testing.

As part of our analysis, we investigate to what extent the models learn modality-specific or shared features. We do this by training classifiers with only a subset of features, ranked by the normalized cross-modality correlation (suggested in [21]). We start with the feature that has the most similar values across modalities and gradually add more, training a new random forest for each subset.

Our networks were implemented using Keras [22] and Theano [23]. We used stochastic gradient descent for 100 epochs on the OAI dataset and 50 epochs on the BRATS dataset, which was sufficient for the networks to converge to a stable state. The minibatch size was 64 patches, the learning rate was 0.3 and the learning rate decay was 0.000002. We used Scikit-learn [24] with the default settings to train random forest classifiers with 30 trees.

We compare the results of our axial CNNs with those of two baseline methods. Both baselines use the same layer architecture as our axial networks (Fig. 2), but instead of learning multiple modality-specific transformations, the baseline methods learn only a single transformation that is shared by all modalities. In this way, they resemble normal autoencoders that encode and decode a single input patch and optimize its reconstruction error.

The two baselines use different training data to learn a common representation. The first baseline method is trained to reconstruct the training modality from itself, which produces a transformation that we also apply to the testing modality. For example, in a cross-modality classification experiment with modalities A and B, the first baseline method learns its representation only from patches of modality A, but the same representation is used to compute the features for the patches from modality B at test time. The second baseline learns the transformation from all modalities combined. In the example with A and B, this baseline would learn its representation from a mixture of patches from A and patches from B, without knowing which modality each patch is from.

We report results obtained in five-fold cross-validation. For each dataset, we divided the paired scans in five random subsets of approximately equal size, making sure that all scans of a subject were kept in the same subset. Using each subset in turn for testing, we first trained the axial neural networks on the remaining four subsets and used these networks to compute features for the training and test samples. For each subset, we trained the random forest classifiers on data from the training set and evaluate it on the test set. We report the mean accuracy over all five folds.

We used a slightly modified procedure for the experiments with subsets of most-correlated features, since these cross-modality correlations need to be computed on data that was not used to learn the representation. For these experiments, we introduced a second, two-fold cross-validation step to compute the results: we split each test subset in two halves, ensuring that all data from the same subject is in the same half, and in turn use one half to select the features and the other half to evaluate the classifier. We report the mean results over all 5×2 subsets, covering all samples in the dataset. If all features are selected, this is equivalent to the normal five-fold cross-validation.

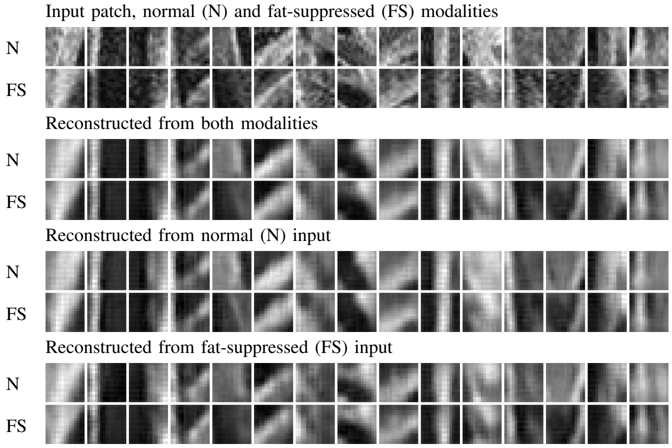

Input patch, normal (N) and fat-suppressed (FS) modalities
N
FS
Reconstructed from both modalities
N
FS
Reconstructed from normal (N) input
N
FS
Reconstructed from fat-suppressed (FS) input
N
FS

Fig. 5. Original input and reconstructions for 15 patches from the OAI dataset, showing the center slice from each 3D patch. The reconstructions are generated by an axial neural network trained with modality dropout, per-feature normalization and a weight $\alpha = 0.1$ for the similarity term. The first reconstruction is generated from the central representation computed from both input modalities. The second and third reconstructions are computed using the central representation from the normal input or the fat-suppressed input modality only.

V. Results

This section presents the results of our experiments. First, Section V-A presents the same-modality and cross-modality classification accuracy for the various models. This provides a overview of the performance of the proposed methods and that of the baseline methods. We present the results for both datasets, averaged over all five cross-validation folds.

