# Deep Multimodal Classification of Image Types in Biomedical Journal Figures

Anonymous

No Institute Given

**Abstract.** This paper presents a robust method for the classification of medical image types in figures of the biomedical literature using the fusion of visual and textual information. A deep convolutional network is trained to discriminate among 31 classes including compound figures, diagnostic image types and generic illustrations, while another shallow convolutional network is used for the analysis of the captions paired with the images. Various fusion methods are analyzed as well as data augmentation approaches. The proposed system is validated on the ImageCLEF 2013 classification task, largely improving the currently best performance from 83.5% to 93.7% accuracy.

## 1 Introduction

The information contained in an image and the methods employed to extract it largely differ depending on its modality, making the latter a crucial aspect of medical image analysis and retrieval, particularly when images of the medical literature are used, where image type information is not available. An image type classification is, therefore, a useful preliminary filtering step prior to further analysis [11]. Besides, the modality is a relevant information to be determined for medical image or document retrieval, allowing clinicians to filter their search to a particular modality, often specific to a diagnosis or organ of interest. Various modality classification tasks, among others, have been released through the ImageCLEF challenges [4]. We focus this work on the 2013 ImageCLEF modality classification task, as it offers multimodal text and image data. The database is publicly available and the results fully reproducible. The database also originates from the PubMed Central database (it is a small subset), allowing us to classify this large database for further processing and analysis. Much of medical knowledge is stored in the medical literature, thus making the content, including images, accessible for research could help in various tasks.

Multimodal analysis is commonly used to extract and fuse information from multiple modalities [7]. In this work, images and captions contain complementary information fused to boost the classification accuracy. Many methods have been used to extract high-level features from text and images independently and to fuse them. Convolutional Neural Networks (CNNs) have obtained the state of the art in most computer vision and biomedical image analysis tasks. It is also well suited for text analysis [8]. This paper, therefore, introduces several late fusion methods to combine powerful visual and textual CNNs.

## 2 Related Work

Multimodal textual and visual analysis has been widely studied for applications including annotation and captioning [7], image generation from text, text and image feature fusion for retrieval and classification [4]. A total of 51 runs from eight groups were presented in [4] for the ImageCLEF 2013 modality classification challenge. The best results (81.7% classification accuracy) were obtained by visual and textual fusion from the IBM Multimedia Analytics group [1] (see Table 1). A set of color and texture, local and global descriptors (including color histogram, moments, wavelets, Local Binary Patterns (LPB) and Scale-Invariant Feature Transform (SIFT)) were extracted as visual descriptors and fused with multiple textual descriptors. The best results were obtained using a maximum late fusion with a classifier built on top of modality tailored keywords (with a hand-selected vocabulary that likely improved the performance) and a two-level Support Vector Machine (SVM) classification. The methods developed by other teams reported in [4] include various types of similar hand-crafted visual and textual descriptors combined by multiple fusion methods.

In [3], the authors build upon [1] to develop a more complex system. An ensemble of SVM models is trained on top of similar visual features, while the text is analyzed by scoring based on the detection of manually-selected patterns from the captions and from sentences in the body of the article. A weighted score average trained on a subset of the training data was used for fusing the visual and textual information. The best current system reached an accuracy of 83.5%.

Another set of hand-crafted visual and textual features are combined in [12]. The visual features include local and global texture and color features, while Bag-of-Words (BoW) features are used to analyze the captions. More recently in [11], modality images are classified by finetuning (only visual) pretrained deep CNNs and combining their outputs in an ensemble classifier. A major drawback of combining multiple CNNs is the increase of computational complexity and redundancy of features to obtain only a limited accuracy improvement.

## 3 Methods

### 3.1 ImageCLEF 2013 Modality Dataset

The goal of this task is to classify the images into medical modalities and other images types. Three main categories, namely compound figures, diagnostic images and generic illustrations are divided into 31 sub-categories [4]. The modality hierarchy and more details on the dataset can be found in [4]. A total of 2879 training and 2570 test images are provided. The classes are highly imbalanced, reflecting the distribution of the images in the data (PubMed Central[1]) containing a large proportion of compound figures.

