Traditional Feature Engineering and Deep Learning Approaches at Medical Classification Task of ImageCLEF 2016

FHDO Biomedical Computer Science Group (BCSG)

Sven Koitka[1,2] and Christoph M. Friedrich[1]

[1] University of Applied Sciences and Arts Dortmund (FHDO)
Department of Computer Science
Emil-Figge-Strasse 42, 44227 Dortmund, Germany
sven.koitka@fh-dortmund.de and christoph.friedrich@fh-dortmund.de
http://www.inf.fh-dortmund.de
[2] TU Dortmund University
Department of Computer Science
Otto-Hahn-Str. 14, 44227 Dortmund, Germany

Abstract. This paper describes the modeling approaches used for the *Subfigure Classification* subtask at *ImageCLEF 2016* by the *FHDO Biomedical Computer Science Group (BCSG)*. Besides traditional feature engineering, modern *Deep Convolutional Neural Networks (DCNN)* were used, trained from scratch and using a transfer learning scenario. In addition *Bag-of-Visual-Words (BoVW)* were computed in Opponent color space, since some classes in this subtask can be distinguished by color. To remove unimportant visual words the *Information Gain* is used for *Feature Selection*. Overall BCSG achieved top performance for all three types of features: textual, visual and mixed.

Keywords: bag-of-visual-words, bag-of-words, deep convolutional neural network, deep learning, feature engineering, medical imaging, non-negative matrix factorization, principal component analysis, subfigure classification, support vector machine, transfer learning, visual features

1 Introduction

In this paper the participation of the *FHDO Biomedical Computer Science Group (BCSG)* at the *ImageCLEF 2016 Medical Task* [20, 45] is described. The task consists of five different subtasks, namely *Compound Figure Detection*, *Multi-Label Classification*, *Figure Separation*, *Subfigure Classification* and *Caption Prediction*. BCSG participated in the subfigure classification subtask and different methods reaching from traditional feature engineering to modern *Deep Convolutional Neural Networks (DCNN)* were applied.

2 Subfigure Classification Task

The goal of the subfigure classification task is to automatically predict the modality of a medical image. Similar to *ImageCLEF 2015 Medical Task* [19], the class structure is hierarchical and contains 30 classes in total with two main groups, as illustrated in Figure 1.

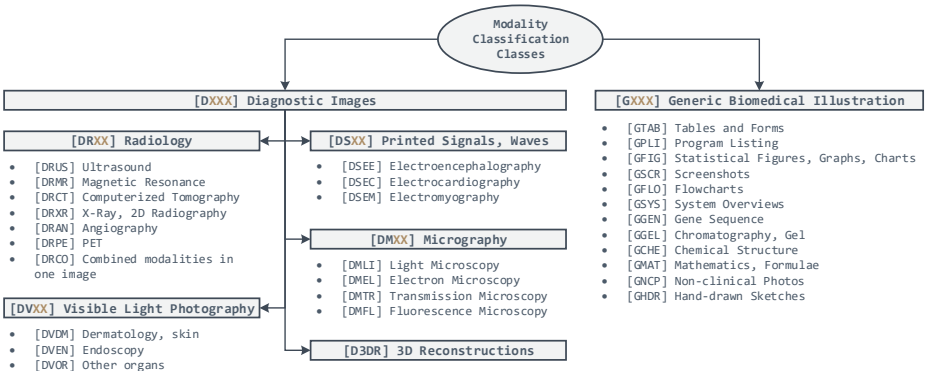

Fig. 1: Class hierarchy of the subfigure classification task (derived from [19])

While the class *GFIG* is very dominant and covered with about 44% of all training images in the ImageCLEF 2016 dataset, a few other classes like *DSEE*, *DSEM* and *GPLI* are underrepresented with less than 10 images. Therefore the dataset is enhanced with the *ImageCLEF 2013 Medical Task* dataset [18], excluding the *Compound Figure (COMP)* category. Subsequent references to the *training set* always include the ImageCLEF 2013 dataset.

An analysis of the class distributions from the dataset of the subfigure classification subtask of ImageCLEF 2015 Medical Task showed a discrepancy between the class distribution of the training and test set. For example the training set contains 0.13%/6 *GGEN* and 0.56%/25 *GSYS* images, whereas the test set includes 7.71%/173 *GGEN* and 2.94%/66 *GSYS* images. Therefore the test set was used as one validation set to incorporate this finding into the model selection process.

