Detecting Bone Lesions in Multiple Myeloma Patients using Transfer Learning

Matthias Perkonigg[1], Johannes Hofmanninger[1], Björn Menze[2], Marc-André Weber[3], and Georg Langs[1]

[1] Computational Imaging Research Lab
Department of Biomedical Imaging and Image-guided Therapy,
Medical University of Vienna
matthias.perkonigg@meduniwien.ac.at,
www.cir.meduniwien.ac.at
[2] Institute of Biomedical Engineering, Image-based Biomedical Modelling,
Technical University of Munich
[3] Institute of Diagnostic and Interventional Radiology,
University Medical Center Rostock

Abstract. The detection of bone lesions is important for the diagnosis and staging of multiple myeloma patients. The scarce availability of annotated data renders training of automated detectors challenging. Here, we present a transfer learning approach using convolutional neural networks to detect bone lesions in computed tomography imaging data. We compare different learning approaches, and demonstrate that pretraining a convolutional neural network on natural images improves detection accuracy. Also, we describe a patch extraction strategy which encodes different information into each input channel of the networks. We train and evaluate our approach on a dataset with 660 annotated bone lesions, and show how the resulting marker map high-lights lesions in computed tomography imaging data.

1 Introduction

Multiple myeloma (MM) is a cancer of plasma cells in the bone marrow. The most common symptom for MM are bone lesions. Bone lesions can be detected in computed tomography (CT) scans. An automated detection of lesions in CT scans is desirable, because it would accelerate reading images and could help during diagnosis and staging of multiple myeloma patients. The detection is difficult, and until now, no algorithms for automatic lesion detection in CT data have been developed. Deep learning approaches such as convolutional neural networks (CNN) are a promising direction for this problem. However, a difficulty arising with MM is the limited availability of annotated training data. The numbers of examples are smaller than those typically used for training CNNs, and the representation capacity suffers correspondingly.

Here, we demonstrate two approaches to perform lesion detection in MM using CNN architectures. First, we evaluate transfer learning as a means of

improving the performance of our approach by transferring knowledge from a natural image classification task. Secondly, we explore two ways of representing visual input data for CNN training: a single channel approach, and an approach which distributes ranges of different hounsfield unit to different channels. We compare these approaches on a data set containing overall 660 annotated bone lesions.

Related work Convolutional neural networks [3] and transfer learning are used in medical imaging for a variety of applications. Transfer learning aims to transfer knowledge learned in a source task to improve learning in a target task [7]. Our approach uses a network pre-trained as a classifier on the natural image database ImageNet [1] (the source task) and then applies it as a detector of bone lesions (the target task). Fine-tuning on images extracted from CT scans is applied to adapt to the target task. Shin et al. describe a similar approach in [5]. They compare different architectures and learning protocols, transfer learning and random initalization, for lymph node detection and interstitial lung disease classification [5]. In [4] the authors show how convolutional neural networks can be used to reduce false positives while detecting sclerotic bone lesions in computer aided detection (CAD) tools. In the most closely related work, Xu et al. use a novel neural network architecture to detect bone lesions in multiple myeloma patients in multimodal PET/CT scans [8].

The proposed method differs from previous approaches in several aspects. It does not need a prior candidate detection stage before using the CNN [4], instead we use the CNN to detect lesions directly. Our approach operates on single 2D patches, while previous work used ensembles of neural networks and extracted multiple patches at one volume location [5, 4]. Finally, we use volumes of a single modality (CT), instead of using a multimodal approach [8].

2 Method

We treat lesion detection as a classification task. We extract local image patches, and train a convolutional neural network to classify patches into *lesion* and *non-lesion*. We compare two ways of extracting image information and encoding it in image patches used by the CNN: a single channel approach, and a three channel approach. Furthermore, we compare two learning protocols to evaluate if the transfer of parameters of pre-trained models is superior to random initialization.