Then, we take a closer look at the feature representations learned by each model. Section V-B shows the standard deviation, the cross-modality correlation, and the mutual information scores of the individual features. These metrics provide an insight into how the modality dropout, per-feature normalization and the similarity term influence the feature learning process. Since each network is initialized randomly and has a different feature representation, it is not possible to average the measurements for individual features over multiple cross-validation folds. We present the plots for one fold on the OAI dataset, but found similar results for the other OAI folds and on the BRATS dataset.

Finally, Section V-C tries to identify whether models learn mostly shared or mostly modality-specific features, We show the classification accuracy obtained using subsets of features with the highest cross-modality correlation. This section shows the results for the OAI dataset, averaged over all five folds.

A. Same-modality and cross-modality classification accuracy

Table I shows the classification accuracy for each combination of methods, measured on both datasets, as well as the

TABLE I
CLASSIFICATION ACCURACY OF RANDOM FORESTS TRAINED WITH FEATURES LEARNED USING THE AXIAL NEURAL NETWORK OR THE BASELINE MODELS, COMPARING THE ACCURACY IN SAME-MODALITY AND CROSS-MODALITY CLASSIFICATION. A COMBINATION OF MODALITY DROPOUT, PER-FEATURE NORMALIZATION AND THE SIMILARITY TERM GIVES THE BEST CROSS-MODALITY CLASSIFICATION PERFORMANCE. THE TABLE REPORTS THE MEAN CLASSIFICATION ACCURACY OVER FIVE CROSS-VALIDATION FOLDS AND OVER ALL MODALITY COMBINATIONS (NORMAL AND FAT-SUPPRESSED SCANS FOR OAI, T1/T1+C/T2/FLAIR FOR BRATS).

	Same-modality classification				**Cross-modality classification**			
OAI knee dataset	*Weight of the similarity term α*				*Weight of the similarity term α*			
Axial neural network	0.0	0.1	0.2	0.5	0.0	0.1	0.2	0.5
No modality dropout, no per-feature normalization	80.0	79.1	79.2	78.9	33.5	52.5	53.7	57.0
No modality dropout, with per-feature normalization	80.4	79.7	79.3	78.6	43.6	65.2	63.1	62.2
Modality dropout, no per-feature normalization	81.8	81.1	80.7	80.3	43.4	63.2	67.1	70.7
Modality dropout and per-feature normalization	81.6	81.5	81.1	80.7	77.0	78.7	78.7	78.2
Baseline network								
Features from all modalities	79.6				70.0			
Features from the training modality only	79.4				69.0			
	Same-modality classification				**Cross-modality classification**			
BRATS brain tumor dataset	*Weight of the similarity term α*				*Weight of the similarity term α*			
Axial neural network	0.0	0.1	0.2	0.5	0.0	0.1	0.2	0.5
No modality dropout, no per-feature normalization	73.0	71.9	71.5	69.8	50.2	51.8	51.7	51.4
No modality dropout, with per-feature normalization	72.6	73.4	73.9	72.9	52.2	55.9	57.5	60.0
Modality dropout, no per-feature normalization	77.5	76.9	76.8	76.5	51.7	55.3	55.9	57.4
Modality dropout and per-feature normalization	77.5	77.5	77.2	76.8	65.8	66.8	67.0	67.9
Baseline network								
Features from all modalities	69.4				54.9			
Features from the training modality only	70.0				55.1			

performance of the baseline methods on the same data. The table shows the average results over all modality pairs: the exact performance depends on which modalities are combined, because some modalities have more in common than others. However, the general pattern and the ordering of the methods were similar for all modality pairs.

The results show that the axial neural network with the additions discussed in this paper can provide much better cross-modality results than the baseline methods that do not take cross-modality differences into account. On both datasets, the baseline methods achieve a much lower accuracy in cross-modality classification than in same-modality classification. The axial neural network also shows a drop in performance going from same-modality to cross-modality classification, but this drop is much smaller. On the knee dataset, the best-performing axial neural network obtains a cross-modality accuracy that is very close to its same-modality accuracy. On the brain tumor dataset, the performance drop is larger, but the axial neural network still performs much better on cross-modality classification than the baseline methods.

Table I shows the results for axial networks with all combinations of the three techniques. The best cross-modality accuracy was obtained with a combination of modality dropout, per-feature normalization and the similarity term. Removing the similarity term from this combination of methods decreased the cross-modality performance only a little, suggesting that modality dropout and per-feature normalization are the most important.

