OFF-THE-SHELF CONVOLUTIONAL NEURAL NETWORK FEATURES FOR PULMONARY NODULE DETECTION IN COMPUTED TOMOGRAPHY SCANS

Bram van Ginneken, Arnaud A. A. Setio, Colin Jacobs, Francesco Ciompi

Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands

ABSTRACT

Convolutional neural networks (CNNs) have emerged as the most powerful technique for a range of different tasks in computer vision. Recent work suggested that CNN features are generic and can be used for classification tasks outside the exact domain for which the networks were trained. In this work we use the features from one such network, OverFeat, trained for object detection in natural images, for nodule detection in computed tomography scans. We use 865 scans from the publicly available LIDC data set, read by four thoracic radiologists. Nodule candidates are generated by a state-of-the-art nodule detection system. We extract 2D sagittal, coronal and axial patches for each nodule candidate and extract 4096 features from the penultimate layer of OverFeat and classify these with linear support vector machines. We show for various configurations that the off-the-shelf CNN features perform surprisingly well, but not as good as the dedicated detection system. When both approaches are combined, significantly better results are obtained than either approach alone. We conclude that CNN features have great potential to be used for detection tasks in volumetric medical data.

Index Terms— Nodule detection, computed tomography, convolutional neural networks

1. INTRODUCTION

Lung cancer is by far the most lethal cancer among men and women. Early detection, when the cancer is still in a treatable stage, appears the most promising avenue to reduce the burden of lung cancer. The National Lung Screening Trial in the United States has shown that screening with low-dose computed tomography (CT) significantly reduces both lung cancer and overall mortality among individuals at high risk for developing lung cancer [1].

Visual inspection of CT scans for the presence of pulmonary nodules, which could develop into lung cancer, is a time-consuming and tedious task and therefore many computer-aided detection (CAD) systems have been developed, both by academic groups and industry [2]. These systems follow the typical pipeline of CAD systems [3]: 1) defining a volume of interest, in this case the lungs; 2) extracting candidate regions that could potentially represent lung nodules with dedicated image processing steps, for example filtering operations that enhance bright and round structures, and possibly using a small set of features and a classifier; 3) false positive reduction with a large set of dedicated features and classifying these with a classifier trained with a large data set of nodules and non-nodule candidate.

The last few years a different approach to pattern recognition has achieved impressive results on a number of popular benchmark tasks in computer vision. This approach forgoes the step of designing and extracting a particular hand-crafted set of features, tailored to the task at hand, but instead feeds image data directly into 'deep' networks. These consist of many convolutional layers, interspersed with pooling layers that reduce the dimensionality of the input signal, and usually a few fully connected layers and a final classification layer. The convolutional layers in such networks can be thought of as a feature extraction subsystem, not designed or selected by algorithm developers, but learned specifically for the task at hand during the training process. Some researchers, however, have also reported good results with networks where the features (i.e. the convolution kernels) were pre-trained on different data, for example with unsupervised techniques [4]. This implies that a good set of kernels for solving visual tasks may be somewhat universal. Recently, Raza et al. [5] have demonstrated that the output of the intermediate layers of a big publicly available pre-trained convolutional neural network called OverFeat, fed directly in a linear support vector machine, can be used to obtain excellent performance for a range of computer vision tasks quite distinct from the task for which OverFeat was trained (object detection in natural images). They concluded that "any algorithm for a particular recognition task must be compared against the baseline approach of using generic deep features and a simple classifier."

In this work we investigate how this 'baseline approach' performs for a task that is clearly very different from object detection in 2D color photographs: the detection of pulmonary nodules in 3D computed tomography scans. While this approach at first may seem absurd, we note that radiologist who perform this task in clinical practice, rely on their primary and secondary visual cortex which has been trained with natural images and contain generic detectors with clear similarities to the kernels found in trained convolutional neural networks.

978-1-4799-2374-8/15/$31.00

We compare the OverFeat approach to a commercially available state-of-the-art nodule detection system, and use this system also to extract nodule candidates. We present various classification configurations with sagittal, coronal and axial patches, consider early and late fusion strategies, and investigate combinations of the score of the commercial CAD system and the approach with generic deep features.

2. DATA

All experiments are performed using the publicly available[1] LIDC database of 1,018 CT scans from different subjects from seven institutions with a wide range of acquisition protocols. All scans were annotated by four radiologists in a two-stage reading process. In the first round, all radiologists independently annotated nodules. In the second round, each reader categorized findings by any reader in the first round as either not a lesion, a non-nodule lesion, a nodule $<$ 3mm, or a nodule $>$ 3mm. In this work we considered all lesions accepted as nodules $>$ 3mm by 3 or 4 out of the 4 readers as nodules (positives). We ignore marks on the other nodular findings (i.e. small nodules and nodules accepted by a minority of the readers; as these are doubtful lesions, we do not want to consider these as false positive findings). We excluded scans with a section thickness of more than 2.5mm and scans with inconsistent or invalid DICOM. This resulted in a data set of 865 CT scans with 1,147 pulmonary nodules, and 3,271 excluded doubtful lesions.