Model selection was performed using a combination of both the validation set as defined above and *Bootstrapping* ($n = 8$), based on the idea of the *.632 Estimator* [12]:

$$Err = 0.368 \cdot Err_{val} + 0.632 \cdot \overline{Err}_{boot} \tag{1}$$

With Err_{val} denotes the error on the validation set and $\overline{Err}_{boot}$ the mean of the bootstrapping errors. Contrary to the .632 estimator the validation error was used instead of the training error.

2.1 Textual Features

Textual features can be extracted from the figure captions and the paper full texts, which were both distributed with the image datasets. Both of them are strong features for classification tasks. Furthermore they are complementary to the features extracted from the images itself, which has been shown previously in [32, 33]. In this participation the *Bag-of-Words (BoW)* approach was used to build the textual features.

Two dictionaries were generated from both captions and full texts from the training set. The R Package *tm* was used for text processing [22]. Each caption and full text was transformed using the following operations: lower case folding, number and punctuation removal, whitespace stripping, stopword deletion, *Porter's Stemming* [35]. The resulting words were tested using information gain for association with the target class and only the top 500 words were selected for each dictionary. An overview of the top terms for both dictionaries is given in Table 1. For further improvement of the classification results, the *Okapi BM25* [36] term weighting approach was used:

$$W(TF_i) = \underbrace{log\frac{N}{n_i+1}}_{\text{Inverse Document Frequency}} \cdot \underbrace{\frac{TF_i \cdot (k_1+1)}{TF_i + k_1 \cdot \left((1-b)+b \cdot DL/\overline{DL}\right)}}_{\text{BM25 Term Frequency Component}} \quad (2)$$

With TF_i denotes the i-th term of the document-term matrix, DL the document length and $\overline{DL}$ the average document length. N denotes the total number of documents and n_i the number of documents in which the i-th term is present. The parameters k_1 and b were set to 1.25 and 0.75 respectively, as recommended in [36].

Before training a classifier the BoW matrices were reduced separately using the *Principal Component Analysis (PCA)* [31] to 40 principal components each. Therefore a PCA model was computed on the training set and encoding matrices were predicted using this model. Using the concatenated set of features from training and validation set for computing the principal components was investigated by [32, 33], but did not produce better results during the development

Table 1: First 30 terms of both generated dictionaries, ordered descending by information gain value.

Dictionary	**Terms**
Captions	cell, stain, cebcm, bar, express, green, red, imag, use, valu, mean, scan, magnif, data, scale, arrow, electron, radiograph, structur, gene, control, plot, mri, sequenc, protein, show, microscopi, analysi, repres, antibodi, ...
Full Texts	express, use, differ, data, shown, cell, stain, analysi, contain, protein, cbc, patient, incub, gene, valu, similar, antibodi, number, result, cebcm, compar, studi, experi, buffer, indic, set, wash, observ, yearold, determin, ...

stage. Both different dictionary sizes and numbers of principal components were evaluated in an iterative fashion using the validation set.

As reported in [32], captions can be truncated to the relevant parts using the subfigure ID. By searching for delimiter pairs the relevant part of the caption for a subfigure can be extracted. Further investigation had shown that a lot of text formatting issues prevent a distinct identification of the delimiter pairs. Another problem is the different usage of subfigure identifiers: as prefix, suffix, ranges, comma-delimited, multiple occurrences and so on. With only half of the captions truncated successfully the classification accuracy was not improved and therefore this approach was dismissed.

Another information source is the *Medical Subject Headings (MeSH)*[3] database, which contains expert annotated meta information for PubMed articles utilizing a carefully chosen vocabulary. However about 26% of the overall training set does not contain any MeSH information (about 49% for the 2013 dataset and 14% for the 2016 training set). Therefore this approach was dismissed due to worse results on the validation set.

2.2 Visual Features

Several visual descriptors describe an image with color, texture and shape information. During development different combinations of visual descriptors were tested, resulting in the following set of used features. Most of the visual features were extracted using the *Lucene Image Retrieval (LIRe)*[4] library [28], which implements many state-of-the-art descriptors.