The overview of the developed networks and fusion approaches is illustrated in Figure 1. The components are described in more details in the following sections.

[1] https://www.ncbi.nlm.nih.gov/pmc/

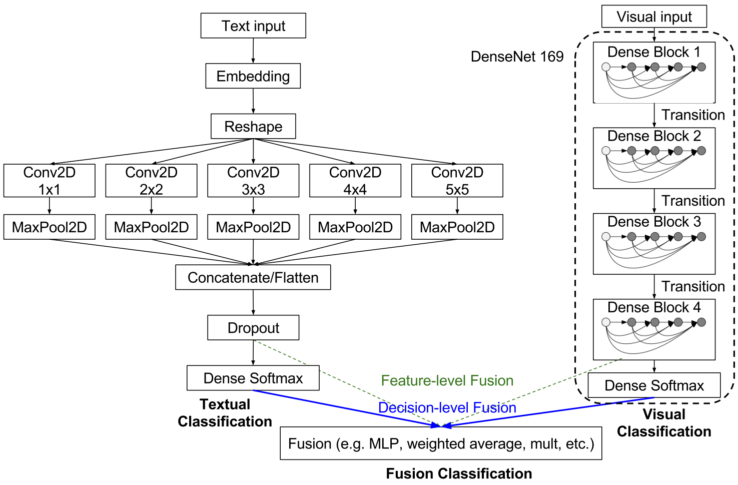


Fig. 1: Overview of the proposed deep learning visual and textual fusion method.

### 3.2 Visual Analysis

DenseNet [6] is a CNN having each layer connected to every other layer (within a dense block, see Figure 1). This architecture obtained excellent results on various image classification datasets while reducing the number of parameters and computation (floating point operations) as compared to other commonly used networks (e.g. AlexNet, VGG, GoogleNet and ResNet). DenseNet169 obtained the best results, as compared to DenseNet121, ResNet50 and 152, VGG19.

The training data is limited (2789 images without data augmentation) and transfer learning is required to obtain a robust image classification. We use networks pre-trained on ImageNet and replace the last fully connected softmax activated dense layer by a layer of 31 neurons, equivalent to the number of classes in the ImageCLEF 2013 dataset.

To increase the visual training data, we explore two complementary data augmentation strategies. The first strategy is to use extra training images from the ImageCLEF 2016 subfigure classification task [5] which are cropped from compound figures. We also use images from the ImageCLEF 2016 compound figure detection task as the compound class is not represented by the subfigure classification set. We ensure that no image is present twice in the training set or in both the training and test set. The second data augmentation strategy is to apply a set of random transformations to the training images including horizontal and vertical flips, width and height shift in the range of $[0, 0.1]$ of the total width and height respectively and a rotation in the range $[0°, 5°]$.

### 3.3 Text Analysis

We develop a CNN on top of word embeddings as inspired by [8]. The words are embedded into a low (300) dimensional space using fastText word embedding [2] pretrained on Wikipedia data[2]. Similar results were obtained with a Global Vectors for word representation (GloVE) pretrained on Wikipedia 2014 and Gigaword 5, while pretraining on biomedical text performed worse.

A maximum number of tokens was set to 20,000 and the maximum length of a caption was set to 200. The embedding layer is finetuned together with the network in order to adapt the embedding to this particular caption classification task and domain. Recurrent networks may seem more intuitive than CNNs and better suited for natural language processing since local features captured by convolution filters are not as evident in texts as they are in images. Captions, however, offer a relatively structured and controlled domain in which words are often organized into meaningful features. CNNs are also better at detecting key phrases or combinations of words than RNNs, which is a useful asset for the evaluated task since the modality is often described by a single sentence or group of words. Besides this, the speed of convolution computations is an important aspect in the choice of an architecture. We also experimented with a 1D convolutional network, a Long Short-Term Memory (LSTM) network and a stacked LSTM network, resulting in a lower accuracy.

