Learning from Experts: Developing Transferable Deep Features for Patient-Level Lung Cancer Prediction

Wei Shen[1], Mu Zhou[2], Feng Yang[3(✉)], Di Dong[1], Caiyun Yang[1], Yali Zang[1], and Jie Tian[1(✉)]

[1] Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China
wei.shen@ia.ac.cn, tian@ieee.org
[2] Stanford University, Stanford, CA, USA
[3] Beijing Jiaotong University, Beijing, China
fengyang@bjtu.edu.cn

Abstract. Due to recent progress in Convolutional Neural Networks (CNNs), developing image-based CNN models for predictive diagnosis is gaining enormous interest. However, to date, insufficient imaging samples with truly pathological-proven labels impede the evaluation of CNN models at scale. In this paper, we formulate a domain-adaptation framework that learns transferable deep features for patient-level lung cancer malignancy prediction. The presented work learns CNN-based features from a large discovery set (2272 lung nodules) with malignancy likelihood labels involving multiple radiologists' assessments, and then tests the transferable predictability of these CNN-based features on a diagnosis-definite set (115 cases) with true pathologically-proven lung cancer labels. We evaluate our approach on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset, where both human expert labeling information on cancer malignancy likelihood and a set of pathologically-proven malignancy labels were provided. Experimental results demonstrate the superior predictive performance of the transferable deep features on predicting true patient-level lung cancer malignancy (Acc = 70.69 %, AUC = 0.66), which outperforms a nodule-level CNN model (Acc = 65.38 %, AUC = 0.63) and is even comparable to that of using the radiologists' knowledge (Acc = 72.41 %, AUC = 0.76). The proposed model can largely reduce the demand for pathologically-proven data, holding promise to empower cancer diagnosis by leveraging multi-source CT imaging datasets.

1 Introduction

Lung cancer is one of the leading causes of cancer death with a dismal 5-year survival rate at 15–18 % [9]. Computed Tomography (CT) sequences at varying

W. Shen and M. Zhou—These two authors contributed equally.

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 124–131, 2016.
DOI: 10.1007/978-3-319-46723-8_15

stages of patients have been fast-evolving over past years. Therefore, developing image-based, data-driven models is of great clinical interest for identifying predictive imaging biomarkers from multiple CT imaging sources.

Recently, Convolutional Neural Networks (CNNs) [3,11] emerge as a powerful learning model that has gained increasing recognition for a variety of machine learning problems. Approaches have been proposed for improving computer-aided diagnosis with cascade CNN frameworks [6–8]. However, these studies are limited at building CNN models for a single diagnostic data source, without considering the relationship across various diagnostic CT data of the disease. To verify the learned CNN model, a related but different diagnostic data can be always served as a good benchmark. Therefore, we ask two specific questions: Can CNN-based features generalize to other sets for image-based diagnosis? How do these features transfer across different types of diagnostic datasets?

We address these questions with an application in lung cancer malignancy prediction. More specifically, we define two malignancy-related sets: (1) *DiscoverySet* (source domain): CT imaging with abundant labels from only radiologists' assessments; (2) *DiagnosedSet* (target domain): CT imaging with definite, follow-up diagnosis labels of lung cancer malignancy. It is reasonable to assume that radiologists' knowledge in assessing risk factors is a helpful resource, but currently lacking a quantitative comparison with definite diagnostic information. Bridging the disconnection between them would accelerate diagnostic knowledge sharing to help radiologists refine follow-up diagnosis for patients. The challenge, however, remains as how can we develop a transferable scheme to fuse the cross-domain knowledge with growing availability of CT imaging arrays nowadays.

To overcome the obstacle, we propose a new, integrated framework to learn transferable malignancy knowledge for patient-level lung cancer prediction. The proposed model, called CNN-MIL, is composed of a convolutional neural network (CNN) model and a multiple instance learning (MIL) model. They are respectively trained on the DiscoverySet (2272 lung nodules) and the DiagnosedSet (115 patients). We achieve the purpose of knowledge transfer by sharing the learned weights between the built CNN and the instance networks (see Fig. 1). The proposed approach draws inspiration from a recent study [5] suggesting the feasibility of the CNN architecture in transfer learning. A difference between such work and ours is that the knowledge adaptation is achieved via the instance networks where the nodule-to-patient relationship is defined, and the layers of target network is deeper than that in the source domain network.