- **ACC**: *Auto Color Correlogram* [21] incorporates the spatial correlation of colors in an input image, as well as the global distribution of local spatial correlations.
- **BoVW**: *Bag-of-Visual-Words* [9] is a well known technique for image representation and highly customizable. For creation the *VLFeat* library [44] was used to extract the relevant features, to create the codebook and finally to build the encoding matrices. The complete creation process is described in section 2.3. As term weighting scheme the Okapi BM25 weighting scheme from Section 2.1 was used.
- **CEDD**: *Color and Edge Directivity Descriptor* [7] is a low-level feature which combines color and textural information. Two fuzzy systems with *Fuzzy Linking* [25] are used to encode the colors to histogram bins.
- **CENTRIST**: *CENsus TRansform hISTogram* [46] is originally designed for scene classification. It mainly encodes the global structure of an image, but suppresses detailed textural information.
- **EHD**: *Edge Histogram Descriptor* [41] is part of the MPEG-7 standard. An input image is divided four times in each dimension and for each region five

[3] http://www.nlm.nih.gov/mesh/ (last access: 09.05.2016)

[4] https://github.com/dermotte/LIRE/ (last access: 14.05.2016)
Commit 3bf3c4ebd2aafaa3b4703b36a65ec65a13166b03 with custom modifications

different edge detectors are applied on each 2×2 pixel block, resulting in a 80-bin histogram.
- **FCTH**: *Fuzzy Color and Texture Histogram* [8] is again a low-level feature similar to CEDD which combines color and textural information. There are three fuzzy systems with fuzzy linking involved in the creation of the FCTH descriptor, one for the textural and two for the color information.
- **FOH**: *Fuzzy Opponent Histogram*, as implemented in LIRe, is a 64-bin *Fuzzy Color Histogram* [14] using the *Opponent Color Space* [38].
- **LBF**: The *LIRe Basic Feature (LBF)* [28] contains global features of an image: brightness, clipping, contrast, hueCount, saturation, complexity, skew and energy. Furthermore an additional boolean attribute was appended to indicate a chromatic image.
- **PHOG**: *Pyramid Histogram of Oriented Gradients* [3] is an extension of the *Histogram of Oriented Gradients (HOG)* [10], which additionally encodes the spatial distribution. As implemented in LIRe the PHOG descriptor is a joined histogram of 1×1, 2×2 and 4×4 HOG, 27 individual histograms in total.
- **RILBP**: *Rotation Invariant Local Binary Patterns (RILBP)* [1] is an extension of the *Local Binary Patterns (LBP)* [30]. The key idea is to map patterns, which are just rotated variants, to one base pattern.
- **Tamura**: *Tamura Features* are six textural features, which were evaluated with psychological measurements [43]: *coarseness*, *contrast*, *directionality*, *line likeness*, *regularity* and *roughness*.

2.3 Bag of Visual Words (BoVW)

Bag-of-Visual-Words (BoVW) [9], also known as *Bag-of-Keypoints (BoK)*, is a technique which involves a local keypoint detector. A keypoint detector determines important locations in an image, which are invariant to small changes and also contain much information. In this context the *Scale Invariant Feature Transform (SIFT)* [27] descriptors are extracted on a dense grid at different scales, also called *dense SIFT (DSIFT)* [10]. It has been shown in [2] that DSIFT is more suitable for classification tasks, whereas SIFT is more appropriate for object recognition.

BoVW involves two separate computational processes: First a codebook with visual words is created from an image dataset. Second the images of the training and test set are encoded using the codebook, resulting in one histogram vector per image.

Codebook: An illustration of the codebook creation process is given in Figure 2. A dataset of images is used to compute the representative visual words. For the submissions of this participation, the ImageCLEF 2013 dataset [18], as well as the ImageCLEF 2016 training set [20] were used. All extracted SIFT descriptors of one image were grouped into 150 clusters using a *k-Means* algorithm [15]. The overall set of clustered descriptors was clustered into 10000 visual words, which is the resulting codebook.

