Deformable Gabor Feature Networks for Biomedical Image Classification

Xuan Gong\textsuperscript{1}\textsuperscript{†}, Xin Xia\textsuperscript{2}\textsuperscript{†}, Wentao Zhu\textsuperscript{3}, Baochang Zhang\textsuperscript{2}, David Doermann\textsuperscript{1}, Li’an Zho\textsuperscript{2}

\textsuperscript{1}University at Buffalo \textsuperscript{2}Beihang University \textsuperscript{3}Kwai Inc.

\{xuangong, doermann\}@buffalo.edu \{xiaxin, bczhang, lianzhuo\}@buaa.edu.cn wentaozhu91@gmail.com

Abstract

In recent years, deep learning has dominated progress in the field of medical image analysis. We find however, that the ability of current deep learning approaches to represent the complex geometric structures of many medical images is insufficient. One limitation is that deep learning models require a tremendous amount of data, and it is very difficult to obtain a sufficient amount with the necessary detail. A second limitation is that there are underlying features of these medical images that are well established, but the black-box nature of existing convolutional neural networks (CNNs) do not allow us to exploit them. In this paper, we revisit Gabor filters and introduce a deformable Gabor convolution (DGConv) to expand deep networks interpretability and enable complex spatial variations. The features are learned at deformable sampling locations with adaptive Gabor convolutions to improve representiveness and robustness to complex objects. The DGConv replaces standard convolutional layers and is easily trained end-to-end, resulting in deformable Gabor feature network (DGFN) with few additional parameters and minimal additional training cost. We introduce DGFN for addressing deep multi-instance multi-label classification on the INbreast dataset for mammograms and on the ChestX-ray14 dataset for pulmonary x-ray images.

1. Introduction

Automated medical imaging techniques for cancer screening are widely used for lesion analysis \cite{8}, but the traditional pipeline for computer aided diagnosis is typically built based on hand-crafted features \cite{25}. These features are not flexible and have poor generalization on unseen data. Deep features, however, are data-driven and are becoming the approach of choice in medical image analysis. Deep learning has achieved great success on skin cancer diagnostics \cite{6}, organs at risk delineation for radiotherapy \cite{32} and pneumonia detection from chest x-ray images \cite{21} for example.

One challenge for deep learning is that it is data hungry and often requires expensive and detailed annotation \cite{10, 24}. For cancer screening training and validation data in medical images, image-level description of the clinical diagnosis may not be sufficient to train for clinical diagnosis \cite{34}. Another challenge arises from CNN itself. CNNs are widely considered black boxes and difficult to interpret. This becomes a greater challenge for weekly supervised learning in biomedical image analysis, whose performance depends highly on powerful representations to handle complicated spatial variations, such as lesion sizes, shapes and viewpoints.

Gabor wavelets \cite{7} are widely considered the state-of-the-art hand-crafted feature extraction method, enhancing the robustness of the representation to scale and orientation changes in images. The advantage of Gabor transforms for specific frequency analysis makes them suitable to interpret and resist to dense spatial variations widely existing in biomedical images. Recently, Gabor convolutional networks (GCNs) \cite{17} have used Gabor filters to modulate convolutional filters and enhance representation ability of CNNs. \cite{17} only consider rigid transformations of kernels, however, and not deformable transformations on features that are required for medical image analysis. Thus the robustness of Gabor filters to spatial variations has not been fully investigated to facilitate feature extraction in CNNs.

On the other hand, deformable convolutional networks (DCNs) \cite{8} augment spatial sampling locations and provide generalized transformations such as anisotropic aspect ratios, demonstrating effectiveness on sophisticated vision tasks such as object detection. We will show that the tailored combination of Gabor filters and deformable convolutions in a dedicated architecture can better characterize the spatial variations and enhance feature representations to facilitate medical image analysis.

