Fusion of Handcrafted and Deep Features for Medical Image Classification

Dong Liu\textsuperscript{1}, Yaohui Liu\textsuperscript{*}, Shanglin Li\textsuperscript{1}, Weiqing Li\textsuperscript{1} and Luda Wang\textsuperscript{1}

\textsuperscript{1}School of software and communication engineering, Xiangnan University, Chenzhou, Hunan, 423300, China

\textsuperscript{*}Corresponding author’s e-mail: aidenle@foxmail.com, xnxlyh@163.com

Abstract. Medical image classification has recently attracted increased attention. Effective feature extraction and learning are key means to improve classification performance. However, in the current study, handcrafted feature are mostly designed with intuitive mode, while the deep feature depends on a large amount of training samples and has weak interpretability. To capture more discriminative features for medical image, a novel feature fusion approach, termed multi-layer visual feature fusion (MLVSF), has been proposed on the basis of low-level, mid-level and deep features. More specifically, by fusing the handcrafted and deep features generated by local binary pattern variant, bag-of-visual-words, convolutional neural network, respectively, MLVSF can effectively enhance the discriminating power of features for medical image recognition. Experimental results on two medical image datasets show that MLVSF can improve convolutional neural networks, and achieve a better classification accuracy in comparison with some state-of-the-art methods.

1. Introduction

Medical images have played an extraordinary role in modern clinical medicine. Image classification technology can deal with large amounts of medical images, which has become an important means for computer-aided medical diagnosis and treatment [1]. Due to the subtle difference and higher ambiguity in medical images, the medical image classification task is more challenge than general image recognition. Moreover, the medical image involves numerous semantic information which also is crucial to clinicopathologic feature analysis.

In terms of visual feature research applied in medical image classification, it can generally be divided into three categories: low-level feature extraction, mid-level feature representation, and deep feature learning. Traditionally, the low-level feature extraction methods mainly describe image content in the aspect of color, shape, texture, and local pixel distribution, such as color vector patterns [2], SIFT[3] and local binary patterns (LBP)[4]. These methods have clear concept and simple calculation, but they usually lose semantic information and cannot achieve satisfactory performance with single kind of feature. The mid-level feature representation methods further make statistic, encoding or reconstitution based on low-level feature extraction, which can obtain more discriminative features and carry semantic information to some extent. For example, the well-known bag-of-visual-words (BoFW) model [6] is still an effective feature representation technique for scene image classification. Recently, some deep learning models have been developed for image classification, which automatically learn deep feature and improve the classification accuracy. Unlike the handcrafted features, convolutional neural networks (CNNs) [7], as one of the most popular deep learning model, can directly extract feature from original image pixels without any priori knowledge. Despite the deep
learning models have achieved remarkable results in scene classification, there still lies many challenges for deep learning in medical image classification task. Firstly, CNNs are data-driven deep learning methods depends on a large number of labeled samples, which restricts its performance in medical image dataset with few samples. Secondly, the features captured by CNNs generally have poor interpretability and also fail to carry semantic information. Finally, the CNNs are prone to "over-fitting" and require massive computational consumption.

In order to overcome the respective weakness of low-level, mid-level and deep features, we propose a handcrafted and deep feature fusion framework for medical image classification. Firstly, the local binary pattern variants are utilized as low-level visual descriptors to capture the global texture features. Then, we employ BoVW model as a mid-level approach to encode the visual feature with distribution patterns. Next, we use a pre-trained CNN with transfer learning to extract deep features from medical images. Finally, we fuse the handcrafted and deep features to train a support vector machine (SVM) classifier. To summarize, our contribution is threefold: (1) a novel handcrafted and deep feature fusion framework is proposed to capture more discriminative features; (2) Different visual descriptors and deep-nets are evaluated and selected for medical image classification; (3) Two medical image datasets are employed to illuminate the superiority.

The remainder of this paper is organized as follows. In Section 2, we introduce the feature fusion of handcrafted and deep features for medical image classification. Section 3 shows the experimental results and comparisons with some other methods. Finally, the conclusion is made in Section 4.

2. Feature fusion of handcrafted and deep features

2.1. Texture features extraction

Suppose, for a given image, we employ local binary pattern (LBP) [4] to extract global texture feature. LBP is a local texture descriptor with powerful discrimination and low computational complexity, which encode the local spatial patterns by using a circularly symmetric neighbour set of $P$ on a circle of radius $R$ and a central pixel $(x_c, y_c)$, which is denoted as:

$$LBP_{p,r}(x_c, y_c) = \sum_{i=0}^{P-1} s(g_i - g_c)2^i,$$

where $g_i$ is the gray value of $(x_i, y_i)$, and $g_c$ is the gray value of the neighbour set. In its standard form, 8 neighbourhood pixels of a central pixel have been selected, i.e. $P=8, R=1$. As shown in Fig.1, different configurations of $(P, R)$ can be selected based on specific applications.