### 3.4 Decision-Level Fusion

The decision-level fusion combines the visual and textual predictions. We first train the visual and textual networks independently, then combine the class probabilities, i.e. outputs of the softmax layers. Simple fusions are used including (1) a weighted sum, (2) a maximum probability decision and (3) a product of probabilities (elementwise product of probability vectors). Equation 1 summarizes the class prediction of these three fusion methods.

$$\begin{aligned} c_{sum} &= amax(\alpha y_v + (1-\alpha) y_t), \\ c_{max} &= amax(max(y_v, y_t)), \\ c_{prod} &= amax(y_v \circ y_t), \end{aligned} \tag{1}$$

where $c_{sum}$, $c_{max}$, $c_{prod}$ are the class predictions from (1), (2) and (3) respectively, $y_v$ and $y_t$ are the probability vectors of the visual and textual networks respectively. The weight $\alpha \in [0, 1]$ is used to balance the importance of the visual and textual parts. Another fusion method is to train a single layer Multi-Layer Perceptron (MLP) on top of the prediction layer. We freeze all previous layers to train only the last added layer. Using a two-layer MLP results in similar performance, yet increases the complexity.

### 3.5 Feature-Level Fusion

The feature-level fusion fuses the outputs of intermediate layers from the visual and textual networks. We first train the two networks independently, then add one layer on top of the last layer before the softmax activated one and train similarly to the decision-level MLP.

[2] https://dumps.wikimedia.org

# 4 Experimental Results

## 4.1 Network Setups

The networks are trained with an Adam [9] optimizer. The textual, visual and fusion MLP networks are trained for $N = 100$, $N = 25$ and $N = 50$ iterations respectively. The initial learning rate is set to $10^{-4}$ for finetuning the visual network and $10^{-3}$ for the textual network and MLP from scratch, average decays $\beta_1$ and $\beta_2$ are 0.9 and 0.999 respectively, the learning rate decay is $\frac{0.1}{N}$ and the batch size 32. Due to the high class imbalance in the training set, class weights are used during training for weighting the loss function as: $w_i = n_{max}/n_i$, where $n_{max}$ and $n_i$ are the number of training samples of the most represented class and of class $i$ respectively. For the visual network, class weights are not needed when artificial data augmentation is used.

## 4.2 Classification Results

The results are reported and compared with the best current systems in Table 1[3]. Best results of the 51 runs submitted by eight groups [4] are reported as well as the best results in the literature [3]. In [1, 3], vocabularies and text patterns were manually selected, and in [3], the text in the body of the article was also used.

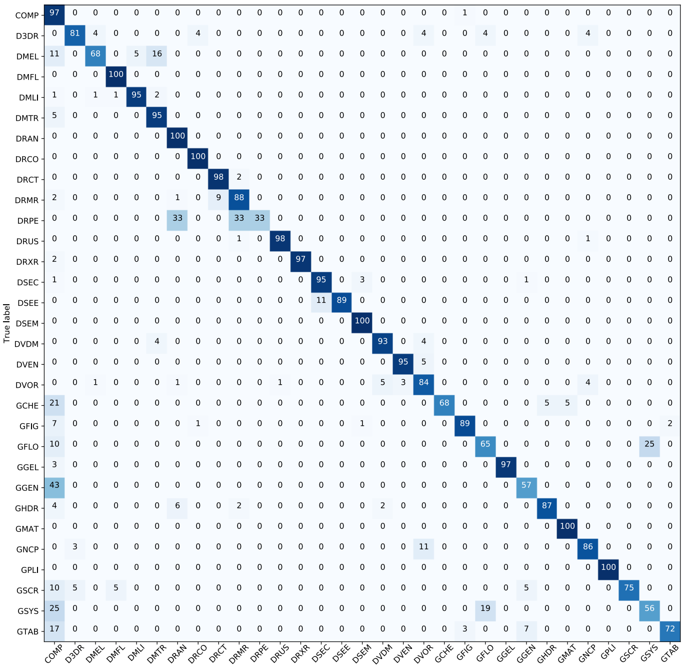
| Modality | Method | Accuracy |
|---|---|---|
| Textual | IBM_modality_run1 [1] | 64.2% |
| | IBM textual [3] | 69.6% |
| | textual CNN | **71.9%** |
| Visual | IBM_modality_run4 [1] | 80.8% |
| | IBM visual [3] | 82.2% |
| | DenseNet169 w/o data augm. w/o extra training | 83.8% |
| | DenseNet169 w/ data augm. w/o extra training | 84.5% |
| | DenseNet169 w/ data augm. w/ extra training | **86.8%** |
| Fusion | IBM_modality_run8 [1] | 81.7% |
| | IBM fusion [3] | 83.5% |
| | weighted average fusion w/ extra training | 89.2% |
| | maximum fusion w/ extra training | 89.4% |
| | product fusion w/ extra training | 91.8% |
| | Decision-level MLP fusion | 86.0% |
| | Feature-level MLP fusion | **93.7%** |

