Adopting Semantic Information of Grayscale Radiographs for Image Classification and Retrieval

Obioma Pelka[1,2], Felix Nensa[3] and Christoph M. Friedrich[1]
[1]*Department of Computer Science, University of Applied Sciences and Arts Dortmund,*
Emil-Figge-Strasse 42, 44227 Dortmund, Germany
[2]*Faculty of Medicine, University of Duisburg-Essen, Germany*
[3]*Department of Diagnostic and Interventional Radiology and Neuroradiology,*
University Hospital Essen Hufelandstrasse 55, 45147 Essen, Germany

Keywords: Biomedical Imaging, Deep Learning, Keyword Generation, Machine Learning, Multi-modal Representation, Transfer Learning, Radiographs.

Abstract: As the number of digital medical images taken daily rapidly increases, manual annotation is impractical, time-consuming and prone to errors. Hence, there is need to create systems that automatically classify and annotate medical images. The aim of this presented work is to utilize Transfer Learning to generate image keywords, which are substituted as text representation for medical image classification and retrieval tasks. Text preprocessing methods such as detection and removal of compound figure delimiters, stop-words, special characters and word stemming are applied before training the keyword generation model. All images are visually represented using Convolutional Neural Networks (CNN) and the Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) Show-and-Tell model is adopted for keyword generation. To improve model performance, a second training phase is initiated, where parameters are fine-tuned using the pre-trained deep learning network Inception-ResNet-V2. For the image classification tasks, Random Forest models trained with Bag-of-Keypoints visual representations were adopted. Classification prediction accuracy was higher for all classification schemes and on two distinct radiology image datasets using the proposed approach.

1 INTRODUCTION

Due to advances in software, hardware, and digital imaging in the medical domain, the number of images taken per patient scan has rapidly increased (Rahman et al., 2007; Tagare et al., 1997). To decrease the burden on radiologists and maintain the maximum interpretation of these radiology images, there is need to create automatic computer-aided interpretation, which can be further applied for image annotation and semantic information extraction.

An important criteria for creating an effective classification system, is the selection and combination of features for an adequate representation of the images. As shown in (Pelka and Friedrich, 2015; Codella et al., 2014; Valavanis et al., 2016; Kalpathy-Cramer et al., 2015; Pelka and Friedrich, 2016), multi-modal representation achieves higher classification rates in biomedical annotation tasks. This is the combination of visual and text representations which sufficiently represents these biomedical images.

However, for real clinical cases and some image classification tasks such as ImageCLEF2009 Medical Annotation Task (Tommasi et al., 2009) and ImageCLEF 2015 Medical Clustering Task (Amin and Mohammed, 2015), corresponding text representations are not available. In this paper, Transfer Learning (Pan and Yang, 2010) is utilized to generate keywords (Pelka and Friedrich, 2017), which are combined with visual features to obtain multi-modal image representations. These text features are further adopted for the medical image classification tasks mentioned above and semantic tagging.

As deep learning techniques (LeCun et al., 2015) have improved prediction accuracies in object detection (Huang et al., 2017), speech recognition (Hinton et al., 2012) and in domain application such as medical imaging (Abrao et al., 2007; Xu et al., 2014), a deep learning architecture is used to create the keyword generation model. Deep Convolutional Neural Networks (dCNN) (Szegedy et al., 2017) are applied to encode the medical images to a feature rep-

Pelka, O., Nensa, F. and Friedrich, C.
Adopting Semantic Information of Grayscale Radiographs for Image Classification and Retrieval.
DOI: 10.5220/0006732301790187
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 2: BIOIMAGING*, pages 179-187
ISBN: 978-989-758-278-3

resentation which is decoded using a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) based Recurrent Neural Network (RNN) (Bengio et al., 1994) to generate appropriate keywords for a given image.

For the image classification tasks, Random Forest (Breiman, 2001) models trained with visual and text representations were adopted. Bag-of-Keypoints (Csurka et al., 2004; Lazebnik et al., 2006) computed with dense SIFT descriptors (Dalal and Triggs, 2005) were combined with text representations obtained by vector quantization on a Bag-of-Words (Salton and McGill, 1983) codebook. The codebook was created with words from the keyword generation model.

