Bridging Computational Features Toward Multiple Semantic Features with Multi-task Regression: A Study of CT Pulmonary Nodules

Sihong Chen[1], Dong Ni[1], Jing Qin[2], Baiying Lei[1], Tianfu Wang[1], and Jie-Zhi Cheng[1(✉)]

[1] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University, Shenzhen, China
jzcheng@ntu.edu.tw
[2] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Abstract. The gap between the computational and semantic features is the one of major factors that bottlenecks the computer-aided diagnosis (CAD) performance from clinical usage. To bridge such gap, we propose to utilize the multi-task regression (MTR) scheme that leverages heterogeneous computational features derived from deep learning models of stacked denoising autoencoder (SDAE) and convolutional neural network (CNN) as well as Haar-like features to approach 8 semantic features of lung CT nodules. We regard that there may exist relations among the semantic features of "spiculation", "texture", "margin", etc., that can be exploited with the multi-task learning technique. The Lung Imaging Database Consortium (LIDC) data is adopted for the rich annotations, where nodules were quantitatively rated for the semantic features from many radiologists. By treating each semantic feature as a task, the MTR selects and regresses the heterogeneous computational features toward the radiologists' ratings with 10 fold cross-validation evaluation on the randomly selected LIDC 1400 nodules. The experimental results suggest that the predicted semantic scores from MTR are closer to the radiologists' rating than the predicted scores from single-task LASSO and elastic net regression methods. The proposed semantic scoring scheme may provide richer quantitative assessments of nodules for deeper analysis and support more sophisticated clinical content retrieval in medical databases.

Keywords: Multi-task regression · Lung nodule · CT · Deep learning

1 Introduction

The semantic features like the "spiculation", "lobulation", etc., are commonly used to describe the phenotype of a pulmonary nodule in the radiology report. For the differential diagnosis of pulmonary nodules in the CT images, the semantic spiculation feature and the high-level texture feature of nodule solidness are suggested to be important factors for the identification of malignancy in several diagnostic guidelines

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 53–60, 2016.
DOI: 10.1007/978-3-319-46723-8_7

[1, 2]. In the context of computer-aided diagnosis (CAD), several methods also attempted to computationally approximate some high-level semantic features to achieve the classification tasks [3–6]. For examples, in [4] the partial goal of the bag-of-frequencies descriptor was to classify 51 spiculated and 204 non-spiculated nodules in the CT images, whereas the nodule solidness categorization method was developed in [5] based on the low-level intensity features. In general, most of these works simply focused on the elaboration of single semantic feature for the discrete trichotomous/dichotomous nodule classification of malignancy, spiculation and solidness [3–6]. There is scarcely any work that has ever attempted to quantify the degrees of these high-level features to support deeper nodule analysis. Since a pulmonary nodule can be profiled with several semantic features, there may exist some kinds of relation among the semantic features. In this paper, we aim to address two specific problems: (1) degree quantification of the semantic features and (2) jointly mapping the computational image features toward the multiple semantic features. Distinct from the traditional CAD scheme that only suggests malignancy probability [3, 6], the proposed nodule profiling scheme may provide broader quantitative assessment indices in the hope to get closer to the clinical usage.

The thoracic CT dataset from the Lung Image Database Consortium (LIDC) [7] is adopted here for the rich annotation resources from many radiologists across several institutes in U.S.A. A nodule with diameter larger than 3 mm was annotated by radiologists to give their ratings for the semantic features of "spiculation", "lobulation", "texture", "calcification", "sphericity", "subtlety", "margin", "internal structure", and "malignancy". The exemplar nodules of each semantic features are shown in Fig. 1. The "malignancy" is excluded in this study as it relates to diagnosis. Most semantic features were scored in the range of 1–5, excepting the "internal structure" and "calcification" that were scored in the ranges of 1–4 and 1–6, respectively.

