Synergistic Combination of Learned and Hand-Crafted Features for Prostate Lesion Classification in Multiparametric Magnetic Resonance Imaging

Davood Karimi(✉) and Dan Ruan

Department of Radiation Oncology, University of California, Los Angeles, USA
davoodkarimi@ucla.edu

Abstract. In this paper, we propose and evaluate a new method for classifying between malignant and benign prostate cancer lesions in multiparametric magnetic resonance imaging (MRI). We show that synergistically combining automatically-learned and handcrafted features can significantly improve the classification performance. Our method utilizes features extracted from convolutional neural networks (CNNs), texture features learned via a discriminative sparsity-regularized approach, and hand-crafted statistical features. To assess the efficacy of different feature sets, we use AdaBoost with decision trees to classify prostate cancer lesions using different sets of features. CNN-derived, texture, and statistical features achieved area under the receiver operating characteristic curve (AUC) of 0.75, 0.68, and 0.70, respectively. Augmenting CNN features with texture and statistical features increased the AUC to 0.84 and 0.82, respectively. Combining all three feature types led to an AUC of 0.87. Our results indicate that in medical applications where training data is scarce, the classification performance achieved by CNNs or sparsity-regularized classification methods alone can be sub-optimal. Alternatively, one can treat these methods as implicit feature extraction mechanisms and combine their learned features with hand-crafted features using meta-classifiers to obtain superior classification performance.

1 Introduction

Detection and classification of prostate carcinoma is important. In recent years, multi-parametric magnetic resonance imaging (MRI) has been utilized to provide comprehensive knowledge for such decision. Differentiation between benign and malignant tumors is complex and combines clinical clues from the size, shape, intensity, texture, and the zonal location of the lesion [1]. Computer-aided methods have been intensively investigated and a wide range of image pre-processing, feature extraction, and classification approaches have been proposed [5].

Most published methods compute features that describe the intensity distribution, shape, texture, and other characteristics of the lesion, and train a classifier based on these features [5]. Recent studies have shown that the deep learning

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 391–398, 2017.
DOI: 10.1007/978-3-319-66179-7_45

networks can achieve state of the art results in cancer detection and classification [2,8]. In the absence of prior knowledge, these networks rely on large datasets to automatically and implicitly learn the features that are relevant to the detection/classification task without the need for feature engineering. However, very few studies have applied these methods for prostate cancer diagnosis in MRI.

In this paper, we use convolutional neural networks (CNNs) as an automatic feature extraction tool, as opposed to its native role of a standalone classifier. In addition, we suggest using a discriminative sparsity-regularized method for learning texture features. We show that with small datasets, such as the one here, these methods on their own do not achieve a high classification performance. To improve their performance, we suggest training a meta-classifier on these features as well as statistical features. We show that this approach leads to superior classification results.

2 Materials and Methods

2.1 Data

The data used in this study consisted of T2w, apparent diffusion coefficient (ADC), and dynamic contrast enhanced (K^{trans}) images of 203 patients. A detailed description of the data can be found in [6]. This dataset was released for PROSTATEx Challenge as part of the 2017 SPIE Medical Imaging Symposium. The winning method in that challenge achieved an area under the receiver operating characteristic curve (AUC) of 0.87 on a separate test dataset. A total of 330 lesions (254 benign and 76 malignant) are marked on these images. We used 232 lesions (182 benign and 50 malignant) for training and 98 lesions (72 benign and 26 malignant) for test. The assignment of each of the 330 lesions to test/train sets was random, except that we tried to have an equal distribution of lesions in terms of zonal location in test/train sets.