![Figure 1. different configurations of (P, R) for LBP](image)

2.2. Mid-level features based on bag-of-visual-words

The medical image classification task based on bag-of-visual-word (BoVW) model can be formulated in the following manner. Given an image dataset $\{I_i\}$ and a set of medical image categories $c = \{c_1, c_2, ..., c_m\}$, we first sample local patches $\{p_i\}$ of the image in a sliding grid manner, and then SIFT features are extracted on each patches. A BoVW vocabulary $V = \{v_1, v_2, ..., v_k\}$ is learned by using $K$-means clustering on all local features. Then, by assigning the visual word to the local patches, we can obtain a feature representation $F_{BoVW}$, which is a vector that indicates the distribution of visual words in an image. The image classification flowchart based on BoVW model is shown in Figure 2.
2.3. Deep features based on Convolution neural network

As a popular deep learning method, convolution neural networks (CNNs) can directly obtain features from original image without prior knowledge, and have demonstrated their great superiority in visual recognition tasks. A typical CNN consists of convolutional layers, pooling layers, fully connected (FC) layers, and a softmax layer [7]. Given an image dataset \( \{ I_n \} \) and its corresponding label \( \{ C_n \} \), a CNN model can be formulated as follows:

\[
\begin{align*}
A_{\text{architecture}} &= F(I_n, C_n) \\
W_{\text{right}} &= \text{Init}(A_{\text{architecture}}) \\
\text{minimize} & \quad D(I_n, W_{\text{right}}, C_n)
\end{align*}
\]  

(2)

where \( F(\cdot) \) denotes the network architecture selecting function, such as AlexNet[7], VGGNet[8], GoogleNet[9] and ResNet[10]; \( \text{Init}(\cdot) \) denotes the initialization process of connection weights \( W_{\text{right}} \) and \( D(\cdot) \) computes the differences between the predicted label and true label. In the training phase of CNN, it is performed to minimize \( D(I_n, W_{\text{right}}, C_n) \) and at last \( W_{\text{right}} \) are iteratively determined. A brief introduction of CNN is shown in Fig.3, the output of last pooling layer is flattened in to a feature vector and fed into the FC layer, which can be considered as deep features of the whole image. Compared with the low-level and the mid-level visual feature, CNNs feature can usually achieve higher image classification accuracy, but it depends on a vast amount of labeled training samples. Furthermore, it is difficult to capture the semantic characteristics for CNNs feature, since it has weak interpretability.

2.4. Feature fusion for medical image classification

To combine the respective advantages of both handcrafted and deep features, an effective multi-layer visual feature fusion (termed MLVSF) framework has been prudently proposed for medical image classification. A schematic flowchart is shown in Fig. 4. More specifically, to obtain low-level features, we employ LBP or its variant CoALBP[5] to extract the global texture features (denoted as \( F_t \)), since the texture information is significant in medical image analysis. To encode the frequent local features, BoVW model with 4-neighbor soft weighting assign strategy [12] is adopted to generate mid-level
features (denoted as \( F_m \)). Then, we use pre-trained CNN with transfer learning to extract the first fully connection layer as deep features (denoted as \( F_d \)) for medical images. In next experiments, we select AlexNet and VGG-16 network respectively for CNN feature learning. Therefore, the fused features can be denoted as \( F_f = (F_i, F_m, F_d) \), which generated from multi-level visual features and provide more powerful discriminant capacity. Moreover, we use reliefF [11] to perform feature selection for \( F_f \).

Finally, the selected feature \( F_f \) are used to training a SVM classifier with histogram intersection (HI) kernels [13] as follows.

\[
K(F_f, F'_f) = \sum_{t} \min(x_{f,t}, x'_{f,t})
\]

where \( F_{f,t} \) is the \( t \)-th element of the feature vector. 

![Figure 4. The MLVSF framework for medical image classification](image)

### 3. Experiments and analysis

#### 3.1. Dataset and experimental protocol

To evaluate the performance of the proposed feature fusion framework, experiments are designed on two medical image datasets. (1) Lymphoma dataset [14], which includes three different patterns for lymphoma types, i.e. CLL with 113 images, FL with 139 images, and MCL with 122 images. The size of each image is equal to \( 1388 \times 1040 \), and the experiments are performed on the image patches with \( 227 \times 227 \times 3 \), i.e. each image is divided into 24 non-overlapped patches. (2) Histology dataset [15], which contains 960 samples and has been divided into 20 categories. Fig. 5 and Fig.6 shows image samples of two medical datasets, respectively.

For LBP feature extraction, we set the configuration of \( P=8, R=1 \). For BoVW model, we use a dense sampling strategy, i.e. the patch size and spacing are set to \( 16 \times 16 \) pixels and 8 pixels, respectively. The visual dictionary size of BoVW is set to 1000. To fine-tuning deep nets, we perform data augmentation for the histology dataset with 50% random affine transformation, rotation and translation before training the CNN model. The images in dataset are all resized to required input for transfer learning with pre-trained deep nets.