Figure 1 shows the complete workflow for the approach presented in this paper. **PART 1** displays necessary steps for creating a keyword generator. This is a distinct and stand-alone process which does not need **PART 2** for application. The keyword generator can be further adopted for several purposes, such as image classification and retrieval. In **PART 2**, the keyword generator is used to create keywords for medical datasets that lack text representations. This second part is dependable on the first part. Two datasets containing grayscale radiographs were utilized in **PART 2**, however there are no restrictions, as the keyword generator was created using biomedical literature figures. The rest of this paper is structured as follows: Section 2.1 lists the adopted dataset for keyword generation and the two datasets used for evaluating the proposed approach. In subsections 2.2 and 2.3, applied deep learning networks, visual representation and machine learning methods for keyword generation and image classification are described. The achieved results are stated in section 3. Finally, results are discussed in section 4 and conclusions are drawn in section 5.

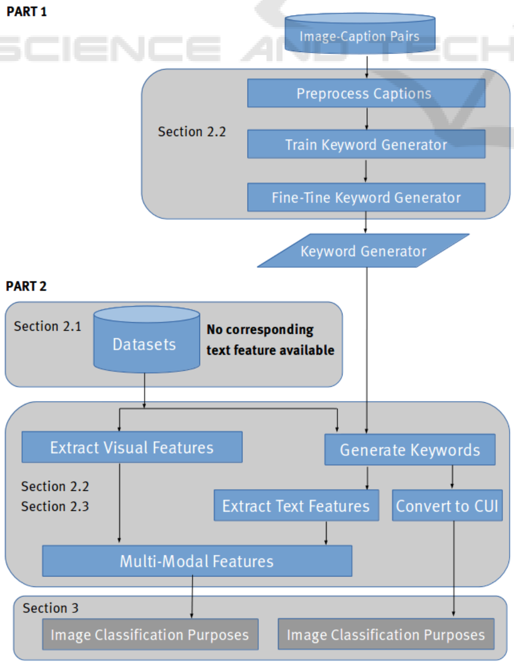


Figure 1: Overview of the proposed approach workflow.

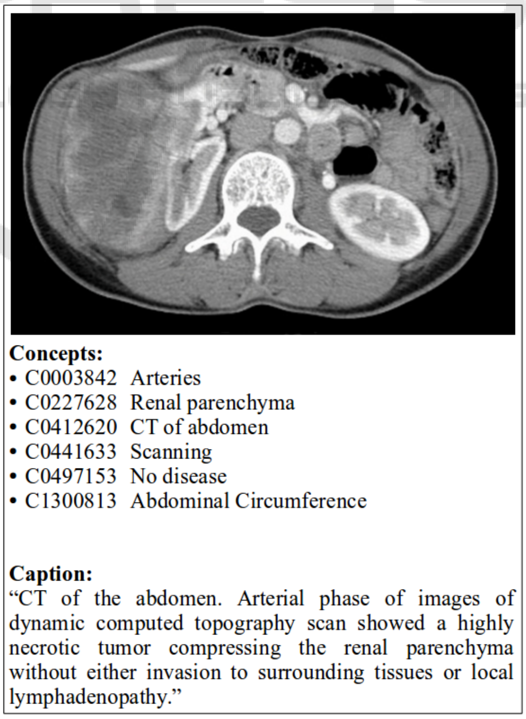
2 MATERIAL AND METHOD

2.1 Datasets

Three datasets are applied for this proposed work:

- ImageCLEFcaption Prediction 2017: For keyword generation model
- ImageCLEF 2009 Medical Image Annotation Task: For image classification and evaluation
- ImageCLEF 2015 Medical Clustering Task: For image classification and evaluation

ImageCLEFcaption Prediction 2017. This dataset was distributed at the ImageCLEFcaption 2017 Task (Eickhoff et al., 2017). ImageCLEFcaption 2017



Concepts:

- C0003842 Arteries
- C0227628 Renal parenchyma
- C0412620 CT of abdomen
- C0441633 Scanning
- C0497153 No disease
- C1300813 Abdominal Circumference

Caption:
"CT of the abdomen. Arterial phase of images of dynamic computed topography scan showed a highly necrotic tumor compressing the renal parenchyma without either invasion to surrounding tissues or local lymphadenopathy."

Figure 2: Example of a medical image with corresponding caption and Concept Unique Identifiers (CUI). The computer tomography scan was randomly chosen from the validation set of the ImageCLEFcaption 2017 Task.

consisted of two subtasks: Concept Detection and Caption Prediction. All distributed figures originate from open access biomedical journal articles published in PubMed Central (PMC) (PubMed Central, 2017). The objective of the concept detection task was to retrieve clinical concepts present in the medical images whereas for the caption prediction task, meaningful captions to the images had to be predicted (Eickhoff et al., 2017). Figure 2 shows an example of an image with the corresponding information provided in the distributed dataset. The same datasets were distributed for both Concept Detection and Caption Prediction tasks and includes a variety of content and situations, ranging from ultrasound images and x-rays to charts and clinical photographs, which can be seen in figure 3.