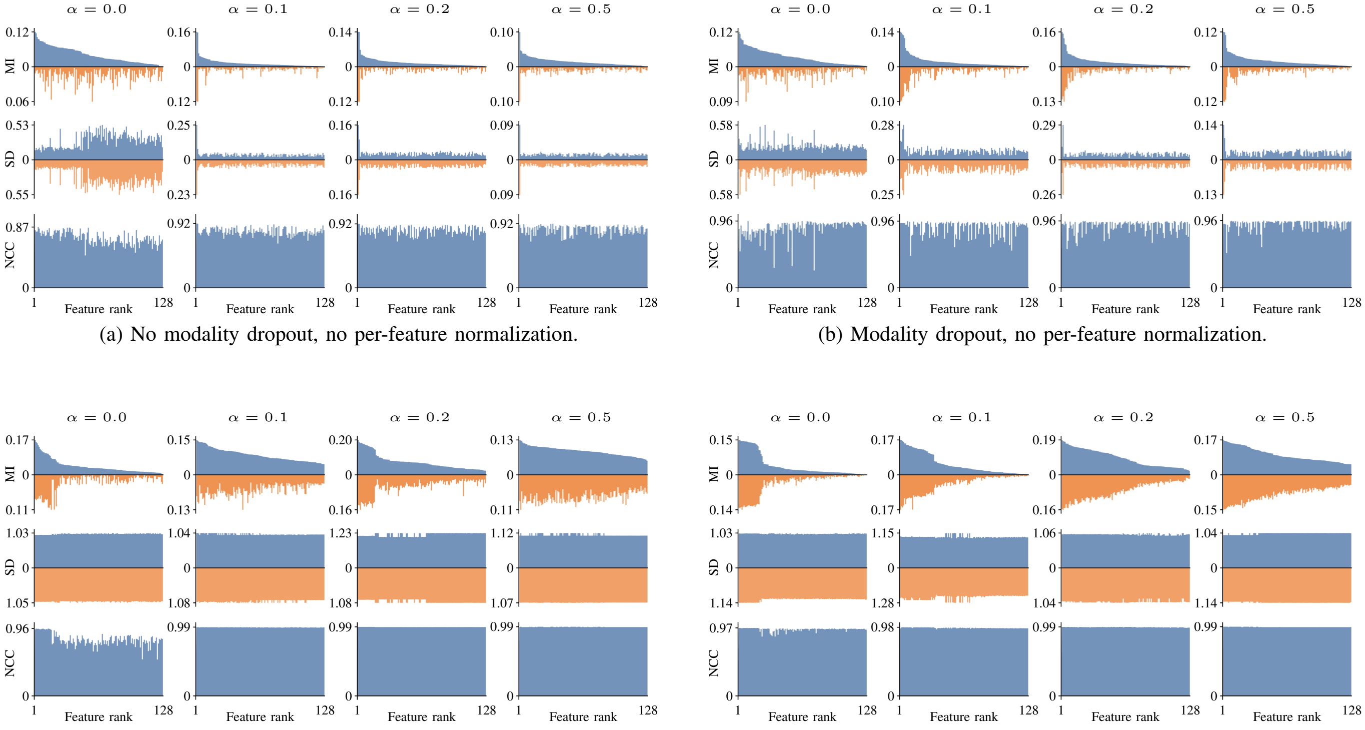
Comparing individual techniques over all different combinations, both modality dropout and per-feature normalization consistently provide an improvement of the classification accuracy. The contribution of the similarity term is less clear: it can give an important improvement if either modality dropout or per-feature normalization is missing, but if both are present the additional improvement of the similarity term is small. However, while the improvement from adding the similarity term might be large or small, it is usually positive: adding the similarity term with an appropriate weight never lead to a large decrease in same-modality or cross-modality performance.

To illustrate the reconstruction part of the network, Fig. 5 shows some of the reconstructions produced by the best-performing network for the OAI dataset. These reconstructions are not used for classification, which is based only on the central feature representation, but it is still useful to see that the network is able to reconstruct the main structures in the image and can also reproduce some of the inter-modality differences.

B. Feature characteristics

The second part of our investigation considers the information content of individual features. For each feature in each modality, Fig. 6 shows the mutual information score between the feature value and the class label, the standard deviation, and the normalized cross-modality correlation for each feature. The features are sorted by mutual information in the first modality: from the most informative feature on the left to the least informative on the right.

We interpret these plots by comparing the values of each feature in the two modalities: a cross-modality feature will



(a) No modality dropout, no per-feature normalization.

(b) Modality dropout, no per-feature normalization.

(c) No modality dropout, with per-feature normalization.

(d) Modality dropout, with per-feature normalization.

