ISIC CNN Documentation

Description

Deep Neural Network trained for the task of "Skin Lesion Analysis Towards Melanoma Detection".

This is a challenge proposed by the International Skin Imaging Collaboration (ISIC), an international effort to improve melanoma diagnosis, sponsored by the International Society for Digital Imaging of the Skin (ISDIS). ISIC has developed an international repository of dermoscopic images, for both the purposes of clinical training, and for supporting technical research toward automated algorithmic analysis. At the time of writing, the challenge is running. Two tasks are available for participation: 1) classify dermoscopic images without meta-data, and 2) classify images with additional available meta-data. We considered the first task, i.e. using only images as input since this is closer to the initial WP7.6 health use case.

The goal for ISIC 2019 is to classify dermoscopic images among nine different diagnostic categories:

- MEL: Melanoma
- NV: Melanocytic nevus
- BCC: Basal cell carcinoma
- AK: Actinic keratosis
- BKL: Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis)
- DF: Dermatofibroma
- VASC: Vascular lesion
- SCC: Squamous cell carcinoma
- UNK: None of the others / out-of-distribution (OOD)

Dataset used for the evaluation and data preprocessing

The dataset is available at: https://challenge2019.isic-archive.com/

It is composed of

- a training set of 25,331 JPEG images of skin lesions + category
- a test set of 8,238 JPEG images of skin lesions

For the experiments, we made use of the training dataset for which ground truth is directly available. From this dataset, we used 80% for training and 20% for validation. We selected the best model in terms of evaluation on the validation set.

For the evaluation, we use the same evaluation protocols (https://github.com/ImageMarkup/ isic-challenge-scoring) as the submission system. The scores are automatically calculated and reported.

Data preprocessing

The images themselves have different resolutions. We preprocess them as follows:

- For the training process, the images are randomly rescaled, rotated and cropped to generate the input to the network. Note that such preprocessing does not deform the lesions in the image. Resolution of the preprocessed images is 224x224.
- For the validation and test process, each image is firstly rescaled to 256x256 according to the shorter edge, then cropped at the center into a 224x224 image.

The “None of the others” category corresponds to a reject option and is not available in the training set. We then focused on the 8 other categories.

Deep NN model

For the classification, we used a classical ResNet architecture, pretrained on Imagenet. We replace the last classification layer by the new one adapted to the number of classes in the diagnosis. We train this layer and fine-tune the rest of the network on the medical images dataset.

The 50-layer ResNet architecture is the following:

Module Name	Output Size	50-layer ResNet
conv1	112x112	7×7, 64, stride 2
conv2_1	56x56	3×3 max pool, stride 2
		1×1, 64, stride 1 3x3, 64, stride 1 1x1, 256, stride 1
conv_2_2, conv_2_3		1×1, 64, stride 1 3x3, 64, stride 1 1x1, 256, stride 1
conv3_1	28x28	1×1, 128, stride 2 3x3, 128, stride 1 1x1, 512, stride 1
conv3_2 to conv3_4		1×1, 128, stride 1 3x3, 128, stride 1 1x1, 512, stride 1
conv4_1	14x14	1×1, 256, stride 2 3x3, 256, stride 1 1x1, 1024, stride 1
conv4_2 to conv4_6		1×1, 256, stride 1 3x3, 256, stride 1 1x1, 1024, stride 1
conv5_1	7x7	1×1, 512, stride 2 3x3, 512, stride 1 1x1, 2048, stride 1
conv5_2, conv5_3		1×1, 512, stride 1 3x3, 512, stride 1 1x1, 2048, stride 1
fc	1x1	average pooling, 9-output fully connected layer, sigmoid

The network is composed of 18 modules sequentially combined together, including 1 conv1, 3 conv2_x, 4 conv3_x, 6 conv4_x, 3 conv5_x and 1 fc. The first module conv1 is composed of a convolution layer. For conv2_x to conv5_x, each module is a residual block including 3 convolution layers in the residual branch. The output of the block is the sum of the input and the output of the convolution layers. The module fc is the newly trained prediction layer. The spatial size is reduced at the first layer of conv3_1, conv4_1 and conv5_1.

