How Much Data is Enough? Benchmarking Transfer Learning for Few Shot ECG Image Classification

Sathvik Bhaskarpandit

Scarce Data for Artificial Intelligence in Healthcare Workshop @ IJCAI 2022

Introduction and Motivation

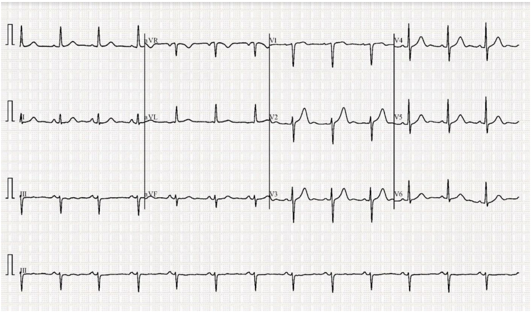
- Over the past couple of decades, numerous research works have been conducted to study and detect abnormalities from ECG signals.
- Although many of these deep learning approaches utilize ECG signals as input, only a handful use *images of patients' ECGs* themselves, that are often stored in hospitals and diagnostic centres.
- Almost no previous works exist that study ECG images under the few-shot learning scenario, i.e. the scenario in which a classifier is made to generalize to unseen data given only a handful of labelled training examples.
- This work aims to study the effectiveness of transfer learning for few-shot ECG image classification, and how classification performance varies with the amount of training data available.

Experimental Details - Dataset and Preprocessing

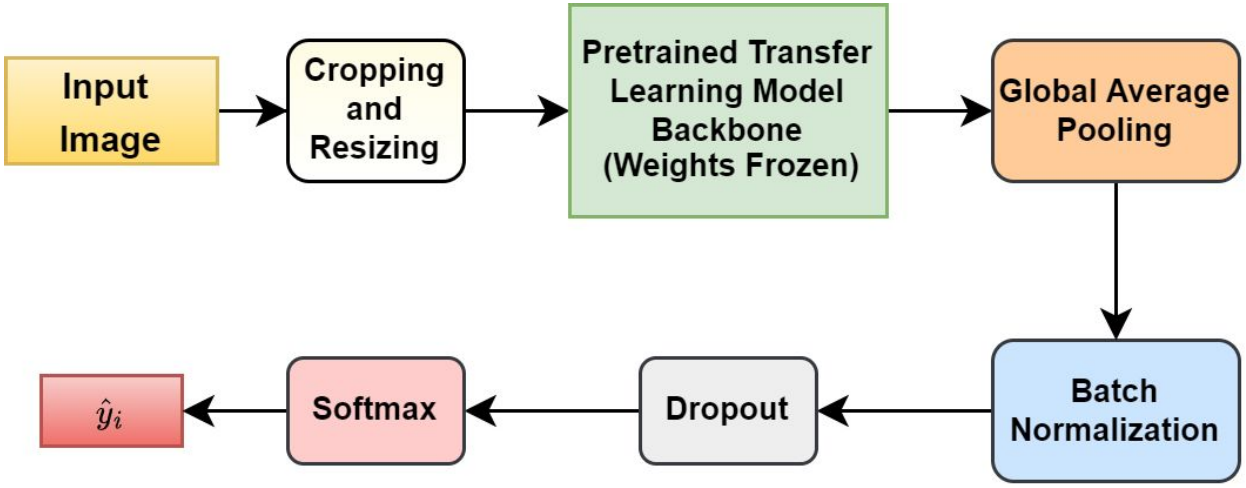
- The data used for the study is taken from the ECG Images dataset of Cardiac and COVID-19 Patients (Khan et al., 2021), which consists of ECG images collected from different health care institutes across Pakistan.
- Each of the images belongs to one of three classes: normal patient, patient with abnormal heartbeat, or patient with myocardial infarction, with each class having approximately 250 images.
- Each image contains metadata such as patient name, ID, height, weight, time of recording, etc. Such details are cropped out of the images to retain only a grid with the snapshots of the ECG signal recordings. The images are then resized to a size of (128, 128, 3) using image anti-aliasing.