In this paper, we investigate deeply into Gabor wavelets with deformable transforms to enhance the networks interpretability and robustness to complex data variations. Unlike previous hand-crafted filters, the newly designed module learns Gabor filters end-to-end, thus improving

\begin{center}
\textsuperscript{†}Equal contribution.
\end{center}
Figure 1. The framework of our deformable Gabor Convolution (DGConv).

its adaptiveness to the input data. As illustrated in Figure 1, our deformable Gabor convolution (DGConv) includes deformable convolutions and adaptive Gabor convolutions that share the same modulation information. The deformable convolutions are endowed with local offset transforms to make the feature sampling locations learnable. The adaptive Gabor convolutions further facilitate the capture of visual properties such as spatial localization and orientation selectivity of the input objects, enhancing the generated deformable Gabor features with various dense transformations. To balance the performance and model complexity, we only employ deformable Gabor convolution (DGConv) to extract high level deep features. We integrate this new Gabor module into deep multi-instance multi-label networks, leading to deformable Gabor feature networks (DGFNs) to deal with large variations of objects in medical images. The contributions of this work are summarized as follows:

- Deformable Gabor feature network (DGFN) exploits deformable features and learnable Gabor features in one block to improve the interpretability of CNNs. The noise-resistant property inherited from Gabor features is successfully validated on CIFAR-10 with a 2% accuracy improvement over the baseline method.

- DGFN features both the adaptiveness to deformation and robustness to generalize spatial variations common in natural images. Their enhanced representative ability are shown to be beneficial for medical image analysis.

- The proposed Gabor module is generic and flexible, which can be easily applied to existing CNNs, such as ResNet and DenseNet.

2. Related Work

2.1. Deformable Convolutional Networks

CNNs have achieved great success for visual recognition but are inherently limited to spatial variations in object size, pose and viewpoint [16][28]. One method that has been used to address this problem is data augmentation which adds training samples with extensive spatial variations using random transformations. Robust features can be learned from the data but at the cost of an increased number of model parameters and additional training resources. Another method is to extract spatial invariant features with learned transformations. Ilse et al. [14] first proposed spatial transformer networks to learn invariance to translation, scale, rotation and generic warping, giving neural networks the ability to actively and spatially transform feature maps. Deformable convolutional networks (DCNs) [3] introduced offset learning to sample the feature map in a local and efficient manner which can be trained end-to-end.

2.2. Gabor Convolution Networks

Gabor wavelets [7] exhibit strong characteristics of spatial locality, scale and orientation selectivity, and insensitivity to illumination change. The recent rise of deep learning has lead to the combination of Gabor filters and convolution neural networks. Previously Gabor wavelets were only used to initialize deep networks or used in the preprocessing [15][31]. [22] replaced selected weight kernels of CNNs with Gabor filters to reduce training cost and time. Recent work has integrated Gabor filters into CNNs intrinsically to enhance the resistance of deep learned features to spatial changes [17]. However, the receptive field of the integrated Gabor filters is fixed and known, and such prior knowledge characterizes limited spatial transformations thus impeding the generalization of complicated spatial variations and new unknown tasks. In this work, we
go further by tailoring Gabor filters with learnable modulation masks and deformable transforms. The steerable property of Gabor filters is therefore inherited into the deformable convolutions and its representativeness to spatial variations is fully exploited.

2.3. Multi-Instance Learning for Weakly Supervised Image Analysis

There have been a number of previous attempts to utilize weakly supervised labels to train models for image analysis [23]. Papandreou et al. [20] proposed an iterative approach to predict pixel-wise labels in segmentation using image-level labels. Different pooling strategies were proposed for weakly supervised localization and segmentation respectively [22, 2]. Wu et al. [29] combined CNN with multi-instance learning (MIL) for image auto-annotation. Deep MIL with several efficient inference schemes was proposed for lesion localization and mammogram classification [33]. Attention based MIL further employed neural attention mechanisms as the inference [13]. Wan et al. [26] proposed a min-entropy latent model for weakly supervised object detection, which reduces the variance of positive instances and alleviates the ambiguity of the detectors. Unlike previous methods, our method uses a novel feature representation network to handle large variations of objects in medical images and improve overall image classification.