2.2 Deep Learning Features Derived from CNN

With large number of parameters, deep CNNs require large amounts of training data and sophisticated training methods to avoid overfitting. This poses a serious difficulty for medical image analysis applications where there is usually limited training data. Various strategies such as data augmentation, transfer learning, and regularization are used to mitigate this problem. In this study, we used a CIFAR-net [4], which we pre-trained on the CIFAR-10 dataset. We adopted the 2.5D-view representation [9] to fit into the three RGB input channels of the CIFAR-net. For data augmentation, we used translation (by shifting the sampling grid by 2, 4, and 6 mm in the three orthogonal directions) and scaling (by using patches of side length equal to 18, 21, 24, 27, and 30 mm). These patch sizes were selected based on the typical size of prostate lesions, which is 5–20 mm [6], and the fact that in typical CNNs center pixels have a greater influence on the network output than border pixels do. All patches were 25×25 pixels.

We modified the network pre-trained on CIFAR-10 dataset such that the number of neurons in the fully-connected layers were 64 and 10, and then fine-tuned it for prostate lesion classification. We chose the sizes of the fully-connected layers empirically after experimenting with different sizes. The output layer had two units, for benign/malignant probabilities. During fine-tuning, we used a learning rate of 0.001 for the fully connected layers, which we learned from scratch, and 0.0001 for the convolutional layers, which were initialized from those learned on the CIFAR-10 dataset. Regularization on the network weights had little effect on the classification performance, but dropout significantly improved the classification results with the best results obtained using a drop-out rate of 85%. Three separate CNNs were trained on T2w, ADC, and K^{trans} respectively.

To utilize CNN as a feature extraction vehicle rather than its conventional capacity as an end-to-end classifier, we used the values of the 10 neurons on the last fully-connected layers. It is reasonable to believe that these features encode the critical information for classification, since a conventional CNN effectively outputs a linear combination of them. We hypothesize that by combining these features with potentially other sources of knowledge and employing a stronger assembly classifier, better performance may be achieved.

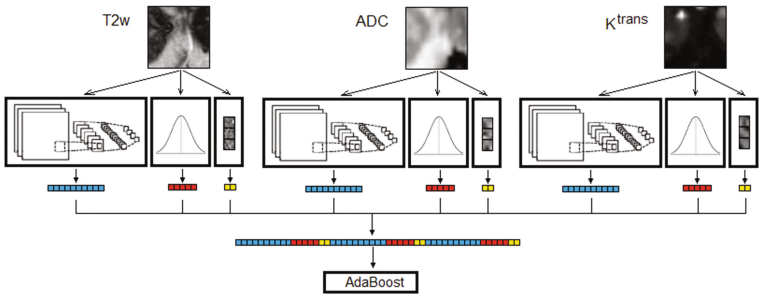

Fig. 1. A schematic depiction of how a feature vector is created from the three images of a lesion. The vector of CNN-derived (blue), statistical (red), and texture (yellow) features is used by the AdaBoost classifier to predict the probability of lesion malignancy. For each lesion, the final malignancy probability is computed as the average of the probabilities computed for its 1715 feature vectors for different scales and translations.

2.3 Hand-Crafted Statistical Features

Local image statistics in multiparametric MRI are known to be clinically useful for prostate cancer diagnosis [1,3,5]. We computed the mean, standard deviation, and 25, 50, and 75 percentiles of image intensity in cubes of side length equal to 12, 14, 16, 18, and 20 mm, and for shifts of the cube center with respect to the lesion center equal to 2, 4, and 6 mm in all three directions.

2.4 Texture Features Learned Using Discriminative Dictionaries

Various types of texture features such as wavelet-based, fractal, and Haralick features have been used for prostate cancer diagnosis in MRI [5]. We propose to design features with classification in mind and use a discriminative sparsity-regularized method. This is a more sound approach than using off-the-shelf texture features since the features are learned such that they have a high discriminative power. We followed an approach similar to that in [7]. Given sets of training patches from benign and malignant classes, denoted as C_b and C_m respectively, we learn dictionaries D_b and D_m for sparse representation of these patches by solving the following optimization problem:

