Feature learning based on visual similarity triplets in medical image analysis: A case study of emphysema in chest CT scans

Silas Nyboe Ørting[1], Jens Petersen[1], Veronika Cheplygina[2], Laura H. Thomsen[3], Mathilde M W Wille[4], and Marleen de Bruijne[1,5]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark, silas@di.ku.dk
[2] Medical Image Analysis (IMAG/e), Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[3] Department of Internal Medicine, Hvidovre Hospital, Copenhagen Denmark
[4] Department of Diagnostic Imaging, Bispebjerg Hospital, Copenhagen, Denmark
[5] Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC - University Medical Center Rotterdam, The Netherlands

Abstract. Supervised feature learning using convolutional neural networks (CNNs) can provide concise and disease relevant representations of medical images. However, training CNNs requires annotated image data. Annotating medical images can be a time-consuming task and even expert annotations are subject to substantial inter- and intra-rater variability. Assessing visual similarity of images instead of indicating specific pathologies or estimating disease severity could allow non-experts to participate, help uncover new patterns, and possibly reduce rater variability. We consider the task of assessing emphysema extent in chest CT scans. We derive visual similarity triplets from visually assessed emphysema extent and learn a low dimensional embedding using CNNs. We evaluate the networks on 973 images, and show that the CNNs can learn disease relevant feature representations from derived similarity triplets. To our knowledge this is the first medical image application where similarity triplets has been used to learn a feature representation that can be used for embedding unseen test images.

Keywords: Feature learning, Similarity triplets, Emphysema assessment

1 Introduction

Recent years have demonstrated the enormous potential of applying convolutional neural networks (CNNs) for medical image analysis. One of the big challenges when training CNNs is the need for annotated image data. Annotating medical images can be a time-consuming and difficult task requiring a high level of expertise. A common issue with annotations is substantial inter- and intra-rater variability. There are many sources of rater variability in annotations, for

example, level of expertise, time-constraints and task definition. A common approach to defining annotation tasks is to ask raters for an absolute judgment, "segment the tumor", "count number of nodules", "assess extent of emphysema". Evidence from social psychology suggests humans in some cases are better at making comparative ratings than absolute ratings [11, 3, 5]. Redefining annotation tasks in terms of relative comparisons could improve rater agreement.

An annotation task that is especially prone to rater variations and may be better suited for comparative ratings is visual assessment of emphysema extent in chest CT scans. Emphysema is a pathology in chronic obstructive pulmonary disease (COPD), a leading cause of death worldwide [4]. Emphysema is characterized by destruction of lung tissue and entrapment of air. The appearance of emphysema in CT scans can be quite varied and in many cases it is difficult to precisely define where healthy tissue starts and emphysema stops. Current visual scoring systems for assessing emphysema extent are coarse yet still subject to considerable inter-rater variability [2, 15]. Emphysema assessment based on visual similarity of lung tissue could improve rater agreement while also improving the granularity of ratings and because it is not limited by current radiological definitions, it could be used to uncover new patterns.

Current practice for visual assessment of emphysema is to consider the full lung volume and decide how much is affected by emphysema [2, 15]. Comparing visual similarity of several 3D lung volumes simultaneously could be a difficult and time-consuming task, leading to worse rater agreement compared to assessing each volume by itself. Comparing visual similarity of 2D slices is a much easier task that could even be performed by non-experts with a little instruction. Simplifying the task to this degree opens the possibility of substituting medical experts with crowdworkers, leading to dramatic reductions in time consumption and costs. Crowdsourced image similarities have successfully been used for fine-grained bird classification [12], clustering of food images [14] and more recently as a possibility for assessment of emphysema patterns [6].

There is a growing body of recent work on learning from similarities derived from absolute labels [13, 8] illustrating that learning from similarities can be better than learning directly from labels. The triplet learning setting used in these works is for learning from visual image similarity where ratings for a triplet of images (x_i, x_j, x_k) are available in the form of "x_i is more similar to x_j than to x_k".

In this work we also consider similarity triplets derived from absolute labels in the form of expert assessment of emphysema extent. However, our focus is on investigating the feasibility of learning in this setting, with the future goal of learning from actual visual similarity assessment of lung images. We aim to learn descriptive image features, relevant for emphysema severity assessment, directly on the basis of visual similarity triplets. We investigate if CNNs can extract enough relevant information from a single CT slice to learn a disease relevant representation from similarity triplets. In our previous work on crowdsourcing emphysema similarity triplets [6] we did not learn a feature representation that could be used for unseen images. We believe this work is the first medical image

application where similarity triplets has been used to learn a feature representation for embedding unseen images.

