Surpassing Traditional Feature Engineering with Deep Learning Transfer Scenarios at Medical Classification Task of ImageCLEF 2016

Biomedical Computer Science Group (BCSG)

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

[1] University of Applied Sciences and Arts Dortmund, Department of Computer Science
[2] TU Dortmund University, Department of Computer Science

Introduction

Dataset

- Contains very few images for the given classification problem, 6776 in training set for 30 classes
- In the past some classes were underrepresented (see Figure 1)
- For training the dataset was enhanced with the ImageCLEF 2013 without *COMP* category, yielding 10140 images for training versus 4166 images for testing

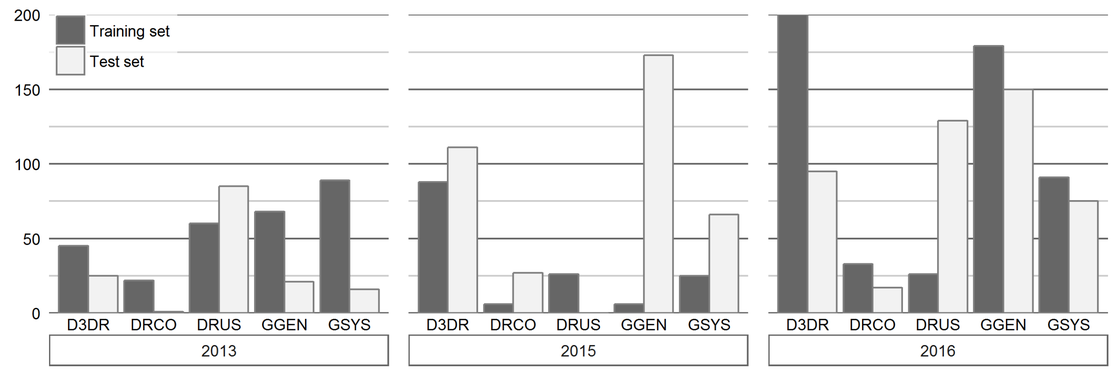

Figure 1: Class distribution of problematic cases for ImageCLEF 2013, 2015 & 2016 datasets.

Model Selection

- Error of both the ImageCLEF 2015 test set, as one validation set, and bootstrapping was combined:

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

- Approach is similar to the .632 Estimator, but uses validation error instead of training error
- Helps to take class imbalances into account

Traditional Feature Engineering

Textual Features using Bag-of-Words

- Preprocessing using the *R* package *tm* involved lower case folding, number and punctuation removal, whitespace stripping, stop word deletion and Porter's stemming
- Two *Bag-of-Words (BoW)* were computed for both image captions and paper full texts
- *BM25* term weighting produced best results during development
- Dictionaries were truncated to top 500 words each using information gain for ranking
- Feature reduction to 40 features for each BoW was applied

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, lectron, radiograph, structur, gene, control, plot, mri, sequenc, protein, show, microscopi, analysi, repres, antibodi, ...
Full Texts	express, use, dier, data, shown, cell, stain, analysi, contain, protein, cbc, patient, incub, gene, valu, similar, antibodi, number, result, cebcm, compar, studi, experi, buer, indic, set, wash, observ, yearold, determin, ...

Visual Features

- Ten visual descriptors extracted by the *Lucene Image Retrieval (LIRe)* library (customized v1.0)
- Additionally *Bag-of-Visual-Words (BoVW)* using *Dense SIFT (DSIFT)*
  - DSIFT extraction in *Opponent* color space was performed using the *VLFeat* library (v0.9.20)
  - Two-stage clustering approach
    - DSIFT features of one image were clustered into k_1 clusters
    - Joined set of clustered features was clustered into k_2 visual words
  - BM25 term weighting produced best results during development again
  - Information gain for evaluation and ranking of visual words was used
  - Feature reduction to 50 features was applied

Feature Reduction

- Reduction was performed using the training set for computation of new data projections
- Both *Principal Component Analysis (PCA)* using the R package *stats* and *Non-negative Matrix Factorizations (NMF)* using the R package *nmfgpu4R* were used

Figure 2: Visualization of the reduced number of attributes per feature.

Deep Learning and Transfer Learning

- Used frameworks: *caffe*, *nvidia-caffe* and *Deep Learning GPU Training System (DIGITS)*
- In **run 7** a *Deep Convolutional Neural Network (DCNN)*, namely a modified *GoogLeNet,* was trained from scratch with *Xavier* initialization and *Parametric Rectified Linear Units (PReLU)* using a *Stochastic Gradient Descent (SGD)* solver
- *Transfer Learning* describes the process of using a pre-trained neural network, probably from a different domain, which might be necessary if the amount of data is too small to train a generalizable network
- In **run 5** features from the last layer of a pre-trained *Residual Network (ResNet)* were extracted
  - Pre-trained ResNets are available at the caffe *Model Zoo* Github page
  - *ResNet-152* was used for feature extraction, which is trained on the ImageNet dataset
  - Layer fc1000 with 1000 attributes was reduced using PCA to 20 principal components and fused with features from **run 3**
- In **run 8** the pre-trained ResNet-152 was used as a classifier
  - A network layer with 30 linear neurons was trained on top using the *Projection Learning Rule*

$$W = (X^T X)^{-1} X^T Y$$

  - Inference is performed using the following equation

$$Y' = X'W$$

  - Class with the largest distance to the separating hyperplane is chosen
  - Works good for small datasets
  - Larger datasets suffer from numerical instability due to the *Moore-Penrose pseudoinverse*