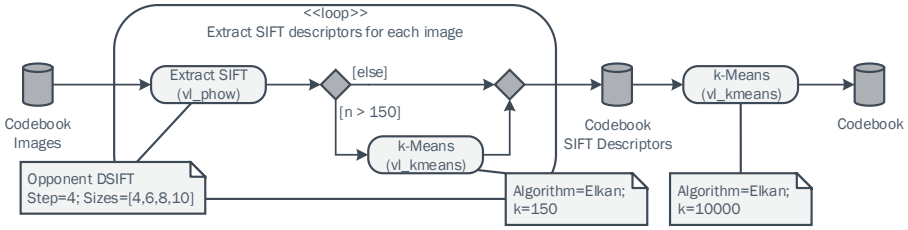

Fig. 2: Codebook creation process for Bag-of-Visual-Words using dense SIFT and a two-layer clustering approach using VLFeat.

Several benchmarks were performed to evaluate different codebook creation strategies. Due to the fact that some classes in this classification problem can be distinguished by color information, it was found that color SIFT descriptors (3x128 attributes) are more powerful than grayscale SIFT (1x128 attributes). Furthermore, SIFT descriptors extracted from images in *Opponent* color space [38] yielded better results than those in *Hue Saturation Value (HSV)* color space.

Encoding: The encoding process is shown in Figure 3. For each image the extracted SIFT features are matched against the codebook using a kd-tree. The resulting indices of matched visual words are then encoded into a histogram vector, which is the final BoVW vector.

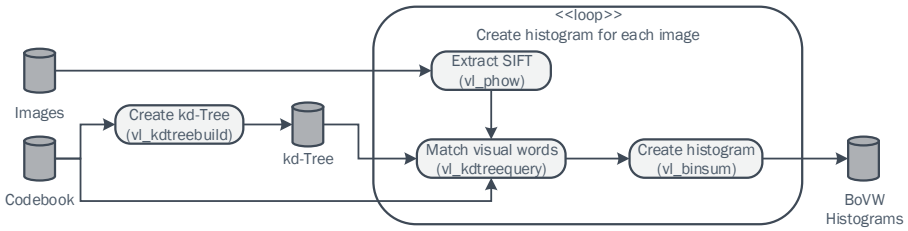

Fig. 3: Encoding process of the Bag-of-Visual-Words histograms using VLFeat.

Information Gain: As already described in Section 2.1, the information gain can be computed to evaluate the importance of a specific word. In the context of BoVW the importance of visual words should also be calculated. Hence removing visual words with a relative low information gain value is a *Feature Selection*, which truncates the dictionary in a similar way. For this participation all visual words with a value below 0.05 were removed, resulting in a reduction from 10000 to 7971 visual words. On the validation set this method improved the overall accuracy by about 1%.

2.4 Submitted Runs

In the following, the ten runs submitted for evaluation are shortly described:

- **Run 1:** A *Support Vector Machine (SVM)* was trained using the *e1071* package [29] for R, which uses *LibSVM* [6] internally. For training the SVM a RBF kernel was used and the cost parameter was set to 2. Adjusting the gamma parameter (default: $\gamma = 1/n_{attributes}$) was investigated, but did not produce any better results. As input all visual features from Section 2.2 were reduced blockwise using the PCA and fused after reduction (see also Table 2).
- **Run 2:** Similar to run 1, a SVM with RBF kernel with default γ and $C = 2$ was trained using only textual features, which are described in Section 2.1.
- **Run 3:** A fusion of the features from run 1 and 2 was trained following the same classifier setup as in run 1 and 2.
- **Run 4:** Referring to Figure 1, the classification problem can be split on the top level of the class hierarchy. Therefore three classifiers were trained on the features from run 3, one for the top level split *DXXX / GXXX* and one for each of the subproblems. It is important to note that the PCA has to be applied also three times in total.
- **Run 5:** Features from run 3 were extended by *Deep Convolutional Activation Features (DeCAF)* [11] from a *Residual Network (ResNet)* with 152 layers [16], named *ResNet-152*. A ResNet is a *Deep Convolutional Neural Network (DCNN)*, which is much deeper than for example other winning networks like GoogLeNet (22 layers) [42]. This network was trained on the ImageNet dataset [37] and won the ImageNet 2015 competition. In this context the network is only used as a feature extractor, which has been previously shown to yield good results [11, 40]. Pretrained networks have been made public by the authors [16] as *caffe* [23] models on Github[5]. Prior feature fusion the DeCAF were reduced to 20 principal components. Using a pretrained network from a different domain is also called *Transfer Learning* [11, 47].
- **Run 6:** Four SVM classifiers, as described in run 1, were trained, each of them for a disjoint set of features. The sets consist of $F_1 = \{\text{BoW}\}$, $F_2 = \{\text{BoVW}\}$, $F_3 = \{\text{ACC, CEDD, FCTH, FOH, LBF}\}$ and $F_4 = \{\text{CENTRIST, EHD, PHOG, RILBP, Tamura}\}$. Final predictions were calculated using the mean of the top-3 probabilities.
- **Run 7:** A modified *GoogLeNet* [42] was trained on the training set using *caffe* [23] and the Nvidia *Deep Learning GPU Training System (DIGITS)*[6]. To achieve higher accuracy, the *Rectified Linear Unit (ReLU)* operations were replaced by the *Parametric Rectified Linear Unit (PReLU)* [17] and the network initialization was changed from gaussian random initialization to *Xavier* initialization [13]. Optimization of the network was performed by a *Stochastic Gradient Descent (SGD)* solver [4] with 100 epochs in total, a base learning rate $\eta = 0.01$ and *Step Down* as policy with 33% as step size and $\gamma = 0.1$. The model used for this run was a snapshot at epoch 60.