Fig. 6. Characteristics of the 128 features learned for one fold of the OAI knee data, for different network configurations. The features are sorted by decreasing mutual information scores. First row: mutual information (MI) between the feature values and the class label, when feature values are computed from normal (up/blue) or fat-suppressed (down/orange) scans. Second row: standard deviation (SD) of each feature. Third row: normalized cross-correlation (NCC) between the values computed from both sources. Without modality dropout or per-feature normalization (a) there is a large difference in values between modalities. Combining all three methods (d) produces features that are much more similar, which suggests that they might be more useful for cross-modality classification.

have a similar meaning in both modalities, and will show similar values in these plots. Conversely, if the plots show a large difference between the values for both modalities, the feature is unlikely to be useful for cross-modality classification.

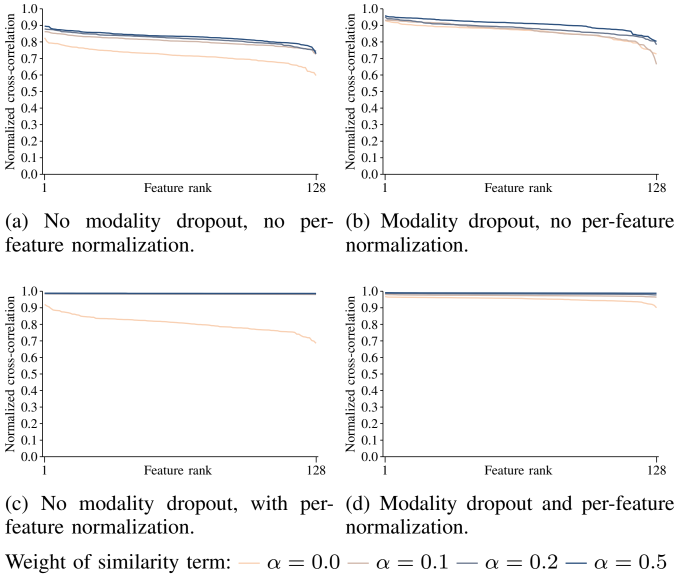
The most basic model, without modality dropout, per-feature normalization or similarity term, produces features that have very different standard deviations and mutual information scores in each modality (Fig. 6a). This suggests that this basic model learns some modality-specific features that are informative for one modality, but not for the other. Adding modality dropout, per-feature normalization and the similarity term makes the features much more similar across domains. The plots for the combination of all methods (Fig. 6d) show very similar values for the features in both modalities. This suggests that this model learns many cross-modality features, which is consistent with the good performance of this model observed in Section V-A.

The plots of the standard deviations in Fig. 6 also provide a further insight into the interaction between the similarity term and per-feature normalization. Because the similarity term attempts to reduce the difference between feature values for different modalities, it encourages the model to reduce the absolute feature values. This is visible in the plots of the standard deviation, which show that the similarity term reduces the standard deviation of the features. This reduction does not necessarily improve cross-modality correlation, but it does decrease the similarity term of the learning objective. Applying per-feature normalization prevents this problem: the improved normalization brought the standard deviation reasonably close to 1 for all features.

C. Classification accuracy for feature subsets

In the final part of our investigation, we look at the classification accuracy obtained using subsets of features, sorted by decreasing normalized cross-modality correlation. Figure 7 shows the cross-modality correlation of individual features, sorted in decreasing order, for the various models. Figure 8 shows the classification accuracy obtained using subsets of features with the highest cross-modality correlation. We show the results for the knee dataset only, but the results for the brain tumor dataset show similar patterns.

From Fig. 7 it becomes clear how the three techniques influence the cross-correlation of the features. For the basic model without modality dropout, without per-feature normalization and with a zero weight for the similarity term (Fig. 7a), the feature representation contains a combination of features with a reasonably high cross-modality correlation (0.9), as well as features that are less correlated across modalities (0.6). The optimal model that combines modality dropout, per-feature normalization and a non-zero weight for the similarity term (Fig. 7d), produces a feature representation in which all features have a high cross-modality correlation (values close to 1 for all features). The results for other models in Fig. 7



(a) No modality dropout, no per-feature normalization.

(b) Modality dropout, no per-feature normalization.

(c) No modality dropout, with per-feature normalization.

(d) Modality dropout and per-feature normalization.

Fig. 7. Normalized cross-correlation (NCC) of features on the OAI knee dataset, sorted from high to low correlation and averaged over five folds. The correlation is computed between corresponding features from both modalities. Combining modality dropout and per-feature normalization produced the most similar features.

show that all three techniques individually can improve the cross-modality correlation of the feature representation.