(a) Before cropping

(a) After cropping



Transfer Learning Model Structure



The base transfer learning (TL) model backbones include VGG16, DenseNet, InceptionV3, ResNet and EfficientNet Models B0 to B7

Experimental Details - Sampling of Tasks

- *k* images randomly sampled per class for the support set.
- *q* images randomly sampled per class for the query set.
- Each TL model first fine-tuned using the support set and evaluated using the query set.
- Adam optimizer, batch size of 16 for training. Early stopping of loss function is used as the criterion for stopping training. Value of q set to 100.
- Value of k varied from 1 to 50. For each value of k, task repeated 20 times (with different initial random seed) and average metrics are reported.

Results

Table 1: Accuracy and F1 Scores of the various transfer learning models with different pretrained base models. Values are in percentage. Underlined values represent the model with the best accuracy or F1 score for the given value of k. The table headings B0 to B7 represent EfficientNet model variants.

k	Metric	VGG16	DenseNet	InceptionV3	ResNet	B0	B1	B2	B3	B4	B5	B6	B7
1	Accuracy	39.67	41.33	47.33	50.07	64.80	58.60	59.10	52.13	40.07	51.87	41.73	45.73
	F1	29.01	35.53	32.95	46.92	61.01	53.39	53.74	42.92	30.48	47.07	32.78	41.69
2	Accuracy	65.03	61.33	47.09	60.33	79.93	79.47	80.63	57.33	57.87	65.21	44.53	41.40
	F1	63.96	61.24	36.52	58.89	79.40	79.11	79.97	49.85	51.59	64.29	34.38	40.24
3	Accuracy	55.67	61.03	53.67	72.93	92.82	87.43	86.21	73.07	73.47	68.80	51.33	60.73
	F1	51.67	56.08	51.41	72.71	92.79	87.4	86.04	70.66	72.17	67.82	45.21	62.88
4	Accuracy	53.01	77.33	56.50	83.4	93.98	92.80	93.10	83.07	82.85	78.67	75.73	78.25
	F1	46.27	75.86	51.16	83.2	93.97	92.78	93.05	82.83	82.21	78.45	75.55	80.32
5	Accuracy	71.33	86.33	64.33	86.00	96.07	92.67	94.73	84.67	84.80	83.00	78.40	82.40
	F1	70.65	86.33	57.09	85.94	96.06	92.65	94.71	84.53	84.88	82.99	78.14	84.30
10	Accuracy	67.33	87.83	81.00	90.73	97.15	97.00	95.60	92.07	90.80	91.53	88.80	90.13
	F1	63.82	87.75	80.88	90.74	97.16	96.98	95.58	92.03	90.78	91.59	88.81	90.36
20	Accuracy	85.33	92.67	88.33	95.53	98.40	98.63	99.20	96.93	96.00	96.27	94.87	95.41
	F1	85.38	92.68	88.11	95.54	98.39	98.63	99.22	96.91	95.98	96.24	94.85	95.45
30	Accuracy	86.33	94.33	89.67	97.67	99.03	99.01	99.37	98.93	96.87	97.87	95.62	96.86
	F1	86.29	94.33	89.65	97.65	99.01	99.02	99.35	98.93	96.85	97.86	95.59	96.90
40	Accuracy	86.67	96.07	93.17	98.93	99.08	99.43	99.57	98.80	98.33	98.13	96.73	97.66
	F1	86.55	95.99	93.16	98.91	99.10	99.37	99.57	98.74	98.33	98.19	96.72	97.69
50	Accuracy	88.67	97.33	93.33	99.61	99.28	99.57	99.67	99.61	98.13	98.73	96.87	98.53
	F1	88.53	97.32	93.26	99.60	99.28	99.53	99.66	99.60	98.12	98.71	96.86	98.59

Results (cont.)