[5] https://github.com/KaimingHe/deep-residual-networks (last access: 24.04.2016)
[6] https://github.com/NVIDIA/DIGITS (last access: 24.04.2016)

– **Run 8:** In this run the ResNet-152 was used again for transfer learning [11, 47]. Since this network was trained on a dataset with 1000 classes, the last network layer *fc1000* was extracted and on top of these activation values a custom network layer with 30 linear neurons was trained using the *Pseudo-Inverse* method, also called *Projection Learning Rule* [34]:

$$W = \left(X^T X\right)^{-1} X^T Y \tag{3}$$

In Equation 3 the weights $W \in \mathbb{R}^{(m+1)\times 30}$ for the linear neuron layer are trained, where $X \in \mathbb{R}^{n\times(m+1)}$ denotes the training set and $Y \in \mathbb{R}^{n\times 30}$ a binary label matrix for the training set. Note that the input data has to be extended by a bias column full of ones.

$$Y' = X'W \tag{4}$$

In Equation 4 the test data $X' \in \mathbb{R}^{n'\times(m+1)}$ is tested against the trained linear classifiers, the class with the largest distance to the separating hyperplane is chosen.

– **Run 9:** The *Non-negative Matrix Factorization (NMF)* [26] is a matrix factorization technique, which computes a purely additive factorization of a non-negative data matrix. However an exact approximation does not necessarily yield a discriminative solution for learning algorithms. Therefore algorithms like the *Gradient Descent Constrained Least Squares (GDCLS)* [39] enforce sparsity in the encoding matrix, leading to more local discriminating features. For this run the GDCLS implementation in the R package *nmfgpu4R* [24] was used with $\lambda = 0.1$ as regularization parameter instead of the PCA for feature reduction. The NMF was applied blockwise with the same dimensions used for the PCA reduction.
– **Run 10:** Each of the previous classifiers has its own issues in classifying each class correctly. But if they are combined, then the results can be stabilized by a certain amount. For this run an ensemble of the predictions from run 3, 5, 8 and 9 was used. As the outputs of different classifiers are not calibrated, a simple voting scheme was used for classifier combination. In this scheme 5, 3 and 1 point(s) for the 1st, 2nd and 3rd predicted class were assigned.

According to [32] the usage of a *Random Forest* [5] classifier has been investigated but produced a major drop in terms of accuracy on the validation set. In the same way a SVM with a linear kernel also produced worse results.

2.5 Results

Biomedical Computer Science Group (BCSG) achieved top performance in all three categories (textual, visual and mixed), as visualized in Figure 4. Overall top performance of 88.43% was achieved by run 10, which is similar to the development results. It is interesting that run 8 achieved a higher performance than run 1. As described earlier in Section 2.4, run 8 is a pre-trained deep convolution neural network, which was trained on the ImageNet dataset [37].