2.1 Extracting image patches

We extract a set of image patches $\mathbf{P}$ for training and testing from a set of whole body CT volumes $\{\mathbf{V}_1, \ldots, \mathbf{V}_n\}$. Due to the anisotrophy of the volumes in axial direction as well as to match the input channels of the CNN after transferring from a 2D natural image task, the image patches are extracted in 2D along the axial axes. For each volume $\mathbf{V}_i \in \{\mathbf{V}_1, \ldots, \mathbf{V}_n\}$ the center positions of all lesions

$\{\mathbf{x}_1^{l_i}, \ldots, \mathbf{x}_j^{l_i}\}$ are annotated. Additionally a bone mask $\mathbf{M}_i$ is provided for each $\mathbf{V}_i$. All patches $\mathbf{p} \in \mathbf{P}$ are extracted with a size of 15x15 millimetres.
For each $\mathbf{V}_i$ and each lesion $\mathbf{x}_m^{l_i}$ a positive patch $\mathbf{p}_m^{p_i}$ is extracted centred around $\mathbf{x}_m^{l_i}$. $\mathbf{p}_m^{l_i}$ is augmented by random rotations and mirroring. This results in a set of positive patches $\mathbf{P}_{p_i}$ for each $\mathbf{V}_i$. A set of negative patches $\mathbf{P}_{n_i} = \{\mathbf{p}_1^{r_i}, \ldots, \mathbf{p}_m^{r_i}\}$ is extracted from $\mathbf{V}_i$ at m random positions $\mathbf{x}_m^{r_i}$ inside $\mathbf{M}_i$, with the restriction that they do not overlap with the extracted patches in $\mathbf{P}_{p_i}$. The negative patches $\mathbf{p}_m^{r_i}$ are not augmented. The final set of patches used for training and testing is given by $\mathbf{P} = \{\mathbf{P}_{p_i} \cup \mathbf{P}_{n_i}\}$ for all $i = 1 \ldots n$.

Figure 1 shows two ways of extracting a patch and representing the information for CNN training:

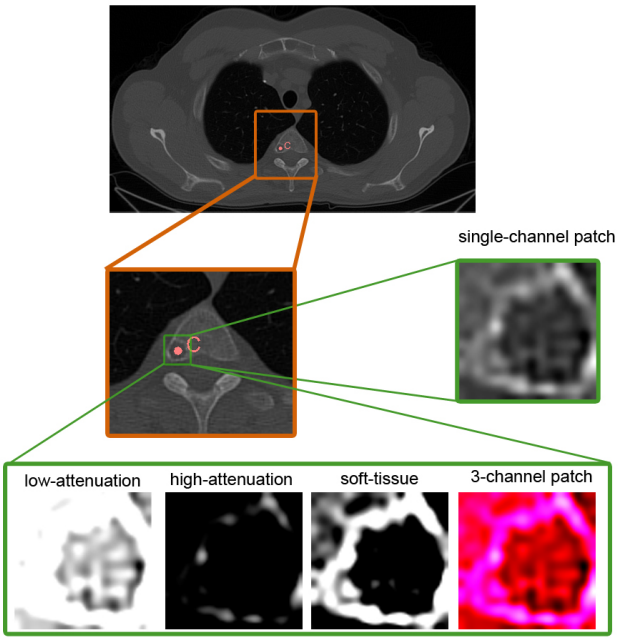

Fig. 1. Positive patch extraction: A patch is extracted along the axial axes around the center of a lesion. Single-channel patch extraction extracts gray-scale patches. 3-Channel patch extraction encodes information of a low-attenuation, high-attenuation, soft tissue window, into a combined three channel image.

1. Single channel patches: In the first approach a set of gray-scale patches $\mathbf{P}^G$ is extracted as described above. To implement transfer learning on pre-trained networks, and to match the input size of the network, the patches are rescaled to 64x64 and the same patch is fed into each of the three input channels.

2. Three channel patches: The second approach exploits the quantitative character of CT images, enabling the splitting of value ranges in a consistant manner across examples. We use three channels to encode different information extracted from the image. By this decoupling, the network can potentially find more meaningful features in different ranges corresponding to specific anatomical characteristics. We extract 3-Channel patches $\mathbf{P}^{3C}$ by assigning different ranges of Hounsfield Units (HU) to three different channels. The first channel focuses on a low-attenuation window of values smaller than 100 HU, the second on a high-attenuation window (>400 HU) and the third on a soft tissue window [100-400 HU]. The patches in $\mathbf{P}^{3C}$ are rescaled to 64x64 to match the input size of the network.