Figure 8 shows the same-modality and cross-modality classification accuracy for subsets of features with the highest cross-modality correlation: from only the most correlated feature on the left, to all features on the right. For same-modality classification (Fig. 8a–d, left column), the accuracy for most methods increases monotonically with the number of features. Adding more features improves the results, although the improvement becomes fairly small after a sufficient number of features have been added.

For cross-modality classification (Fig. 8a–d, right column), the accuracy does not increase monotonically, but first increases and then decreases again as features with a lower correlation are added. The low cross-modality correlation indicates that these features have a different meaning in each modality, which will confuse a cross-modality classifier. However, the proposed techniques can alleviate this problem. For the combination of modality dropout and per-feature normalization (Fig. 8d), the cross-modality classification accuracy increases monotonically with the number of features. Including the similarity term leads to an earlier peak in the classification accuracy. This is consistent with the high cross-modality correlations of all features (Fig. 7d), which indicates that this combination of methods learns mostly cross-modality features. For the other models, there is a larger range of cross-modality correlations (e.g., Fig. 7a), which together with the decrease in accuracy suggests that these models learn a mixture of modality-specific and shared features.

VI. Discussion

We evaluated three strategies to improve cross-modality feature learning in an axial neural network: modality dropout, per-feature normalization, and a similarity term. The best results were obtained using a combination of all three methods (Table I). For both of our datasets, the features learned using this combination of techniques resulted in the best cross-modality classification accuracy, without affecting the same-modality classification accuracy too much. The cross-modality classification accuracy obtained using this combination of methods was higher than that obtained with the baseline method, a similar feature-learning model that used the same transformation for all modalities.