For the Caption Prediction Task, a training set containing 164,614 image - caption pairs and an additional validation set of 10,000 biomedical image - caption pairs for evaluation purposes in the development stage were distributed. Official evaluation was computed using BLEU scores (Papineni et al., 2002) on a test set with 10,000 biomedical images. For the Concept Detection Task, a set of UMLS® (Unified Medical Language System) Concept Unique Identifiers (CUIs) (Bodenreider, 2004) were provided for each of the 164,614 biomedical images. These UMLS concepts were identified using the QuickUMLS library (Soldaini and Goharian, 2016) from the original captions published with the images (Eickhoff et al., 2017). F_1-Score was used as evaluation metric.

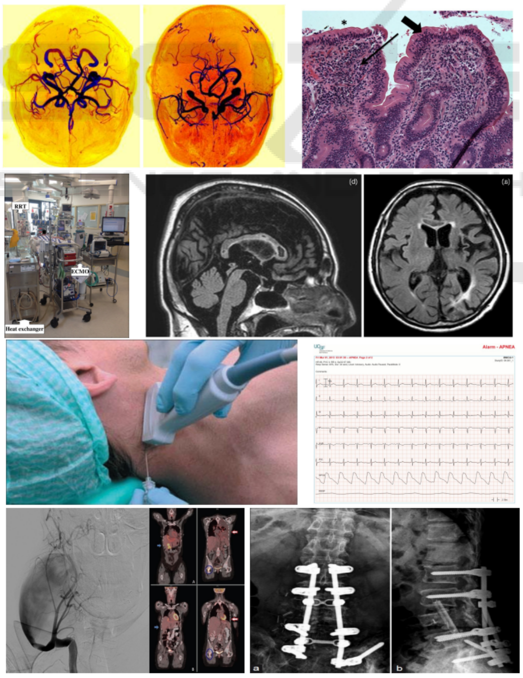


Figure 3: Examples of biomedical images showing variety of content and situation. All images were randomly chosen from the validation set of the ImageCLEFcaption 2017 Task.

ImageCLEF 2015 Medical Clustering Task. This dataset was distributed at the Medical Clustering Task held at ImageCLEF 2015 and contains high resolution x-ray images collected from a hospital in Dhaka, Bangladesh (Amin and Mohammed, 2015). X-rays of both male and female patients aged 6 months to 72 years were present in the distributed dataset (Amin and Mohammed, 2015). The training set included 500 images and the test set distributed for evaluation contained 250 images. For each of the classes, 'Body', 'Head-Neck', 'Upper-Limb, 'Lower-Limb' and 'True-Negative', 100 images were present in the training set. An example of the x-rays is shown in figure 4.

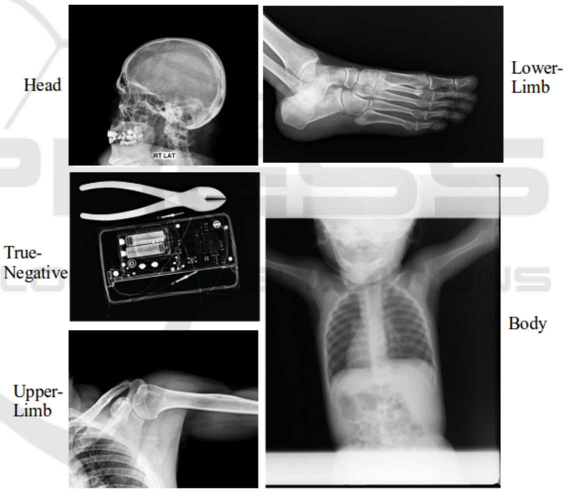


Figure 4: Examples of x-rays with corresponding classification. All images were randomly chosen from the training set of the ImageCLEF 2015 Medical Clustering Task.

ImageCLEF 2009 Medical Image Annotation Task. The dataset was distributed at the ImageCLEF 2009 Medical Annotation task (Tommasi et al., 2009). The training set consists of 12,671 grayscale x-rays and the official evaluation set has 1,732 grayscale x-rays. Each radiograph in the training set is annotated with a 13 character string denoting the Image Retrieval in Medical Applications (IRMA) (Lehmann et al., 2004) classification code.