2.2 Network architecture

We use the VGG-16 architecture [6] as a base for our network. This enables the comparison of networks trained only on our data to transfer learning using networks trained on substantially larger sets of natural images. We use 64x64x3 as input size. Except the fully connected layers and the classification layer at the end of the network, the architecture remains unchanged to the original VGG-16 network. These layers are exchanged to fit the detection task resulting in a single output value. The final model used is depicted in Figure 2. It consists of five convolutional blocks with two, respectively three convolutional layers separated by a max pooling layer with a stride of 2. In the end of the network three fully-connected layers are used. Rectified Linear Units (ReLU) are used as activation function for all hidden layers. A sigmoid activation function is used for the output unit to produce a probability value of seeing a lesion. Depending

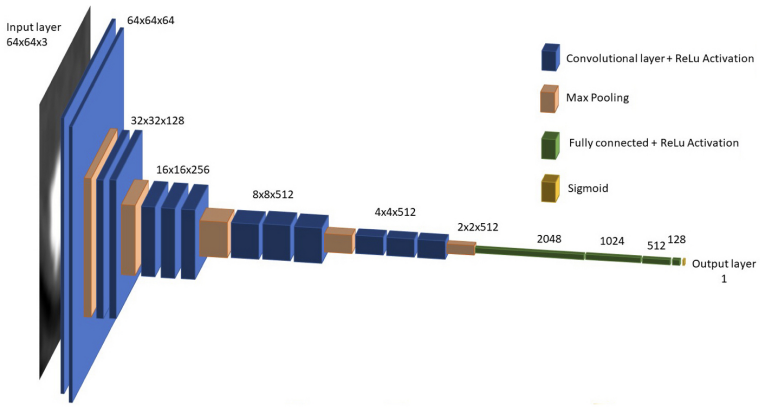

Fig. 2. The CNN architecture as used in our approach. It is based on the architecture of VGG-16 [6], the input shape and the fully connected layers at the end of the network are adapted to fit our detection task.

on the learning protocol used, we input different image patches to this network. The single channel patches $\mathbf{P}^G$ are rescaled from size 15x15 to 64x64 and the same patch is given to each channel of the input layer. The three channel patches $\mathbf{P}^{3C}$ are rescaled from 15x15x3 patches to 64x64x3 and each channel of the patch is used as input to one of the three input channels.

2.3 Four models

Two different learning protocols, transfer learning and random initialization as well as two different patch extraction strategies, single channel patches and 3-channel patches, are used. For all four models trained the architecture of the network shown in Figure 2 remains unchanged.
1. Transfer Learning: The weights of our network are transfered from pre-training a VGG-16 network on the natural image dataset ImageNet [1]. The custom fully-connected layers are initialized randomly. For the transfer learning approach the first six layers shown in Figure 2 are frozen and the neural network is fine-tuned on $\mathbf{P}^G$. We fine-tune the model with stochastic gradient descent and use binary-crossentropy as loss function. The network is fine-tuned for 30 epochs. This approach will be called ***TL-approach***.
2. Random Initialization: Instead of transferring from a pre-trained VGG-16 model all weights and biases are initialized randomly. The whole CNN is trained with stochastic gradient descent from scratch. Training is done for 40 epochs. Only $\mathbf{P}^G$ is used for training. We will call this approach ***RI-approach***.
3. 3-Channel Transfer Learning approach: For this approach the transfer learning protocol is used as described above. The only difference is that we use 3-channel patches $\mathbf{P}^{3C}$ for training and evaluating the model. The approach will be denoted as ***3C-TL-approach***.
4. 3-Channel Random Initialization approach: The random initialization learning protocol is used in combination with 3-channel patches $\mathbf{P}^{3C}$. The approach will be denoted as ***3C-RI-approach***.

2.4 Volume parsing

After training the network, we apply the detection to a volume of the test set $\mathbf{V}_j \in \{\mathbf{V}_1, \ldots, \mathbf{V}_t\}$, which was not part of the training process. At every position $\mathbf{x}_i^j$ within $\mathbf{M}_j$ an image patch $\mathbf{p}_i^j$ of size 15x15 millimetres along the axial axes is extracted. Depending on the model used, single channel or three channel patches are extracted, rescaled and used as input to the network. The output is a probability value $P(\mathbf{p}_i^j)$ that the patch is showing a lesion. The probabilities are visualized as a probability map of the same size as $\mathbf{V}_j$.

3 Evaluation

Data For training and evaluation a subset of the VISCERAL Detection Gold Corpus [2] is used. We use a set of 25 volumes for which manually annotated

Table 1. Number of samples in the dataset

	lesion	non-lesion
Training set	2153	2124
Validation set	299	294
Test set	538	530
	2990	2948

lesions and organ masks are provided. Three of those volumes, with a total of 62 lesions, are used for the evaluation of volume parsing in whole CT scans. In the 22 CT volumes used for training and validating the CNN, a total of 598 lesions are annotated. 5938 image patches are extracted and split randomly into a training (72%), validation (10%) and test (18%) set. The training and validation set are used during the training phase of the networks. The test set is used for the evaluation of the models. Table 1 gives an overview of the dataset used for the supervised fine tuning, respectively training of the network.