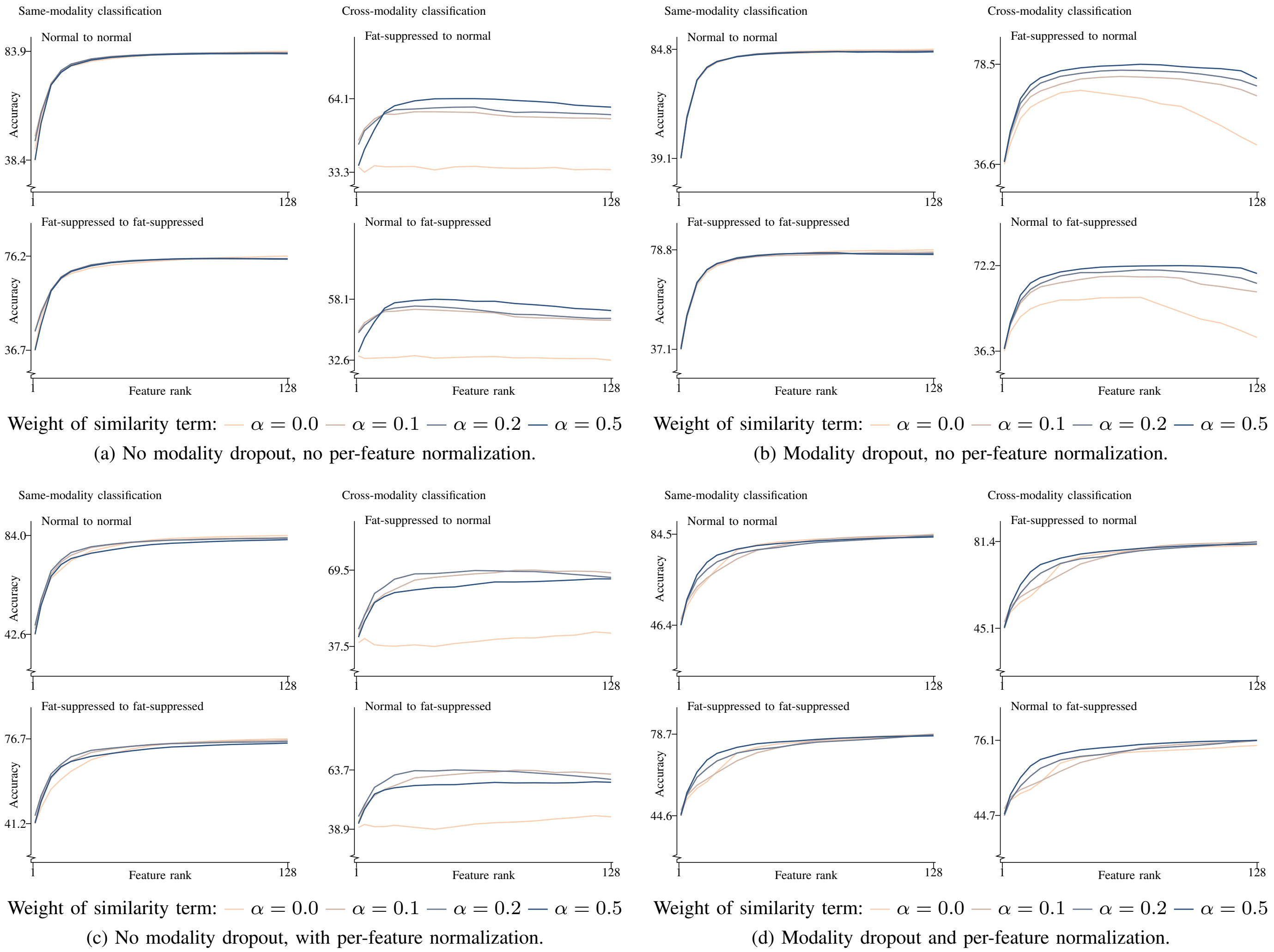
A. Comparing the three techniques

Modality dropout improved the accuracy of the axial neural network in both same-modality and cross-modality classification experiments (Table I), perhaps because it explicitly trains the model to work well in both scenarios. Modality dropout forces the model to reconstruct the target modality from itself, which is useful for same-modality classification. It also forces the model to reconstruct the target modality from another modality, which helps the cross-modality case because it forces the network to learn representations that encode sufficient cross-modality information and prevents it from depending too much on one modality.

The second important factor was per-feature normalization (Table I). Forcing the features to have a similar mean and standard deviation in all modalities turns out to be an effective way of minimizing cross-modality representation differences (Fig. 6). It prevents the network from learning features that are used for one modality but not for another, and it simplifies the optimization by removing part of the cross-modality differences. The features learned with per-feature normalization also had higher mutual information scores, suggesting that they contained more discriminative information.

Adding a similarity term in the objective function had a positive influence on the cross-modality classification accuracy, but the strength of this influence depended on whether it was combined with the other techniques (Table I). For the combination of modality dropout and per-feature normalization, the additional effect of the similarity term was fairly small: the results of this combination were only slightly better if the similarity term was included. This was different for all other combinations: there, increasing the weight of the similarity term produced more similar features and a better cross-modality classification accuracy. This suggests that the combination of modality dropout and per-feature normalization is powerful enough to remove most of the need for the extra similarity term, but that the term can still have a positive effect in other cases. It is important, however, to limit the weight of the similarity term: setting it too close to 1 can cause the model to learn trivial, non-informative and non-discriminative features [2].

Because the similarity term tries to minimize the absolute difference between feature representations, it also tends to reduce the absolute feature values. This is visible in Fig. 6: the features learned with the similarity term have lower standard deviations than the features learned without the term. Per-feature normalization counters this side-effect and stabilizes the feature values, improving the accuracy in the process.



(a) No modality dropout, no per-feature normalization.

(b) Modality dropout, no per-feature normalization.

(c) No modality dropout, with per-feature normalization.

(d) Modality dropout and per-feature normalization.

Fig. 8. Classification accuracy on the knee dataset, for models trained with and without modality dropout and per-feature normalization. The horizontal axes indicate the number of selected features, starting with the features with the highest cross-correlation. All plots use the same scale; vertical tick marks indicate the minimum and maximum accuracy in each plot.

It is important to note that, while the three methods greatly improved the cross-modality performance, they also maintained most of the original same-modality performance (Table I). This is a useful property if the same representation is used for both same-modality and cross-modality classification.

B. Modality-specific vs. shared features

One of the hypotheses behind our experiments was that the models might learn a combination of modality-specific and shared features. Shared features are good for cross-modality classification, but the models might still learn modality-specific features to preserve crucial modality-specific information. To investigate this further, we tried to separate shared and modality-specific features by sorting the features based on the cross-modality correlation (Fig. 7) and training on subsets of highly correlated features. Our approach produced different results for each of the methods (Fig. 8). For the best combination of modality dropout, per-feature normalization and the similarity term, we found that almost all features had a very strong cross-modality correlation, and that the classification performance improved monotonically with the number of features. This suggests that this combination of methods learned mostly cross-modality features. The other methods produced both highly correlated features and features that had a lower cross-modality correlation. In these cases, although the same-modality accuracy increased as we added more features, we obtained the best cross-modality accuracy by training on a smaller subset of highly correlated features. This suggests that these representations contained not only shared but also modality-specific features, which help same-modality classification but can harm the cross-modality case.