$$\underset{D_b, D_m}{\text{minimize}} \sum_{\substack{j \in \{b,m\} \\ i \in C_j}} \left(\beta r_j^i + \log \left(e^{\lambda(r_j^i - r_b^i)} + e^{\lambda(r_j^i - r_m^i)} \right) \right), \tag{1}$$

where r_j^i is the residual of sparse representation of training patch x_i in dictionary D_j, computed as:

$$r_j^i = \|x_i - D_j \gamma_j^i\|_2^2 \quad \text{where} \quad \gamma_j^i = \underset{\gamma}{\text{argmin}} \, \|x_i - D_j \gamma\|_2^2 \quad \text{s.t.:} \quad \|\gamma\|_0 \leq T. \tag{2}$$

In the above equation, T is the sparsity level. The first term in the sum in Eq. (1) penalizes the residual of sparse representation of each patch in its corresponding dictionary. The second term encourages "discriminative" dictionaries: it gives preference to a dictionary that has high residual in representing patches from the other class. In Eq. (1), λ determines the sharpness of the discriminative function. A larger λ will increase the penalty for misclassified training patches, but also makes the optimization less stable by increasing the maximum slope of the cost function. The regularization parameter β balances the constructive and discriminative terms of the cost function. We minimize the cost function alternately with respect to the dictionaries and sparse coefficients. For a stable learning, we begin the optimization with a small $\lambda = 10^{-6}$ and large $\beta = 10^{-1}$ to give more weight to the representation aspect of the dictionaries in the early iterations, and gradually increase λ and decrease β such that by iteration 100 (last iteration) we have $\lambda = 10^{-4}$ and $\beta = 10^{-2}$.

We used 7×7-pixel patches at different scales with physical side lengths of 12, 14, 16, 18, and 20 mm sampled on the axial slice, and for shifts of patch center with respect to the lesion center of 2, 4, and 6 mm, similar to the CNN and statistical features. Our trained dictionaries each had 100 atoms and we used a sparsity level of $T = 5$. A patch sampled from a malignant lesion will, in general, have a smaller residual in D_m than in D_b. We use the residuals of a patch in the two dictionaries (i.e., r_b^i and r_m^i) as texture features.

2.5 Combining Different Types of Features

For each lesion, $5 \times 7^3 = 1715$ feature vectors were computed, for 5 different scales and 7 different translations in the three orthogonal directions. Each

feature vector contained $(10 + 5 + 2) \times 3 = 51$ features, for 10 CNN-derived, 5 statistical, and 2 texture features, each computed on all three images. Figure 1 shows schematically how a feature vector is created. We used AdaBoost with decision trees as a meta-classifier to predict the probability of malignancy for each feature vector. For each lesion, the malignancy probability is obtained as the average of the probabilities of all its 1715 feature vectors. A total of 200 decision trees, each with a maximum depth of 5, were used.

3 Results and Discussion

Table 1 shows the AUC for AdaBoost classifiers built using different sets of features. For benchmark, we performed conventional CNN classification and the AUCs were 0.64, 0.70, and 0.69, respectively, for T2w, ADC, and K^{trans}. Table 1 shows the results for each MR image separately and also for the combination of all images. In addition, Fig. 2 shows the receiver operating characteristic (ROC) curve for the AdaBoost classifiers trained on certain combinations of the feature sets. The results presented in this section, including those in Table 1 and Fig. 2 are for classification of the 98 lesions in the test set. The classification results for the training set were in general slightly better than for the test set.

Table 1. Lesion-based AUC obtained with AdaBoost classifiers trained on different sets of feature.

