A Cross-Modality Neural Network Transform for Semi-automatic Medical Image Annotation

Mehdi Moradi(✉), Yufan Guo, Yaniv Gur, Mohammadreza Negahdar, and Tanveer Syeda-Mahmood

IBM Almaden Research Center, San Jose, CA, USA
mmoradi@us.ibm.com

Abstract. There is a pressing need in the medical imaging community to build large scale datasets that are annotated with semantic descriptors. Given the cost of expert produced annotations, we propose an automatic methodology to produce semantic descriptors for images. These can then be used as weakly labeled instances or reviewed and corrected by clinicians. Our solution is in the form of a neural network that maps a given image to a new space formed by a large number of text paragraphs written about similar, but different images, by a human expert. We then extract semantic descriptors from the text paragraphs closest to the output of the transform network to describe the input image. We used deep learning to learn mappings between images/texts and their corresponding fixed size spaces, but a shallow network as the transform between the image and text spaces. This limits the complexity of the transform model and reduces the amount of data, in the form of image and text pairs, needed for training it. We report promising results for the proposed model in automatic descriptor generation in the case of Doppler images of cardiac valves and show that the system catches up to 91 % of the disease instances and 77 % of disease severity modifiers.

1 Introduction

The availability of large datasets and today's immense computational power have resulted in the success of data driven methods in traditional application areas of computer vision. In such applications, it is fairly inexpensive to label images based on crowd sourcing methods and create datasets with millions of categorized images or use the publicly available topical photo blogs. One hurdle for fully utilizing the potential of big data in medical imaging is the expensive process of annotating images. Crowd-sourcing in simple annotation tasks has been reported in the past [7,10]. However, the expert requirements for certain medical labeling and annotation tasks limit the applicability of crowd sourcing. More importantly, privacy concerns and regulations prohibit the posting of some medical records on crowd sourcing websites even in anonymized format.

Electronic medical records (EMR) are the natural sources of big data in our field. One potential solution for establishing ground truth labels such as disease type and severity for images within EMR is automatic concept extraction from

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 300–307, 2016.
DOI: 10.1007/978-3-319-46723-8_35

unstructured sources such as clinician reports stored with images. This is a very active and mature area of work [13]. In many situations, however, the clinical reports are not available. In other situations, a clinical record consists of many images and only one report. In an echocardiography study of cardiac valves, for example, there is usually many continuous wave (CW) Doppler images of four different cardiac valves. Typically these are stored as short videos. Only some patient records also include a cardiologist report (less than half in our dataset). Even when the report is available, there is no matching between each image and passages of the text. For low level algorithm development tasks, such as learning to detect a specific disease from CW Doppler, we need individually annotated images.

Our work here addresses a scenario in annotation of a set of medical images where we also have access to a rather large set of text reports from clinical records, written by clinicians based on images of the same modality from other patients. This could be a text data dump from the EMR. We do not have access to the images matched to these reports. In fact, the lack of a huge set of images and text reports that are matched with each other separates our scenario from some of the work in the machine learning community in the area of automatic image captioning [4,5].

Our goal is to speed up the process of labeling images for semantic concepts such as the imaged valve, disease type and severity by providing a fairly accurate initial automatic annotation driven by the text reports of similar images written by clinicians. To this end, we propose a learned transform between the image and text spaces. We use a multilayer perceptron (MLP) neural network which acts in the role of a universal function approximator, as opposed to a classifier. This transform network receives a fixed length representation of an image and outputs a vector in the space defined by fixed length representations of text reports. The key to success is that we have separated the process of learning the quantitative representation of images and texts from the process of learning the mapping between the two. The former relies on rather large datasets and deep learning, while the latter uses a small neural network and can be trained by using a small set of paired images and text. We show the practical value of this innovation on a clinical dataset of CW Doppler images. This methodology can significantly speed up the process of creating labeled datasets for training big data solutions in medical imaging.

2 Method

The general methodology involves three networks: a transform network that acts as a mapping function and requires a fixed length feature vector describing the image as input and outputs a fixed length text vector as output; and two deep networks that act in the capacity of feature generators to map images and text paragraphs to their corresponding fixed length spaces. We will describe the proposed methodology in the context of fast annotation of CW Doppler echocardiography images for the most common valvular diseases, namely regurgitation and stenosis, and the severity of these conditions. CW Doppler images