The IRMA-code describes the modality of the images, orientation of the image, examined body region and the biological system investigated. This classification scheme contains 193 distinct classes. Figure 6 shows two radiographs with annotations *1121-127-732-500* and *1121-410-620-625*, representing ”xray overview image; coronal anteroposterior; middle right abdomen; gastrointestinal system” and ”xray analog low beam energy; other oblique orientation; left breast; Reproductive female system breast”.

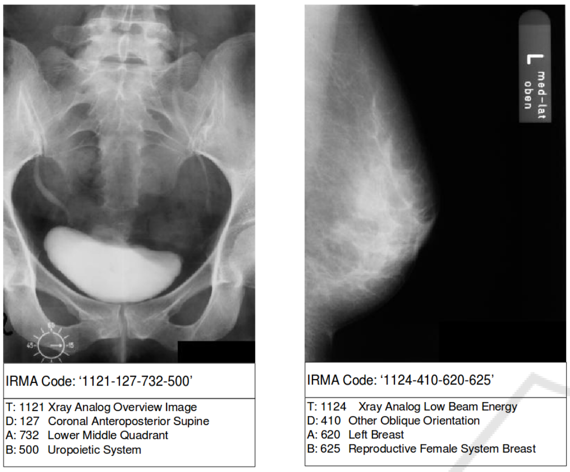

IRMA Code: '1121-127-732-500'
T: 1121 Xray Analog Overview Image
D: 127 Coronal Anteroposterior Supine
A: 732 Lower Middle Quadrant
B: 500 Uropoietic System
L
med-lat
oben
IRMA Code: '1124-410-620-625'
T: 1124 Xray Analog Low Beam Energy
D: 410 Other Oblique Orientation
A: 620 Left Breast
B: 625 Reproductive Female System Breast

Figure 6: Example of two grayscale radiographs annotated with the 13-digit IRMA classification code. Both images are from the ImageCLEF 2009 Medical Annotation Task Training Set.

2.2 Keyword Generation

For keyword generation, a combination of encoding and decoding using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) based Recurrent Neural Networks (RNN) (Bengio et al., 1994) is adopted. This approach, also known as Show-And-Tell model was proposed in (Vinyals et al., 2015) and further improved in (Vinyals et al., 2017).

To produce rich visual representations of the images, CNN is used as an image encoder by pretraining it for an image classification task. The LSTM-RNN utilized as caption decoder generates the image keywords, using the CNN last hidden layer as input (Vinyals et al., 2015).

Figure 5 shows the keyword generation model training setup. In the first training phase, the LSTM is trained using a corpus of paired image and captions generated from the biomedical figures in the ImageCLEF 2017 Caption Prediction Task Training Set (Eickhoff et al., 2017). No further dataset was used for training. In the second phase, parameters of the image model and LSTM are fine-tuned using the deep learning network Inception-ResNet-V2 (Szegedy et al., 2017). The parameters for the image keyword generation model are:

