Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Medical image classification using synergic deep learning

Jianpeng Zhang[a,b], Yutong Xie[a,b], Qi Wu[b], Yong Xia[a,c,*]

[a] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[b] School of Computer Science, University of Adelaide, SA 5005, Australia
[c] Centre for Multidisciplinary Convergence Computing (CMCC), School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China

Doesn't let model learn properly

- interclass similarity of med images

intra class variation of Medical images

ARTICLE INFO

Article history:
Received 24 September 2018
Revised 21 January 2019
Accepted 15 February 2019
Available online 18 February 2019

Keywords:
Medical image classification
Intra-class variation
Inter-class similarity
Synergic deep learning model

ABSTRACT

Novel solution

Architecture

Results

Problem

The classification of medical images is an essential task in computer-aided diagnosis, medical image retrieval and mining. Although deep learning has shown proven advantages over traditional methods that rely on the handcrafted features, it remains challenging due to the significant intra-class variation and inter-class similarity caused by the diversity of imaging modalities and clinical pathologies. In this paper, we propose a synergic deep learning (SDL) model to address this issue by using multiple deep convolutional neural networks (DCNNs) simultaneously and enabling them to mutually learn from each other. Each pair of DCNNs has their learned image representation concatenated as the input of a synergic network, which has a fully connected structure that predicts whether the pair of input images belong to the same class. Thus, if one DCNN makes a correct classification, a mistake made by the other DCNN leads to a synergic error that serves as an extra force to update the model. This model can be trained end-to-end under the supervision of classification errors from DCNNs and synergic errors from each pair of DCNNs. Our experimental results on the ImageCLEF-2015, ImageCLEF-2016, ISIC-2016, and ISIC-2017 datasets indicate that the proposed SDL model achieves the state-of-the-art performance in these medical image classification tasks.

1. Introduction

The significance of digital medical imaging in the modern healthcare has led to the indispensable role of medical image analysis in the clinical therapy (Ghosh et al., 2011; de Bruijne, 2016; Kalpathy-Cramer et al., 2015). Medical image classification, a fundamental step in medical image analysis, aims to distinguish medical images according to a certain criterion, such as clinical pathologies or imaging modalities. A reliable medical image classification system is able to assist doctors in the fast and accurate interpretation of medical images.

Medical image classification has been thoroughly studied during the past decades with a huge number of solutions in the literature (Baloch and Krim, 2007; Song et al., 2013; Koitka and Friedrich, 2016), most of which are based on handcrafted features. Despite the success of these methods, it is usually difficult to design handcrafted features that are optimal for a specific classification task. In recent years, deep learning techniques (Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016; Huang et al., 2017; Li et al., 2017), especially deep convolutional neural networks (DCNN), have led to significant breakthroughs in medical image classification (Koitka and Friedrich, 2016; Xu et al., 2016; Shen et al., 2017; Esteva et al., 2017; Personnaz et al., 2017; Kumar et al., 2017; Yu et al., 2017b), and medical image segmentation (Dong et al., 2017; Soltaninejad et al., 2017). However, although these methods are more accurate than handcrafted feature-based approaches, they have not achieved the same success on medical image classification (Sirinukunwattana et al., 2016; Xie et al., 2018) as they have done in the ImageNet Challenge (Deng et al., 2009; Krizhevsky et al., 2012). The suboptimal performance is attributed mainly to two reasons.

→ ML bad

→ DL good

The performance on medical images is not as good as ImageNet.

Reasons for suboptimal performance

First, deep models may overfit the training data, which is far from adequate, as there is usually a small dataset in medical image analysis and this relates to the work required in acquiring the image data and then in image annotation (Weese and Lorenz, 2016). To address this issue, pre-trained deep models have been adopted, since it has been widely recognized that the image representation ability learned from large-scale datasets, such as ImageNet (Deng et al., 2009), can be efficiently transferred to generic visual recognition tasks, where the training data is limited

Reason 1 → Overfitting (small data)
→ Can be solved using Transfer Learning

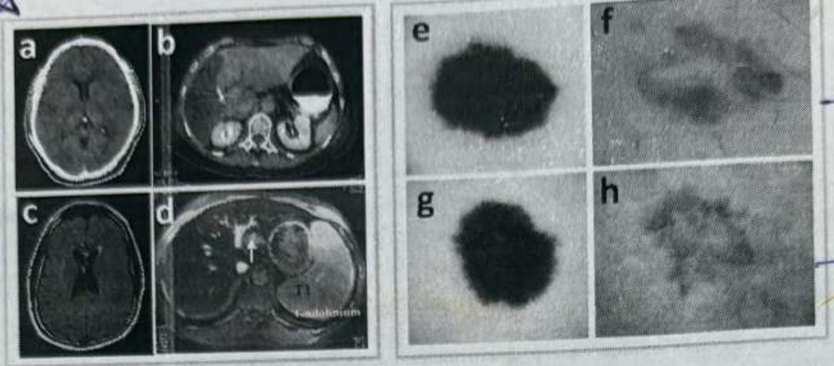
Conflicts of Interest Statement: The authors declare no conflict of interest.

[*] Corresponding author.
E-mail address: yxia@nwpu.edu.cn (Y. Xia).

https://doi.org/10.1016/j.media.2019.02.010
1361-8415/© 2019 Elsevier B.V. All rights reserved.

Reason 2 - Intra class ~~o~~ variation
Inter class similarity

CT →
difficult to distinguish
MRI →



→ benign skin lesion
→ malignant skin lesion

Fig. 1. Examples show the intra-class variation and inter-class similarity in modality-based medical image classification (a–d) and clinical pathology-based medical image classification (e–h): (a) a brain CT image, (b) a pleural CT image, (c) a brain MR image, (d) a pleural MR image, (e, f) benign skin lesions, and (g, h) malignant skin lesions.

(Zhou et al., 2017; Ravishankar et al., 2017; Oquab et al., 2014; [illegible] et al., 2016).

Second, and more significantly, the intra-class variation and inter-class similarity pose an even greater challenge to the classification of medical images (Song et al., 2015). As an example shown in Fig. 1(a)–(d), the separation of images from computed tomography (CT) and magnetic resonance (MR) imaging scanners is difficult in: (1) both CT and MR images provide the anatomical information about the body parts that are imaged, and hence share many visual similarities and non-professionals can have difficulty in separating them (see Fig. 1(a) vs (c), or Fig. 1(b) vs (d)); and (2) images from the same modality will differ depending upon the anatomical location and individual variability (see Fig. 1(a) vs (b), or Fig. 1(c) vs (d)). Another example shown in Fig. 1(e)–(h) is the separation of malignant skin lesions from benign ones. It reveals that there is a big visual difference between the benign skin lesions (e) and (f) and between malignant ones (g) and (h). Nevertheless, the benign skin lesions (e) and (f) are even more similar to the malignant lesions (g) and (h), respectively, in the shape and color. To address this challenge, human observers focus more on the ambiguity caused by hard cases, which may provide more discriminatory information than easy ones (Bengio et al., 2009). The pair-wise learning strategy is an effective technique that learns from pairs of samples and captures more information in favor of distinguishing hard cases.