Official Evaluation Results

- Evaluation results are visualized in Figure 3, the best run of each category is described below:
  - **Textual**:
    - **Run 2** with 72.22% overall accuracy was the best submitted textual classifier
    - *Support Vector Machine (SVM)* with two joined Bag-of-Words
  - **Visual**:
    - **Run 8** with 85.38% overall accuracy was the best submitted visual classifier
    - Transfer Learning of a ResNet-152 using the Projection Learning Rule
    - Outperformed second best visual classifier, **run 1** with 84.46%, which used traditional feature engineering
  - **Mixed**:
    - **Run 10** with 88.43% overall accuracy was the best submitted mixed classifier
    - Aggregation of **run 3, 5, 8 and 9** using a weighted voting scheme with the following weights: 1st = 5 points, 2nd = 3 points, 3rd = 1 point
    - Results of aggregation were stabilized in terms of noise regarding the confusion matrices

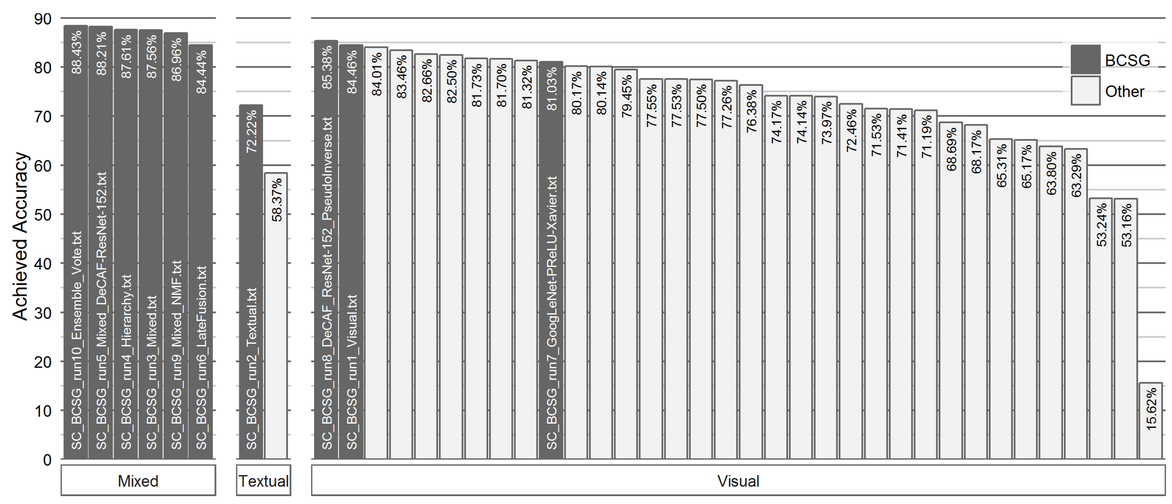

Figure 3: Official evaluation results for the submitted run files of the subfigure classification task.

Ex-post Evaluation Findings

Findings based on the evaluation of **run 3**:

- Bag-of-Words and Bag-of-Visual-Words were the strongest features, whereas CENTRIST and CEDD had a negative impact on accuracy (see Figure 4)
- PCA computation on the combined training and test set reduced the accuracy by 0.33%
- Training the classifier with only the ImageCLEF 2016 training set reduced the accuracy by 2.9%

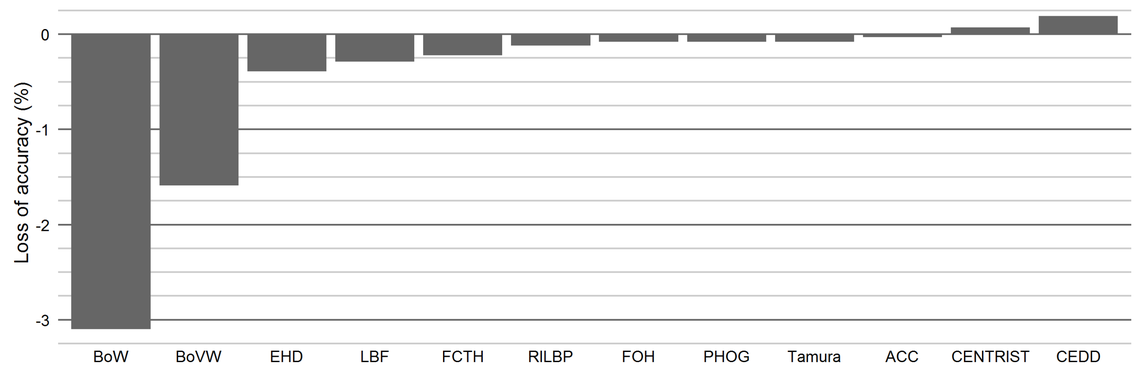

Figure 4: Ex-post evaluation of the loss of accuracy when omitting one descriptor from run 3 (Mixed).

Conclusions

- Transfer learning from a different domain (ImageNet) yielded remarkable classification accuracy
- Could be even improved if models are fine-tuned or trained from scratch
- Dataset needs more samples for effective DCNN approaches, when learning from scratch
- Textual features are independent and give an significant boost in terms of accuracy
- Good Bag-of-Visual-Words in conjunction with DSIFT and Opponent color space

Contact Information

Sven Koitka
Emil-Figge-Straße 42
44227 Dortmund
Germany

☎ +49 (0)231 755-6720
✉ sven.koitka@fh-dortmund.de

Christoph M. Friedrich
Emil-Figge-Straße 42
44227 Dortmund
Germany

☎ +49 (0)231 755-6796
✉ christoph.friedrich@fh-dortmund.de