We adopt a binary cross entropy loss for each class, so that the problem is considered as 8 individual one-vs-rest binary classification problems. We reject the sample and label it as UNK if the output of every class is less than 0.5.

Evaluation criteria

The official evaluation criterion for the challenge is what they call “Normalized (or balanced) multi-class accuracy”. It is defined as the average of recall obtained in each class. The best value is 1 and the worst is 0. This metric makes all the classes equally important.

Auxiliarily, other metrics in the evaluation of ISIC 2019 are also provided, including: Threshold Metrics and Integral Metrics. These metrics consider the problem for each class as a one-vs-rest binary classification.

Threshold Metrics (with threshold at 0.5):

- recall (sensitivity),
- specificity,
- accuracy,
- F1 score,
- positive predictive value (PPV), and
- negative predictive value (NPV).

Integral Metrics:

- area under the receiver operating characteristic curve (AUC),
- AUC integrated between 80% to 100% sensitivity (AUC80), and
- average precision.

Performance

In terms of balanced multi-class accuracy (mean value of categorical recalls), the trained model achieved **0.838** on the validation set.

Detailed performance on the validation set:

Category Metrics		Mean Value	Diagnosis Categories							
			MEL	NV	BCC	AK	BKL	DF	VASC	SCC
Threshold Metrics	Recall	**0.889**	0.818	0.927	0.942	0.890	0.847	0.963	0.886	0.836
	Specificity	**0.902**	0.926	0.882	0.960	0.964	0.955	0.988	0.996	0.969

	Accuracy	**0.953**	0.906	0.906	0.958	0.961	0.944	0.987	0.995	0.966
	F1	**0.725**	0.763	0.910	0.845	0.624	0.753	0.630	0.772	0.505
	PPV	**0.631**	0.716	0.894	0.766	0.480	0.678	0.469	0.684	0.362
	NPV	**0.980**	0.957	0.918	0.991	0.995	0.982	0.999	0.999	0.996
Integral Metrics	AUC	**0.978**	0.948	0.967	0.990	0.980	0.966	0.996	0.997	0.978
	AUC80	**0.963**	0.907	0.940	0.984	0.970	0.946	0.996	0.997	0.967
	AP	**0.880**	0.844	0.967	0.941	0.802	0.848	0.910	0.917	0.810

We report also the performance on the official test set. The score is **0.488,** which is potentially impacted by the presence of out-of-distribution samples (UNK). The model was not tuned for rejection.

Detailed performance on the test set:

Category Metrics		Mean Value	Diagnosis Categories								
			MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK
Threshold Metrics	Recall	**0.584**	0.782	0.830	0.834	0.503	0.517	0.622	0.634	0.529	0.00496
	Specificity	**0.902**	0.834	0.855	0.773	0.888	0.898	0.969	0.970	0.939	0.994
	Accuracy	**0.872**	0.826	0.847	0.781	0.869	0.865	0.965	0.965	0.931	0.804
	F1	**0.383**	0.601	0.779	0.499	0.281	0.399	0.303	0.332	0.246	0.00965
	PPV	**0.317**	0.489	0.733	0.356	0.195	0.324	0.200	0.225	0.160	0.175
	NPV	**0.949**	0.950	0.913	0.969	0.971	0.951	0.995	0.995	0.989	0.807
Integral Metrics	AUC	**0.852**	0.887	0.918	0.896	0.846	0.827	0.937	0.888	0.875	0.595
	AUC80	**0.716**	0.769	0.844	0.782	0.694	0.644	0.888	0.745	0.737	0.339
	AP	**0.457**	0.684	0.852	0.573	0.232	0.379	0.410	0.471	0.283	0.226