MRI & CT

skin lesions

1.1. Related work

Handcrafted feature-based medical image classification: The descriptors for color, texture, and shape and combined descriptors have been widely used in medical image classification. [illegible] and Krim (2007) proposed a flexible skew-symmetric shape model to capture shape variability within a certain neighborhood and account for all potential variability. Song et al. (2013) designed a novel texture descriptor to represent rich texture features by integrating multi-scale Gabor filters and local binary patterns (LBP) histograms for lung tissue classification. Koitka and Friedrich (2016) extracted up to 11 handcrafted visual descriptors and jointly used them for modality based medical image classification. Compared with these handcrafted feature-based methods, the proposed SDL can learn the discriminative feature representation from data adaptively and effectively.

Deep learning-based medical image classification: DCNN models provide a unified feature extraction-classification framework to free human users from the troublesome handcrafted feature extraction for medical image classification. Xu et al. (2014) adopted a DCNN to minimize manual annotation and produced good feature representations for histopathological colon cancer image classification. Shen et al. (2017) proposed a multi-crop pooling strategy and applied it to a DCNN to capture object salient information for lung nodule classification on chest CT images. Esteva et al. (2017) trained a DCNN using 129,450 clinical images for diagnosing the most common and deadliest skin cancers and achieved the performance that matches the performance of 21 board-certified dermatologists. Koitka and Friedrich (2016) extracted the output of the last fully connected layer in a pre-trained ResNet-152 model and adopted them to train a custom network layer using the pseudo-inverse method (Personnaz et al., 1986). Kumar et al. (2017) integrated two different pre-trained DCNN architectures and combined them into a stronger classifier. Yu et al. (2017b) presented an ensemble of multiple pre-trained ResNet-50 and VGGNet-16 models and multiple fully-trained DCNNs by calculating the weighted sum of predicted probabilities. In our previous work (Zhang et al., 2018a; Xie et al., 2018), we jointly used deep and handcrafted visual features for medical image classification and found that handcrafted features were able to complement the image representation learned by DCNNs on small training datasets. Different from these networks, the proposed SDL model simultaneously takes multiple images as input, and thus enables multiple DCNN components mutually improve each other for learning better discriminative representation. → USP for ~~De~~ SDL

Impossible

Pair-wise learning: In the past decade, the pair-wise learning strategy has been applied to various perception tasks, such as signature verification (Bromley et al., 1994), face verification (Chopra et al., 2005), speech analysis (Kamper et al., 2015; 2016; Renshaw et al., 2015), and natural language processing (Mueller and Thyagarajan, 2016). Bromley et al. (1994) described a Siamese neural network for verification of signatures written on a pen-input tablet by comparing the distance which is cosine of the angle between an extracted feature vector and a stored feature vector. Chopra et al. (2005) presented a general discriminative method for learning a similarity metric from data pairs by minimizing a discriminative loss function that enlarges the metrics of pairs of faces from the same person and narrows the pairs from different persons. Recent years have witnessed the widespread applications of pair-wise learning in unsupervised speech feature learning. Kamper et al. (2015) proposed an unsupervised deep auto-encoder feature extractor for zero-resource speech processing by using weak top-down supervision from word pairs obtained by an unsupervised term discovery system. Kamper et al. (2016) also used word pairs to train a Siamese DCNN that takes a pair of speech segments as input and uses a hinge loss to classify same-word pairs and different-word pairs. Renshaw et al. (2015) claimed that guiding the representation learning using word pairs provides a major benefit over standard unsupervised methods. Pair-wise learning has also been applied to natural language processing. Mueller and Thyagarajan (2016) presented a Siamese recurrent

USP → ① ~~[illegible]~~ takes in multiple images as input multiple (probably for 2 DCNN branches)

USP → (2) No distance loss func^n, uses cross-entropy loss

|| comp.? ← {Simultaneously learns multiple image pairs

DCNN branch

chitecture which is trained on paired examples to learn a highly structured space of sentence representations that captures rich semantics for learning sentence similarity. Different from the traditional pair-wise learning, the SDL model avoids handcrafted design of tricky distance metric loss functions for optimization, and automatically learns whether image pairs belong to the same category or not by using a cross-entropy loss function. Besides, the SDL model supports the simultaneous learning of multiple image pairs, which works with multiple DCNN components on the premise of not sharing parameters such that the model can benefit from an ensemble of multiple networks.

→ USP of SDL

→ Parameters are not shared

1.2. Outline of our work

Reason 2 In this paper, we propose a synergic deep learning (SDL) model to learn the discriminative representation simultaneously from pairs of images, which include both similar images in different categories and dissimilar images in the same category, for medical image classification. The SDL model consists of n pre-trained DCNNs and C_n^2 synergic networks. Each DCNN learns image representation and classification, and each pair of DCNNs has their learned image representation concatenated as the input of a synergic network, which has a fully connected structure, to predict whether the pair of input images belongs to the same class or not. Thus, the SDL model can be trained in an end-to-end fashion under the supervision of both the classification error from each DCNN and the synergic error from each pair of DCNNs. We have evaluated the proposed model on the 2015/2016 Image Cross Language Evaluation Forum (ImageCLEF) subfigure classification challenge datasets, and the 2016/2017 International Skin Imaging Collaboration (ISIC) skin lesion classification challenge datasets. Our results suggest that the SDL model achieves the state-of-the-art performance on these four medical image classification tasks.

Arch

2 losses

The main contributions of this paper are three-fold. First, we propose the SDL model that learns the discriminative feature representation from multiple images simultaneously including both similar inter-class images and dissimilar intra-class images. Second, we enable each pair of DCNNs in the SDL model to mutually facilitate each other during the learning process, since, if one DCNN makes correct decision, the mistake made by the other DCNN may lead to a synergic error that serves as an extra force to learn the discriminative representation. Finally, we achieve the state-of-the-art performance on the ImageCLEF-2015, ImageCLEF-2016 Subfigure Classification datasets, ISIC-2016 and ISIC-2017 Skin Lesion Classification datasets.

Reason 2

Algo

Read A pilot data of this work was presented in MICCAI 2018 (Zhang et al., 2018b). In this paper, we have substantially revised and extended the conference paper. The main extension includes that (1) the SDL model was generalized from a special version SDL^2 which has only two DCNN components, to a generalized version SDL^n with n DCNNs, and the generalization leads to improved performance in medical image classification; and (2) the proposed model was evaluated not only on pathology-based image classification datasets (i.e. ISIC-2016 and ISIC-2017 datasets), but also on modality-based image classification datasets (i.e. ImageCLEF-2015 and ImageCLEF-2016 datasets).

{Extension of Zhang et al 2018b, this paper) (or generalisation)

2. Material and method

2.1. Datasets

For this study, we use four medical image classification datasets, including two modality-based medical image classification datasets, i.e. ImageCLEF 2015 (de Herrera et al., 2015) and ImageCLEF 2016 (de Herrera et al., 2016) datasets, and two pathology-based medical image classification datasets, i.e. ISIC-2016 (Gutman et al., 2016) and ISIC-2017 (Codella et al., 2018) datasets.

ImageCLEF-2015, ImageCLEF-2016: Recognizing the increasing complexity of images in biomedical literatures, ImageCLEF collected medical figures with sub-figures that produced by multiple imaging modalities and illustrations drawn from analysis of medical data from the PubMed Central (PMC) (Müller et al., 2012). The ImageCLEF-2015 dataset consists of 4532 training images and 2244 testing images, whereas the ImageCLEF-2016 dataset contains 6776 training images and 4166 testing images. The images in both datasets are divided into 30 categories, including 18 categories of medical diagnostic images such as CT, MRI, and PET images, and 12 categories of illustrations such as figures, tables, and flow charts.