3. Deformable Gabor Convolution

Without loss of generality, the convolution operation described here is in 2D.

3.1. Deformable and Adaptive Gabor Convolution

To extract highly representative features, we combine the deformable convolution (DConv) with an adaptive Gabor convolution (GConv) by sharing modulation information. As illustrated in Figure 2, both the deformable convolution and Gabor transforms are adjusted with the learned masks.

Deformable Convolution: We are given $U$ standard convolution filters of size $H \times H$, which after being modulated by $V$ scale kernels of size $H \times H$, result in $U \times V$ modulated convolution filters of size $H \times H$. We define:

$$D_{u,v} = C_u \circ S_v,$$

where $D_{u,v}$ indicates the deformable convolution filter, $\circ$ is element wise product operation, $C_u$ is the $u^{th}$ convolution filter, and $S_v$ is the $v^{th}$ kernel to modulate the convolution filter. In our implementation, the deformable transforms [3] augment $D_{u,v}$ with translated offsets which are learned from the preceding feature maps through additional convolutions.

Consider a $3 \times 3$ kernel convolution, $R = \{(-1,-1), \cdots, (1,0), (1,1)\}$, with a dilation of 1, for example. Given $r_0$ as the 2D position of output feature and $r_n$ as the location of $R$, the deformable convolution filter $\hat{D}$ can be operated on as follows:

$$F_y(r_0) = \sum_{r_n \in R} \hat{D}(r_n) \times F_x(r_0 + r_n + \Delta r_n)$$

where $F_x$ and $F_y$ indicate the input and output feature respectively. The learned offset $\Delta r_n$ updates the offset location to $r_n + \Delta r_n$ and adjusts the receptive field of input $F_x$ on which $\hat{D}$ is implemented.

Adaptive Gabor Convolution: Adaptive Gabor filters are generated from $U$ Gabor filters of size $H \times H$ with $V$ learned kernels of size $H \times H$, where $U$ indicates the number of orientations of Gabor filters. We have:

$$\hat{G}_{v,u} = S_v \circ G_u,$$

where $G_u$ is the Gabor filter with orientation $u$, and $\hat{G}_{v,u}$ is the adaptive Gabor filter corresponding to the $u^{th}$ orientation and the $v^{th}$ scale. For DGConvs, different layers share the same Gabor filters $G = (G_1, \cdots, G_{TV})$ with various orientations but are adjusted with different information from the corresponding deformable convolution features.

If the dimensions of the weights in traditional convolution are $M_0 \times N_0 \times H \times H$, the dimensions of the learned convolution filters are $M \times N \times U \times H \times H$ in DGConv, where $U$ represents the number of additional orientation channels, $N(N_0)$ and $M(M_0)$ represent the channel number of the input and output respectively. In DGConv we set $N = N_0/\sqrt{U}$ and $M = M_0/\sqrt{U}$ to maintain similar amount of parameters with traditional convolution. Additional parameters in DGConv include $V \times H \times H$ parameters of mask and $(2 \times H \times H) \times N \times U \times H \times H$ parameters for offset learning, where $2 \times H \times H$ is the channel of offset fields and means that each position of input feature corresponds to an offset of size $2 \times H \times H$ for deformable convolution.

In DGConv, the number of orientation channels in the input and output feature needs to be $U$. So the number of orientation channels in the first input feature must be extended to $U$. For example, if the dimension of original input feature is $1 \times N \times W \times W$ where $W \times W$ is the size of input feature, it will be $U \times N \times W \times W$ after duplicating and concatenating. Thus the new module is light weight and can easily be implemented with a small number of additional parameters.