3. METHOD

To generate candidate locations for nodules we used an existing state-of-the-art FDA approved commercially available CAD system (MeVis Medical Solutions AG, Bremen, Germany). This system produces a list of locations that could represent pulmonary nodules and a score that is monotonically related to the likelihood that the location is a nodule. From this score a free receiver operating characteristic (FROC) curve can be constructed and this system is indicated as CAD. In order to combine CAD with the systems described below, we converted this score to a probability estimate that the lesion represents a nodule, using a calibration procedure described previously [2, 6].

At every possible nodule location we extracted a patch in the sagittal (x), coronal (y) and axial (z) plane. The size of the patch is 50 $\times$ 50 mm (the size in pixels varies because the resolution of the scans varies) and rescaled to an 8-bit grayscale 221 $\times$ 221 pixels using Hounsfield unit rescaling and linear interpolation.

As in [5], we used the publicly available[2] trained convolutional neural network termed OverFeat [7], an implementation of the network of Krizhevsky et al. [8]. The network (see [7, 8] for details) uses a 221$\times$221 RGB image patch as input. It consists of convolutional layers containing 96 to 1024 kernels of size 3$\times$3 to 7$\times$7. Half-wave rectification, and max pooling kernels of size 3$\times$3 and 5$\times$5 are used at different layers to build robustness to intra-class deformations. OverFeat was trained for the image classification task of ImageNet 2013 with 1.2 million images hand labeled for presence or absence of objects in 1000 classes. In 2012, the network from [8] won both ImageNet classification and localization competitions by a large margin. In 2013, OverFeat won the ImageNet object localization competition.

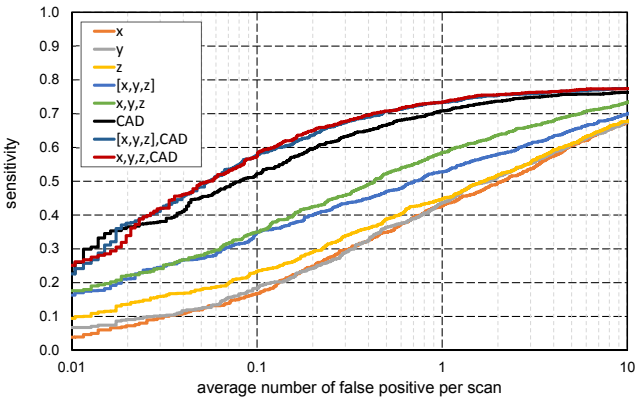

Fig. 1. Free response operating receiver curves for the eight nodule detection configurations.

We use the 4096 features from the first fully connected layer of OverFeat as the input for a linear SVM classifier (liblinear[3]) with C optimized in cross-validation. Training the SVM classifier was done in 10-fold scan-based cross-validation.

We construct three separate systems for patches from each orthogonal plane, designated x, y, and z. To fuse these results and use more than only 2D information, we consider an early and a late fusion approach. In early fusion, we concatenate the features from each plane and classify feature vectors $[x, y, z]$ of length 12,288. In late fusion, we use the output of the three systems x, y, z and use a second stage classifier with these three inputs to estimate the probability that the candidate is a nodule. We experimented with different classifiers and obtained similar results with several of them. All results reported in this work use a linear SVM for all classification stages.

Finally we can use the CAD probability with the output of the x, y, z systems to construct two combined approaches, early fusion plus CAD with two input features ($[x, y, z]$,CAD) and late fusion plus CAD with four input features (x, y, z,CAD).

[1] http://ncia.nci.nih.gov/

[2] https://github.com/sermanet/OverFeat

[3] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

systems	1/8	1/4	1/2	1	2	4	8	mean
x	0.19	0.28	0.35	0.43	0.49	0.57	0.65	0.42
y	0.20	0.26	0.36	0.44	0.51	0.59	0.65	0.43
z	0.24	0.32	0.39	0.45	0.52	0.59	0.66	0.45
$[x, y, z]$	0.36	0.41	0.47	0.53	0.58	0.63	0.68	0.52
x, y, z	0.38	0.45	0.52	0.58	0.64	0.68	0.72	0.57
CAD	0.54	0.62	0.67	0.71	0.74	0.75	0.76	0.68
$[x, y, z]$,CAD	0.60	0.66	0.71	0.73	0.75	0.76	0.77	0.71
x, y, z,CAD	0.60	0.66	0.71	0.73	0.76	0.77	0.77	0.71

Table 1. Sensitivity of the eight configurations considered in this study at an average number of $2^n, n = -3, -2, \ldots, 3$ false positive detections per scan.