Although shared representations may be best for cross-modality classification, preserving modality-specific information is important for the same-modality performance. This is somewhat reflected by the results of our baseline methods (Table I), which show that features learned for one specific modality give a slightly better same-modality accuracy. In applications where the same-modality performance is as important as the cross-modality performance, it may be useful to give the model a way to preserve modality-specific features without including them in the shared representation. One way to do this could be to reserve a separate, modality-specific part of the representation that is only used for a single modality.

C. Data requirements

The approach discussed in this paper makes some assumptions about the data and the problem to which the methods are applied. Firstly, the approach assumes that data is available for all modalities and that at least some of this (unlabeled) data is registered with a voxelwise correspondence. The axial neural network learns its feature representation from corresponding patches, which represent the same physical area in each modality. For models with more than two modalities, it is not strictly necessary to have all modalities available for all subjects: as the modality dropout method shows, it is possible to train with patches for which only a subset of modalities is available.

Secondly, learning a shared representation for multiple modalities assumes that the modalities have something in common. Because our model learns a separate transformation for each modality, it can handle large differences between modalities. However, the shared representation can only preserve and transfer information that is available in all modalities: if a modality provides information that is not visible in the other modalities, this information can not be used in cross-modality classification. The performance of the proposed methods depends therefore on the type of problem and on the differences between the modalities. If the modalities are very different and the modality-specific information has important discriminative value, removing it from the shared representation may reduce the same-modality classification performance.

In our experiments, the modalities in the knee dataset have more in common than the modalities in the brain tumor dataset. The knee images have a different resolution and have different intensities for some of the structures, but the image structures that are important for classification are recognizable in both images. As a result, the cross-modality classification performance on this dataset comes fairly close to that in the same-modality case. This suggests that transfer learning could be successful in this scenario. In the brain tumor dataset, the four modalities have larger differences, and some tumor structures are clearly visible in some images but not in others (Fig. 4b). In our cross-modality classification experiments, this meant that the cross-modality classification accuracy was noticeably lower than the same-modality accuracy. The performance differed per modality: in our classification task, T2 and FLAIR gave much better results than T1 and contrast-enhanced T1. This preference for T2 and FLAIR is most likely an artifact of how we grouped the tumor components into a single class, and would be different when classifying other components (for example, contrast-enhanced T1 would be important for identifying the necrotic core, which we grouped with the other tumor components in our experiments). While our experiments clearly show the potential of our method as a transfer learning method, accurate tumor classification in this dataset will require the use of multiple modalities. However, it might be possible to use transfer learning between pairs of modalities that together contain sufficient information (e.g., T1/T2 and T1+c/FLAIR).

D. Remaining thoughts

In this paper, we used autoencoder-like models to learn features without discriminative training. The advantage of this approach is that the representation learning does not require labels, only paired scans. Labels are required for training the classifier, but they can also come from unpaired scans from only one of the modalities. A disadvantage of this unlabeled feature learning is that the representations may contain some features that have no discriminative value, but are needed to compute the reconstruction. An alternative network that combines feature learning and classification might be able to obtain a better performance by focusing only on discriminative features. Although this is outside the scope of this paper, the approaches discussed here could also be applied to such classification networks.

The axial neural network discussed in this paper learns a separate transformation for each modality, as opposed to models that use a single tranformation for all modalities (such as our baseline methods). Single-transformation models essentially learn to extract modality-invariant features with transformations that are insensitive to the source modality, which limits them to features that can be extracted in the same way from all modalities. In contrast, multi-transformation models such as ours learn a shared feature representation by learning modality-specific transformations. This is a more flexible approach that can, in theory at least, extract any information that is common to all modalities.

Since this paper is focused on analyzing cross-modality classification, and not on finding the best knee cartilage or brain tumor segmentation segmentation method per se, it is difficult to compare our results with those of state-of-the-art approaches. Many knee cartilage segmentation methods use shape-based post-processing methods [25]. Brain tumor segmentation methods, such as those for the BRATS challenge [17], generally use multi-modal information to get good classification results. The results of these more specialized methods are better than those presented in this paper.