2 Materials & Method

In this section we define the triplet learning problem and present a CNN based approach for learning a mapping from input images to a low dimensional representation that reflects the characteristics of the visual similarity measurements.

2.1 The triplet learning problem

Let $\mathbf{X}$ be an image space and $x_i \in \mathbf{X}$ an image. We define a similarity triplet as an ordered triplet of images (x_i, x_j, x_k) such that the ordering satisfies the triplet constraint, given by

$$\delta(x_i, x_j) \leq \delta(x_i, x_k) \tag{1}$$

where δ is some, potentially unknown, measure of dissimilarity. Let $\mathbf{T} \subseteq \mathbf{X}^3$ be a set of ordered triplets that satisfies (1). We want to find a mapping from image space to a low dimensional embedding space, $h^* : \mathbf{X} \to \mathbb{R}^d$, that minimizes the expected number of violated triplets

$$h^* = \arg\min_h \mathbb{E}_{(i,j,k)\in\mathbf{T}} \left[\mathbb{1}\{\tilde{\delta}(h(x_i), h(x_j)) \leq \tilde{\delta}(h(x_i), h(x_k))\} \right]. \tag{2}$$

where $\mathbb{1}$ is the indicator function and $\tilde{\delta} : \mathbb{R}^d \to \mathbb{R}$ is a known dissimilarity.

2.2 Learning a mapping

End-to-end learning using CNNs is a convenient and powerful method for learning concise representations of images. Optimization of CNNs is based on gradient descent and we cannot optimize (2) directly, because the subgradient is not defined. A commonly used approach is to define a loss function based on how much a triplet is satisfied or violated

$$L((x_i, x_j, x_k)) = \max\{0, \tilde{\delta}(h(x_i), h(x_j)) - \tilde{\delta}(h(x_i), h(x_k) + C)\} \tag{3}$$

where C is a fixed offset used to avoid trivial solutions and encourage over-satisfying triplet constraints. Large violations can dominate the loss (3) and force the optimization to focus on outliers. Since we expect some inconsistencies in the similarity triplets, we consider a variant of (3) that bounds the loss on both sides

$$L((x_i, x_j, x_k)) = \text{clip}_{l,u}(\tilde{\delta}(h(x_i), h(x_j)) - \tilde{\delta}(h(x_i), h(x_k))) \tag{4}$$

where

$$\text{clip}_{l,u}(x) = \begin{cases} 0 & \text{if } x < l \\ 1 & \text{if } x > u \\ \frac{x-l}{u-l} & \text{otherwise} \end{cases} \tag{5}$$

We consider two CNN architecture setups loosely based on VGGnet [9], one with increasing and one with a fixed number of filters in each layer. In both cases a layer is comprised of zeropadding, 3x3 convolution and maxpooling. After the final layer we add a global average pooling layer, and d fully connected units to obtain a d-dimensional embedding of the input. We use squared Euclidean distance as dissimilarity, i.e $\tilde{\delta} = ||\cdot||_2^2$.

2.3 Data

We use CT scans of 1947 subjects from a national lung cancer screening study [7] with visual assessment of emphysema extent [15] and segmented lung masks. Emphysema is assessed on a six-point extent scale for six regions of the lung: the upper, middle and lower regions of the left and right lung. Here we restrict our attention to the upper right region, defined as the part of the right lung lying above the carina. The six-point extent scale is defined by the intervals {0, 1-5%, 6-25%, 26-50%,51-75%,76-100%}. Distribution of emphysema scores is skewed towards 0% with about 73% having 0% and only about 13% having more than 1-5%. Example slices with varying emphysema extent are shown in Figure 1.

