Performance evaluation of various CNN-based architectures on automated skin-lesion diagnosis

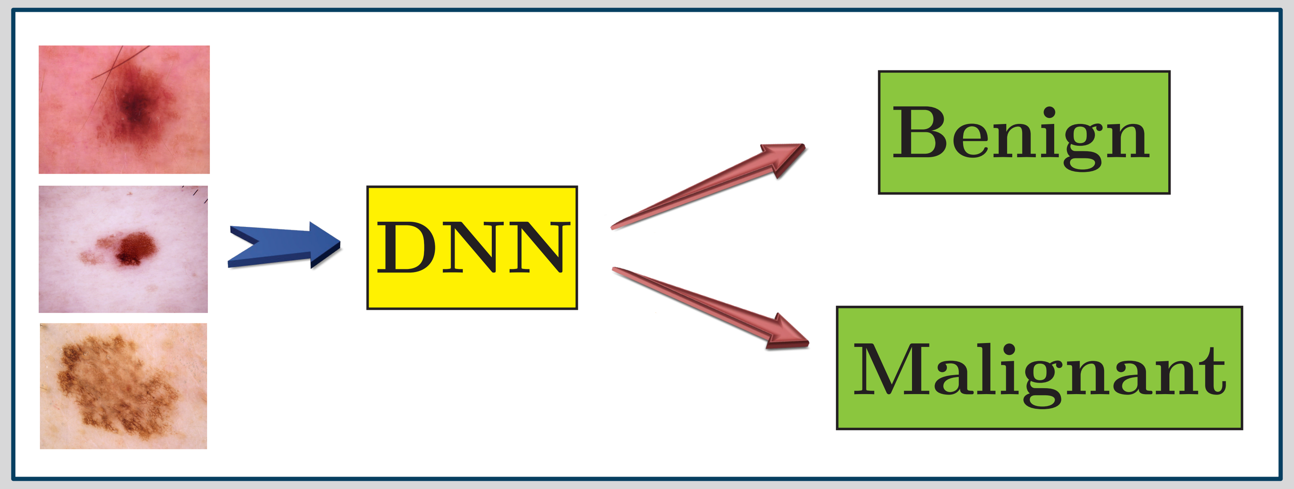
Balavignesh Vemparala Narayana Murthy

Department of Mechanical & Aerospace Engineering, The Ohio State University

Introduction

Automated Skin-lesion Diagnosis

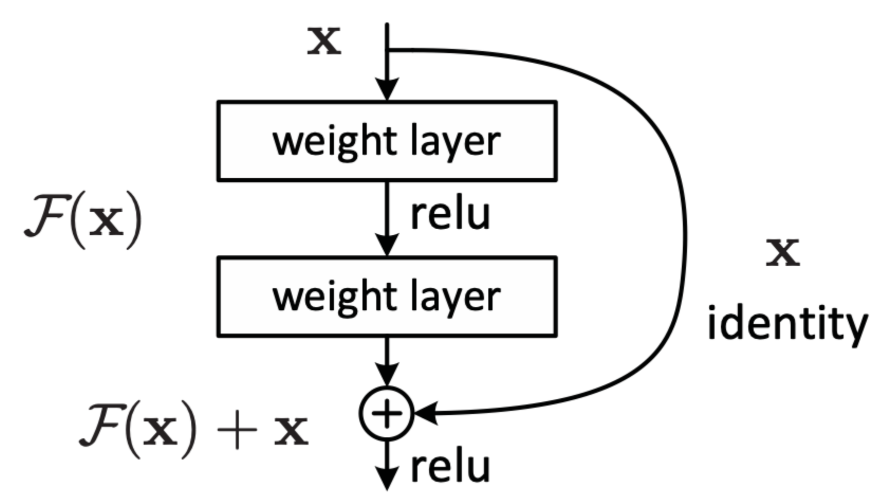
- Skin-lesion diagnosis is important as it can be used to pre-screen patients to identify the seriousness of the disease