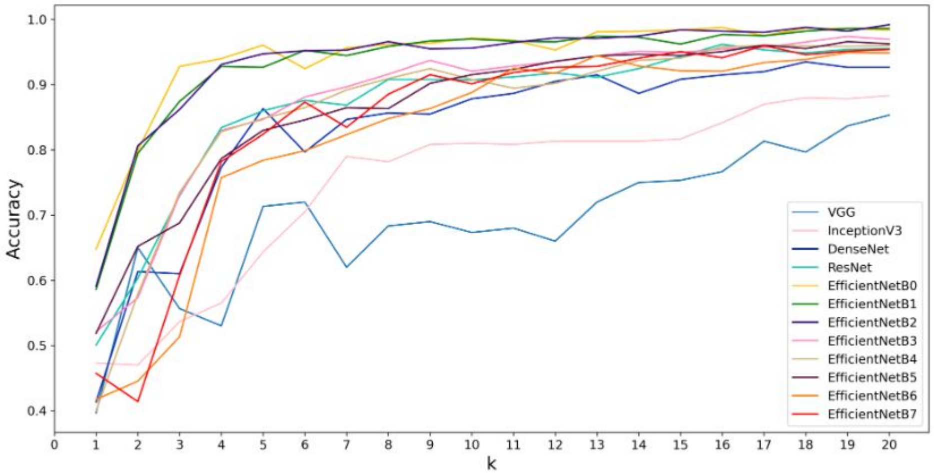


Figure 1: Plot of accuracy as a function of k for k=1 to k=20

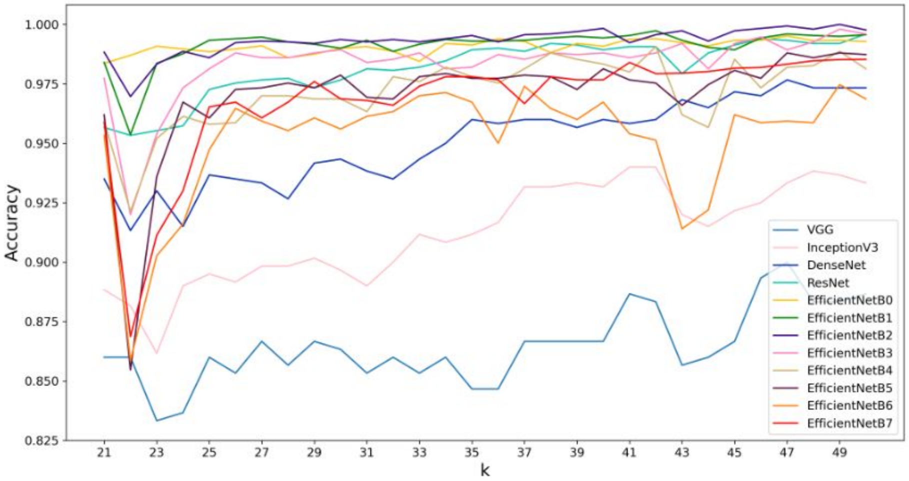


Figure 2: Plot of accuracy as a function of k for k=21 to k=50

Observations and Discussion

- Medical diagnoses carry a high degree of responsibility and medical research strives to make such diagnoses impeccable.
- How much data is required to obtain a certain threshold accuracy δ?
  - For δ = 99%, k_{min} = 24 using EfficientNetB2
  - For δ = 99.5%, k_{min} = 40 using EfficientNetB2
- 40 examples per class can be retrieved, but if disease being observed is rarer, it may be even more difficult to collect such images.

Conclusion and Future Work

- Transfer learning using popular image classification architectures is a promising direction for few-shot ECG image classification.
- With around 20 images per class available for training, models such as ResNet and EfficientNet are able to achieve accuracies of at least 99%.
- When the training set comprises 5 images per class or fewer, simple transfer learning fails to classify ECG images with high accuracy.
- Future direction:
    - Comparison with other FSL approaches.
    - Better transfer learning and fine-tuning methods.
    - Techniques to deal with class-imbalance and very low-shot learning.
    - Experimentation with data from different ethnicities and regions.

Thank you!