Table 1: Comparison of our methods with the best runs in ImageCLEF 2013.

The best fusion results are obtained with the feature-level MLP (93.7%). The confusion matrix of this best method is shown in Figure 2. MLPs are trained

[3] 34 images were removed from the dataset since the original submission due to their presence in both training and test sets.

without data augmentation as the visual augmented data is not paired with text inputs. A solution can be sought in future work to overcome this basic limitation and expect higher accuracy.

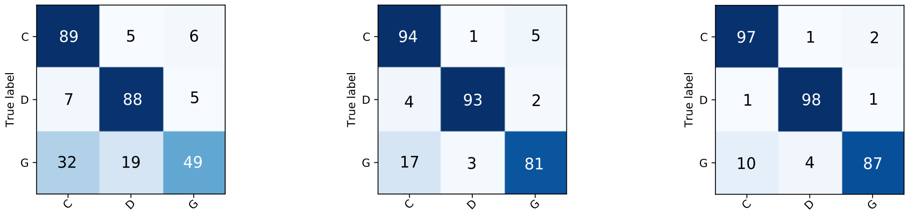


| | COMP | D3DR | DMEL | DMFL | DMLI | DMTR | DRAN | DRCO | DRCT | DRMR | DRPE | DRUS | DRXR | DSEC | DSEE | DSEM | DVDM | DVEN | DVOR | GCHE | GFIG | GFLO | GGEL | GGEN | GHDR | GMAT | GNCP | GPLI | GSCR | GSYS | GTAB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| COMP | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D3DR | 0 | 81 | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| DMEL | 11 | 0 | 68 | 0 | 5 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMFL | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMLI | 1 | 0 | 1 | 1 | 95 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMTR | 5 | 0 | 0 | 0 | 0 | 95 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DRAN | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DRCO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DRCT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 98 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DRMR | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 9 | 88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DRPE | 0 | 0 | 0 | 0 | 0 | 0 | 33 | 0 | 0 | 33 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DRUS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 98 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| DRXR | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DSEC | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 95 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DSEE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 89 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DSEM | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DVDM | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 93 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DVEN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 95 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DVOR | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 5 | 3 | 84 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| GCHE | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 68 | 0 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 0 | 0 |
| GFIG | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 89 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| GFLO | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 |
| GGEL | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GGEN | 43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 57 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GHDR | 4 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 87 | 0 | 0 | 0 | 0 | 0 | 0 |
| GMAT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 |
| GNCP | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 86 | 0 | 0 | 0 | 0 |
| GPLI | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| GSCR | 10 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 75 | 0 | 0 |
| GSYS | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 56 | 0 |
| GTAB | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 72 |

Fig. 2: Normalized confusion matrix (%) of the feature-level fusion method.

The most relevant classes are the diagnostic images as they offer more potential in clinical applications such as retrieval. The confusion matrices for the three main categories (compound, diagnostic and generic illustrations) are illustrated in Figure 3. It shows that our approach performs an excellent discrimination between diagnostic (e.g. MRI, CT, histopathology) and other images with less importance (e.g. compound figures, diagrams and maps).