ISIC-2016, ISIC-2017: Both datasets were leveraged by ISIC, which is an international effort to improve melanoma diagnosis. The ISIC-2016 dataset is made up of 900 training and 379 test dermoscopy images which were screened for both privacy and quality assurance. Lesions in these images are all paired with a gold standard (definitive) malignancy diagnosis, i.e. benign or malignant. The ISIC-2017 dataset contains 2000 training, 150 validation and 600 test dermoscopy images. Similarly, each skin lesion is paired with a gold standard diagnosis, i.e. melanoma, nevus and seborrheic keratosis. Actually, this dataset contains two binary classification sub-tasks melanoma classification (i.e. melanoma vs. others) and seborrheic keratosis classification (i.e. seborrheic keratosis vs. others).

2.2. Method

The proposed SDL model, denoted by SDL^n, consists of three major modules: an image pair input layer, n DCNN components and C_n^2 synergic networks (see Fig. 2(a)). A special case SDL^2 is shown in Fig. 2(b). The input of the SDL model is a group of randomly selected images, instead of a single image. Each DCNN component of any network structure serves to independently learn representation from images under the supervision of true labels of input images. A synergic network, which has a fully connected structure, is used to verify whether the input pair belongs to the same category or not, and give the corrective feedback if a synergic error exists. We then delve into each of the three modules of the SDL model.

2.2.1. Pair input layer

3 labels (2) ↑

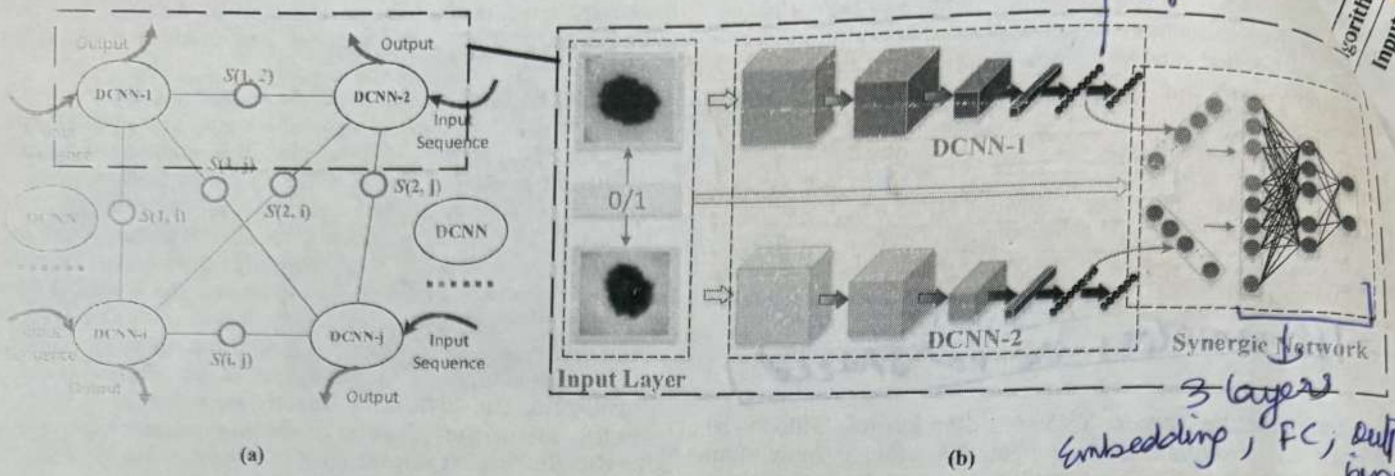
Different from traditional DCNNs, the proposed SDL^n model simultaneously accepts n input images that are randomly selected from the training set. Each image, together with its class label, is input into a DCNN component, and each pair of images has a corresponding synergic label that will be used by a synergic network. In order to unify the image size, we resize each image to 224 × 224 × 3 using the bicubic interpolation.

2.2.2. DCNN components → Vanilla transfer learning

Due to the strong representation capability of the famous residual network (He et al., 2016), we employ a pre-trained 50-layer residual neural network (ResNet-50) as the initialization of each DCNN component which is denoted by DCNN-i (i = 1, 2, ..., n). However, it is worth noting that any DCNN, such as AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2014) and GoogLeNet (Szegedy et al., 2015), can be embedded in the SDL^n model as a DCNN component. Each DCNN component is trained using an image sequence $\boldsymbol{X} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)}\}$ and a corresponding class label sequence $\boldsymbol{Y} = \{y^{(1)}, y^{(2)}, \ldots, y^{(M)}\}$

[1] http://www.isdis.net/index.php/isic-project/.

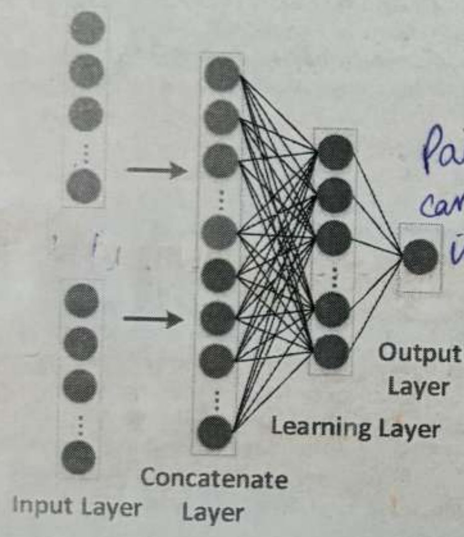
DCNNs → independent → 1 if $y_1 = y_2$
3 labels → y_1, y_2, $y_{synergic}$
3 loss → L_1, L_2, $L_{synergic}$



second last layer

3 layers
Embedding, FC, output binary

Fig. 2. (a): Architecture of the SDL^n model which has n DCNN components and C_n^2 synergic networks. DCNN-i, $(i = 1, \ldots, n)$, represents i-th DCNN component, and $S(i, j)$ represents the synergic network between DCNN-i and DCNN-j. DCNN-i serves to independently learn representation from images under the supervision of true labels of input images. $S(i, j)$ of a fully connected struture is used to verify whether the input pair belongs to the same category or not, and give the corrective feedback if a synergic error exists. (b): Architecture of the SDL^2 model of dual DCNNs and a synergic network.



Pairs can be inter or intra class

Fig. 3. Diagram of the synergic network.

aiming to find a set of parameters $\boldsymbol{\theta}$ that minimizes the following cross-entropy loss

$$l(\theta) = -\frac{1}{M}\left[\sum_{i=1}^{M}\sum_{j=1}^{K} 1\{y^{(i)} = j\} \log \frac{e^{z_j^{(i)}}}{\sum_{l=1}^{K} e^{z_l^{(i)}}}\right] \tag{1}$$

where K is the number of classes, $\mathbf{Z}^{(i)} = \mathcal{F}(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ represents the forward computing. This optimization problem can be solved by using the mini-batch stochastic gradient descent (mini-batch SGD) algorithm. The obtained parameter set for DCNN-i is denoted by $\theta^{(i)}$, and the parameters are not shared among different DCNN components. → independent DCNNs

2.2.3. Synergic network

To further supervise the training of each DCNN component with the synergic label of each pair of images, we design a synergic network, which consists of an embedding layer, a fully connected learning layer and an output layer (see Fig. 3).