Our contributions of this paper can be summarized as follows: (i) We demonstrate that the knowledge defined from radiologists can be effectively learned by a CNN model and then transferred to the domain with definite diagnostic CT data. (ii) We present experimental evidence that knowledge adaptation can improve the accuracy of patient-level lung cancer prediction from a baseline model. (iii) The proposed CNN-MIL largely reduces the demand for pathologically-proven CT data by incorporating a referenced discovery set, holding promise to empower lung cancer diagnosis by leveraging multi-source CT imaging datasets.

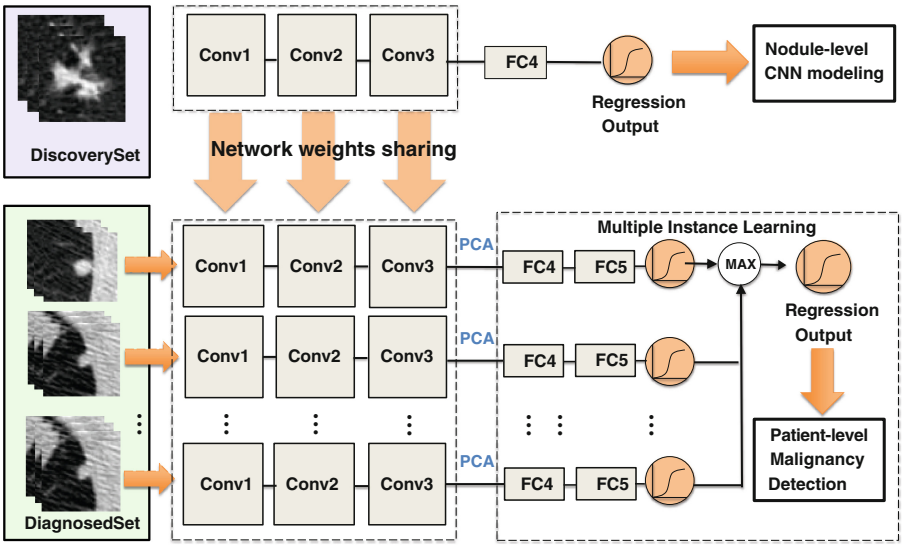

Fig. 1. Illustration of the proposed framework. **Upper part**: a nodule-level CNN model (CNN_{nodule}) is firstly trained on the DiscoverySet to extract the radiologist's knowledge (Sect. 2.1). It has three convolutional layers (Conv1-3) and the output layer has one neuron that estimates the malignancy rating of the input nodule. The number of hidden neurons in FC4 is 32. The network weights (Conv1-3) learned from DiscoverySet will be directly applied into the DiagnosedSet. **Lower part**: Multiple Instance Learning (MIL) models the nodule-to-patient relationship towards patient-level cancer prediction on the DiagnosedSet. Notably, the dimension of Conv3 feature from the instance network is reduced to 32 via Principal Component Analysis. The number of hidden neurons in FC5 is 4. The output of the MIL model is the aggregated output of the instance network, estimating true lung cancer malignancy (Sect. 2.2).

2 Methods

2.1 Knowledge Extraction via the CNN Model

As seen in Fig. 1, we firstly build a nodule-level CNN model to learn the radiologist's knowledge in estimating nodule malignancy likelihood from the source domain. The proposed CNN model is composed of three concatenated convolutional layers (with each comes with a Rectified Linear Unit plus a max-pooling layer). The followed two fully-connected layers (FC4 layer and regression layer) are used to determine the malignancy rating distribution over nodules. The used layers here follow the standard structure introduction in CNN structure, more details are referred to [10]. The input of our CNN model is the raw nodule patches with size $64 \times 64 \times 64$ voxels centering around the nodule shape. Each convolutional layer has 64 convolutional kernels with size 3×3. The pooling window size is 4×4 in the first max-pooling layer and 2×2 in remained layers. The loss function is the L-2 norm loss between the predicted rating and the malignancy rating:

$$L = \frac{1}{N}\sum_{i=1}^{N}(R_i - P_i)^2, \tag{1}$$

where N is the number of nodule patches in the DiscoverySet. The R_i and P_i are the ith nodule rating from radiologists and our model. The loss function of Eq. 1 is minimized via stochastic gradient descent.