	CNN-derived	Statistical	Texture	All features
T2w	0.65	0.54	0.66	0.71
ADC	0.72	0.65	0.59	0.74
K^{trans}	0.72	0.68	0.56	0.75
All three images	0.75	0.70	0.68	0.87

As shown in Table 1, features learned by CNN achieved better results than statistical and texture features, which have been used by most published studies. The first column in Table 1 using CNN-derived features with Adaboost showed slightly better performance than the native CNN, benefiting from the stronger classifier as we conjectured. Using AdaBoost on CNN+statistical and CNN+texture features resulted in AUCs equal to 0.82 and 0.84, respectively. Combining all three feature types resulted in an AUC of 0.87, which is much better than 0.75 obtained when using only CNN-derived features. This indicates that when the training data is limited, the classification performance of CNNs can be greatly improved by introducing additional features. It is interesting to note that even though texture features on their own achieved a lower AUC than statistical features (0.68 versus 0.70), they complement CNN-derived features better (0.84 versus 0.82).

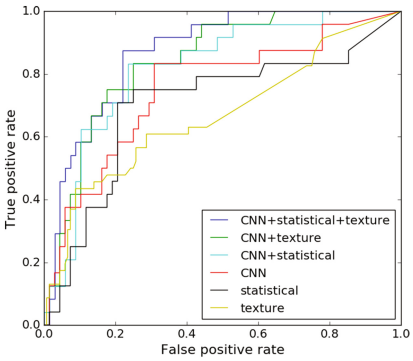

Fig. 2. The lesion-based receiver operating characteristic curve for AdaBoost classifiers built using different sets of features.

Figure 3 shows the dictionaries learned from benign and malignant lesions for each of the three images. It also shows the histograms of the difference between the residuals of 1000 benign and malignant patches in the two dictionaries (i.e., $r_b - r_m$). Since the dictionaries are overcomplete (each dictionary is composed of 100 atoms of size 49), there is much visual similarity between them. For a quantitative comparison, we computed the following dictionary similarity measure, which is closely related to the dissimilarity measure proposed in [10].

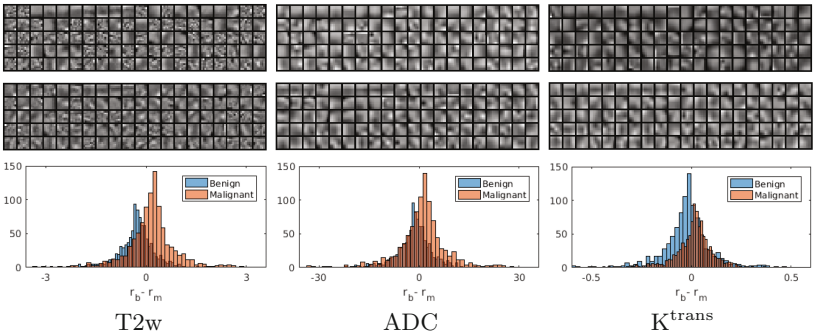

Fig. 3. The dictionaries learned for benign (top) and malignant (middle) prostate lesions. The bottom row shows the histograms of the difference in the residual of 1000 benign and malignant patches in the two dictionaries.

$$\text{similarity}(D_b, D_m) = \frac{1}{n_b + n_m}\Big(\sum_{i=1}^{n_b} \text{s}(D_m, d_b^i) + \sum_{i=1}^{n_m} \text{s}(D_b, d_m^i)\Big), \tag{3}$$

where $\text{s}(D_m, d_b^i)$ is the inner product between the i^{th} atom of D_b and its most similar atom in D_m, and similarly for $\text{s}(D_b, d_m^i)$. All dictionary atoms have unit

Euclidean norm. n_b and n_m denote the size of the dictionaries. The similarity of D_b and D_m for T2w, ADC, and K^{trans}, were 0.64, 0.80, and 0.78, respectively. In comparison, dictionaries learned without the dicsriminative term (i.e., only using the first term in Eq. (1)) had similarities of 0.74, 0.85, and 0.86, respectively.