- Batch size = [1. Trainingphase = 32; 2. Training-

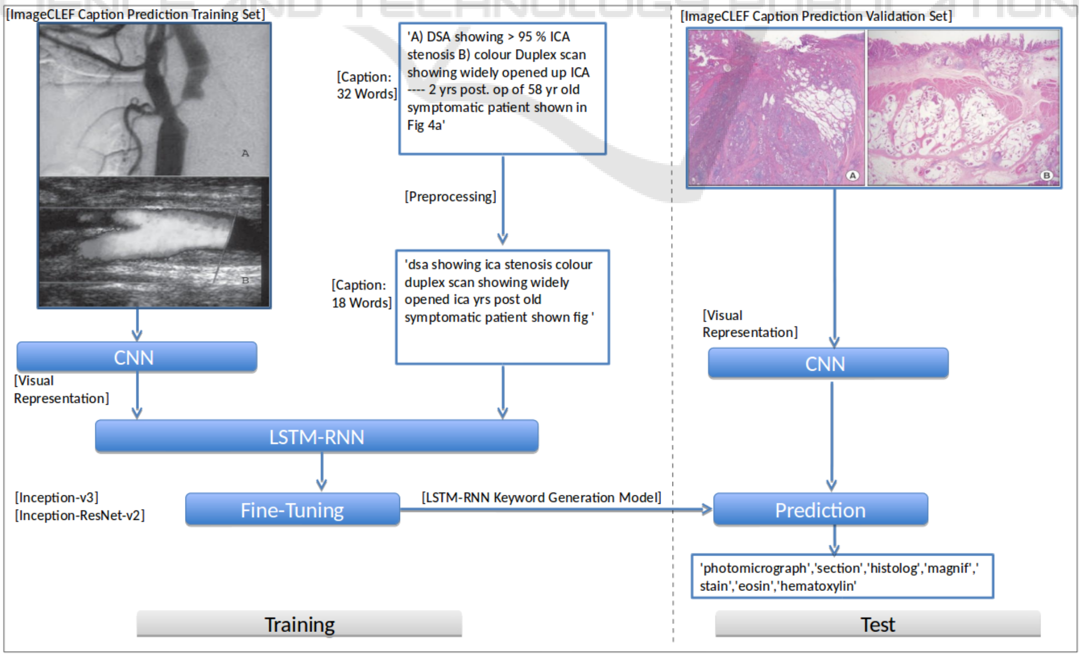

[ImageCLEF Caption Prediction Training Set]
A
B
[Caption: 32 Words]
'A) DSA showing > 95 % ICA stenosis B) colour Duplex scan showing widely opened up ICA ---- 2 yrs post. op of 58 yr old symptomatic patient shown in Fig 4a'
[Preprocessing]
[Caption: 18 Words]
'dsa showing ica stenosis colour duplex scan showing widely opened ica yrs post old symptomatic patient shown fig '
CNN
[Visual Representation]
LSTM-RNN
[Inception-v3]
[Inception-ResNet-v2]
Fine-Tuning
[LSTM-RNN Keyword Generation Model]
Training
[ImageCLEF Caption Prediction Validation Set]
A
B
[Visual Representation]
CNN
Prediction
'photomicrograph','section','histolog','magnif','stain','eosin','hematoxylin'
Test

Figure 5: Overview of Long Short-Term Memory based Recurrent Neural Network Model applied for biomedical image keyword generation.

phase = 32]

- Number of Epochs = [1. Trainingphase = 194; 2. Trainingphase = 583]
- Vocabulary size = 23,000 {Minimum word occurrence ≥ 4}
- Initial learning rate = 2
- Model optimizer = stochastic gradient descent
- Learning rate decay factor = 0.5
- Number of epochs per decay = 8
- Inception learning rate = 0.0005
- Inception model initialization = Inception-ResNet-V2
- LSTM embedding size = 512
- LSTM units number = 512
- LSTM initializer scale = 0.08
- LSTM dropout keep probability = 0.7

For all other parameters not mentioned above, the default values as proposed in (Vinyals et al., 2015) and implemented in the Tensorflow-Slim **im2txt**-model (Abadi et al., 2015; Shallue, 2017) were adopted.

Several text preprocessing methods such as reduction of image captions to nouns and adjectives, removal of stopwords (Bird et al., 2009) and special characters, and word stemming (Porter, 1980) were performed. These text preprocessing steps are further detailed in (Pelka and Friedrich, 2017).

2.3 Classification

Visual Representation. For whole-image classification tasks, the Bag-of-Keypoints (BoK) (Csurka et al., 2004) approach has achieved high classification accuracy results (Lazebnik et al., 2006; Zhang et al., 2006). BoK is based on vector quantization of affine invariant descriptors of image patches (Csurka et al., 2004). The simplicity and invariance to affine transformation are advantages that come with this approach.

All functions applied for visual representation computation are from the *VLFEAT* library (Vedaldi and Fulkerson, 2010). Dense SIFT (dSIFT) (Li and Perona, 2005) applied at several resolutions were uniformly extracted with an interval of 4 pixels using the *VL-PHOW* function. Computational time was sped up by computing k-means clustering with Approximated Nearest Neighbor (ANN) (Indyk and Motwani, 1998) on randomly chosen descriptors using the *VL-KMEANS* function. This partitions the observations into k clusters so that the within-cluster sum of square is minimized.

A maximum number of 20 iterations was defined to allow the k-means algorithm converge and cluster centers were initialized using random data points (Hartigan and Wong, 1979). A codebook containing 1,000 keypoints was generated as k = 1,000. Using the *VL-KDTREEBUILD* function, the codebook was further optimized by adapting a kd-tree with metric distance L_2 for quick nearest neighbor lookup.

Text Representation. Utilizing the keyword generation model described in subsection 2.2, keywords were generated for all radiology images in both ImageCLEF 2009 Medical Annotation Task and ImageCLEF 2015 Medical Clustering Task datasets. Figures 7 and 8 show generated keywords for randomly chosen radiographs from the ImageCLEF 2009 Medical Annotation Task Training Set and ImageCLEF 2015 Medical Clustering Task Training Set, respectively. No further text preprocessing methods were applied to the generated keywords, as this was done before creating the keyword generation model.