The accuracies are reported through the 5-fold cross-validation. For the classifier, we employ the SVM from the public LIBSVM library. All experimental results are obtained on a personal computer with Intel Core i9-7900X CPU, an NVIDIA TITAN XP GPU, and 16 GB of RAM. The operating system is Windows 10, and the implementation environment is under MATLAB 2019a with deep learning toolbox.
3.2. Result and discussion

In proposed multi-layer visual feature fusion framework (MLVSF), we use AlexNet to extract CNN features. And we employ LBP and CoLBP to extract low-level feature, respectively, and then two variant MLVSF-L and MLVSF-C can be generated with different texture descriptors. The classification accuracies of different methods are presented in Table 1.

| Methods     | LBP  | CoLBP | BoVW | AlexNet | VGG16 | MLVSF-L | MLVSF-C |
|-------------|------|-------|------|---------|-------|---------|---------|
| Lymphoma    | 72.34| 88.23 | 64.21| 91.12   | 92.23 | 93.22   | 95.02   |
| Histology   | 79.27| 85.56 | 89.16| 90.41   | 91.11 | 92.31   | 93.49   |

From Table 1, we can make the following conclusions. Firstly, both MLVSF-L and MLVSF-C improve AlexNet on both two medical datasets. Secondly, the CoLBP applied to the proposed MLVSF excels the basic LBP, which means CoLBP can provide more useful compensation for mid-level and deep features. Moreover, MLVSF-C consistently outperforms all compared methods, including VGG16 that has more number of layers and time complexity. Note that the BoVW achieve competitive classification accuracy in Histology dataset, but it has low accuracy in Lymphoma dataset, this may be due to the fact BoVW is unable to distinguish the local nuance of images in Lymphoma dataset. However, when BoVW applied to MLVSF, the performance has been greatly improved.

To discuss, we also employ VGG16 instead of AlexNet to proposed MLVSF framework, the proposed MLVSF-C can also achieve 3.18% and 2.92% improvement on basis of VGG16 in Lymphoma and Histology datasets, respectively.

The experimental result indicates that the proposed fusion of the global texture feature, the local mid-level feature, and the deep feature can provide a representative description for medical images.

4. Conclusion and future work

In this paper, we design a novel multi-level visual feature fusion method for medical image classification, which can efficiently integrate the respective advantages of handcrafted and deep features. Based on the experiments on two medical datasets, we can conclude that the proposed framework is indeed a superior model with discriminative features for medical image classification. However, in mid-level feature extraction, the spatial information description is inadequate, this may be the reason that BoVW has low accuracy in Lymphoma dataset. In future work, we will further explore the spatial relationship representation based on multi-level feature fusion.

Acknowledgments

This work was supported by the Natural Science Foundation of Hunan Province (No. 2018JJ3479, No. 2018JJ2370, 2017JJ3287, 2019JJ50564) and Scientific Research Fund of Hunan Provinicial Education (No.18B504, No, 17A200, No, 16B244)
References

[1] Mohan G, Subashini M. (2018) MRI based medical image analysis: Survey on brain tumor grade classification [J]. Biomedical Signal Processing and Control, 39: 139-161.

[2] Häfner M, Liedlgruber M, Uhl A. (2012) Color treatment in endoscopic image classification using multi-scale local color vector patterns [J]. Medical image analysis, 16(1): 75-86.

[3] Xia G S, Delon J, Gousseau Y. (2010) Shape-based invariant texture indexing [J]. International Journal of Computer Vision, 88(3): 382-403.

[4] Ojala T, Pietikäinen M, Mäenpää. T. (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary pattern, IEEE Trans. Patt. Anal. Mach. Intell., 24(7):971–987.

[5] Nosaka, Ryusuke, Yasuhiro O, and Kazuhiro F. (2012) Feature extraction based on co-occurrence of adjacent local binary patterns. Advances in Image and Video Technology. Springer Berlin Heidelberg, 82-91.

[6] Yang J, Jiang Y.G, Hauptmann A.G, et al. (2007) Evaluating bag-of-visual-words representations in scene classification[C]. Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 197-206.

[7] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. (2012) Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, CA, USA, pp. 1097–1105

[8] Simonyan K, Zisserman A. (2014) Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556.

[9] Szegedy C, Liu W, Jia Y, et al. (2015) Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition,1-9.

[10] He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 770-778.

[11] Robnik-Šikonja M, Kononenko I. (2003) Theoretical and empirical analysis of ReliefF and RReliefF[J]. Machine learning, 53(1-2): 23-69.

[12] Jiang Y G, Ngo C W, Yang J. (2007) Towards optimal bag-of-features for object categorization and semantic video retrieval[C]. Proceedings of the 6th ACM international conference on Image and video retrieval. ACM, 494-501.

[13] Maji S, Berg A C, Malik J. (2008) Classification using intersection kernel support vector machines is efficient[C]. IEEE Conference on Computer Vision and Pattern Recognition, 1-8.

[14] Shamir L, Orlov N, Eckley D M, et al. (2008) IICBU 2008: a proposed benchmark suite for biological image analysis [J]. Medical & biological engineering & computing, 46(9): 943-947.

[15] Kumar M D, Babaie M, Zhu S, et al. (2017) A comparative study of cnn, bovw and lbp for classification of histopathological images[C]. IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 1-7.