Deep learning networks are meant to automatically learn the most relevant features for the task at hand. In addition to a proper design of the network architecture and optimization algorithm, the success of deep learning methods depends on the availability of large amounts of training data, which poses a major challenge in applying these methods in medical image analysis. Using data augmentation and transfer learning, some studies have recently fine-tuned larger CNNs such as AlexNet and GoogLeNet for cancer lesion classification [8]. However, even with these strategies, AlexNet did not achieve better results than CIFAR-net in this study. This indicates that the limited data sizes as in this project may prevent CNNs from achieving their full capacity in prostate lesion classification, unlike general-purpose image processing tasks. Furthermore, such gap may not be addressed with the use of larger networks. On the other hand, our results show that an effective approach to improving the CNN performance is to consider the CNN as no other than a feature-supplying module and augment the CNN-derived features with other prior-knowledge bearing features.

Our results agree with the published research that multi-parametric MRI can achieve much better classification results than any single MR image [1,5]. Our classification method achieved slightly better results with ADC or K^{trans} images than with T2w image, but the classification performance was much better when all three images were used. An unbiased comparison of our results with previously published methods is difficult. A recent review of the literature found that the AUCs reported by different studies ranged from 0.71 to 0.97, and depended on many factors including the number and type of MR images used, data size, and scanner type (1.5 Tesla or 3.0 Tesla) [5]. In addition, it would be intriguing to investigate the performance of our proposed method in determining the tumor grade. Unfortunately, this information was not available for this dataset at the time of writing of this manuscript and would be a future research topic.

It may be possible to further improve the performance of our proposed method. One way is by including the information regarding the zonal location of the lesion. The appearance of malignant and benign prostate lesions is known to depend on the prostate zone [1,3]. In the current study we did not learn separate classifiers for different zones because the data was unbalanced in their distribution of zonal location. Since we used AdaBoost with decision trees as our final classifier, we hoped that the difference between zones could be taken into account. Nonetheless, we think that better results may be achievable if enough training data is available such that a different classifier is trained for each prostate zone or the zonal location is included in classification as a feature.

4 Conclusion

Deep CNNs have achieved state of the art results in many applications in computer vision and image processing when training data is abundant. However,

their application to medical images is challenged by the scarcity of labeled training data. The same is true, albeit to a lesser degree, for methods that rely on learning overcomplete dictionaries, such as the method that we used in this study for learning texture features. In this paper we showed that, given the current size of available training data, these methods could not achieve a high performance in classification of prostate cancer lesions on multiparametric MRI. To tackle this problem, we suggested synergistically combining the information contained in these features as well as additional statistical features using a meta-classifier. Our results showed that an AdaBoost classifier trained on these combined features could achieve a significantly higher classification performance. The proposed method may be useful for many similar applications in medical image analysis where training data is limited.

References

1. Barentsz, J.O., Richenberg, J., Clements, R., Choyke, P., Verma, S., Villeirs, G., Rouviere, O., Logager, V., Fütterer, J.J.: ESUR prostate MR guidelines 2012. Eur. Radiol. **22**(4), 746–757 (2012)
2. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017)
3. Kitajima, K., Kaji, Y., Fukabori, Y., Yoshida, K.L., Suganuma, N., Sugimura, K.: Prostate cancer detection with 3T MRI: comparison of diffusion-weighted imaging and dynamic contrast-enhanced MRI in combination with T2-weighted imaging. J. Magn. Reson. Imaging **31**(3), 625–631 (2010)
4. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
5. Lematre, G., Mart, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F.: Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. Comput. Biol. Med. **60**, 8–31 (2015)
6. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H.: Computer-aided detection of prostate cancer in MRI. IEEE Trans. Med. Imaging **33**(5), 1083–1092 (2014)
7. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Discriminative learned dictionaries for local image analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
8. Roth, H.R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., Kim, L., Summers, R.M.: Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Trans. Med. Imaging **35**(5), 1170–1181 (2016)
9. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 520–527. Springer, Cham (2014). doi:10.1007/978-3-319-10404-1_65
10. Skretting, K., Engan, K.: Learned dictionaries for sparse image representation: properties and results. In: SPIE Optical Engineering+Applications, p. 81381N. International Society for Optics and Photonics (2011)