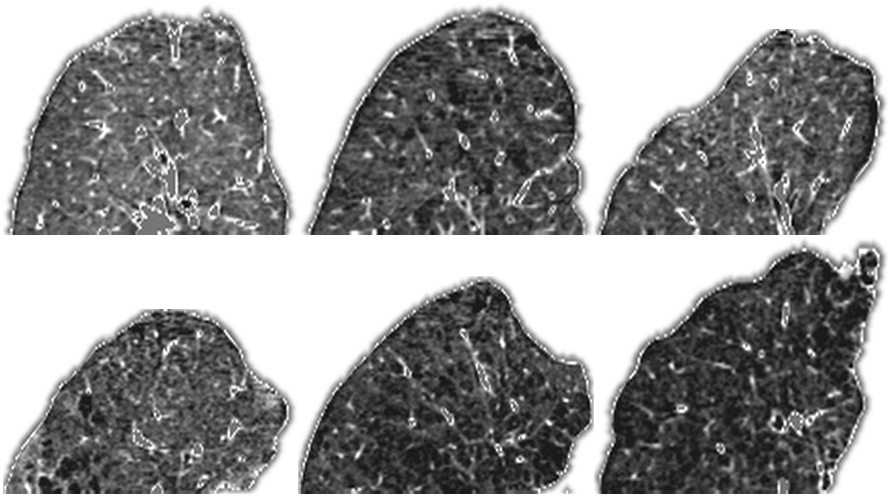

Fig. 1. Example slices. From top left, visually assessed emphysema extent is 0%, 1-5%, 6-25%, 26-50%, 51-75% and 76-100%. Window level -780HU, window width 560HU.

3 Experiments & Results

We split subjects randomly into a training group of 974 subjects and a test group of 973 subjects. For each experiment we then split the training group

randomly in half and use one half for training and the other half for validation. Each experiment was run 10 times and we report median statistics calculated over these 10 runs. We use the same clip function for all experiments, with $[l, u] = [-0.01, 0.1]$.

3.1 Preprocessing

A single coronal slice was extracted from the center of the upper right region. Bounding boxes were calculated for the lung mask of each extracted slice in the training data and all images were cropped to the size of the intersection of these bounding boxes (57×125 pixels). Pixels outside the lung mask were set to -800HU to match healthy lung tissue. This aggressive cropping was introduced to avoid background pixels dominating the input data. Finally all pixel intensities were scaled by $\frac{1}{1000}$ resulting in values roughly in the range $[-1, 0]$.

3.2 Selecting training triplets

For 974 images there are close to 10^6 possible triplets. Many of these triplets will contain very little information, and choosing the right strategy for selecting which triplets to learn from could result in faster convergence and reduce the required number of triplets needed. When class labels are available they can be used to select triplets as suggested in [8]. However, we are primarily interested in the setting where we do not have class labels. To understand the importance of triplet selection we compare uniform sampling of all possible triplets to sampling based on emphysema extent labels.

When selecting triplets based on emphysema extent we pick the first image uniformly at random from all images, the second uniformly at random from all images with the same emphysema extent as the first, and the third from from all images with different emphysema extent. For the third image we sample images with probability proportional to the absolute difference between the labels.

3.3 Simulating similarity assessment

We use visually assessed emphysema extent to simulate similarity assessment of image triplets. For a triplet of images (x_i, x_j, x_k) with emphysema extent labels (y_i, y_j, y_k) the ordering of the triplets satisfies

$$|y_{\sigma(1)} - y_{\sigma(2)}| \leq |y_{\sigma(1)} - y_{\sigma(3)}| \tag{6}$$
$$|y_{\sigma(1)} - y_{\sigma(2)}| \leq |y_{\sigma(2)} - y_{\sigma(3)}| \tag{7}$$
$$|y_{\sigma(1)} - y_{\sigma(3)}| \leq |y_{\sigma(2)} - y_{\sigma(3)}| \tag{8}$$

This corresponds to asking a rater to order images based on similarity.

3.4 CNN selection

We implemented all CNNs in Keras [1] and used the default Adam optimizer. We searched over networks with $\{3, 4, 5\}$ convolution layers. We used 16 filters for the setup with a fixed number of filters, and 8,16,32,64,128 for the setup with an increasing number of filters. We used a batch size of 15 images and trained the models for 100 epochs or until 10 epochs passed without decrease in triplet violations on the validation set. We then selected the weights with the lowest triplet violations on the validation set. We expect an untrained network with randomly initialized weights will show some degree of class separation and include it as a baseline. Table 1 summarizes median validation triplet violations of the selected models and the median number of epochs used for training. Triplet selection based on emphysema results in somewhat faster convergence and slightly fewer violations compared to uniform triplet selection. The difference in median epochs between uniform triplet selection and extent based triplet selection corresponds to 7500 extra training triplets for uniform selection.