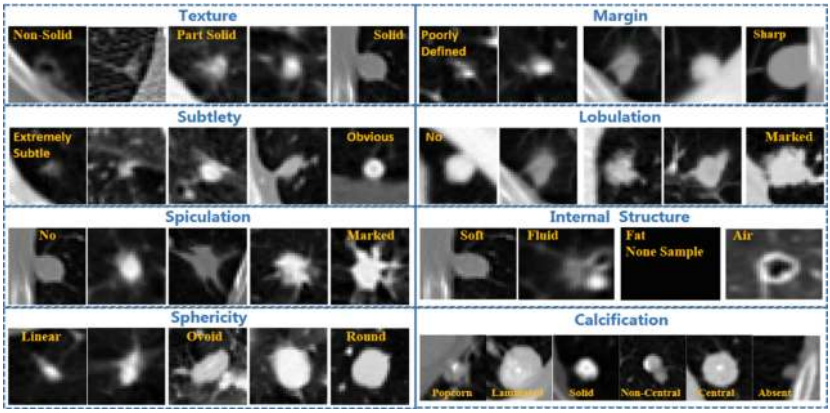

Fig. 1 Illustration of nodule patterns for the 8 semantic features.

As shown in Fig. 1, our goal is challenging. The appearances and shapes of nodules are very diverse for each semantic feature, and the surrounding tissues of nodules may further complicate the image patterns of nodules. Therefore intensive elaboration on the

extraction and selection of effective computational image features for each semantic feature is needed. In this study, we leverage the techniques of stacked denoising autoencoder (SDAE) [8], convolutional neural network (CNN) [9], and Haar-like feature computing [10] along with multi-task regression (MTR) framework to approximate our predicted scores to radiologists' ratings. With the SDAE, CNN and Haar-like features, the MTR can automatically exploit the sharable knowledge across the semantic features and select useful computational features for each of them. Here, each semantic feature is treated as an individual task.

2 Method

The training of nodule scoring scheme for the 8 semantic features is based on 2D nodule ROIs to avoid direct 3D feature computing from the LIDC image data with anisotropic resolution between x-y and z directions. The slice thickness variation is quite high (1.25–3 mm) in the LIDC dataset. At testing, the predicted scores for a nodule are derived with the averaged scores over all its member slices. Each nodule ROI is defined as the expanding bounding boxes of radiologists' outlines with offset of 10 pixels to include more anatomical contexts. For training and testing, all ROIs are resized as 28 × 28 for efficiency. The flowchart of our scheme is shown in Fig. 2.

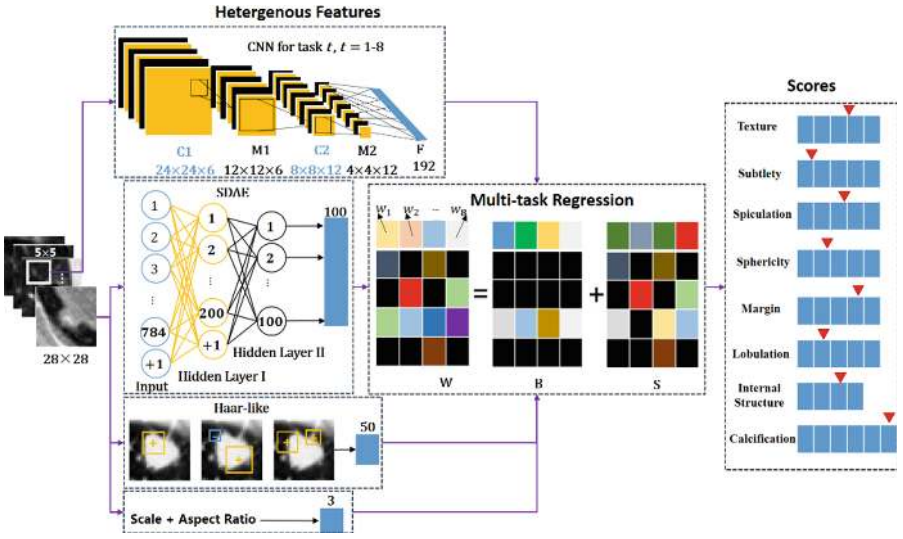

Fig. 2 Flowchart of the proposed scheme.