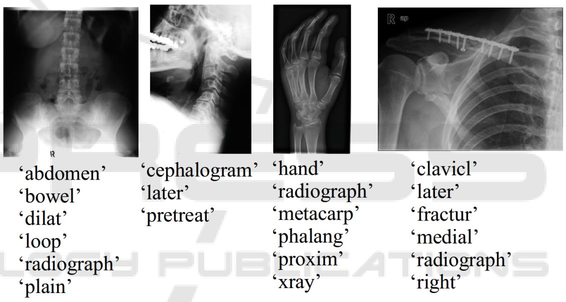


Figure 7: Examples of keywords generated. All images are from the ImageCLEF 2009 Medical Annotation Task Training Set. Keyword generation model was created using the ImageCLEFcaption Prediction Task Training Set.

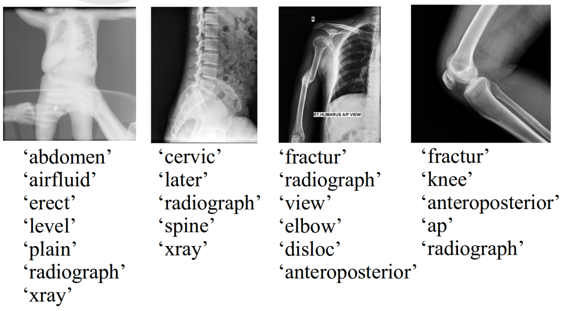


Figure 8: Examples of keywords generated. All images are from the ImageCLEF 2015 Medical Clustering Task Training Set. Keyword generation model was created using the ImageCLEFcaption Prediction Task Training Set.

For text representation, the Bag-of-Words (BoW) (Salton and McGill, 1983) approach was adopted. The basic concept is to extract features by counting the frequency or presence of words in the text to be classified. These words have to be first defined in a dictionary or codebook. The fifty words with the highest occurrence were used as dictionary, hence obtaining a feature vector size of 50. Several dictionary sizes as well the benefit of Information Gain (Guyon and Elisseeff, 2003) were investigated, but not further used as no relevant advantage was detected.

Random Forest. Random forest (RF) (Breiman, 2001) models with 1,000 deep trees were created as image classifiers. These RF-models were trained using multi-modal image representation. This is the concatenation of visual features derived with the BoK approach and the text features computed using BoW. To reduce computational time, feature dimension and noise, Principal Component Analysis (PCA) (Jolliffe, 2011) was applied on the visual representation. Parameters used to tune BoK and RF are:

- Codebook size: 1,000
- Number of descriptors extracted: 1,000
- Visual representation size: 4,000 (2x2 grid)
- Feature size reduction: 4000 to 100 (Principal Component Analysis)
- Number of trees (RF): 1,000
- Ensemble method (RF): Bag

Support Vector Machines. Multi-class Support Vector Machine (SVM) (Burges, 1998) image classification models were created for comparison. These SVM-models were trained using the same multi-modal image representation applied with the Random Forest models. Parameters used to tune the SVM-Models are:

- Kernel type: Radial basis function
- Cost parameter: 10
- Gamma: 1 / Number-Of-Features

Classification Schemes. For the ImageCLEF 2009 Medical Annotation Task, 4 different classification schemes were used for evaluation. These schemes are derived by using the complete IRMA code, mentioned in subsection 2.1, as well splitting the code to its' four axes.

- (T) Technical-Code: 6 classes
- (D) Directional-Code: 34 classes
- (A) Anatomical-Code: 97 classes
- (B) Biological-System-Code: 11 classes

For the ImageCLEF 2015 Medical Clustering Task, a classification scheme of four (4) classes was used.

3 RESULTS

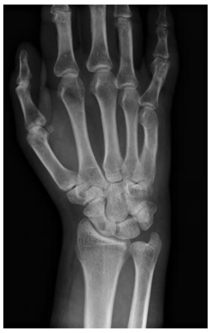
Table 2 shows generated keywords grouped to the ImageCLEF 2015 Medical Clustering Task classification scheme. Figure 9 displays the keywords and corresponding UMLS CUIs, and prediction performance of the random forest classification model is shown in table 1.

Table 2: Keywords frequently generated for radiology images of the ImageCLEF 2015 Medical Clustering Task Training Set. The keywords are grouped to the classification scheme accordingly.