3.2. Forward Propagation

We use deformable Gabor convolutions (DGConvs) to produce deformable Gabor features. Given the input features $F$, the output Gabor features $\hat{F}$ are denoted:

$$\hat{F} = \text{DGConv}(F, \hat{D}, \hat{G}),$$

*The subscript is omitted for easy presentation.
The loss function of the network $L$, which will be described in the next section, is differentiable within a neighborhood of a point, which will be illustrated in the next section. We design a novel back propagation (BP) scheme to update parameters:

$$\delta_S = \frac{\partial L}{\partial S} = \frac{\partial L}{\partial G} \circ \sum_{u=1}^{V} G_u, \quad S \leftarrow S - \eta_1 \delta_S, \quad (6)$$

where $G_u$ is the Gabor filter with orientation $u$ and $\eta_1$ denotes the learning rate for $S$. We then fix $S$ and update parameters $D$ of deformable convolution filters:

$$\delta_C = \frac{\partial L}{\partial C} = \frac{\partial L}{\partial D} \circ \sum_{v=1}^{V} S_v, \quad D \leftarrow D - \eta_2 \delta_C, \quad (7)$$

where $S_v$ is the $v^{th}$ learned kernel and $\eta_2$ denotes the learning rate of convolution parameters.

4. Biomedical Image Analysis

There are many different ways to formulate problems in biomedical image analysis. Two of the most common are to classify an entire image as either having a particular condition or not (a binary-label task) and to associate the image with several labels (a multi-label task). To test our deformable Gabor feature network (DGFN), we have identified two representative datasets, the INbreast dataset [18] and the ChestX-ray14 dataset [27].

4.1. The INbreast Dataset

The INbreast Dataset [18] is a dataset of mammogram images consisting of 410 images from a total of 115 cases, of which 90 cases are from women with both breasts (4 images per case) and 25 cases are from mastectomy patients (2 images per case) [18]. The dataset includes four types of lesions: masses, calcifications, asymmetries, and distortions.

4.2. The ChestX-ray14 Dataset

As one of the largest publicly available chest x-ray datasets, ChestX-ray14 consists of 112,120 frontal-view x-ray images scanned from 32,717 patients including many patients with advanced lung diseases [27]. Each image is labeled with one or multiple pathology keywords, such as atelectasis, or cardiomegaly. This dataset consists of complicated diseases which may have interrelations which can be challenging for the classification task. The ChestX-ray14...
acquire the last deformable Gabor features

4.3.1 Multi-Instance Learning for Mammograms

In our DGFNs for Chest X-Rays dataset, we define a fourteen-dimensional label vector \( y_n = [y_n^1, y_n^2, \cdots, y_n^C] \) for \( n^{th} \) image \( I_n \), where \( C = 14 \) with binary values, representing either the absence (0) or the presence (1) of a pathology. The \( y_n^c \) indicates the presence of an associated pathology in the \( n^{th} \) image where \( c = \{1, 2, \cdots, C\} \), while a zero vector \([0, 0, \cdots, 0]\) represents the current x-ray image without any pathology. We consider each pathology as an independent multi-instance learning problem, which is the same as the mammogram classification, to solve the weakly supervised multi-label classification problem. We consider each patch as an instance and the problem can be formulated using equation (10). If there is no explicit priors on these labels, we can derive the loss function as:

\[
\mathcal{L} = - \sum_{n=1}^{N} \sum_{c=1}^{C} w(c) \log(p(y = y_n^c | I_n)),
\]

where \( w(c) = \frac{N}{\sum_{n=0}^{N} \mathbb{I}(y_n^c = 0)} \) and \( \mathbb{I}() \) is an indicator function for \( y_n^c \) being label \( c \).

Figure 3. Histogram of label frequencies on ChestX-ray14 dataset. The ChestX-ray14 dataset is imbalanced.
5. Experiments

Our deformable Gabor feature networks (DGFNs) are evaluated on the two medical image datasets described above and CIFAR-10 dataset. To balance the performance and training complexity, we use traditional convolution in the first two blocks and deploy deformable Gabor feature convolution in the following high level features.