In order to evaluate the complementarity of the visual and textual information, we measured the overlap of correct classification. With the best method previously described, the percentage of images correctly classified by both the visual and textual networks is 64.3%. 22.5% of the test set is correctly classified by the visual network but incorrectly classified by the textual and, vice-versa, 7.6% is correctly classified using the caption but incorrectly classified using visual information. These results suggest, as confirmed by the fusion results in Table 1, that the visual and textual analyses offer some degree of complementarity to boost the final classification accuracy.



(a) Textual (b) Visual (c) Feat.-level MLP

Fig. 3: Normalized confusion matrices for the three main categories: Compound, Diagnostic and Generic illustrations.

The accuracy obtained with multiple values of $\alpha$ in Equation 1 is illustrated in Figure 4. The best results with this weighted sum fusion are obtained with a contribution of the visual analysis slightly larger than the textual one ($\alpha = 0.51$), although a gradually reducing, yet neat, improvement from the single modality results is obtained with $\alpha$ values in the range $[0.51, 0.99]$.

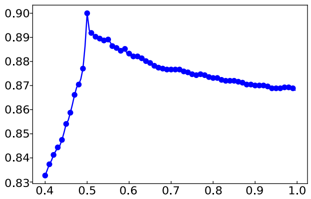


Fig. 4: Accuracy of the average fusion method for various weights $\alpha$.

The networks are implemented in Keras with TensorFlow backend. The computational training and test times are reported in Table 2 using a Titan Xp GPU.

| **Method** | Train time (nb. images) | Test time (nb. images) |
|---|---|---|
| Textual CNN | 1,079 s (2789) | 1.1 s (2570) |
| DenseNet169 w/o extra training | 3,233 s (2789) | 14.6 s (2570) |
| DenseNet169 w/ extra training | 22,510 s (25880) | 14.6 s (2570) |
| Feat.-level MLP w/ extra training | 2,626 s (25880) | 15.2 s (2570) |

Table 2: Computational time of the various networks.

## 5 Discussions and Future Work

As illustrated with the experiments, the proposed approach largely outperforms the state of the art [3] (93.7% vs. 83.5%). The results demonstrated the major importance of the visual analysis in the developed method (86.8% accuracy), in line with the results and conclusion from the literature [4, 10, 1, 3]. The visual data augmentation had an expected positive impact on the results with an increase of accuracy of 3%. The complementarity of textual and visual information was demonstrated by the series of experiments and analyses.

The proposed robust image modality classification enables to classify the large dataset from PubMed Central with over five million publicly available images and captions in 2017 and to use it as training or semi-supervised data for various medical image and text analysis tasks.

## References

[1] Abedini, M., Cao, L., Codella, N., Connell, J., Garnavi, R., Geva, A., Merler, M., Nguyen, Q., Pankanti, S., Smith, J., et al.: IBM research at ImageCLEF 2013 medical tasks. In: Proc. Workshop CLEF 2013 Working Notes. vol. 1179 (2013)

[2] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information (2017)

[3] Codella, N., Connell, J., Pankanti, S., Merler, M., Smith, J.R.: Automated medical image modality recognition by fusion of visual and text information. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 487–495. Springer (2014)

[4] 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) (September 2013)

[5] García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum) (September 2016)

[6] Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)

[7] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3128–3137 (2015)

[8] Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751 (2014)

[9] Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: Proc. of the 3rd International Conference on Learning Representations (ICLR) (2015)

[10] Kitanovski, I., Dimitrovski, I., Loskovska, S.: FCSE at medical tasks of ImageCLEF 2013. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum) (September 2013)

[11] Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D.: An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE journal of biomedical and health informatics 21(1), 31–40 (2017)

[12] Pelka, O., Friedrich, C.: Modality prediction of biomedical literature images using multimodal feature representation. GMS Medical Informatics, Biometry and Epidemiology (MIBE) (2016)