Sampling scheme	Model type	Median epochs	Median violations
Untrained	F3	–	46.80 ± 0.94
Uniform	I4	23.0 ± 7.0	40.84 ± 0.71
Extent	F4	18.0 ± 5.0	39.30 ± 0.58

Table 1. Validation set performance. The letter in model type indicates **F**ixed or **I**ncreasing number of filters and the digit indicates number of convolution layers.

3.5 Triplet prediction performance

Selecting test triplets Because we simulate similarity assessments from class labels, the selection of test triplets will have a large influence on the interpretation of performance metrics. In our case about 71% of subjects in the test set do not have emphysema. This implies that selecting triplets uniformly at random results in about 36% of the triplets having no emphysema images. We choose to ignore these same-class triplets when measuring test performance.

In addition to the issue of same-class triplets, we are also faced with a dataset where more than 50% of those subjects that have emphysema only have 1-5% extent. Ignoring this issue will lead to performance metrics dominated by the ability of the network to distinguish subjects with very little emphysema from those without emphysema. This is a difficult task even when given access to the full volume. To more fully understand how well the network embeds images with varying levels of emphysema extent, we calculate test metrics under five different test triplet selection schemes. (1) two images with same extent and one image with different extent, (2) two images without emphysema and one with emphysema, (3) two images with 0-5% and one with $> 5\%$, (4) two images with 0-25% and one with $> 25\%$, (5) two images with $0 - 50\%$ and one with $> 50\%$.

Table 2 summarizes the results. As expected we see that the networks are much better at distinguishing between subjects with moderate to severe emphysema versus mild and no emphysema (0-5%), than subjects with emphysema versus subjects with no emphysema (0%). We also see that the untrained network provides decent separation of images with severe emphysema versus moderate to no emphysema (0-50%). In all cases we see that using information about emphysema extent for generating training triplets leads to better performance compared with uniform sampling of triplets.

	Test triplet selection method				
Sampling scheme	All	0%	0-5%	0-25%	0-50%
Uniform	41.0	40.2	30.0	19.0	11.6
Extent	39.3	39.0	26.4	14.6	9.4
Untrained	48.5	48.9	44.3	37.2	29.2

Table 2. Median triplet violations on test set for the selected models from Table 1 using different schemes for selecting test triplets. See text for explanation of column names.

An example embedding of the test set is shown in Figure 2. We used the models with best performance on the validation set to generate the embedding. Although we see significant overlap between subjects with and without emphysema, both of the trained embeddings have a reasonably pure cluster of subjects with emphysema. There is a clear tendency towards learning a one dimensional embedding. We hypothesize that several factors contribute to this tendency, (1) clipping at $[-0.01, 0.1]$ encourages small distances, (2) pairwise distances for uniformly distributed points increase as the dimensionality is increased, (3) the underlying dissimilarity space, emphysema extent, is one dimensional and all triplets can in principle be satisfied by embedding unto the real line.

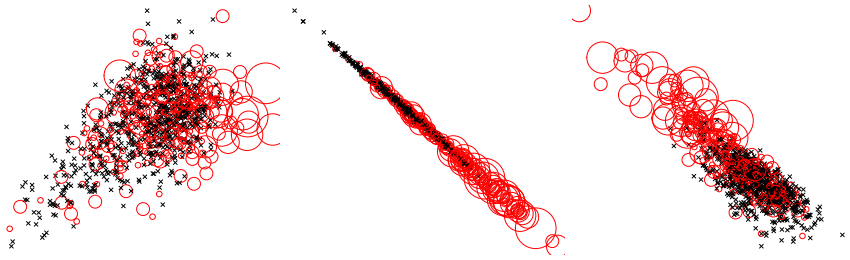

Fig. 2. Example embedding of test data. Black crosses are subjects without emphysema, red circles are subjects with emphysema. Size of circle correspond to emphysema extent. From left: Untrained (48.3% testset violations), uniform (39.5% testset violations), visual (38.8% testset violations).

4 Discussion & Conclusion

We formulated assessment of emphysema extent as a visual similarity task and presented an approach for learning an emphysema relevant feature representation from similarity triplets using CNNs. We derived similarity triplets from visual assessment and investigated the importance of selecting informative triplets.