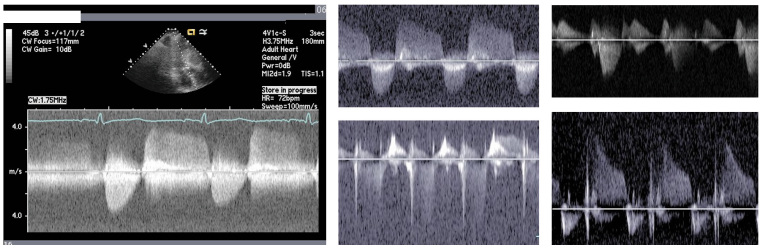

Fig. 1. Examples of the CW Doppler images: left panel shows a full CW image from the aortic valve. Right: region of interest CW images of aortic (top left), mitral (top right), tricuspid (bottom left) and pulmonic (bottom right) valve.

are routinely used for the study of mitral, tricuspid, pulmonic, and the aortic valves (Fig. 1). In the context of this specific problem, our solution also includes a fourth neural network that acts as a classifier to label the CW images for the valve. The motivation to separate this step is to limit the search space for the closest text paragraph in the final stage to only those text paragraphs that describe the relevant valve.

2.1 Learning a Fixed Length Vector Representation of Text Paragraphs

The text data was from the EMR of a local hospital network and included 57,108 cardiac echocardiography reports. The first step in our text pipeline was to isolate paragraphs focused on each of the four valve types. This was fairly trivial as the echo reports routinely include paragraphs starting with "Aortic valve:" and alike for mitral, pulmonic and tricuspid valves. With this simple rule, we isolated 10,253 text paragraphs with a valve label.

Traditionally, text can be represented as a fixed-length feature vector, composed of a variety of lexical, syntactic, semantic, and discourse features such as words, word sequences, part-of-speech tags, grammatical relations, and semantic roles. Despite the impressive performance of the aforementioned features in many text analytics tasks, especially in text classification, vector representations generated through traditional feature engineering have their limits. Given the complexity and flexibility within natural languages, features such as bag of words or word sequences usually result in a high dimensional vector, which may cause the data sparsity issues when the size of training data is incomparable to the number of features. Moreover, in a traditional feature space, words such as "narrowing", "stenosis", and "normal" are equally distant from each other, regardless of meaning.

In this work, we used a neural network language model proposed in [6] to generate distributed representations of texts in an unsupervised fashion, in the absence of deliberate feature engineering. This network is often referred to as

Doc2Vec in the literature[1]. The input of the neural network includes a sequence of observed words (e.g. "aortic valve peak"), each represented by a fixed-length vector, along with a text snippet token, also in the form of a dense vector and corresponding to the sentence/document source for the sequence. The concatenation or average of the word and paragraph vectors was used to predict the next word (e.g. "velocity") in the snippet. The two types of vectors were trained on the 10,253 paragraphs. Training was performed using stochastic gradient descent via backpropagation. At the testing stage, given an unseen paragraph, we freeze the word vectors from training time and just infer the paragraph vector.

The fixed length of the text feature vector m is a parameter in *Doc2Vec* model. In our application, since the length of the paragraphs is typically only two to three sentences, we prefer a short vector. This also helps with limiting the complexity of the transform network as it defines the number of output nodes. We report the results for $m = 10$.

2.2 Image Vectors

We rely on transfer deep learning to create a vector of learned features to represent each image. Pre-trained large deep learners such as the convolution network designed by the Visual Geometry Group (VGG) of the University of Oxford [2] have been widely used in the capacity of "feature generator" in both medical [1,11] and non-medical [9] applications, as an alternative to computation and selection of handcrafted features. We use the VGG implementation available through the MatConvNet Matlab library. This network consists of 5 convolution layers, two fully connected layers and a SoftMax layer with 1000 output nodes for the categories of the ImageNet challenge [3]. We ignore this task-specific SoftMax layer. Instead, we harvest a feature vector at the output of the fully connected layer (FC7) of the network.

The VGG network has several variations where FC7 layer has between 128 and 4096 nodes. We run each CW image through the pre-trained VGG networks with both FC7 size of 128 and 4096. The former is used for the transform network training, and the latter is used for valve type classification network. The choice of the smaller feature vector size for the transform network is due to the fact that it defines the size of the input layer. Given the small size of the dataset used to train the transform network, we keep the size of the image vectors to 128 to minimize the number of weights. For the valve classifier network, we use the 4096 dimensional representation of the images since the size of the dataset is larger and the output layer is also only limited to the number of valve classes which is four.