Body	Head-Neck	Lower-Limb	Upper-Limb
bodi	massiv	ray	posterior
obstruct	right	ulna	acromion
level	neck	union	ulna
lesion	effus	humerus	shoulder
plain	pleural	shaft	carpus
abdomen	disc	radius	end
distal	fractur	embrochag	midshaft

Table 1: Prediction accuracies obtained using the random forest classification and support vector machine models. Column 'Visual Features' shows accuracy with visual representation whereas 'Multi-Modal' shows performance accuracies obtained with the combination of visual and text representations.

Classification Scheme	Visual Features		Multi-Modal		Test Set
	Random Forest	SVM	Random Forest	SVM	
Technical-Code	97.00%	86.84%	**97.75%**	95.35%	1,732
Directional-Code	61.64%	55.31%	**62.41%**	61.26%	1,732
Anatomical-Code	51.15%	54.16%	54.62%	**57.79%**	1,732
Biological-Code	90.76%	81.12%	**91.74%**	82.47%	1,732
Medical Clustering	65.60%	66.40%	**70.40%**	68.80%	250



'bone':	**C0262950** Skeletal bone
'hand':	**C0018563** Hand
'metacarp':	**C0025526** Metacarpal bone
'phalang':	**C0223792** Phalanx of hand
'proxim':	**C0205107** Proximal
'xray':	**C1306645** Plain Xray

Figure 9: Generated keywords and the corresponding Unified Medical Language Systems (UMLS) Concept Unique Identifiers (CUI). The radiograph used for demonstration was randomly chosen from the ImageCLEF 2009 Medical Annotation Task Training Set.

4 DISCUSSION

It can be seen from table 1, that substituting the generated keywords as text representation improves the classification prediction accuracy in both datasets and all classification schemes. This positive increase is obtained regardless of the classification method, as both RF-models and SVM-models predicted better with multi-modal representations.

The prediction performance varies to classification scheme and method. The hierarchical approach of splitting the IRMA-code to its four axes, proved to be the better way to address this image annotation task.

As Deep Convolutional Neural Networks have proven the obtain improved prediction accuracies, an approach combining generated keywords with features extracted from the activation of a deep convolutional network (DeCAF) (Donahue et al., 2014) is intended. Positive results regarding biomedical image classification using DeCAF were reported in (Koitka and Friedrich, 2016).

The presented approach can be utilized for image structuring and tagging of semantic information. The generated keywords can be transformed to Unified Medical Language Systems (UMLS) Concept Unique Identifiers (CUIs), which is displayed in figure 9. The conversion was obtained by applying QuickUMLS (Soldaini and Goharian, 2016). The converted CUIs are valuable and essential in terms of image retrieval purposes.

5 CONCLUSION

As multi-modal image representation has proven to obtain higher prediction results and some image dataset lack text representation, an approach to generate keywords utilizing transfer learning was proposed. To create a keyword generation model, image-caption pairs of 164,614 biomedical figures distributed at the ImageCLEFcaption 2017 Caption Prediction Task was adopted to train Long Short-Term Memory based Recurrent Neural Network models. The image captions were preprocessed by removing compound figure delimiters, single digits, special characters, word stemming and reducing the captions to nouns and adjectives.

Utilizing the keyword generation model, text representation were created for two distinct radiology datasets: ImageCLEF 2009 Medical Annotation Task and ImageCLEF 2015 Medical Clustering Task. The ImageCLEF 2015 Medical Clustering Task training set contains 500 high resolution radiographs, 250 in the test set and has a classification scheme with 4 classes. The ImageCLEF 2009 Medical Annotation Task has 12,671 radiographs in the training set, 1,732 radiographs in the test set, and four classification schemes with 5, 34, 97, 11 classes, respectively.

The generated keywords were further applied for image classification purposes. In both image datasets and all classification schemes, the prediction accuracies obtained with multi-modal image representation outperformed those achieved using just visual features. Using these generated keywords, semantic information in form of Unified Medical Language Systems (UMLS) Concept Unique Identifiers (CUIs) can be tagged to the images, which is beneficial and of assistance to image retrieval solutions. The proposed work can be further enhanced by extracting image visual representation using Deep Convolutional Neural Networks and optimized Bag-of-Words.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Abrao, M. S., Gonçalves, M. O. d. C., Dias Jr, J. A., Podgaec, S., Chamie, L. P., and Blasbalg, R. (2007). Comparison between clinical examination, transvaginal sonography and magnetic resonance imaging for the diagnosis of deep endometriosis. *Human Reproduction*, 22(12):3092–3097.

Amin, M. A. and Mohammed, M. K. (2015). Overview of the ImageCLEF 2015 Medical Clustering Task. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*.

Bengio, Y., Simard, P. Y., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly, first edition.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database-Issue):267–270.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery*, 2(2):121–167.