VII. Conclusion

Differences in appearance make it difficult to apply a classifier trained on data from one source to data from another. The proposed representation learning method attacks this problem by transforming data from different sources to a shared feature representation. We found that this yields both modality-specific and cross-modality features. The basic axial neural network architecture can be extended with three methods that further improve cross-modality performance. Modality dropout trains the network by randomly removing some modalities during training, which forces the model to learn cross-modality reconstructions. Per-feature normalization improves cross-modality similarity by normalizing all features to zero mean and unit standard deviation. A similarity term explicitly adds cross-modality similarity to the learning objective of the network. Based on our experiments on two different datasets, we found that modality dropout and per-feature normalization are crucial to maximize the number of cross-modality features and obtain the best cross-modality classification results. The similarity term has a strong influence in models without either modality dropout or per-feature normalization, but has only a minor positive contribution if both other techniques are used.

Acknowledgments

This research is financed by the Netherlands Organization for Scientific Research (NWO). We used supercomputer facilities sponsored by NWO Exact and Natural Sciences and used public data from the Osteoarthritis Initiative (OAI) and the Brain Tumor Segmentation Challenge (BRATS).

References

[1] J. Ngiam, A. Khosla, and M. Kim, "Multimodal deep learning," in *ICML*, 2011.

[2] G. van Tulder and M. de Bruijne, "Representation Learning for Cross-Modality Classification," in *MICCAI Workshop on Medical Computer Vision*, 2016.

[3] ——, "Why Does Synthesized Data Improve Multi-sequence Classification?" in *MICCAI*, 2015, pp. 531–538.

[4] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-Modal Image Segmentation," in *MICCAI*, vol. 9900, 2016, pp. 469–477.

[5] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, "Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities," in *MICCAI*, vol. 9900, 2016, pp. 478–486.

[6] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-Modal Retrieval via Deep and Bidirectional Representation Learning," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.

[7] S. Rastegar, M. S. Baghshah, H. R. Rabiee, and S. M. Shojaee, "MDL-CW: A Multimodal Deep Learning Framework with Cross Weights," in *CVPR*, 2016, pp. 2601–2609.

[8] L. Zhao, Q. Hu, and W. Wang, "Heterogeneous Feature Selection With Multi-Modal Deep Neural Networks and Sparse Group LASSO," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1936–1948, Nov 2015.

[9] F. Feng, X. Wang, R. Li, and I. Ahmad, "Correspondence Autoencoders for Cross-Modal Retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 1s, pp. 1–22, 2015.

[10] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *Advances in Neural Information Processing*, 2012.

[11] V. Vukotić, C. Raymond, and G. Gravier, "Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16*. New York, New York, USA: ACM Press, 2016, pp. 343–346.

[12] M. Moradi, Y. Guo, Y. Gur, M. Negahdar, and T. Syeda-Mahmood, "A Cross-Modality Neural Network Transform for Semi-automatic Medical Image Annotation," in *MICCAI*, vol. 9901, 2016, pp. 300–307.

[13] Z. Zhang, L. Yang, and Y. Zheng, "Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network," in *CVPR*, 2018.

[14] K. Kamnitsas, C. Baumgartner, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *IPMI*, 2017.

[15] J. Folkesson, E. B. Dam, O. F. Olsen, P. C. Pettersen, and C. Christiansen, "Segmenting articular cartilage automatically using a voxel classification approach," *IEEE Transactions on Medical Imaging*, vol. 26, no. 1, pp. 106–115, 2007.

[16] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *MICCAI*, vol. 8150, 2013, pp. 246–253.

[17] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, R. Meier, D. Precup, S. J. Price, T. Riklin-Raviv, S. M. S. Reza, M. Ryan, L. Schwartz, H.-C. Shin, J. Shotton, C. a. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct 2015.

[18] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle, "Brain tumor segmentation with Deep Neural Networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[19] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *ICML*, 2015.

[20] C. G. Peterfy, E. Schneider, and M. Nevitt, "The Osteoarthritis Initiative: Report on the design rationale for the magnetic resonance imaging protocol for the knee," *Osteoarthritis and Cartilage*, vol. 16, no. 12, pp. 1433–1441, 2008.

[21] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. Hopcroft, "Convergent Learning: Do different neural networks learn the same representations?" in *ICLR*, 2016.

[22] F. Chollet, "Keras," 2017. https://keras.io

[23] The Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," Tech. Rep., 2016.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct 2011.

[25] A. Aprovitola and L. Gallo, "Knee bone segmentation from MRI: A classification and literature review," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 2, pp. 437–449, 2016.