2.3 Valve Recognition Network

Since the text paragraphs are trivially separated based on the valve, we can reduce the errors and limit the search space in the final stage of the pipeline by

[1] Open source code: http://deeplearning4j.org/doc2vec.html.

first accurately classifying the images for the depicted valve exclusively based on the image features. In most cases, the text fields on the image (left side of Fig. 1) include clues that reveal the valve type and can be discerned using optical character recognition (OCR). In this work, however, we opt for a learning method. The classifier used in this work is an MLP network that uses the 4096 dimensional feature vector from VGG FC7 as input, has a single hidden layer, and four SoftMax output nodes each for one type of valve.

To train this valve classifier, we created an expert-reviewed dataset of 496 CW images, each labeled with one of the four valve types. The network was optimized in terms of the number of nodes in the hidden layer using leave-one-out cross-validation. The results are reported for a network with 128 nodes in the hidden layer.

2.4 The Transform Network: Architecture and Training

Universal approximation theorem states that a feedforward neural network with a hidden layer can theoretically act as a general function approximator, given sufficient training data. The transform network used in our work is designed based on this principle. This is the only network in our system that requires images and clinical text paragraph pairs.

Since this network acts as a regressor as opposed to a classifier, the output layer activation functions were set to linear as opposed to SoftMax. To optimize the number of hidden nodes of this network and train the weights, we used a dataset of 226 images and corresponding text reports, in a leave-one-out scheme. We optimized the network with the objective of minimizing the mean Euclidean distance between the output vector and the target text vector for the image. The optimal architecture had four nodes in the hidden layer.

2.5 Deployment Stage and the Independent Test Data

Given an image, we first determine the valve type using the valve classifier network. The remaining steps to arrive at the disease descriptors are depicted in Fig. 2. The given image is first passed through the VGG network. The output is fed to the transform network to obtain a vector in the text space. Then we search for the closest matches to this vector in the text dataset. The closest match, or top few, in terms of Euclidean distance are used for extraction of semantic descriptors of the image. Note that the use of the valve classifier reduces the cost of the search step by a factor of four as we only search the text paragraphs written for the same type of valve. The extraction of the semantic descriptors from the retrieved paragraphs is performed by a propriety concept extractor that accurately identifies given descriptors in the text only when they are mentioned in the positive sense [12].

The overall performance of the model is investigated on a holdout dataset of CW images that has not been used in the training or cross validation of the transform network or the valve classifier network. This consists of 48 CW images with corresponding text reports which were used only to validate the semantic

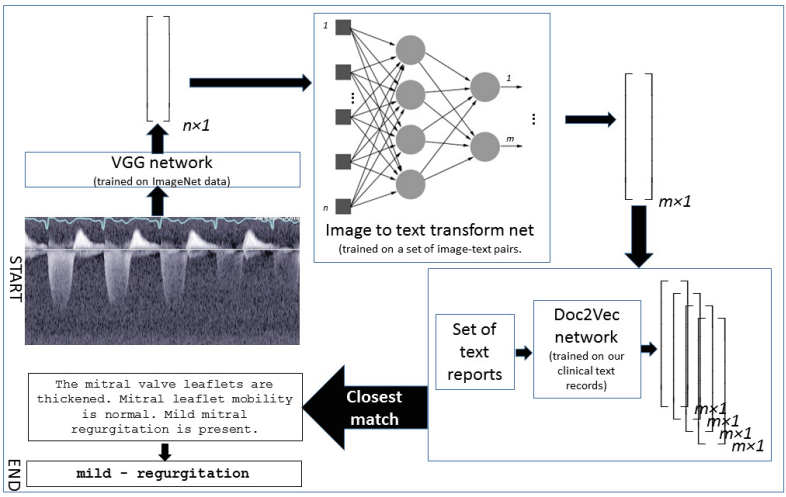

Fig. 2. The workflow of identifying a text segment as the source for semantic descriptors for a given image. The valve classifier network is not depicted in this illustration.

labels extracted for the image using our model. This test set includes 14 CW images of mitral, aortic, and tricuspid valves and six of the pulmonic valve.

3 Results

Result of classification for valve type: The optimized automatic valve classifier achieved an accuracy of 96 % on the test set, mis-classifying only two of the 48 test samples, both in case of tricuspid valve. Note that the OCR can potentially improve this by correct determination of the valve type without the need for classification, when the information is recorded on the image.

Evaluation of the quality of identified semantic labels: Table 1 provides examples of real text segments and retrieved text segments using our proposed methodology for a number of images in the test set. It is notable that the retrieved paragraphs often include anatomic information about the valve that are not discernible from CW images. For example, the thickening of the valve is visible in B-mode images. Since we use only CW images, we evaluate the results only for descriptors related to these images. It should be noted that the system is not designed as a diagnostic tool and the results here should be evaluated in light of the intended application.