2.1 Extraction of Heterogeneous Computational Features

Referring to Fig. 1, the semantic features (tasks) cover the high-level description about the shape and appearance of pulmonary nodules, and the effective computational

features for each task is generally unknown. Therefore, we firstly compute heterogeneous features as diverse as possible, and then use the MTR framework to seek the suitable features for each task. The SDAE, CNN and Haar-like features are computed as the heterogeneous features. The SDAE and CNN are deep learning models that can automatically learn spatial patterns as features. The learnt SDAE and CNN features may encode both appearance and shape characteristics of nodules. SDAE features are derived from unsupervised phase and thus are general features. The training of CNN requires the sample labels, and hence CNN features are more task-specific. The Haar-like features aim to characterize low-level image contextual cue of nodules. To compensate the ROI resizing effect, we add the scaling factors of x and y directions and aspect ratio as three extra features; see Fig. 2.

SDAE model is constituted of unsupervised and supervised training phases. Here, we only take output neurons of the unsupervised phase as the SDAE features. At the unsupervised phase, the SDAE architecture is built by stacking the autoencoders in a layer-by-layer fashion. A layer of autoencoder can be constructed by seeking the coding neurons with the minimization of the reconstruction error of

$$\|x - \sigma(W'\sigma(W\tilde{x} + b) + b')\|^2, \tag{1}$$

where x is the input data and $\tilde{x}$ is the corrupted input data for better performance. The data corruption is conducted with random 0.5 zero masking. W, b, $W^{'}$, and $b^{'}$ are the synaptic matrices and biases of coding and reconstruction neurons, respectively, and σ is the sigmoid function. There are totally 100 SDAE features for MTR.

A typical CNN model is composed of several pairs of convolutional (C) and max-pooling (M) layers and commonly ended with fully-connected (F) and soft-max layers. We train 8 CNN models for the 8 tasks and adopt the neural responses at the fully-connected layer as the CNN features. Therefore, the CNN features are task-specific. The number of CNN features for each task is 192.

The Haar-like feature for a nodule ROI, Z, is computed with two blocks cropped from the resized ROI as:

$$\frac{1}{(2s_1+1)^2}\sum\nolimits_{\|p-c_1\|\le s_1} Z(p) + \frac{\varepsilon}{(2s_2+1)^2}\sum\nolimits_{\|p-c_2\|\le s_2} Z(q), \tag{2}$$

where c_1, s_1 and c_2, s_2 are the center and half-size of two square blocks, respectively, and $Z(p)$ is the HU value of p. ε can be 1, 0, and −1 and is randomly determined. The center and half-sizes of blocks are randomly set to generate 50 Haar-like features from. The half-size can be 1, 2, or 3.

2.2 Multi-task Regression

The 8 semantic features (tasks) describes nodule shape and appearance and hence may relate to each other in semantic meaning. The relation among the 8 tasks are generally unknown, and some tasks may share some computational features whereas some other tasks may not. To exploit the inter-task relation, we apply a MTR scheme with the

constraints of block and element sparsity [11]. Specifically, the cost function of the MTR is expressed as:

$$\sum_{t=1}^{8} \left\| X_t^T W_t - Y_t \right\|_F^2 + \lambda_B \|B\|_{1,\infty} + \lambda_S \|S\|_{1,1}; W = B + S, \tag{3}$$