4. RESULTS

The CAD system generated 37,262 candidate locations (on average 43.1 per scan). 78% of all true nodules were among the candidates, so this is the maximally achievable sensitivity in this study.

Table 1 summarizes the results. FROC curves are provided in Fig. 1 and Fig. 2 shows example cases. Here the patches at the size and resolution as processed by the OverFeat network are displayed.

As expected, the 2D approaches x, y, and z that attempt to classify if a candidate is a nodule from only a single patch, have the lowest scores, but they still perform reasonably well at higher false positive levels. Axial patches perform slightly better than coronal and sagittal patches, probably because the resolution of the scans is better in the axial plane for the vast majority of scans in the LIDC data set.

Fusion of the analysis of the three 2D patches substantially improves performance. Late fusion outperforms early fusion, but by a small margin only. Nevertheless, the OverFeat approaches all are clearly inferior to the performance of the commercial CAD system.

However, when CAD is combined with OverFeat classification results, a significant boost in performance is observed. The best configuration achieves an average sensitivity of 71% at all seven operating points, and especially at the highly specific operating points, the combination outperforms CAD. Statistical analysis with bootstrapping showed that the difference is highly significant.

5. DISCUSSION & CONCLUSION

The remarkable result of this work is that it is possible to obtain good results on a medical image analysis problem – nodule detection in chest CT – using a highly complex feature extraction system trained with images from a completely different domain (2D natural color images), for a completely different task (detection of the presence of one of 1000 different real-world objects). This confirms the findings in [5] where a similar approach was applied successfully to a range of computer vision classification tasks. Apparently the 4096 features extracted by OverFeat provide a generic image descriptor.

The data used in this study is 3D, and our approach of analyzing various 2D projections may be seen as a limitation.

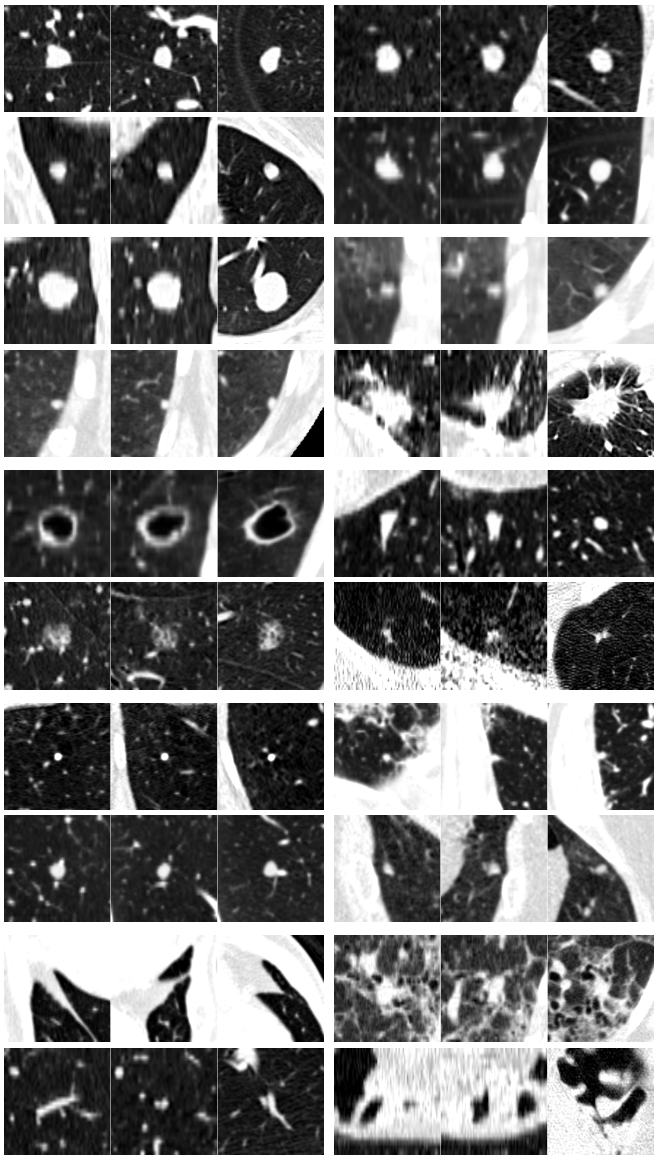

Fig. 2. Triples of sagittal, coronal and axial patches of 50×50mm are displayed as illustrative result. Top 4 cases: these achieved highest scores by x, y, z,CAD. They are prototypical solitary round nodules surrounded by lung parenchyma. Cases 5-8: random true positives detection of x, y, z,CAD at 1 FP/scan operating point. Cases 9-12: detected by x, y, z, at 1 FP/scan, but not by CAD at 1 FP/scan. These include nodules of an unusual morphology. Cases 13-16: false positives of x, y, z,CAD at 1 FP/scan. These are in fact nodules that were missed by LIDC readers. Bottom 4 cases: true false positives of x, y, z,CAD at 1 FP/scan.