For the 34 test images of tricuspid, mitral, and pulmonic valves, the most common finding is regurgitation, present in 24 of the images. In 22 cases (91 %) at least one of the top retrieved reports also reported regurgitation. The modifiers in this case included “trace”, “mild”, “moderate” and “severe”. If extracted only from the top retrieved report, the modifier is under-estimated in 9 (41 %) of the positive regurgitation findings. This drops to 5 (23 %) if the semantic descriptors are extracted from the top three retrieved reports.

Table 1. Examples of the text segments retrieved with CW images (right column) and the text segment from the original clinical report for the image (middle column).

Valve imaged	Corresponding report	Retrieved text
Aortic	"Aortic Valve: Aortic sclerosis present with evidence of stenosis. 55.1 mmHg, and the mean gradient is 27.0 mmHg, compatible with moderate stenosis"	"Aortic Valve: Severe aortic stenosis is present. The aortic valve peak velocity is 4.07 m/s, the peak gradient is 66.3 mmHg, and the mean gradient is 35.5 mmHg. The aortic valve area is estimated to be 0.61 cm2. Moderate aortic valve insufficiency"
Mitral	"Mitral Valve: There is marked mitral annular calcification. The mitral valve leaflets are calcified. Mild mitral regurgitation is present"	"Mitral Valve: Mild mitral regurgitation is present"

In case of the aortic valve, the most critical finding is stenosis. Cardiologists report aortic stenosis following the guidelines of the American Heart Association (AHA) with "mild", "moderate" or "severe" modifiers. In our set of 14 independent cases of aortic valve CW images in the final test set, the original corresponding text paragraphs reported stenosis in five cases. In all five cases, the combination of the top three retrieved paragraphs provided "stenosis" as a descriptor. In one case, there was a finding of stenosis in the top retrieved paragraph, but not in the original report. Although further examination revealed that the case was positive based on one measure of stenosis (maximum jet velocity) and negative based on another measure (mean pressure gradient). For modifiers, in four cases the original modifier was "mild" and the true modifier was also either moderate or mild. In one case, the clinician had not reported a modifier and the retrieved paragraph reported "severe".

4 Conclusion

We proposed a methodology for generating annotations, in form of semantic disease related labels, for medical images based on a learned transform that maps the image to a space formed by a large number of text segments written by clinicians for images of the same type. Note that we used a pre-trained convolutional neural network. Handcrafted feature sets such as histogram of gradients can be studied as alternative image descriptors in this framework. However, the CNN based features proved more powerful in our previous work [8] and also here.

While quantitative analysis reported here is limited to stenosis and regurgitation, there is no such limitation in our implementation. Our evidence from over 10,000 text reports show that we can cover a wide range of labels. For example,

we can accurately pick up labels related to deficiencies such as valve thickening, calcification and decreased excursion. As examples in Table 1 show, in many cases the retrieved reports also include values of relevant measured clinical features. As a future step we will explore the idea of expanding the list of top matches and averaging the values to obtain a rough estimate of the measurements for the image of interest. Also, inclusion of B-mode images can improve the value of the retrieved paragraphs that often include features only visible in such images. Finally, a larger user study is under way to understand the practical benefits of the system in terms of cost saving.

References

1. Bar, Y., Diamant, I., Wolf, L., Greenspan, H.: Deep learning with non-medical training used for chest pathology identification. In: SPIE, Medical Imaging 2015 (2015)
2. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: British Machine Vision Conference (2014)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: CVPR (2009)
4. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
5. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Understanding and generating image descriptions. In: CVPR (2011)
6. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: ICML, pp. 1188–1196 (2014)
7. Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H.G., Eisenmann, M., Speidel, S.: Can masses of non-experts train highly accurate image classifiers? In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 438–445. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10470-6_55
8. Moradi, M., et al.: A hybrid learning approach for semantic labeling of cardiac CT slices and recognition of body position. In: IEEE ISBI, pp. 1418–1421 (2016)
9. Park, C.C., Kim, G.: Expressing an image stream with a sequence of natural sentences. In: NIPS (2015)
10. Rodrguez, A.F., Muller, H.: Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In: Proceedings of the ACM Workshop on Crowdsourcing for Multimedia (2012)
11. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**, 1285–1298 (2016)
12. Syeda-Mahmood, T., Chiticariu, L.: Extraction of information from clinical reports 29 Aug 2013. http://www.google.com/patents/US20130226841, US Patent App. 13/408,906
13. Wang, F., Syeda-Mahmood, T., Beymer, D.: Information extraction from multimodal ECG documents. In: ICDAR, pp. 381–385 (2009)