where Y_t and X_t are the data labels and the SDAE + CNN + Haar-like features for the task t, respectively, W_t^T is the feature coefficient matrix of task t, $W = [W_1, \cdots, W_8]$, and $\|\cdot\|_F$ is the Frobenius norm. The regularization terms $\|B\|_{1,\infty}$ and $\|S\|_{1,1}$ assure the block and element sparsity and λ_B and λ_S are their weightings. The $\|S\|_{1,1}$ is defined as $\sum_{i,j} |S_{i,j}|$, and $\|B\|_{1,\infty}$ is computed as $\sum_i \|B_i\|_\infty$, where $\|B_i\|_\infty = \max_j |B_{i,j}|$; i and j are the indices of the rows and columns w.r.t. each matrix. The $\|S\|_{1,1}$ encourages zero elements in the matrix, whereas the $\|B\|_{1,\infty}$ favors zero rows in the matrix. Each column of W carries the feature coefficients of each task, while the coefficients of shared and task-specific features are hold in B and S, respectively, see Fig. 4(right). The minimization of the Eq. (3) is realized by interleavedly seeking proper B and S with the coordinate descent algorithm. The output W can be obtained with the final B and S. As shown in [11], with the constraints in Eq. (3), the coefficients of non-zero rows in B are sparse and distinctive, because the term $\|B\|_{1,\infty}$ may help avoid the situation of nearly-identical elements in the non-sparse rows with the constraint of l_1/l_q-norm. In such case, the MTR can not only exploit the shared features but also reserve the flexibility of coefficient variation of the shared features w.r.t. each task.

3 Experiments and Results

To illustrate the efficacy, the MTR framework are compared with two single-task regression schemes of LASSO [12] and elastic net [13], which can also select sparse features within the linear regression frameworks. The single-task regression schemes use the same set of computational features with MTR, except the CNN features, and perform the regression for each task independently. For MTR, all CNN features from 8 tasks are involved, whereas the single-task regressions only use the CNN features derived from each task. The performances of the two single-task regression schemes for each task are tuned independently and thus the regression parameters are different from task to task. To further show the effect of each type of computational features, we also compare the sole uses of SDAE, CNN, Haar-likes features in all regression schemes. 1400 nodules randomly selected from the LIDC dataset are involved in this study with the 10-fold cross-validation (CV) evaluation (basic unit is nodule). The feature computing and regression use the same data partition in each fold. Each nodule may have more than one annotation instances from different radiologists. In each fold of training, only one instance is utilized for nodules with multiple annotation instances, whereas all instances of the same nodule are involved in the testing in each fold. There are totally 581, 321, 254, 244 nodules with one, two, three and four annotation instances from different radiologists, respectively. We adopt the differences between the computer-predicted and radiologists' scores as the assessment metrics.

Table 1 summarizes the statistics of absolute differences between the computer-predicted (MTR, LASSO, and elastic net) and radiologists' scores over the 10-fold of CV. The performance of sole uses of three heterogeneous features for the three regression schemes are reported in Table 1 to show the effectiveness of the three types of computational features. The absolute differences of inter-observer ratings are also shown in the Table 1 for comparison. The inter-observer variation is computed from all possible pairs of annotation instances of the same nodule. As can be observed, the inter-observation variation is quite close to the variation between MTR scores and radiologists' scores. It may suggest there may exist ambiguity between the scoring degrees of the 8 semantic features that leads to rating disagreements among radiologists, see Fig. 4(left), where two ROIs of a nodule are shown. The nodule in Fig. 4(left) has degree ambiguity in "Subtlety" with scores from 4 radiologists of (2, 4, 5, 3), while the

Table 1. Abosulte distance perfomance. The "Tex", "Sub", "Spi", "Sph", "Mar", "Lob", "IS", and "Cal" stand for of the tasks "Texture", "Subtlety", "Spiculation", "Sphericity", "Margin", "Lobulation", "Internal Structure", and "Calcification", respectively. "IB", "LS" and "EN" indicate the inter-observer variation, LASSO and elastic net, respectively.

