[**Evaluating Deep Neural Networks Trained on Skin Images with the Fitzpatrick 17k Dataset (Full)**](https://www.youtube.com/watch?v=bizJpy5VQmQ) **– Notes**

**Paper that the video was based on:** [**Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset**](https://arxiv.org/pdf/2104.09957)

Motivation

There is a clear lack of imaging data for those of darker skin in the context of skin disease, which means that NNs trained for the purpose of assisting with skin disease diagnostics are likely to be biased. Investigating causes for bias is therefore incredibly important, and yet (in the USA) there is only one publicly available imaging dataset with skin type labels, that being PAD-UFES-20 with 579/1373 patients labelled with Fitzpatrick skin type.

This is why the Fitzpatrick 17k dataset was created – 16 577 clinical images (12 672 from DermaAmin and 3 905 from Atlas Dermatalogico) labelled Fitzpatrick skin types, spanning 114 skin conditions with at least 53 images per condition (with a maximum of 653). 22 skin condition images were excluded from this final selection if the condition was either too broad symptomatically, the images for it were of poor quality or the resulting skin conditions is of an extremely rare hereditary disease presenting with skin afflictions (genodermatosis).

A board-certified dermatologist analysed 504 images (~3% of the data) and found:

|  |  |  |
| --- | --- | --- |
| 69.0% | 348 | Diagnostically relevant |
| 19.2% | 97 | Potentially diagnostic |
| 6.3% | 32 | Characteristic |
| 3.4% | 17 | Wrong |
| 2.0% | 10 | Other |

A second board-certified dermatologist confirmed the error rate, which is consistent amongst other commonly used test datasets for CV, natural language processing and audio processing.

The dataset was annotated by a team at Scale AI.

A screenshot of a graph

AI-generated content may be incorrect.

Annotators are generally good at matching board-certified skin-type labels (, within of trials). Fitzpatrick labelling is subjective, dependent on lighting conditions and level of image details.

A screenshot of a data

AI-generated content may be incorrect.

The dataset has at least one image for all 114 skin conditions it contains for Fitzpatrick-I to III. It is missing some for IV and V, and misses 25 for VI – which shows that more work needs to be done for collection of skin condition image data for darker skin tones.

The VGG16-type CNN Trained to Classify Images from the dataset

A computer screen with text

AI-generated content may be incorrect.

Code for training a VGG16 DNN/CNN model that was used for classifying images from the dataset. Below is the PyTorch implementation for VGG16 by default (but only the classifier layers):

A screen shot of a computer program

AI-generated content may be incorrect.

The code was adapted from its original state – the final dense layer is usually 1000 units, but the sixth level of the classifier was changed:

To

Where for each of the skin disease labels, activated by logarithmic softmax for multi-label classification.

What is VGG16?

VGG16 is an architecture coined by Oxford’s Visual Geomtry Group (hence VGG), containing 16 layers comprised of 13 convolutional-pooling layers and 3 fully connected layers. It is simple compared to more novel architectures, but is still very effective ([Source](https://www.geeksforgeeks.org/computer-vision/vgg-16-cnn-model/)).

Drop-out layers ([Source](https://medium.com/@vishnuam/dropout-in-convolutional-neural-networks-cnn-422a4a17da41))

is a dropout layer with a 40% chance to “drop-out”. Dropout is a regularisation technique to prevent overfitting. It works by removing a certain fraction of neurons in the network at random, which prevents the network from relying on a subset of all neurons available to it.

It is usually applied after dense (fully connected) layers but can be used after convolutional ones. Convolutional layers have regularisation built-in through shared weights and pooling, whereas dense layers are more prone to overfitting from large number of parameters (number of layers, neurons per layer, etc).

This can be seen as a form on ensemble learning – a combination of many uncorrelated weak learners to build one strong predictor/classifier, which reduces variance and/or bias. Ensemble learning improves overall accuracy by averaging noise and uncorrelated errors (variance). Since a proportion of neurons are “dropped out” on each forward pass, this means that every forward pass we are training a different classifier network, which in turn is trained on a different mini-batch of the training data. Since all neurons are part of the larger neural network, this makes dropout similar to homogeneous parallel ensemble learning – each variation of the dropped out network is trained independently per forward pass, and then aggregated by average during testing. This reduces variance, as it combines each of the predictions learned from the different classifiers, which averages away the impact of isolated errors (Source – my notes).

In tandem with using dropout – other types of regularisation such as L2, batch normalisation (demeaning each mini-batch, then scaling/shifting) and early stopping may be used. It may also be necessary to lower the learning rate and increase batch size.

Random image transformations

A computer screen with text and numbers

AI-generated content may be incorrect.

A series of random transformations are applied to images before training to improve training, in order:

* Conversion to Python Imaging Library image object (compatibility)
* Random crop over of the image area, before resizing to
* Random rotation up to
* Random brightness, contrast, saturation and hue shift
* Random left-to-right image flip, chance
* Crop the centre region from the now image to match ImageNet standards. This is because VGG16 architecture by default assumes images are of this dimension
* Converts the now PIL image to a PyTorch tensor, giving it and normalised pixel values in , since any PyTorch model requires a tensor as input
* Normalisation of the image by the top row of means and bottom row of standard deviations, per channel i.e.

A screenshot of a computer code

AI-generated content may be incorrect.

This is in accordance with ImageNet means and standard deviations. Note that Normalize cannot be used on PIL objects, tensors only

Model Evaluation

A table with numbers and a number of objects

AI-generated content may be incorrect.

The model was evaluated using seven different holdouts, with remainders being lumped as the training dataset:

* Verified: Images identified by board-certified dermatologists as diagnostically relevant to the labelled condition were set aside as the test dataset
* Random: Approximately 75:25 random train-test split, stratified by skin condition
* Source A: Images from Atlas Dermatologico are used for testing
* Source B: Images from DermaAmin are used for testing
* Fitzpatrick Holdouts: Specific Fitzpatrick labelled images are used for testing

The model tends to be more accurate on images with closest Fitzpatrick types to the training images e.g. model trained on I and II performed better on III and IV than V and VI.

ITA vs Fitz

Individual typology angle (ITA) is a different annotation for skin type labelling. It is calculated using statistical features of pixels in the image, ideally over non-diseased skin. Instead of obtaining segmentation masks, the paper documents the use of [a YCbCr algorithm that produces skin masks](https://arxiv.org/pdf/1708.02694). Having obtained a skin mask, the ITA is calculated as:

Where L-bar and B-bar are the mean luminance and yellow of non-masked pixels of values within one standard deviation of the mean. The use of ITA is mainly as a method of objective, quantitative and automatic skin labelling.

A group of math equations

AI-generated content may be incorrect.

Two different methods for calculating Fitzpatrick label given ITA are given above – equation 14 comes from thresholds described by [this paper by Kinyanjui et al.](https://arxiv.org/pdf/1910.13268) whilst equation 15 is based on the empirical distribution of ITA scores over Fitzpatrick labels as given below. Although some correspondence can be seen, that being a downward trend per label, a large spread of overlapping ITA values are given for each Fitzpatrick label.

A group of different colored objects

AI-generated content may be incorrect.

The relevant value thresholds were selected with the idea of minimising overall error whilst maintaining cohesion between ranges.

A table with numbers and a few words

AI-generated content may be incorrect.

Above is concordance of ITA with Fitzpatrick, using both full image and YCbCr mask. Discrepancies appear to be driven more by the higher fidelity and thereby variance in the ITA algorithm.

A close-up of several skin diseases

AI-generated content may be incorrect.