Evaluation on patches After training we evaluate the four approaches on a dataset of image patches. We measure true positives, false positives, true negatives, and false negatives. F-Score, precision and recall are computed on the test set. To evaluate if the transfer of parameters is beneficial the *TL-approach* and the *RI-approach*, respectively the *3C-TL-approach* and the *3C-RI-approach* are compared. For the evaluation of the different patch extraction strategies the results of the *TL-approach* and the *3C-TL-approach*, respectively *RI-approach* and *3C-RI-approach* are compared.

Evaluation of volume parsing We parse CT volumes that were not used for training. Bone masks are used to restrict the Region of Interest (ROI) to bones. We predict a probability score for each position in the ROI and generate probability maps. Those probability maps are compared visually to evaluate the ability of the different models to detect lesions.

4 Results

Results on image patches Table 2 compares different performance measures for all four models. All four approaches can classify image patches into lesion and non-lesion with high accuracy. The results show that models using transfer learning achieve a higher F-Score, and AUC, outperforming networks trained only on the CT image patches. Both approaches using transfer learning (TL-approach and 3C-TL-approach) outperform the corresponding random initialization approaches. The 3C-TL-approach has the lowest number of false negatives, which is critical as lesions should not stay undetected.

The comparison of the different patch extraction strategies demonstrates that the models trained with three channel patches performs better. The 3-channel-TL-approach (0.92) has a slightly higher F-Score than transfer learning with gray

Table 2. Comparison of detection performance measures for the four approaches: transfer learning (TL), random initialization (RI) and the 3 channel approaches (3C-TL and 3C-RI).

	Precision	Recall	F-Score	AUC
TL	0.91	0.87	0.89 ± 0.010	0.96 ± 0.006
RI	0.82	0.84	0.83 ± 0.010	0.91 ± 0.008
3C-TL	0.95	0.90	0.92 ± 0.008	0.97 ± 0.004
3C-RI	0.92	0.90	0.91 ± 0.010	0.97 ± 0.004

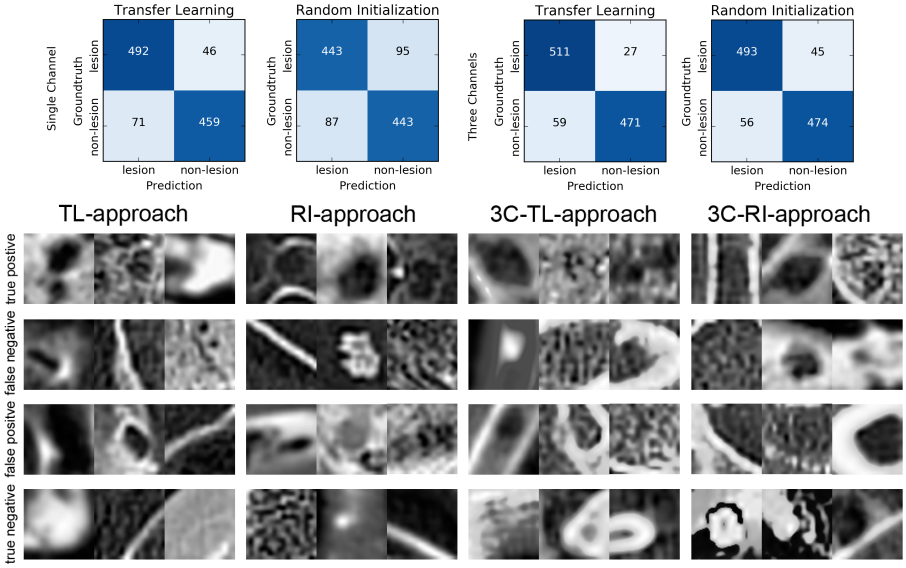

Fig. 3. Results on image patches. In the upper row confusion matrices for all four approaches are given. The lower row show examples for correct classifications and misclassifications of the networks.

scale patches (0.89). Absolute numbers and examples of true/false classifications are given in Figure 3. The increased accuracy of the 3-channel approaches could be due to the decoupling of the different HU ranges enabling a better exploitation of the CNN architecture.

Results of volume parsing Three details of probability maps for axial slices are shown in Figure 4. The TL-approach produces smooth results with a lot of noise, while the RI-approach and the 3C-RI-approach produce more noise and sharper borders between regions classified as lesion/non-lesion. The 3C-TL approach produces the sharpest borders between regions and less noise than both RI-approaches, consistent with its higher quantitative accuracy. The qualitative analysis shows that the 3C-TL-approach outperforms the other approaches.