		Tex	Sub	Spi	Sph	Mar	Lob	IS	Cal	Overall
	IB	**0.46 ± 0.84**	**0.86 ± 0.88**	**0.68 ± 0.95**	**0.83 ± 0.75**	**0.83 ± 0.86**	**0.77 ± 0.93**	**0.02 ± 0.26**	**0.22 ± 0.76**	**0.58 ± 0.78**
All features	LS	1.04 ± 0.53	1.25 ± 0.65	0.89 ± 0.85	1.25 ± 0.90	1.13 ± 0.62	0.95 ± 0.84	0.02 ± 0.19	2.18 ± 0.61	1.09 ± 0.65
	EN	1.24 ± 0.50	1.20 ± 0.63	0.86 ± 0.79	1.09 ± 0.74	0.98 ± 0.91	0.96 ± 0.93	0.14 ± 0.24	1.44 ± 0.84	0.99 ± 0.70
	MTR	**0.58 ± 0.67**	**0.75 ± 0.59**	**0.80 ± 0.66**	**0.81 ± 0.49**	**0.86 ± 0.64**	**0.87 ± 0.66**	**0.04 ± 0.19**	**0.48 ± 0.59**	**0.65 ± 0.56**
CNN	LS	1.06 ± 0.58	1.13 ± 0.66	1.04 ± 1.08	1.29 ± 0.95	1.13 ± 1.00	1.28 ± 1.04	0.02 ± 0.19	2.12 ± 0.66	1.13 ± 0.77
	EN	1.27 ± 0.53	1.68 ± 0.82	1.04 ± 1.08	1.39 ± 0.97	1.31 ± 0.96	1.08 ± 0.95	0.02 ± 0.19	1.89 ± 0.71	1.21 ± 0.78
	MTR	0.74 ± 0.60	0.84 ± 0.56	0.86 ± 0.65	0.83 ± 0.48	0.97 ± 0.64	0.90 ± 0.66	0.04 ± 0.19	0.69 ± 0.56	0.73 ± 0.54
Haar-like	LS	2.42 ± 1.18	2.43 ± 1.09	0.88 ± 1.02	1.50 ± 1.05	1.88 ± 1.16	0.95 ± 1.02	0.02 ± 0.19	4.38 ± 1.13	1.81 ± 0.98
	EN	3.48 ± 1.05	3.08 ± 1.08	0.95 ± 1.01	2.68 ± 0.95	2.71 ± 1.18	1.02 ± 1.00	0.26 ± 0.46	4.61 ± 0.99	2.35 ± 0.97
	MTR	1.70 ± 1.17	1.63 ± 1.11	0.91 ± 1.04	1.65 ± 1.08	1.56 ± 1.09	0.96 ± 1.04	0.07 ± 0.29	2.94 ± 1.37	1.43 ± 1.02
SDAE	LS	1.17 ± 0.54	1.43 ± 0.76	1.17 ± 0.91	1.68 ± 0.88	1.35 ± 0.80	1.16 ± 0.89	0.02 ± 0.19	2.24 ± 0.58	1.28 ± 0.69
	EN	1.23 ± 0.75	1.23 ± 0.70	1.15 ± 0.90	1.58 ± 0.87	1.47 ± 0.90	1.18 ± 0.92	0.55 ± 0.24	1.56 ± 0.95	1.24 ± 0.78
	MTR	0.74 ± 0.74	0.84 ± 0.63	0.95 ± 0.69	0.84 ± 0.50	1.00 ± 0.69	0.97 ± 0.70	0.05 ± 0.19	0.64 ± 0.71	0.76 ± 0.61

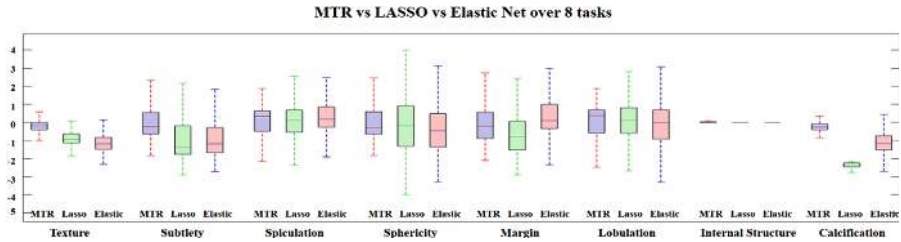

Fig. 3 Boxplots of the signed differences for MTR, LASSO, and elastic net, respectively.

Fig. 4 Annotation ambiguity (left) and illustration of B, S and W at one fold of CV (right).