### 5.1. Experiments on the INbreast Dataset

To prepare the data we first remove the background of the mammograms in a pre-processing step using Otsu’s segmentation method [19]. We then resize the pre-processed mammogram to $224 \times 224$. We use five-fold cross validation with three-fold training, one-fold validation and one-fold testing. We randomly flip the mammograms horizontally, rotate within 90 degree, shift them by 10% horizontally and vertically, and set a $50 \times 50$ box as 0 for data augmentation.

The proposed DGFNs employ AlexNet and ResNet18 as the backbones. We use the Adam optimization [5] algorithm with the initial learning rate of 0.0001 for both $\eta_1$ and $\eta_2$ and weight decay of 0.00005 in the training process. The learning rate decay is set to 10% for every 100 epochs and weight decay of 0.00005 in the training process. The results in Table 1 confirm that our DGFNs outperform CNN on INbreast-Deform.

### 5.2. Experiments on the ChestX-ray14 Dataset

We resize the x-ray images from $1024 \times 1024$ to $224 \times 224$ to reduce the computational cost and normalize them.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Kernel Stages</th>
<th>AUC (%)</th>
<th>#Params (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>16-32-64-128</td>
<td>75.89</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>32-64-128-256</td>
<td>78.26</td>
<td>2.80</td>
</tr>
<tr>
<td>ResNet18</td>
<td>8-16-32-64</td>
<td>76.90</td>
<td>0.70</td>
</tr>
<tr>
<td>(GCNs)</td>
<td>16-32-64-128</td>
<td>79.16</td>
<td>2.80</td>
</tr>
<tr>
<td>ResNet18</td>
<td>16-32-64-128</td>
<td>80.40</td>
<td>0.83</td>
</tr>
<tr>
<td>(DCNs)</td>
<td>32-64-128-256</td>
<td>82.03</td>
<td>3.05</td>
</tr>
<tr>
<td>ResNet18</td>
<td>8-16-32-32</td>
<td>77.59</td>
<td>0.53</td>
</tr>
<tr>
<td>(DGFNs)</td>
<td>16-32-64-128</td>
<td>81.29</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Figure 4. AUC comparison on INbreast-deform. All the networks are of similar model sizes with CNN 0.70M, GCN 0.70M, DCN 0.83M and DGFN 0.98M.
The feature map has $8 \times 8$ patches.

Table 3. Comparisons on INbreast dataset. DGFN with ResNet18 yields the best performance.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Acc (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet+Label Assign. MIL [33]</td>
<td>84.16</td>
<td>76.90</td>
</tr>
<tr>
<td>AlexNet+ DGFN+ MIL</td>
<td>86.22</td>
<td>78.12</td>
</tr>
<tr>
<td>ResNet18+ DGFN+ MIL</td>
<td>88.61</td>
<td>82.19</td>
</tr>
<tr>
<td>Pretrained AlexNet+Sparse MIL [33]</td>
<td>90.00</td>
<td>85.86</td>
</tr>
<tr>
<td>Pretrained AlexNet+ DGFN + MIL</td>
<td>91.34</td>
<td>87.22</td>
</tr>
<tr>
<td>Pretrained ResNet18 + DGFN + MIL</td>
<td>93.18</td>
<td>88.05</td>
</tr>
</tbody>
</table>

Based on the mean and standard deviation of images from the ImageNet training set [4]. In our experiments, we employ a DenseNet121 [12] as the backbone of our DGFN on ChestX-ray14 dataset. We resize the images to $224 \times 224$ and further augment the training data with random rotation and horizontal flipping. During training we use stochastic gradient descent (SGD) with momentum 0.9 and batch size 16. We use initial learning rates of 0.001 that are decayed by a factor of 10 each time when the validation loss has no improvement.