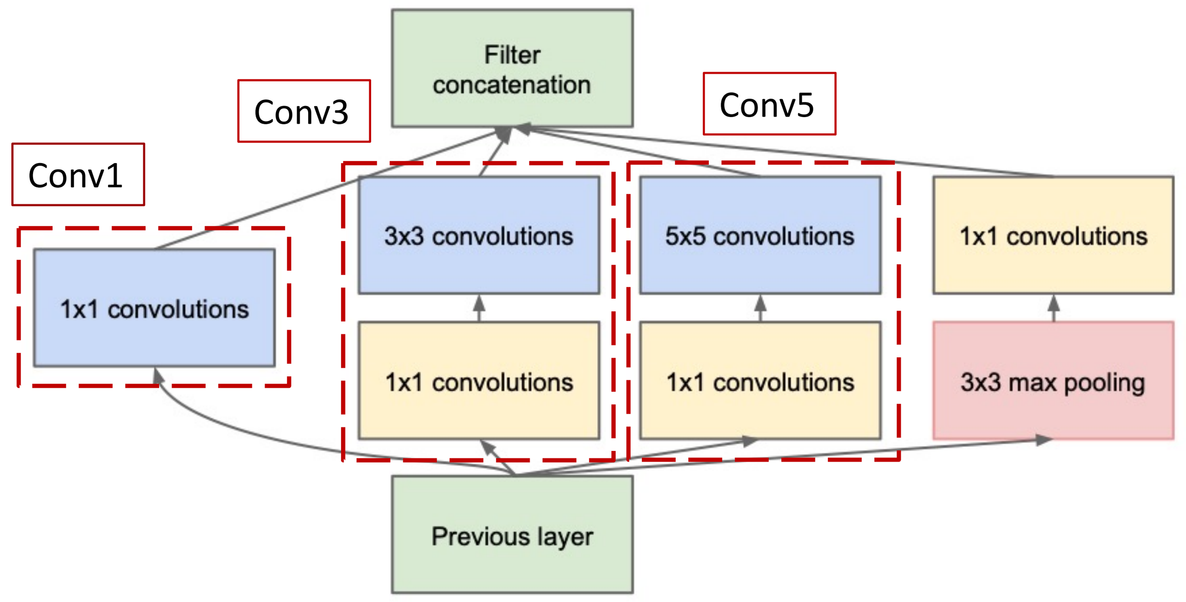
- In the current study, we compare three different architectures : plain CNN, CNN with residual blocks (res-CNN), CNN with inception blocks (incept-CNN)
- Also, we use two different datasets : HAM10000 & Diverse Dermatology Images (DDI). We train the model on HAM10000, and see how well it generalizes to DDI dataset
- Following this, we also try several oversampling methods to balance the training data, and study the effect on performance
- Moreover, we also identify the best performing model and perform an ablation study to identify which model components contribute most to the performance, which can help reduce unnecessary model complexity

Challenges & Motivation

- Lack of balanced datasets – issue with over-representation of some of the classes, which could bias the model predictions
- Lack of thorough studies on model generalization – training and testing on one dataset may not be generalizable to a large population



Residual block (Kaiming He et. al 2015)



Inception block (Szegedy et. al. 2015)

Methodology

Data Pre-processing

- HAM10000 dataset is used to train the model, with each class representing a skin disease (akiec ,bcc,bkl,df,mel,nv,vasc)
- However, DDI dataset has only two labels : Benign and Malignant
- Labels from HAM10000 were classified into Benign (bkl, df) and Malignant (akiec, bcc, mel) and other labels were neglected as (nv,vasc)
- All images were down-sampled to a lower resolution 100×100 to ensure uniformity between datasets as well as avoid memory issues during model training

Oversampling Methods

- Resulting dataset from pre-processing was unbalanced, and hence there was a need to perform some data augmentation
- Traditional data augmentation methods such as flip and rotate transformations, do not introduce any new information to the already existing dataset
- Thus, we used three different methods – random, SMOTE, ADASYN to oversample the datasets

Bayesian Optimization for Hyper-Parameter Optimization

- Moreover, Bayesian Optimization was used for hyper-parameter optimization
- For instance, in case of plain CNN, hyperparameters include number of Convolution filters & Convolutional layers, number of dense layers & dense units, dropout rate, batch size, learning rate
- In res-CNN and incept-CNN architectures on the other hand, we had other hyper-parameters such as number of residual layers, number of inception blocks, etc.

Model training

- All models were trained on a single-CPU

Experiments

Overall Performance

- Incept-CNN model, best the most-heavy, also shows the best performance

Model	Number of parameters	Training time
Best CNN	785,793	0.71hrs
Best res-CNN	3,661,377	1.74hrs
Best incept-CNN	31,376,673	9.6 hrs

Model	HAM10000 test accuracy (%)	DDI dataset accuracy (%)
Best CNN	70.82%	40.85%
Best res-CNN	67.98%	40.55%
Best incept-CNN	71.61%	57.77%

Comparison of Oversampling methods

- Oversampling methods, in general improve model performance and helps generalize them better
- SMOTE oversampling method is the most consistent across all the models

Oversampling Method	Number of Ones	Total samples	Percentage (%)
No oversampling	1564	2534	60.54%
Random	1564	3128	49.04%
SMOTE	1564	3128	49.04%
ADASYN	1564	3104	49.42%

Model	HAM10000 test accuracy (%)	DDI dataset accuracy (%)
Best incept-CNN	71.61%	57.77%
Best incept-CNN + Random Oversampling	74.29%	53.51%
Best incept-CNN + SMOTE Oversampling	74.60%	58.69%
Best incept-CNN + ADASYN Oversampling	73.97%	61.89%

Incept-CNN Ablation study

- Not all convolutions contribute equally
- Conv1 seems to contribute to the performance most, and also has the least training time

Model	HAM10000 test accuracy (%)	DDI dataset accuracy (%)
Best incept-CNN	71.61%	57.77%
Best incept-CNN, no Conv1	69.40%	44.36%
Best incept-CNN, no Conv3	75.08%	49.70%
Best incept-CNN, no Conv5	72.24%	53.81%

Model	Number of parameters	Training time
Best incept-CNN	31,376,673	9.6 hrs
Best incept-CNN, no Conv1	23,615,601	8 hrs
Best incept-CNN, no Conv3	23,490,897	6.86 hrs
Best incept-CNN, no Conv5	23,269,713	6.09 hrs

Conclusions

- Incept-CNN architecture shows best performance for automated skin-lesion diagnosis
- Oversampling can help balance datasets, which may lead to better model performance and generalization
- Future study involves validating the hypotheses on more datasets in the field