Once the training is done, the knowledge of nodule malignancy estimation is learned in terms of the retaining weights in the trained CNN model. Next, the weights of the three convolutional layers are shared with instance networks for knowledge transfer. The weights from the fully-connected layers are not considered for domain transfer as higher layers appear to be more domain-biased which are less transferable [4]. Having learned feature representation for nodules, we detail the patient-level prediction via a MIL model next.

2.2 The MIL Model for Patient-Level Lung Cancer Prediction

We formulate the patient-level cancer prediction as a MIL task shown in Fig. 1. The input to the nodule level CNN is the nodule patch while the input to the MIL network is all the nodule features within a patient case. MIL builds on the concept of *bags* and *instances*, where the label of a bag is positive if it has at least one positive containing instance; the label of the bag is negative if and only if all its containing instances are negative. Thus, in the scenario of two-category malignancy prediction, we similarly define each patient as a bag and each nodule as an instance. Given a patient (O_i), if all his/her nodules are non-malignant, the patient is non-malignant; while if at least one nodule is malignant, the patient is malignant.

Let the patient-level malignancy predictions of m patients be $O = \{O_i | i = 1, 2, 3, ..., m\}$ and patients' malignancy labels $t = \{t_i | i = 1, 2, 3, ..., m\} \in [0, 1]$. Given a patient with n nodules ($1 \leq n \leq 15$ in this study), we denote the output from the jth nodule of the ith patient by $o_{i,j} (i = 1, 2, 3..., m, j = 1, 2, 3..., n\}$. The final output from the regression layer (lower part, Fig. 1) is used to determine the ith patient's malignancy by aggregating nodule instance outputs $o_{i,j}$:

$$O_i = \max(o_{i1}, o_{i2}, o_{i3}, ..., o_{in}), \text{ where } n \in [1, 15], \tag{2}$$

The loss function is also the L-2 norm loss between the prediction O_i and the diagnostic label t_i. As discussed, the weights of the convolutional layers are shared between the instance networks and the nodule-level CNN networks. The weights of the fully-connected layers (FC4 and FC5) will be continuously learned as in [5]. Next, we report experimental results on the DiagnosedSet for patient-level lung cancer prediction.

3 Experiments and Results

Dataset: We use the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset [2]. Nodule samples (>3 mm) are either

included into the DiscoverySet or the DiagnosedSet based on absence or presence of definite diagnosis. In DiscoverySet, the nodule malignancy likelihood was rated by four experienced thoracic radiologists, estimating an increasing degree (i.e., $R_{rad} \in$ [1,5]). The averaged rating report from four radiologists was chosen for determining the final rating of each nodule as in [7]. Overall, there were 2272 nodules included. We further split the DiscoverySet into a training set containing 80 % (1817 nodules) samples and a validation set containing 20 % (455 nodules) samples to observe the CNN model performance. For DiagnosedSet, there are 115 cases with true pathologically-proven diagnostic labels: non-malignant cases (30 cases), malignant cases (85 cases including 40 primary cancer and 45 malignant, metastatic cancer cases).

Model Configuration: For the nodule-level CNN model, the learning rate was 0.0001 and the number of training epochs (one epoch means that each sample has been seen once in the training phase) was 50. For the MIL model, the learning rate was 0.001. To evaluate the performance of the MIL model under different settings, we investigated different number of hidden neurons n_h ($n_h = [4, 8, 16]$) in FC5 layer and the number of the MIL model training epochs n_e ($n_e = [5, 10]$) in Sect. 3.2, while the default values were $n_h = 4$ and $n_e = 10$ in Sect. 3.3. We reported average results of the MIL model from 10 times five-fold cross validation. During each round of cross-validation, there were 92 cases (24 non-malignant and 68 malignant cases) in the training set and 23 cases (6 non-malignant and 17 malignant cases) in the test set. Since the number of the non-malignant cases was much smaller than that of malignant cases in the training set, we fed the non-malignant cases multiple times to our MIL model to make a proximately balanced dataset.