Let a pair of images $(\mathbf{x}_A, \mathbf{x}_B)$ be input into two DCNN components (DCNN-i, DCNN-j), respectively. The output of the second last fully connected layer in a DCNN is defined as the deep image features learned by that DCNN, which can be obtained through forward computing, formally shown as follows

$$\begin{aligned} \boldsymbol{f}_A &= \mathcal{F}(\mathbf{x}_A, \boldsymbol{\theta}^{(i)}) \\ \boldsymbol{f}_B &= \mathcal{F}(\mathbf{x}_B, \boldsymbol{\theta}^{(j)}) \end{aligned} \tag{2}$$

Then, the deep features of both images are concatenated as $\boldsymbol{f}_{A\circ B}$ and input into the synergic network. The corresponding expected output is the synergic label of the image pair, which is defined as follows

$$y_S(\mathbf{x}_A, \mathbf{x}_B) = \begin{cases} 1 & if \quad y_A = y_B \\ 0 & if \quad y_A \neq y_B \end{cases} \tag{3}$$

To avoid the unbalance data problem, the percentage of intra-class image pairs in each batch is about 45% – 55%. It is convenient to monitor the synergic signal by adding another sigmoid layer and using the following binary cross entropy loss

$$l^S(\boldsymbol{\theta}^S) = y_S \log \widehat{y_S} + (1 - y_S)\log(1 - \widehat{y_S}) \tag{4}$$

where $\boldsymbol{\theta}^S$ is the parameters of the synergic network, $\widehat{y_S} = \mathcal{F}(f_{A\circ B}, \boldsymbol{\theta}^S)$ is the forward computing of the synergic network. This synergic network verifies whether the input image pair belongs to the same category or not, and gives the corrective feedback if a synergic error exists. → Penalise

2.2.4. Training and testing

The proposed SDL^n model consists of n DCNN components and C_n^2 synergic networks. During the end-to-end training, the parameters of each DCNN component and each synergic network can be updated as

$$\begin{cases} \boldsymbol{\theta}^{(i)}(t+1) = \boldsymbol{\theta}^{(i)}(t) - \boldsymbol{\eta}(t) \cdot \Delta^{(i)} \\ \boldsymbol{\theta}^{S(i,j)}(t+1) = \boldsymbol{\theta}^{S(i,j)}(t) - \boldsymbol{\eta}(t) \cdot \Delta^{S(i,j)} \end{cases} \tag{5}$$

vanilla

where $\boldsymbol{\eta}(t)$ is the variable learning rate, $S(i, j)$ represents the synergic network between DCNN-i and DCNN-j,

$$\Delta^{(i)} = \frac{\partial l^{(i)}(\boldsymbol{\theta}^{(i)})}{\partial \boldsymbol{\theta}^{(i)}} + \lambda \sum_{j=1, j\neq i}^{n} \frac{\partial l^{S(i,j)}(\boldsymbol{\theta}^{S(i,j)})}{\partial \boldsymbol{\theta}^{S(i,j)}} \tag{6}$$

backprop + SDL penalisation

$$\Delta^{S(i,j)} = \frac{\partial l^{S(i,j)}(\boldsymbol{\theta}^{S(i,j)})}{\partial \boldsymbol{\theta}^{S(i,j)}} \tag{7}$$

and λ represents the trade-off between subversion of classification error and synergic error. Algorithm 1 summarizes the training process of the SDL^2 model, which can be extended to the training of the SDL^n model.

How much Synergic layer effects ~~to~~/penalises the wrong output

for 2DCNNs

Algorithm 1 The training process of the SDL^2 model.

```
Input: Two image batches $X_1 = \{x_1^{(1)}, x_1^{(2)}, \dots, x_1^{(M)}\}$ and $X_2 = \{x_2^{(1)}, x_2^{(2)}, \dots, x_2^{(M)}\}$, initialized parameters of dual DCNNs and synergic network, $\theta^{(1)}$, $\theta^{(2)}$ and $\theta^S$, learning rate $\eta(t)$ and the hyper parameter $\boxed{\lambda}$.

Step 1: Forward propagation:
    $F_1 = \mathcal{F}(X_1, \theta^{(1)})$
    $F_2 = \mathcal{F}(X_2, \theta^{(2)})$          ⇒ vanilla CNN
Step 2: Concatenate $F_1$ and $F_2$ to $F_{1\circ2} = \{f_{1\circ2}^{(1)}, f_{1\circ2}^{(2)}, \dots, f_{1\circ2}^{(M)}\}$ where $f_{1\circ2}^{(i)}$ represents the combination of $f_1^{(i)}$ and $f_2^{(i)}$, and input them to the synergic network. The labels of these three supervisions are $Y_1 = \{y_1^{(1)}, y_1^{(2)}, \dots, y_1^{(M)}\}$, $Y_2 = \{y_2^{(1)}, y_2^{(2)}, \dots, y_2^{(M)}\}$ and $Y_S = \{y_S^{(1)}, y_S^{(2)}, \dots, y_S^{(M)}\}$, where $y_S^{(i)} = 1$ if $y_1^{(i)} = y_2^{(i)}$, otherwise $y_S^{(i)} = 0$.
    $Y_1 = DCNN_1$, $Y_2 = DCNN_2$ $Y_S = SL$
Step 3: Update parameters $\underbrace{\theta^{(1)}, \theta^{(2)}, \theta^S}_{\text{weights}}$ by using back-propagation algorithm.
    Compute loss:
        $l^{(1)}(\theta^{(1)})$, $l^{(2)}(\theta^{(2)})$ and $l^S(\theta^S)$.
    Compute gradient:          Basically classic backprop with λΔS penalisation term
        $\Delta^S = \frac{\partial l^S(\theta^S)}{\partial \theta^S}$,
        $\Delta^{(1)} = \frac{\partial l^{(1)}(\theta^{(1)})}{\partial \theta^{(1)}} + \lambda\Delta^S$,
        $\Delta^{(2)} = \frac{\partial l^{(2)}(\theta^{(2)})}{\partial \theta^{(2)}} + \lambda\Delta^S$          → Synergic layer
    Update parameters:          Classic
        $\theta^{(1)}(t+1) \leftarrow \theta^{(1)}(t) - \eta(t) \cdot \Delta^{(1)}$
        $\theta^{(2)}(t+1) \leftarrow \theta^{(2)}(t) - \eta(t) \cdot \Delta^{(2)}$
        $\theta^S(t+1) \leftarrow \theta^S(t) - \eta(t) \cdot \Delta^S$
```

inference

When applying the trained SDL^n model to the classification of a test image $\boldsymbol{x}$, each DCNN component DCNN-i gives a prediction vector $\mathbf{P}^{(i)} = (p_1^{(i)}, p_2^{(i)}, \dots, p_K^{(i)})$, which is the activations in its last fully connected layer. The class label of this test image can be predicted as

numpy

$$y(\boldsymbol{x}) = \underset{j}{argmax}\left\{\sum_{i=1}^{n} p_1^{(i)}, \dots, \sum_{i=1}^{n} p_j^{(i)}, \dots, \sum_{i=1}^{n} p_K^{(i)}\right\} \tag{8}$$

used as vanilla DCNN during inference

3. Experiments

3.1. Experimental settings

To alleviate the overfitting of deep models, we employed two data argumentation (DA) strategies to enlarge the training dataset. The first strategy (DA1) is to use the ImageDataGenetrator toolbox (Chollet et al., 2015) to apply geometric transformations to training images, including random rotation ([−10°, +10°]), shifts (0 ~ 10% of total width and height), shear (0 ~ 0.1 radians in the counter-clockwise direction), zoom (90% ~ 110% of width and height), and horizontally and vertically flip. The second strategy (DA2) is to add new training data if available. We collected 1796 images used for the modality classification task in the ImageCLEF-2013 Challenge (de Herrera et al., 2013) to enlarge the ImageCLEF-2015 and ImageCLEF-2016 subfigure classification training datasets, and collected 1320 dermoscopy images from the ISIC Archive[2] to enlarge the ISIC-2017 dataset. Meanwhile, we chose the pre-trained ResNet-50 model as the DCNN component, which has been trained on the ImageNet dataset. To adapt it to our datasets, we replaced all of its fully connected layers with a fully connected layer of 1024 neurons, a fully connected layer of K neurons and a softmax layer, and then fine-tuned it using our own medical image dataset. The weights of newly inserted fully connected layers were initialized by sampling a uniform distribution $U(-0.05, 0.05)$. We set the variable learning rate as follows

Table 1
Performance comparison of ResNet-50^n and SDL^n models on the ImageCLEF-2015 classification test set.