It is slightly surprising that a single cropped 2D slice contains enough information for the level of separation illustrated by the embeddings in Figure 2. This shows that learning can be accomplished from simple annotation tasks. However, there are likely instances where the particular slice is not representative for the image as a whole, which may explain why there is a large overlap between subjects with and without emphysema in Figure 2. We suspect that with triplet similarities based on individual slice comparisons, class overlap would be less.

As a proof of concept, in this work we simulated slice similarity assessment from experts' emphysema extent scores. Potentially such triplets could be gathered online via crowdsourcing platforms such as Amazon Mechanical Turk. Our previous results [6] showed that crowdsourced triplets could be used to classify the emphysema type (rather than extent) with a better than random performance. Preliminary results indicate that the crowdsourced triplets are too few or too noisy for training the proposed CNNs. However, we expect that improving the quality and increasing the quantity of crowdsourced triplets will allow CNNs to learn an emphysema sensitive embedding without needing expert assessed emphysema extent for training.

We investigated the importance of triplet selection and found that performance improved slightly when selecting triplets based on emphysema extent, in particularly for subjects with moderate emphysema extent (columns 0-5% and 0-25% in Table 2). While using disease class labels to select triplets is not a viable solution, for medical images we often have access to relevant clinical information that could be used to select triplets. In the context of emphysema, measures of pulmonary function are potential candidates for triplet selection. However, our preliminary results indicate that using pulmonary function measures for triplet selection is not straightforward and can harm performance compared to uniform triplet selection.

We assumed that there is a single definition of visual similarity between the slices. However, this does not have to hold in general. For emphysema it is relevant to consider both pattern and extent as measures of similarity. The idea of having multiple notions of similarity is explored in [10], where different subspaces of the learned embedding corresponds to different notions of similarity. Simultaneously modeling multiple notions of similarity could lead to more expressive feature representations. Additionally, it be useful when learning from crowdsourced triplets, where some raters might focus on irrelevant aspects, such as size and shape of the lung.

In conclusion, we have shown that CNNs can learn an informative representation of emphysema based on similarity triplets. We believe this to be a promising direction for learning from relative ratings, which may be more reliable and intuitive to do, and therefore could allow the collection of large data

sets that CNNs benefit from. The next step is to explore embeddings resulting from directly annotated similarity triplets. We expect such embeddings to show different notions of similarity and it will be interesting to see how these notions compare to current radiological definitions.

References

1. F. Chollet et al. Keras. https://keras.io, 2015.
2. COPDGene CT Workshop Group: R. Graham Barr et al. A combined pulmonary-radiology workshop for visual evaluation of COPD: Study design, chest CT findings and concordance with quantitative evaluation. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 9(2):151–159, 2012.
3. R. D. Goffin and J. M. Olson. Is it all relative? comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, 6(1):48–60, 2011.
4. Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017, 2017.
5. I. Jones and C. Wheadon. Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, 47:93–101, 2015.
6. S. N. Ørting, V. Cheplygina, J. Petersen, L. H. Thomsen, M. M. Wille, and M. de Bruijne. Crowdsourced emphysema assessment. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 126–135. Springer, 2017.
7. J. H. Pedersen, H. Ashraf, A. Dirksen, K. Bach, H. Hansen, P. Toennesen, H. Thorsen, J. Brodersen, B. G. Skov, M. Døssing, J. Mortensen, K. Richter, P. Clementsen, and N. Seersholm. The Danish randomized lung cancer CT screening trial–overall design and results of the prevalence round. *Journal of Thoracic Oncology*, 4(5), 2009.
8. F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
9. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
10. A. Veit, S. Belongie, and T. Karaletsos. Conditional similarity networks. *Computer Vision and Pattern Recognition (CVPR 2017)*, 2017.
11. S. H. Wagner and R. D. Goffin. Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes*, 70(2):95–103, 1997.
12. C. Wah, G. Van Horn, S. Branson, S. Maji, P. Perona, and S. Belongie. Similarity comparisons for interactive fine-grained categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, 2014.
13. J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.
14. M. J. Wilber, I. S. Kwak, and S. J. Belongie. Cost-effective hits for relative similarity comparisons. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

15. M. M. Wille, L. H. Thomsen, A. Dirksen, J. Petersen, J. H. Pedersen, and S. B. Shaker. Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers. *Eur Radiol*, 24(11):2692–2699, Nov 2014.