3.1 Knowledge Extraction

Given the output value $P_{cnn} \in$ [1,5] made by the nodule-level CNN model (CNN$_{nodule}$) and the $R_{rad} \in$ [1,5] given the radiologist's rating on the DiscoverySet. To verify that the radiologist's knowledge of malignancy is properly extracted, the estimation error defined as E $= |P_{cnn} - R_{rad}|, \in$ [0,4] in Fig. 2. We observed that the E $\in$ [0,1] already occupied 90.99 % of test nodules in the validation set from the DiscoverySet, revealing the outputs of the CNN model approximated those of the radiologist' inputs. Once the radiologist's knowledge of malignancy was well preserved by the nodule-level CNN model, we further report its results on patient-level cancer prediction on the DiagnosedSet.

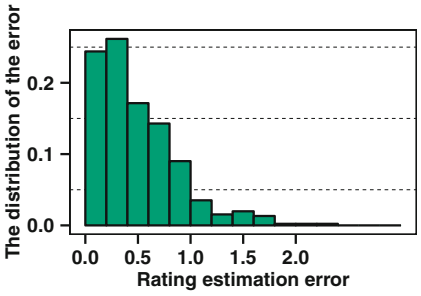

Fig. 2. The estimation error (E) distribution of CNN$_{nodule}$ on the DiscoverySet.

3.2 Patient-Level Lung Cancer Malignancy Prediction

We show the performance of our CNN-MIL model with respect to different configuration values of n_h and n_e on patient-level malignancy prediction. Prediction accuracy (i.e. the ratio of the number of correctly classified patient malignancy O_i over the entire DiagnosedSet) and area under the curve (AUC) score were used to measure the model performance. As shown in Table 1, the performance of our model was insensitive to n_h and n_e. It could be explained that shared weights preserved much nodule information that allows discriminative features fed into final fully-connected layers. We continue to verify the performance of the proposed approach with competing methods next.

Table 1. Mean value and standard deviation of prediction accuracy and AUC score (in parenthesis) of the CNN-MIL model with different n_h and n_e.

	$n_h = 4$	$n_h = 8$	$n_h = 16$
$n_e = 5$	68.80±3.12 %(0.65±0.03)	70.56±2.25 %(0.64±0.02)	68.12±1.97 %(0.62±0.02)
$n_e = 10$	70.69±2.34 %(0.66±0.03)	68.99±1.90 %(0.63±0.02)	68.98±2.10 %(0.62±0.03)

3.3 Methods Comparison

We chose the nodule-level CNN_{nodule} as a baseline model and the reports from the radiologists' ratings (RR) as a reference model. For CNN_{nodule} and RR, all nodule malignancy likelihoods within a patient were combined according to Eq. 2 as the patient-level malignancy score. We also implemented a MI-SVM model [1] and a deep MIL model without knowledge transfer (DMIL) [11]. The features fed to the MI-SVM were also the 32-dimensional CNN features generated from Conv3 (Fig. 1) and the kernel function was the radial basis function. The best parameters for MI-SVM were obtained via grid search and the parameter settings of DMIL were identical with our CNN-MIL except that DMIL did not have PCA operation inside.

Table 2. Average prediction accuracy and AUC score of patient-level cancer prediction using different models.

	Accuracy	AUC
DMIL [11]	59.40 %	0.56
MI-SVM [1]	61.93 %	0.55
CNN_{nodule}	65.38 %	0.63
CNN-MIL	**70.69 %**	**0.66**
RR	72.41 %	0.76

As seen in Table 2, with efficient knowledge transfer, our CNN-MIL outperformed both DMIL and MI-SVM. When comparing the performance of our CNN-MIL model to CNN_{nodule}, our CNN-MIL model integrating transferable-features through shared network weights could bring a boosted performance. Surprisingly, the performance of our CNN-MIL model was only marginally lower than that using the radiologists' ratings, which demonstrated the effectiveness of the proposed method in transferring human knowledge into unknown samples prediction. On the other hand, despite knowing that radiologists' ratings (i.e. RR) may affect our model learning due to the potential mislabelled samples,