(1)

n	Acc (%)					
	Group 1: No DA		Group 2: DA1		Group 3: DA1+DA2	
	ResNet-50^n	SDL^n	ResNet-50^n	SDL^n	ResNet-50^n	SDL^n
1	73.31	/	74.11	/	76.25	/
2	73.98	**75.00**	74.69	**75.04**	76.52	77.58
3	74.47	**75.22**	75.13	**75.27**	76.56	77.76
4	74.55	**75.36**	75.18	**75.53**	76.65	78.21

Very small improvement ?

$$\eta(t) = \frac{\eta(0)}{1 + 10^{-4} \times t} \tag{9}$$

where t is the index of iteration, and the initial learning rate $\eta(0)$=0.0001. We set the maximum iteration number to 100,000 and adopted the mini-batch SGD algorithm with a batch size 32 as the optimizer. To stop the training process when the model falls into overfitting, 10% of training data were randomly selected to form a validation set, which was used to monitor the performance of our model. We empirically set the hyper parameter λ of the SDL model to 3 in our experiments. Since the SDL^n model uses the average predicted score produced by n DCNN components to label each test image, we independently fine-tuned n pre-trained ResNet-50 models and evaluated the SDL^n model against the ensemble of them (ResNet-50^n) for a fair comparison.

Config

3.2. Results on the ImageCLEF-2015 Dataset

(1)

Table 1 gives the classification accuracy (*Acc*) of the ResNet-50^n and SDL^n models on the ImageCLEF-2015 test dataset in three experiments: without using DA, using the DA1 strategy, and using both DA1 and DA2 strategies. It shows that the proposed SDL^n model is steadily more accurate than the ResNet-50^n model in all three experiments. We assumed that the image classification accuracy of ResNet-50^n and SDL^n are random variables X1 and X2, respectively, each following a Gaussian distribution, i.e. $X1 \sim N(\mu_1, \sigma_1^2), X2 \sim N(\mu_2, \sigma_2^2)$. The difference between X1 and X2 is defined as $D = X1 - X2$. We adopted the paired $\boldsymbol{t}$-test to determine whether the accuracy gain obtained by the proposed SDL^n model over ResNet-50^n is statistically significant. Thus, the hypotheses to be tested are H_0: $\mu_D \geq 0$ versus H_1: $\mu_D < 0$. Given the significance level $\alpha = 0.01$, $t_{0.01}(9-1) = 2.896$, and the rejection domain is $W = \{t \leq -2.896\}$. According to Table 1, we had

t-test

~~[illegible]~~ t-test

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \approx -0.79 \tag{10}$$

$$S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2} \approx 0.46 \tag{11}$$

and the value of statistic t_0 is

$$t_0 = \frac{\overline{D}}{S_D/\sqrt{n}} = \frac{-0.79}{0.46/\sqrt{9}} \approx -5.15 < -2.896 \tag{12}$$

Since t_0 belongs to the rejection domain W, we rejected the hypothesis H_0. Therefore, the proposed SDL^n model achieved more accuracy image classification than the ResNet-50^n model on the ImageCLEF-2015 test set, and the performance improvement is statistically significant. It reveals that using more DCNN components

[2] https://isic-archive.com/.

Model config

ResNet 50 + variable η + SGD (Mini-batch) (32) + n=100,000 + 10% validation data + λ=3

{1024 neurons in 1 FC Layer}

↑n → ↑accuracy

Table 2
Classification accuracy of our SDL model and eight best-performed solutions on the ImageCLEF-2015 test dataset.

(1)

Index	Method	Acc(%)
0	**SDL model (SDL^4)**	**78.21**
1	Yu et al. (2017b)	76.87
2	Pelka and Friedrich (2015)	60.91
3	Cirujeda and Binefa (2015)	52.98

Table 3
Performance comparison of ResNet-50^n and SDL^n models on the ImageCLEF-2016 classification test set.

(2)

n	Acc (%)					
	Group 1: No DA		Group 2: DA1		Group 3: DA1+DA2	
	ResNet-50^n	SDL^n	ResNet-50^n	SDL^n	ResNet-50^n	SDL^n
1	84.69	/	85.77	/	86.80	/
2	84.90	**85.93**	86.15	**86.32**	87.11	**87.54**
3	85.21	**86.36**	86.25	**86.61**	87.21	**87.69**
4	85.43	**86.51**	86.39	**86.75**	87.33	**87.97**

Table 4
Classification accuracy of our SDL model and five best-performed solutions on the ImageCLEF-2016 test dataset.

(2)

Index	Method	Acc(%)
0	**SDL model**	**87.97**
1	Yu et al. (2017b)	87.37
2	Zhang et al. (2018a)	85.47
3	Koitka and Friedrich (2016)	85.38
4	Koitka and Friedrich (2016)	84.46
5	Valavanis et al. (2016)	84.01

↑n→↑accuracy

Beats every thing

DA1 + DA2 + 4DCNN = Best SDL Model

Table 5
Performance comparison of the proposed SDL^n model and ResNet-50^n model.

(3)

Methods	No DA			DA1		
	AveP	*Acc*	*AUC*	*AveP*	*Acc*	*AUC*
ResNet-50	0.6102	0.8496	0.7829	0.6224	0.8522	0.7742
ResNet-50^2	0.6115	0.8443	0.7826	0.6308	0.8549	0.7968
SDL^2	**0.6536**	**0.8522**	**0.8139**	**0.6644**	**0.8575**	**0.8179**
ResNet-50^3	0.6201	0.8470	0.7976	0.6323	0.8575	0.7946
SDL^3	**0.6608**	**0.8549**	**0.8149**	**0.6713**	**0.8602**	**0.8371**
ResNet-50^4	0.6283	0.8443	0.8050	0.6318	0.8575	0.7971
SDL^4	**0.6704**	**0.8575**	**0.8287**	**0.6810**	**0.8628**	**0.8224**

Table 6
Performance comparison of the proposed SDL model and five top results listed in competition leaderboard. The evaluation index *AveP* is the only assessment index that all participants were ranked according to this metric.

(3)

Methods	*AveP**	*Acc*	*AUC*
SDL	**0.681**	**0.863**	0.822
CUMED (Yu et al., 2017a)	0.637	0.855	0.804
GTDL	0.619	0.813	0.802
BF-TB Thierno	0.598	0.834	**0.826**
ThrunLab	0.563	0.786	0.796
Jordan Yap	0.559	0.844	0.775

always leads to more accurate classification and the highest accuracy 78.21% was obtained by our SDL^n model when it uses the data augmentation strategies DA1 and DA2 and 4 DCNN components.