MTR score is 3.78. The "Subtlety" scoring is highly subjective and depends on radiologists' experience. Figure 3 shows the box-plots of signed differences between the computer-predicted and radiologists' scores in CV 10 folds.

In Table 1, it can be found that the performances for the task "internal structure" are very good. It is because most nodules were rated as score 1 (1388) where the sample numbers for the scores 2–4 are nearly zeros. Accordingly, the regression for the task "internal structure" will not be difficult. For tasks like "spiculation" and "lobulation", the performance of the two single-task regression methods are not bad. However, it shall be recalled that the performance of these two single-tasks methods require tedious task-by-task performance tuning. It may turn out to be impractical if the task number goes formidably large. On the other hand, the MTR jointly considers the 8 tasks and achieves better performance.

To further insight on meaning and effect of the W separation mechanism in the MTR scheme, the sought B and S at the one fold of CV are shown in Fig. 4(right), where the black and non-black areas suggest zero and non-zero elements respectively. The blue rectangle identify the Haar-like features, while the left and right sides of the rectangle are the CNN and SDAE features, respectively. The selected task-specific features in S are very sparse, and many CNN and SDAE features are sharable across tasks as can be found in B.

4 Discussion and Conclusion

A computer-aided attribute scoring scheme for CT pulmonary nodules is proposed by leveraging the heterogeneous SDAE, CNN and Haar-like features with the multi-task regression (MTR) framework. The yielded scores with the MTR are shown to be more close to the radiologists' ratings, comparing to the scores from the two single-task regression methods. The two single-task methods share similar formulation like Eq. (3) without the consideration of multiple tasks, and are suitable for comparison. Accordingly,

the MTR can help to select useful features for each task with the exploration of inter-task relation. The effectiveness of using all SDAE, CNN, and Haar-like features are also illustrated in Table 1. Therefore, the efficacy of the MTR and the used heterogeneous features shall be well corroborated. Our automatic scoring scheme may help for deeper nodule analysis and support more sophisticated content retrieval of clinical reports and images for better diagnostic decision support [14].

Acknowledgement. This work was supported by the National Natural Science Funds of China (Nos. 61501305, 61571304, and 81571758), the Shenzhen Basic Research Project (Nos. JCYJ20150525092940982 and JCYJ20140509172609164), and the Natural Science Foundation of SZU (No. 2016089).

References

1. Naidich, D.P., et al.: Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. Radiology **266**, 304–317 (2013)
2. Gould, M.K., et al.: Evaluation of individuals with pulmonary nodules: When is it lung cancer?: Diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines. Chest **143**, e93S–e120S (2013)
3. Cheng, J.-Z., et al.: Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci. Rep. **6**, 24454 (2016)
4. Ciompi, F., et al.: Bag-of-frequencies: a descriptor of pulmonary nodules in computed tomography images. IEEE TMI **34**(4), 962–973 (2015)
5. Jacobs, C., et al.: Solid, part-solid, or non-solid?: classification of pulmonary nodules in low-dose chest computed tomography by a computer-aided diagnosis system. Invest. Radiol. **50**(3), 168–173 (2015)
6. Gurney, W., Swensen, S.: Solitary pulmonary nodules: determining the likelihood of malignancy with neural network analysis. Radiology **196**, 823–829 (1995)
7. Armato III, S.G., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
8. Vincent, P., et al.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
9. LeCun, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
10. Gao, Y., Shen, D.: Collaborative regression-based anatomical landmark detection. Phys. Med. Biol. **60**(24), 9377 (2015)
11. Jalali, A., Sanghavi, S., Ruan, C., Ravikumar, P.K.: A dirty model for multi-task learning. NIPS, pp. 964–972 (2010)
12. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B **58**(1), 267–288 (1996)
13. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B **67**(2), 301–320 (2005)
14. Kurtz, C., et al.: On combining image-based and ontological semantic dissimilarities for medical image retrieval applications. Med. Image Anal. **18**(7), 1082–1100 (2014)