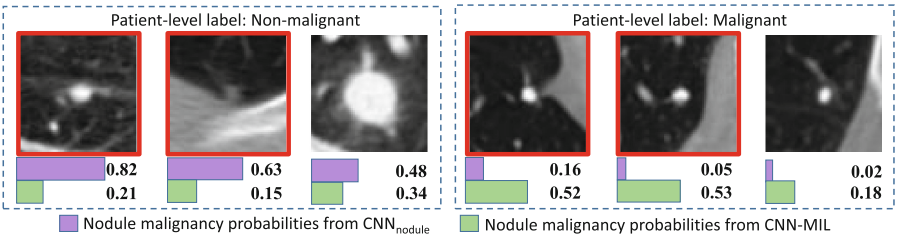

Fig. 3. Illustration of our CNN-MIL for patient-level cancer prediction with a non-malignant patient and a malignant patient. Our CNN-MIL model can make more accurate patient-level prediction than CNN_{nodule} (ratings rescaled to [0,1]) by reassigning nodule malignancy probability (red boxes) from the nodule-level CNN model.

we demonstrate that experts' knowledge, building upon consensus agreement from multiple radiologists, can be captured by our CNN-MIL model to further estimate true nodule malignancies in lung cancer.

As shown in Fig. 3, two patients using our CNN-MIL model and CNN_{nodule} illustrated that transfer learning on DiagnosedSet allowed us to optimize the instance networks for improved patient-level cancer prediction, permitting an error correction from the nodule-level CNN model. Using $p = 0.5$ as a division point ($p<0.5$ as non-malignant and $p>=0.5$ as malignant), CNN-MIL corrected erroneous predictions (red boxes) from CNN_{nodule} on both patients. The success of our model could be attributed to the ability of the CNN to learn rich mid-level image representations (e.g. features derived from the layer Conv3 in CNN) that are proven to be transferable to related visual recognition task [3,5].

Overall, our purpose of this study is not to pursue precise diagnosis for malignancy classification on a single diagnostic CT set, rather, we sought to infer data-driven knowledge across different sets (with different diagnostic labels), which holds promise to reduce the pressing demand of truly diagnosed, labelled data that typically require invasive assessment of biopsy and lasting monitoring of cancer progressions. We developed the domain transfer model based on the fact that the DiscoverySet (with radiologist ratings) is relatively easy-accessed at early stage of diagnosis with ubiquitous CT screening (2272 defined nodules here). Meanwhile, it is not surprising that the DiagnosedSet (with definitive clinical labels) is much difficult to scale due to invasive biopsy testing and surgery for pathological verification with a controlled patient population (115 case here).

4 Conclusion

Multi-source data integration in medical imaging is a rising topic with growing volumes of imaging data. Developing causal inference among different sets would allow better understanding of imaging set-to-set relationships in computer-aided diagnosis, thus enabling alternative biomarkers for improved cancer diagnosis. In this paper, we demonstrate that the transfer learning model is able to learn

transferable deep features for lung cancer malignancy prediction. The empirical evidence supports a feasibility that data-driven CNN is useful for leveraging multi-source CT data. In the future, we plan to expand to a large-scale, multi-model image sets to improve predictive diagnostic performance.

Acknowledgement. This paper is supported by the CAS Key Deployment Program under Grant No. KGZD-EW-T03, the National NSFC funds under Grant No. 81227901, 81527805, 61231004, 81370035, 81230030, 61301002, 61302025, 81301346, 81501616, the Beijing NSF under Grant No. 4132080, the Fundamental Research Funds under Grant No. 2016JBM018, the CAS Scientific Research and Equipment Development Project under Grant No. YZ201457.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 561–568 (2002)
2. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., 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)
3. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
4. Long, M., Wang, J.: Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 (2015)
5. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724 (2014)
6. Roth, H.R., Yao, J., Lu, L., Stieger, J., Burns, J.E., Summers, R.M.: Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications. In: Yao, J., Glocker, B., Klinder, T., Li, S. (eds.) Recent Advances in Computational Methods and Clinical Applications for Spine Imaging, vol. 20, pp. 3–12. Springer, Heidelberg (2015)
7. Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: Ourselin, S., Alexander, D.C., Westin, C.-F., Cardoso, M.J. (eds.) IPMI 2015. LNCS, vol. 9123, pp. 588–599. Springer, Heidelberg (2015). doi:10.1007/978-3-319-19992-4_46
8. Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., Zang, Y., Tian, J.: Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognition (2016)
9. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer Statistics, 2015. CA Cancer J. Clin. **65**(1), 5–29 (2015)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
11. Wu, J., Yinan, Y., Huang, C., Kai, Y.: Deep multiple instance learning for image classification and auto-annotation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3460–3469. IEEE (2015)