We also compared proposed SDL model to three top-ranking algorithms listed in the Challenge Leaderboard. The first algorithm (Yu et al., 2017b) uses an ensemble of five pre-trained ResNet-50 models, five pre-trained VGG models and five fully-trained DCNN models with the help of augmented data from the ImageCLEF-2013 dataset, and achieved a much-improved accuracy over the baseline ResNet-50 model. The second (Pelka and Friedrich, 2015) and third (Cirujeda and Binefa, 2015) algorithms use handcrafted visual features extraction, feature engineering, and classifier construction. The classification accuracy given in Table 2 shows that our SDL^4 model is able to produce substantially more accurate image classification on this dataset than other three algorithms.

3.3. Results on the ImageCLEF-2016 Dataset

(2)

DA1 + DA2 consistent

Table 3 gives the classification accuracy of the ResNet-50^n and SDL^n models on the ImageCLEF-2016 dataset in three experiments. Similarly, no matter using the data augmentation strategy DA1 or DA2 or not using data augmentation, the proposed SDL^n model steadily outperformed the ResNet-50^n model when the number of DCNN components ranges from 2 to 4. The highest accuracy 87.97% was achieved by our model when it uses the data augmentation strategies DA1 and DA2 and 4 DCNN components. This conclusion is consistent with the finding on the ImageCLEF-2015 Dataset.

In Table 4, we compared the accuracy of proposed SDL^n model with that of five best-performed algorithms. The first algorithm (Yu et al., 2017b) uses an ensemble of 15 deep neural networks, the second algorithm (Zhang et al., 2018a) jointly uses the features learned by three deep models and two types of handcrafted features, the third algorithm (Koitka and Friedrich, 2016) extracts the deep features from a pre-trained 152-layer ResNet model and uses them to train a classifier, the fourth algorithm (Koitka and Friedrich, 2016) combines 11 types of handcrafted visual features with feature engineering, and the fifth algorithm (Valavanis et al., 2016) fuses the classical bag of words (BoW) model, and bag of colors (BOC) model. To ensure a fair comparison, all these algorithms, except for the second one, use the ImageCLEF-2013 dataset as additional training data. It shows that our SDL model achieved the highest classification accuracy, which is even higher than the accuracy obtained by using an ensemble of 15 deep models.

3.4. Results on the ISIC-16 Dataset

(3)

outperforms even with 2DCNN

Fig. 4 shows the receiver operating characteristic (*ROC*) curves and area under the *ROC* curve (*AUC*) obtained by applying the ResNet-50^n and SDL^n models without data augmentation to the ISIC-16 test set, respectively. It reveals that the proposed SDL^n model (red curves) outperforms the ResNet-50^n model(blue curves) no matter using 2,3, or 4 DCNN components.

Table 5 gives the quantitative comparison of the performance of both models. The skin lesion classification performance on 379 test images was assessed by the average precision (*AveP*), *Acc*, and *AUC*. It shows that no matter using 2, 3 or 4 DCNN components, the SDL^n model performs steadily better than the ResNet-50^n model on all metrics. It also reveals that using data augmentation obviously improves most of the obtained classification metrics.

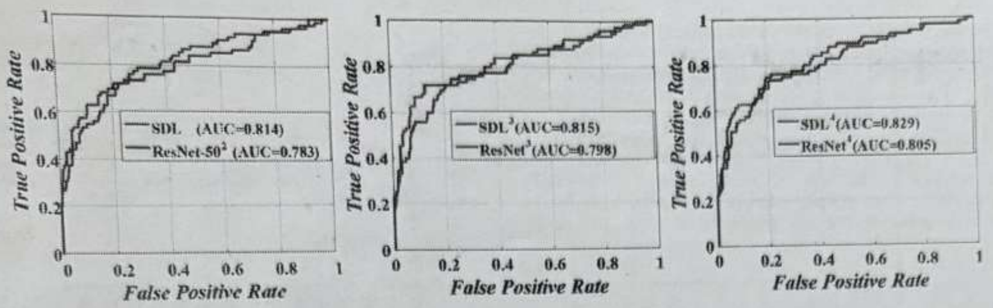
Table 6 shows the performance of the proposed SDL model with 4 DCNN components and the top five challenge records,[3] which were ranked based on *AveP*. Among these six solutions, the proposed SDL model achieves the highest *AveP*, highest *Acc*, and second highest *AUC*. The 1st place method (Yu et al., 2017a) leveraged a segmentation network to extract lesion objects based on the segmented results, for helping the classification network focus on more representative and specific regions. Based on the strength of synergic learning, the proposed SDL model achieved a higher performance in skin lesion classification without using lesion segmentation.

→ Beats prev. segmentation based model even after not segmenting

[3] https://challenge.kitware.com/#phase/5667455bcad3a56fac786791.

Comparison → a) Base Resnet models
b) Top performing algos for that dataset

All datasets give consistent results →
① ↑ Augmentation (DA1+DA2) → ↑ accuracy
② ↑ n → ↑ accuracy



ig. 4. ROC curves of the proposed SDL^n model and ResNet-50^n model. (Left: n=2, middle: n=3, right: n=4). (For interpretation of the references to colour in this figure egend, the reader is referred to the web version of this article.)

Table 7
Performance comparison in the melanoma (M) classification and seborrheic keratosis (SK) classification when applying the proposed SDL^2, SDL^3, and SDL^4 models, baseline ResNet-50, and six top results listed in the ISIC-17 competition leaderboard. The evaluation index '*AveAUC*' is the only assessment index that all participants were ranked according to this metric. (#1(Matsunaga et al., 2017) #2(Díaz, 2017) #3(Menegola et al., 2017) #4(Bi et al., 2017) #5(Yang et al., 2017) #6(DeVries and Ramachandram, 2017)).

Methods	External data	M Classification			SK Classification			*AveAUC*
		AUC	*AveP*	*Acc*	*AUC*	*AveP*	*Acc*	
ResNet-50	1320	0.856	0.590	0.853	0.948	0.794	0.842	0.902
SDL^2	1320	0.868	0.689	0.872	0.955	0.818	0.917	0.912
SDL^3	1320	0.871	0.702	0.878	0.955	0.825	0.918	**0.913**
SDL^4	1320	0.868	0.720	**0.888**	0.958	**0.840**	**0.925**	**0.913**
#1	1444	0.868	0.710	0.828	0.953	0.786	0.803	0.911
#2	900	0.856	**0.747**	0.823	**0.965**	0.839	0.875	0.910
#3	7544	**0.874**	0.715	0.872	0.943	0.790	0.895	0.908
#4	1600	0.870	0.732	0.858	0.921	0.770	0.918	0.896
#5	0	0.830	0.665	0.830	0.942	0.808	0.917	0.886
#6	1341	0.836	0.703	0.845	0.935	0.771	0.913	0.886

best ↑ values

3.5. Results on the ISIC-17 Dataset

Table 7 gives the performance comparison of SDL^2, SDL^3, and SDL^4 to the baseline ResNet-50, and six top ranking results in the ISIC 2017 challenge leaderboard.[4] Note that the average *AUC* (*AveAUC*) of melanoma classification and seborrheic keratosis classification is the gold evaluation metric, according to which all participants were ranked. To improve the performance of DCNNs by using more training data, as done in #1, #2, #3, #4 and #6, we enlarged the training set by using additional 1320 dermoscopy images. The proposed SDL^2, SDL^3, and SDL^4 models have a substantial improvement in all evaluation metrics when compared with the baseline ResNet-50 model, and SDL^3 and SDL^4 are slightly superior to SDL^2. The proposed SDL^4 model got highest *Acc*, second highest *AveP* in melanoma classification and highest *AveP*, *Acc*, and second highest *AUC* in seborrheic keratosis classification. In summary, both SDL^3 and SDL^4 achieved an *AveAUC* of 0.913, higher than the *AveAUC* of top-ranking solutions (Matsunaga et al., 2017) and (Menegola et al., 2017) which uses an ensemble of multiple pre-trained DCNNs and as many as 7544 external images, respectively.