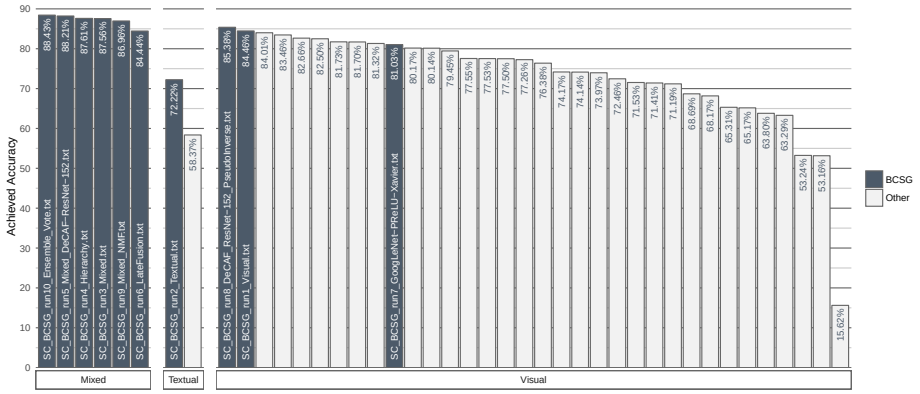

Fig. 4: Official evaluation results for the submitted run files.

However run 1 incorporates 11 visual descriptors and a classifier which was trained on images from the same domain as the test images. In conclusion the generalization capabilities of DCNNs are verified once more.

When analyzing the confusion matrices in Figure 5, it is noticeable that the *GFIG* class is still the major problem. Furthermore it can be observed that *GFIG* is mainly confused with other classes within the *GXXX* class group. Enhancing the training set to provide more information for a correct separation of those classes could help. One more interesting point is that the ensemble contains the least noise whereas the run 8 contains the most noise in the confusion matrices.

2.6 Ex-post Evaluations

Further evaluations were performed after the submission deadline using the official ground truth information. In Table 2 the overall set of visual features from run 3 (Mixed) was analyzed for accuracy gain. Therefore one classifier per feature was trained, but the specified feature was omitted from the configuration.

Similar to the findings in [32] both BoW and BoVW remain the strongest features for the configuration. However the other features' contributions are very low with even two features with a negative impact. For further analysis of the feature vectors, the linear correlation matrix of the fused feature vectors is visualized in Figure 6. It is noticeable that features other than BoW and BoVW are more linearly correlated. In addition it can be seen that both BoW matrices are linearly correlated in the first few principal components but then do explain different information.

Another customizable point is the set of features for principal components calculation. These can either be computed by using only the training set or both the training and validation/test set in a semi-supervised fashion. In [33] further evaluations were performed on the ImageCLEF 2015 dataset [19] and an improvement of about 4% was observed when using both sets combined. In this year the combination of both sets was dismissed as it did not improve the

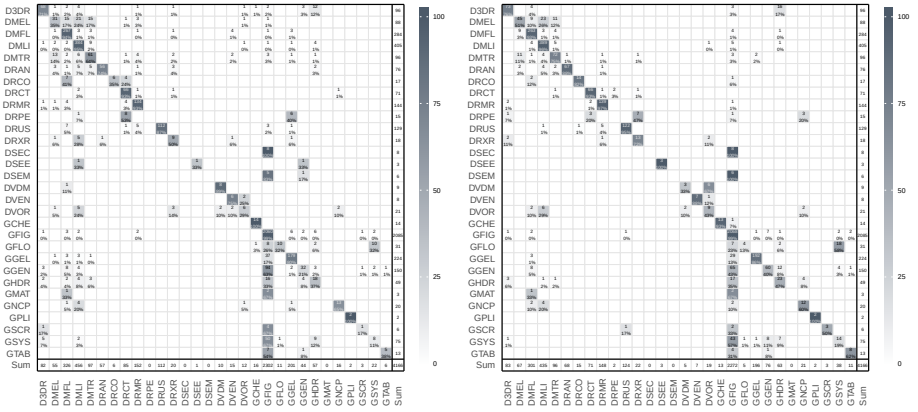

Fig. 5: Confusion matrices for run 8 (left), the transfer learning approach, and run 10 (right), the ensemble of multiple runs. Rows represent the actual classes and columns the predicted classes. These plots are viewed best in electronic form.

accuracy during the development stage. An ex-post evaluation of run 3 (Mixed) lead to an accuracy degradation of 0.33% when using both sets combined.