We used the official split released by Wang et al. [27] with 70% training, 20% testing and 10% validation. While Yao et al. [30] and Chexnet [21] randomly split the dataset and ensure that there is no patient overlap between the splits. Yao et al. [30] noted that there is insignificant performance difference with different random splits. Thus it is a fair comparison. We divide the compared methods into Fine-Tune (FT) and Off-The-Shelf (OTS) based on whether it used additional data for training. Guendel et al. [11] used another fully annotated dataset-PLCO Dataset [9] to facilitate training. While our DGFN and other comparable fine-tuned methods [21, 27, 1] are initialized with ImageNet. Table 4 demonstrates that among the group labeled fine-tune, DGFN with DenseNet121 outperforms [21, 27, 1] on all fourteen pathologies from the ChestX-ray14 dataset. Among the group labeled off-the-shelf, DGFN achieves average AUC of 78.39% and performs better on 11 out of 14 pathologies than other methods [30, 1]. Figure 6 illustrative effectiveness of DGFN to enhance variant representations, which is potentially of great help on automated biomedical image analysis.

5.3. Experiments on the CIFAR-10 Dataset

To verify the effectiveness of DGFN on the natural image dataset, we conduct extensive experiments on CIFAR-10 as well as CIFAR-10 with noise. We generate a noisy version of CIFAR-10 called CIFAR-10-Noise by replacing the pixel value with 255 at a probability of 1% percentage to test the network’s robustness to random Gaussian noise. We train on CIFAR-10 with random flipping and crop as augmentation. We test on CIFAR-10 and CIFAR-10-Noise respectively. We use ResNet18 as the backbone and use SGD optimization with the initial learning rates as 0.05. The batch size is set as 128 and the total number of training epochs is 300. Figure 7 is the comparison of test accuracy on CIFAR10-Noise with CNN, GCN, DCN and DGFN of similar sizes. Table 5 shows that the proposed DGFN...
Table 4. AUC (%) comparisons of DGFN with Off-The-Shelf (OTS) and Fine-Tune (FT) state-of-art methods on ChestX-ray14 dataset. Bold text emphasizes the highest value among each group.

<table>
<thead>
<tr>
<th>Pathology</th>
<th>Off-The-Shelf</th>
<th>Fine-Tune</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atelectasis</td>
<td>73.3</td>
<td>73.2</td>
</tr>
<tr>
<td>Cardiomegaly</td>
<td>85.8</td>
<td>75.9</td>
</tr>
<tr>
<td>Consolidation</td>
<td>71.7</td>
<td>75.3</td>
</tr>
<tr>
<td>Edema</td>
<td>80.6</td>
<td>85.7</td>
</tr>
<tr>
<td>Effusion</td>
<td>80.6</td>
<td>80.6</td>
</tr>
<tr>
<td>Emphysema</td>
<td>84.2</td>
<td>79.8</td>
</tr>
<tr>
<td>Fibrosis</td>
<td>74.3</td>
<td>73.9</td>
</tr>
<tr>
<td>Hernia</td>
<td>77.5</td>
<td><strong>81.9</strong></td>
</tr>
<tr>
<td>Infiltration</td>
<td>67.5</td>
<td>67.0</td>
</tr>
<tr>
<td>Mass</td>
<td>77.8</td>
<td>68.6</td>
</tr>
<tr>
<td>Nodule</td>
<td>72.7</td>
<td>66.5</td>
</tr>
<tr>
<td>Pleural_Thickening</td>
<td>72.4</td>
<td>70.8</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>69.0</td>
<td>68.3</td>
</tr>
<tr>
<td>Pneumathorax</td>
<td>80.5</td>
<td>79.1</td>
</tr>
<tr>
<td>Average</td>
<td>76.1</td>
<td>74.8</td>
</tr>
</tbody>
</table>

Acknowledgements

Baochang Zhang is the corresponding author. This study was supported by Grant NO.2019JZZY011101 from the Key Research and Development Program of Shandong Province to Dianmin Sun.

References