4. Discussion

4.1. Stability interval of hyper parameter λ

We used the SDL^2 model as a case study to evaluate the impact of hyper parameter λ on the classification performance. Fig. 5 shows the accuracy obtained by applying the SDL^2 model with different values of λ to four datasets. It reveals that, when λ takes a value from the range [3, 8], the SDL^2 model achieved good accuracy on all datasets and its performance is relatively robust to the value of λ. Therefore, we suggest taking the value of λ from [3, 8].

4.2. Performance without data augmentation

There are several commonly used data augmentation strategies, including rotation, zooming, shifting, flipping, and adding random noise. The deep learning methods used in our comparative experiments use similar but not the same augmentation strategies (see Table 8). To demonstrate that the performance improvement is mainly contributed from the novel model architecture, instead of using more suitable data augmentation strategies, it is worthwhile to compare the performance of deep models when data augmentation was not used. (Yu et al., 2017b) reported the accuracy of 72.42% and 82.61% on the ImageCLEF-2015 and ImageCLEF-2016 datasets, respectively, without any data augmentation. It shows in Tables 1 and 3 that our SDL^2 model achieved the 75.00% and 85.93% on both datasets, respectively, without any augmentation. Therefore, without the contribution of data augmentation, our SDL^n model is more accurate than the method proposed by (Yu et al., 2017b), which is second in accuracy only to ours in Tables 2 and 4.

4.3. Accuracy vs. time and memory cost

The accuracy and time-cost of our SDL^n model on four datasets versus the number of DCNN components were plotted in Fig. In each subfigure, the Y-axis on the left is the metrics of image classification accuracy such as *Acc* or *AveP*, and the Y-axis on the right is the metric of time-cost, which is defined as the ratio between the training time of SDL^n and the training time of the baseline ResNet-50 model. It is clear that, with the increase of DCNN

[4] https://challenge.kitware.com/#phase/584b0afccad3a51cc66c8e38.

The model does not perform well only because of better augmentation. It does better even without augmentation. Augmentation only boosts the accuracy.

Metric v/s λ → best range [3, 8]

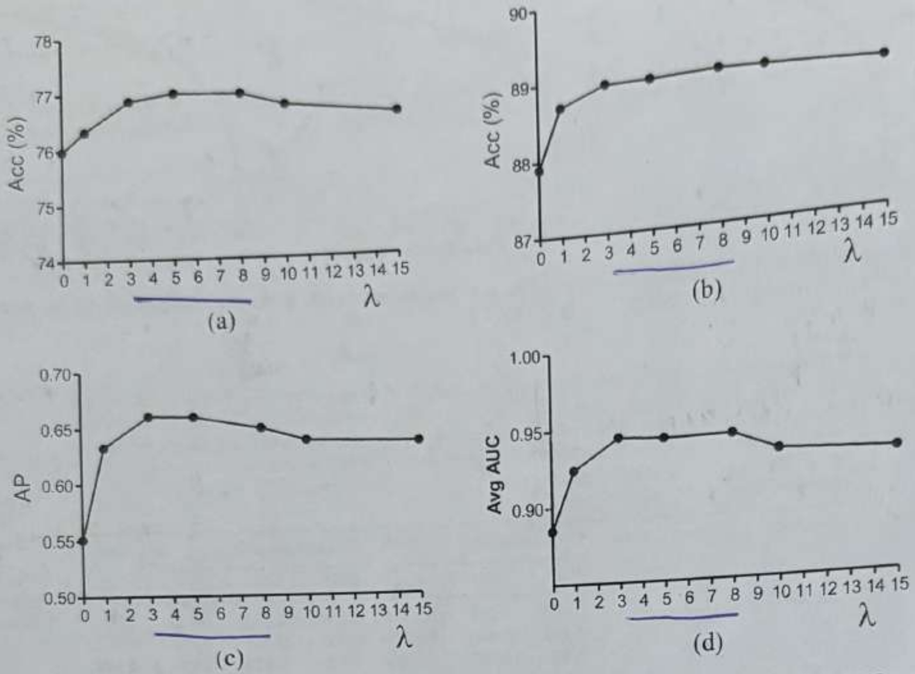


Fig. 5. Validation performance curves of the proposed SDL² model in four datasets ((a) ImageCLEF-2015, (b) ImageCLEF-2016, (c) ISIC-2016, and (d) ISIC-2017) with different hyper parameter λ.

Table 8
Data augmentation strategies used in state-of-the-art deep learning methods..

	Rotation	Zoom	Shift	Flip	Shear	Random noise
Yu et al. (2017b)	✓	✓	✓	✓	×	×
Yu et al. (2017a)	✓	×	✓	×	×	✓
Matsunaga et al. (2017)	✓	✓	✓	✓	×	×
Ours	✓	✓	✓	✓	✓	×

Metric v/s time cost

$$TC = \frac{\Delta T(SDL^n)}{T(Resnet50)}$$

Training time

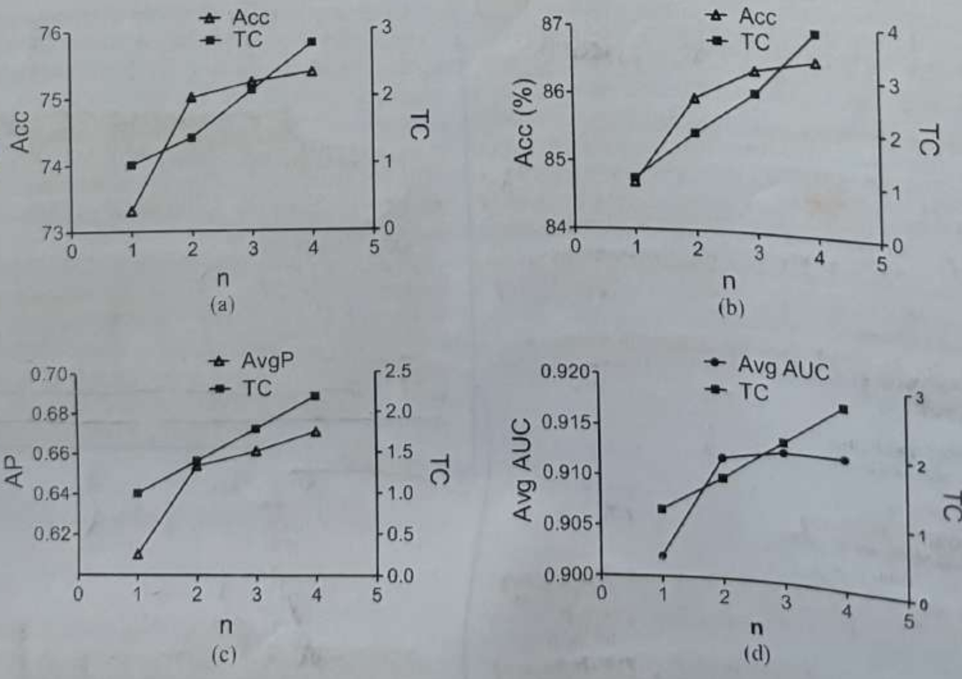


Fig. 6. Performance-time curves of the proposed SDL^n model on the (a) ImageCLEF-2015, (b) ImageCLEF-2016, (c) ISIC-2016, and (d) ISIC-2017 datasets when n changes from 1 to 4 (SDL^1 corresponding to the single ResNet-50 model).

n↑ → TC↑

n↑ → acc↑ and then converges/ stabilizes

→ GPU ☆ SDL^4 is memory 8 time intensive
☆ Even SDL^2 is more memory intensive than ResNet

Table 9
GPU memory cost in training the baseline ResNet-50 model and proposed SDL^n model.