As explained earlier, the dataset was enhanced with the ImageCLEF 2013 Medical Task dataset for the subfigure classification task. If only the ImageCLEF 2016 dataset is used for training the classifier, then the accuracy of run 1 (Visual) drops by 1.89% and of run 3 (Mixed) by 2.9%. Hence collecting more images should further improve the overall accuracy of classifiers.

Using information gain for important visual words selection improved the accuracy by about 1% during the development stage, as described in Section 2.3. For example when computing run 3 (Mixed) with all visual words, without removal of any visual words, the accuracy is reduced by 0.21%. However two different dictionaries were used during development and evaluation stage, since the dictionary for evaluation also includes the validation set. For any reliable and statistical conclusions further experiments have to be done.

3 Conclusions

Several approaches for modality classification of medical images were evaluated for the ImageCLEF 2016 medical task. Especially the transfer learning model was surprisingly strong compared to traditional feature engineering. Fine-tuning the ResNet-152 or even training from scratch with a larger medical database could further improve the accuracy of the DCNN. In addition to this the importance of textual information was verified once again, as this information source is independent from image information. Bag-of-Visual-Words in conjunction with dense SIFT and Opponent color space appeared to be a very strong visual feature and feature selection for visual words further improved the results.

Table 2: Evaluation of the loss of accuracy when omitting one descriptor from run 3 (Mixed).

Descriptor	Original Dimension	Reduced Dimension	Loss of Accuracy (%)
BoW	500 + 500	40 + 40	−3.10
BoVW	10000/7817	50	−1.59
LBF	9	9	−0.27
RILBP	36	8	−0.12
EHD	80	10	−0.39
FOH	576	4	−0.08
Tamura	18	3	−0.08
CEDD	144	10	+0.19
FCTH	192	10	−0.22
ACC	256	10	−0.03
PHOG	630	10	−0.08
CENTRIST	256	4	+0.07

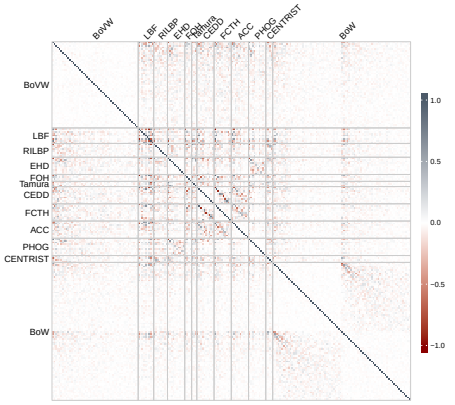

Fig. 6: Visualization of the correlation matrix from features used in run 3 (Mixed).

References

1. Ahonen, T., Matas, J., He, C., Pietikäinen, M.: Rotation invariant image description with local binary pattern histogram fourier features. In: Proceedings of the 16th Scandinavian Conference on Image Analysis. pp. 61–70. SCIA '09, Springer-Verlag, Berlin, Heidelberg (2009)
2. Bosch, A., Zisserman, A., Muñoz, X.: Computer Vision – ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV, chap. Scene Classification Via pLSA, pp. 517–530. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
3. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval. pp. 401–408. CIVR '07, ACM, New York, NY, USA (2007)
4. Bottou, L.: Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers, chap. Large-Scale Machine Learning with Stochastic Gradient Descent, pp. 177–186. Physica-Verlag HD, Heidelberg (2010)
5. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
6. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27:1–27:27 (2011)
7. Chatzichristofis, S.A., Boutalis, Y.S.: CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In: Proceedings of the 6th International Conference on Computer Vision Systems. pp. 312–322. ICVS'08, Springer-Verlag, Berlin, Heidelberg (2008)
8. Chatzichristofis, S.A., Boutalis, Y.S.: FCTH: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In: Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services. pp. 191–196. WIAMIS '08, IEEE Computer Society, Washington, DC, USA (2008)
9. Cula, O.G., Dana, K.J.: Compact representation of bidirectional texture functions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 1041–1047 (2001)