However, radiologists also analyze 3D data in this way, and a similar approach to use a CNN for a medical detection task was recently presented in [10] and also analyzed multiple 2D patches. In that study, orthogonal patches were entered into a CNN as RGB channels, which may not be optimal as a triple of voxels originating from the same location in the three orthogonal patches do not correspond to the same spatial location.

Future work could investigate how many patches are optimal, and what the optimal size for a patch is, and if a multi-scale approach in which features or classifications of patches of multiple sizes are used increases robustness. This approach was used for natural images in [7]. Our work did not use any information on the size of the nodules, which was available from the commercial CAD system and used patches of a fixed 50 mm size.

Despite its good performance, the systems built on CNN features alone perform worse than the highly tuned and optimized CAD system that we used as our starting point and candidate detector. Given the substantial gap between CAD and x, y, z, it is surprising that combining both approaches leads to such clear improvement, especially at a low false positive range. Apparently the two approaches are quite complementary in nature. We have previously shown that fusing CAD systems with very big differences in performance can, in some situations, lead to a combined system that substantially outperforms the best system alone [6].

In this work we used the candidates produced by the commercial CAD system for training and testing of the system. Over 20% of all nodules are missed by this candidate detector. An important reason for this is that the commercial CAD system is only meant to detect solid nodules between 4 and 30 mm in size. Non-solid nodules and nodule outside this size range tend to be missed. The diversity of nodules is illustrated in Fig. 2. In fact, one could use a CNN approach for candidate detection as well. To keep computation times within acceptable limits (we extracted about 40 candidates per scan in this study, but chest CT scans contain in the order of 10^6 voxels), a multi-stage system that uses a more complex network only for the most promising candidates may be needed.

Visual inspection of false positive detections with a high probability of x, y, z,CAD revealed that the LIDC data set, despite being annotated by four radiologists, contains nodules that are not contained in the reference standard. Fig. 2 shows some examples. This shows the potential of using high quality nodule CAD systems in clinical practice. We have noticed this previously [11] and plan to release an update to the LIDC annotations with CAD marks confirmed by radiologists.

Data augmentation has often been shown to improve the performance of CNNs. The fact that we use 3D data provides a large number of possibilities for generating an augmented set of 2D patches. This would be especially relevant when complete CNNs would be trained end-to-end on nodule data, which is a topic for future work.

6. REFERENCES

[1] D. R. Aberle, A. M. Adams, C. D. Berg, et al., "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, vol. 365, pp. 395–409, 2011.

[2] B. van Ginneken, S. G. Armato, B. de Hoop, et al., "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study," *Medical Image Analysis*, vol. 14, pp. 707–722, 2010.

[3] B. van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: How to move from the laboratory to the clinic," *Radiology*, vol. 261, pp. 719–732, 2011.

[4] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *AISTATS*, 2011.

[5] A. S. Razavian, H. Azizpour, J. Sullivan, et al., "CNN features off-the-shelf: An astounding baseline for recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2014, arXiv: 1403.6382.

[6] M. Niemeijer, M. Loog, M. D. Abràmoff, et al., "On combining computer-aided detection systems," *IEEE Transactions on Medical Imaging*, vol. 30, pp. 215–223, 2011.

[7] P. Sermanet, D. Eigen, X. Zhang, et al., "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *International Conference on Learning Representations (ICLR 2014)*, April 2014, arXiv: 1312.6229.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[9] O. Russakovsky, J. Deng, H. Su, et al., "ImageNet Large Scale Visual Recognition Challenge," 2014, arXiv:1409.0575.

[10] H. R. Roth, L. Lu, A. Seff, et al., "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *MICCAI*, 2014, vol. 8673 of *Lecture Notes in Computer Science*, pp. 520–527.

[11] C. Jacobs, B. van Ginneken, S. Fromme, et al., "Benchmarking computer-aided detection of pulmonary nodules on the recently completed publicly available LIDC/IDRI database," in *Annual Meeting of the Radiological Society of North America*, 2013.