Models	GPU Memory Cost (GB)
ResNet-50	3.9
SDL^2	$\boxed{5.1}$
SDL^3	6.9
SDL^4	8.7

→ Can't train on my Legion

components, the time-cost grows steadily, whereas the accuracy first improves and then becomes stable. Meanwhile, Table 9 shows the GPU memory cost in training the baseline ResNet-50^n model and proposed SDL^n model with one NVIDIA GTX Titan XP GPU, when using input of size $224 \times 224 \times 3$ and a batch size of 32. It reveals that the memory cost of our SDL^n model increases steadily with the augment of DCNN components. Fortunately, training a SDL^4 model only requires 8.7 GB GPU memory, which can be performed on a 12 GB TITAN XP GPU. Therefore, taking the computational and spatial complexity into consideration, we suggest using the SDL^2 and SDL^3 models. → Suggestion

5. Conclusion

In this paper, we propose the SDL model to address the challenge caused by the intra-class variation and inter-class similarity for medical image classification. This model simultaneously uses multiple DCNNs with synergic networks to enable those DCNNs to mutually learn from each other. Our results on the ImageCLEF-2015, ImageCLEF-2016, ISIC-2016, and ISIC-2017 datasets show that the proposed SDL model achieves the state-of-the-art performance in these medical image classification tasks. In the future, we will focus on the reinforcement learning algorithms to automatically search the number of DCNNs, model parallel computing and structure optimization to enlarge the scale of the SDL model.

→ FW: RL to find perfect SDL configuration

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 61771397 and 61471297.

References

Baloch, S., Krim, H., 2007. Flexible skew-symmetric shape model for shape representation, classification, and sampling. IEEE Trans. Image Process. 16, 317–328. doi:10.1109/TIP.2006.888348.

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), pp. 41–48. doi:10.1145/1553374.1553380.

Bi, L., Kim, J., Ahn, E., Feng, D., 2017. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv:1703.04197v1.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1994. Signature verification using a" siamese" time delay neural network. In: Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS), pp. 737–744.

de Bruijne, M., 2016. Machine learning approaches in medical image analysis: from detection to diagnosis. Med. Image Anal. 33, 94–97. doi:10.1016/j.media.2016.06.032.

Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., Heng, P., 2017. Dcan: deep contour-aware networks for object instance segmentation from histology images. Med. Image Anal. 36, 135–146. doi:10.1016/j.media.2016.11.004.

Chollet, F., et al., 2015. Keras. GitHub repository (2015).

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 539–546. doi:10.1109/CVPR.2005.202.

Cirujeda, P., Binefa, X., 2015. Medical image classification via 2d color feature based covariance descriptors. In: Proceedings of CLEF (Working Notes).

Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: Proceedings of IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 168–172. doi:10.1109/ISBI.2018.8363547.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. doi:10.1109/CVPR.2009.5206848.

DeVries, T., Ramachandram, D., 2017. Skin lesion classification using deep multiscale convolutional neural networks. arXiv:1703.01402v1.

Díaz, I.G., 2017. Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. arXiv:1703.01976v1.

Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y., 2017. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In: Annual Conference on Medical Image Understanding and Analysis. Springer, pp. 506–517.

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., Thrun, S., 2017. Corrigendum: dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. doi:10.1038/nature21056.

Ghosh, P., Antani, S., Long, L.R., Thoma, G.R., 2011. Review of medical image retrieval systems and future directions. In: Proceedings of 2011 24th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6. doi:10.1109/CBMS.2011.5999142.

Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern A., 2016. Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv:1605.01397v1.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.

de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S.K., Müller, H., 2013. Overview of the imageclef 2013 medical tasks. In: Proceedings of CLEF (Working Notes).

de Herrera, A.G.S., Müller, H., Bromuri, S., 2015. Overview of the imageclef 2015 medical classification task.. CLEF (Working Notes).

de Herrera, A.G.S., Schaer, R., Bromuri, S., Müller, H., 2016. Overview of the imageclef 2016 medical classification task., CLEF (Working Notes).

Kalpathy-Cramer, J., Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Muller, H., 2015. Evaluating performance of biomedical image retrieval systems -an overview of the medical image retrieval task at imageclef 2004 – 2013. Comput. Med. Imaging Graphics 39, 55–61. doi:10.1016/j.compmedimag.2014.03.004.

Kamper, H., Elsner, M., Jansen, A., Goldwater, S., 2015. Unsupervised neural network based feature extraction using weak top-down constraints. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5818–5822. doi:10.1109/ICASSP.2015.7179087.

Kamper, H., Wang, W., Livescu, K., 2016. Deep convolutional acoustic word embeddings using word-pair side information. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4950–4954. doi:10.1109/ICASSP.2016.7472619.

Koitka, S., Friedrich, C.M., 2016. Traditional feature engineering and deep learning approaches at medical classification task of imageclef 2016. In: Proceedings of CLEF (Working Notes).

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105.

Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D., 2017. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J. Biomed. Health Inf. 21, 31–40. doi:10.1109/JBHI.2016.2635663.

Li, R., Zeng, T., peng, H., Ji, S., 2017. Deep learning segmentation of optical microscopy images improves 3d neuron reconstruction. IEEE Trans. Med. Imaging 36, 1533–1541. doi:10.1109/TMI.2017.2679713.

Matsunaga, K., Hamada, A., Minagawa, A., Koga, H., 2017. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv:1703.03108v1.

Menegola, A., Tavares, J., Fornaciali, M., Li, L. T., Avila, S., Valle, E., 2017. Recod titans at isic challenge 2017. arXiv:1703.04819v1.

Mettes, P., Koelma, D.C., Snoek, C.G., 2016. The imagenet shuffle: reorganized pre-training for video event detection. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR), pp. 175–182. doi:10.1145/2911996.2912036.

Mueller, J., Thyagarajan, A., 2016. Siamese recurrent architectures for learning sentence similarity. In: Proceedings of Thirtieth AAAI Conference on Artificial Intelligence (AAAI), 16, pp. 2786–2792.

Müller, H., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., 2012. Creating a classification of image types in the medical literature for visual categorization. In: Proceedings of Medical Imaging 2012: Advanced PACS-based Imaging Informatics and Therapeutic Applications, 8319. International Society for Optics and Photonics, p. 83190P.

Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2014. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1717–1724. doi:10.1109/CVPR.2014.222.

Pelka, O., Friedrich, C.M., 2015. Fhdo biomedical computer science group at medical classification task of imageclef 2015.. In: Proceedings of CLEF (Working Notes).

Personnaz, L., Guyon, I., Dreyfus, G., 1986. Collective computational properties of neural networks: new learning mechanisms. Phys. Rev. A 34, 4217–4228. doi:10.1103/PhysRevA.34.4217.