Codella, N. C. F., Connell, J. H., Pankanti, S., Merler, M., and Smith, J. R. (2014). Automated medical image modality recognition by fusion of visual and text information. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014 - 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II*, pages 487–495.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision, ECCV 2004, Prague, Czech Republic, May 11-14, 2004*, pages 1–22.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA. IEEE Computer Society.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 647–655.

Eickhoff, C., Schwall, I., de Herrera, A. G. S., and Müller, H. (2017). Overview of ImageCLEFcaption 2017 - Image Caption Prediction and Concept Detection for Biomedical Images. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*.

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-a., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, USA, July 22-25, 2017*.

Indyk, P. and Motwani, R. (1998). Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA. ACM.

Jolliffe, I. T. (2011). Principal Component Analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer Berlin Heidelberg.

Kalpathy-Cramer, J., de Herrera, A. G. S., Demner-Fushman, D., Antani, S. K., Bedrick, S., and Müller, H. (2015). Evaluating performance of biomedical image retrieval systems - An overview of the medical image retrieval task at ImageCLEF 2004-2013. *Computerized Medical Imaging and Graphics*, 39:55–61.

Koitka, S. and Friedrich, C. M. (2016). Traditional Feature Engineering and Deep Learning Approaches at Medical Classification Task of ImageCLEF 2016. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 304–317.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, Ney York, USA, June 17-22 2006*, pages 2169–2178.

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep Learning. *Nature*, 521(7553):436–444.

Lehmann, T. M., Güld, M. O., Thies, C., Plodowski, B., Keysers, D., Ott, B., and Schubert, H. (2004). IRMA - Content-Based Image Retrieval in Medical Applications. In *MEDINFO 2004 - Proceedings of the 11th World Congress on Medical Informatics, San Francisco, California, USA, September 7-11, 2004*, pages 842–846.

Li, F.-F. and Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, San Diego, USA, June 20-26, 2005*, pages 524–531.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Pelka, O. and Friedrich, C. M. (2015). FHDO Biomedical Computer Science Group at Medical Classification Task of ImageCLEF 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

Pelka, O. and Friedrich, C. M. (2016). Modality prediction of biomedical literature images using multimodal feature representation . *GMS Medizinische Informatik, Biometrie und Epidemiologie*, 12(2):1345–1359.

Pelka, O. and Friedrich, C. M. (2017). Keyword Generation for Biomedical Image Retrieval with Recurrent Neural Networks. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*

Porter, M. (1980). An algorithm for suffix stripping. *Program-electronic Library and Information Systems*, 14:130–137.

PubMed Central (2017). National institutes of health NIH, US National Library of Medicine. In *https://www.ncbi.nlm.nih.gov/pmc/ Last accessed: 2017-11-30.*

Rahman, M. M., Bhattacharya, P., and Desai, B. C. (2007). A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques With Relevance Feedback. *IEEE Transactions on Information Technology in Biomedicine*, 11(1):58–69.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill computer science series. McGraw-Hill, New York.

Shallue, C. (2017). Im2txt github. In *https://github.com/tensorflow/models/tree/master/research/im2txt, Last accessed: 2017-11-30.*

Soldaini, L. and Goharian, N. (2016). QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *Medical Information Retrieval (MedIR) Workshop in SIGIR Conference on Research and Development in Information Retrieval 2016, Pisa, Italy, July 17-21, 2016.*

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4278–4284.

Tagare, H. D., Jaffe, C. C., and Duncan, J. S. (1997). Synthesis of research: Medical image databases: A content-based retrieval approach. *Journal of the American Medical Informatics Association JAMIA*, 4(3):184–198.

Tommasi, T., Caputo, B., Welter, P., Güld, M. O., and Deserno, T. M. (2009). Overview of the CLEF 2009 Medical Image Annotation Track. In *Multilingual Information Access Evaluation II. Multimedia Experiments - 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, pages 85–93.

Valavanis, L., Stathopoulos, S., and Kalamboukis, T. (2016). IPL at CLEF 2016 Medical Task. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 413–420.

Vedaldi, A. and Fulkerson, B. (2010). VLFEAT: an open and portable library of computer vision algorithms. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 1469–1472.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and Tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2017). Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663.

Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., and Chang, E. I. (2014). Deep learning of feature representation with multiple instance learning for medical image analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 1626–1630.

Zhang, H., Berg, A. C., Maire, M., and Malik, J. (2006). SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, Ney York, USA, June 17-22 2006*, pages 2126–2136.