10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 886–893 (2005)
11. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). pp. 647–655. JMLR Workshop and Conference Proceedings (2014)
12. Efron, B., Tibshirani, R.: Improvements on cross-validation: The 632+ bootstrap method. Journal of the American Statistical Association 92(438), 548–560 (1997)
13. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics May 13-15, 2010, Chia Laguna Resort, Sardinia, Italy. JMLR Workshop and Conference Proceedings, vol. 9, pp. 249–256 (2010)
14. Han, J., Ma, K.K.: Fuzzy color histogram and its use in color image retrieval. IEEE Transactions on Image Processing 11(8), 944–952 (2002)
15. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28(1), 100–108 (1979)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
17. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
18. García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum). CEUR Workshop Proceedings, vol. 1179 (September 2013)
19. García Seco de Herrera, A., Müller, H., Bromuri, S.: Overview of the ImageCLEF 2015 medical classification task. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum). CEUR Workshop Proceedings, vol. 1391 (September 2015)
20. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 Medical Task. In: CLEF 2016 Working Notes. CEUR Workshop Proceedings, vol. 1609. CEUR-WS.org <http://ceur-ws.org>, Évora, Portugal (September 5-8 2016)
21. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In: Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on. pp. 762–768 (1997)
22. Ingo Feinerer, Kurt Hornik, D.M.: Text mining infrastructure in R. Journal of Statistical Software 25(5), 1–54 (2008), http://www.jstatsoft.org/v25/i05/.
23. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia. pp. 675–678. MM '14, ACM, New York, NY, USA (2014)
24. Koitka, S., Friedrich, C.M.: nmfgpu4R: Computation of non-negative matrix factorizations (NMF) using CUDA capable hardware. R Journal (2016), Status: Accepted
25. Konstantinidis, K., Gasteratos, A., Andreadis, I.: Image retrieval based on fuzzy color histogram processing. Optics Communications 248(4–6), 375 – 386 (2005)
26. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)

27. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision. vol. 2, pp. 1150–1157 (1999)
28. Lux, M., Chatzichristofis, S.A.: Lire: Lucene image retrieval: An extensible Java CBIR library. In: Proceedings of the 16th ACM International Conference on Multimedia. pp. 1085–1088. MM '08, ACM, New York, NY, USA (2008)
29. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2015), https://CRAN.R-project.org/package=e1071, R package version 1.6-7
30. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
31. Pearson, K.: On lines and planes of closest fit to system of points in space. Philiosophical Magazine 2, 559–572 (1901)
32. Pelka, O., Friedrich, C.M.: 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. CEUR-WS Proceedings Notes, vol. 1391 (2015)
33. Pelka, O., Friedrich, C.M.: Modality prediction of biomedical literature images using multimodal feature representation. GMS Medical Informatics, Biometry and Epidemiology (MIBE) (2016), Status: Submitted
34. Personnaz, L., Guyon, I., Dreyfus, G.: Collective computational properties of neural networks: New learning mechanisms. Physical Review A (General Physics) 34(5), 4217–4228 (1986)
35. Porter, M.F.: An algorithm for suffix stripping. Program: Electronic Library and Information Systems 40(3), 211–218 (1980)
36. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC–3. In: Overview of the Third Text REtrieval Conference (TREC–3). p. 109–126. Gaithersburg, MD: NIST (1995)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115(3), 211–252 (2015)
38. van de Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582 –1596 (2010)
39. Shahnaz, F., Berry, M.W., Pauca, V., Plemmons, R.J.: Document clustering using nonnegative matrix factorization. Information Processing & Management 42(2), 373–386 (2006)
40. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2014)
41. Sikora, T.: The MPEG-7 visual standard for content description-an overview. IEEE Transactions on Circuits and Systems for Video Technology 11(6), 696–702 (2001)
42. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
43. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics 8(6), 460–473 (1978)

44. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM International Conference on Multimedia. pp. 1469–1472. MM '10, ACM, New York, NY, USA (2010)
45. Villegas, M., Müller, H., García Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Mikolajczyk, K., Puigcerver, J., Toselli, A.H., Sánchez, J.A., Vidal, E.: General Overview of ImageCLEF at the CLEF 2016 Labs. Lecture Notes in Computer Science, Springer International Publishing (2016)
46. Wu, J., Rehg, J.M.: Centrist: A visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8), 1489–1501 (2011)
47. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems. vol. 27, pp. 3320–3328. Curran Associates, Inc. (2014)