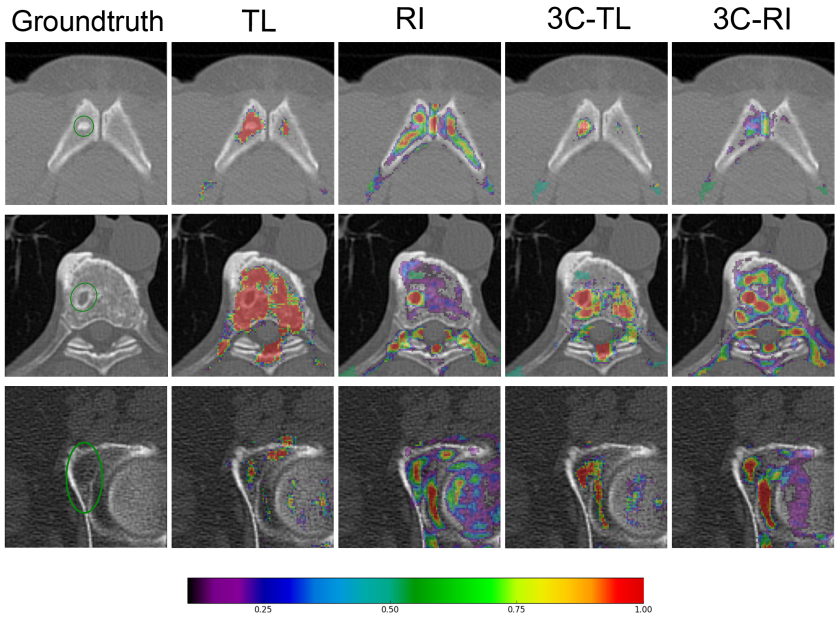

Fig. 4. Details of probability maps for detecting bone lesions in axial slices. Each row depicts the groundtruth and the detection probability of the three approaches.

The probability maps show that all approaches can detect the lesions annotated in the groundtruth. However, the false positive rate is high for all approaches with the 3C-TL approach showing the best performance. This indicates that, due to the limited number of annotated lesions in the training data, the generalization of the network is limited.

5 Conclusion

We propose an algorithm for automatic detection of bone lesions in CT data of multiple myeloma patients. We evaluated two questions: can we transfer models from natural image data to improve accuracy, and does a decoupling of HU ranges in the input representation help classification. We compared four different approaches: a CNN trained on a set of lesion and non-lesion examples of CT imaging data, a CNN pre-trained on natural images, transferred and fine tuned on the CT data, and an alternative 3-channel representation of the image data for both approaches. Results show that classification with high accuracy is possible. Transfer learning, and splitting image information into channels, both improve detection accuracy. Qualitative experiments on calculating marker maps for lesions on full volumes, show that on large volumes the suppression of false positives still needs to be improved. By providing insight the into number of

lesions detected as well as their extend, the proposed method could be used in clinical context as a tool to monitor the progression of the disease.

Acknowledgement

This work was supported by the Austrian Science Fund (FWF) project number I2714-B31.

References

1. Deng, J., Dong, W., Socher, R., et al.: ImageNet: A large-scale hierarchial image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 248–255 (2009)
2. Krenn, M., Grünberg, K., Jimenez-del-Toro, O., et al.: Datasets created in VISCERAL. In: Hanbury, A., Müller, H., Langs, G. (Eds) Cloud-Based Benchmarking of Medical Image Analysis. 69–84 (2017)
3. LeCun, Y., Bottou, L., Bengio, Y., and Haggner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE 86(11). 2278–2324 (1998)
4. Roth, H. R., Lu, L., Liu, J., et al.: Efficient False Positive Reduction in Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation. In: Deep Learning and Convolutional Neural Networks for Medical Image Computing. Advances in Computer Vision and Pattern Recognition. 35–48 (2017)
5. Shin, H., Roth, H., Gao M., et al.: Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. In: IEEE Transactions on Medical Imaging 35(5). 1285–1298 (2016)
6. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICLR (2015)
7. Torrey, L. and Shavlik, J.: Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, 242 (2009)
8. Xu, L., Tetteh, G., Mustafa, M., et al.: W-Net for Whole-Body Bone Lesion Detection on ^{68} Ga-Pentixafor PET/CT Imaging of Multiple Myeloma Patients. In Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment. 23–30 (2017)