Thesis 2019

Australia's Global University | Faculty of Engineering | School of Electrical Engineering and Telecommunications

Qualitative Assessment Metrics For Transfer Learning

Author: Joel Smith

Supervisor: Dr. Beena Ahmed

Abstract

- The 'black-box' nature of transfer learning makes it difficult to assess the performance of applications beyond a few standard, high-level metrics (i.e. accuracy).
- This limits the ability to improve the system without a more qualitative, finer-grade perspective.
- There is a need to develop qualitative assessment metrics to understand the performance of transfer learning applications.
- This would provide further insight into potential errors within the application and areas of improvement, beyond what is perceivable by higher-level metrics.

Aims

- To identify qualitative metrics that can be used to successfully evaluate the performance of transfer learning applications at a finer-grade level than accuracy.
- To show how these metrics can be used to develop insight into improving the application and explain higher-level metrics, such as accuracy.

Background

Deep Learning (DL)
- Machine learning (ML) using neural network frameworks with many hidden layers, producing algorithms to model high level abstractions

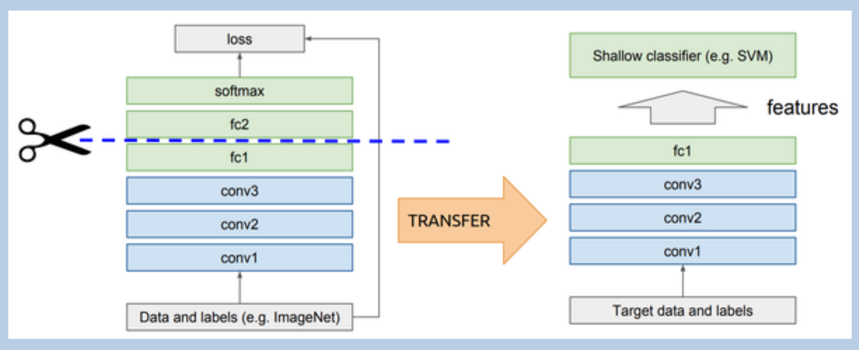

Figure 1: Transfer learning using a pre-trained feature extractor

Transfer Learning (TL)
- ML technique that takes what is learned in one setting and exploits to improve generalization in another.

Typical performance metrics
- Most performance metrics are high-level
- Most common: Accuracy.
- Common: ROC and loss curves
- Sometimes: F1-score, precision and prevalence

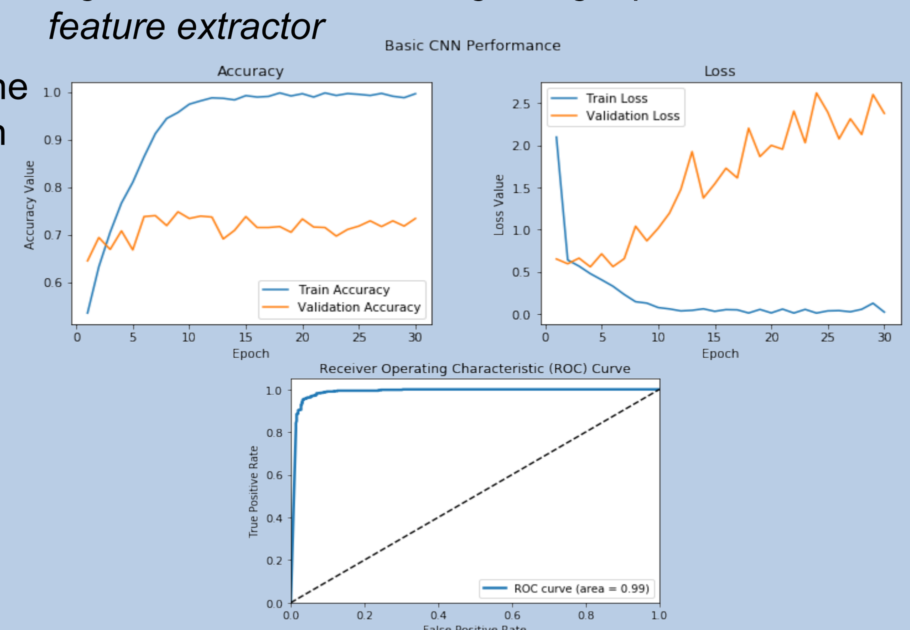

Figure 2: Accuracy, loss, and ROC curve examples

Experimental Setup

Dataset
- Binary image classification is explored using a subset of the famous *Dogs vs. Cat* dataset,
- Training: 3000, validation: 1000, testing: 1000

Pre-trained feature Extractor
- VGG-16 model is a state-of-the-art 16-layer CNN and FC network trained on ImageNet database, built for large-scale image classification (*see figure 3).*

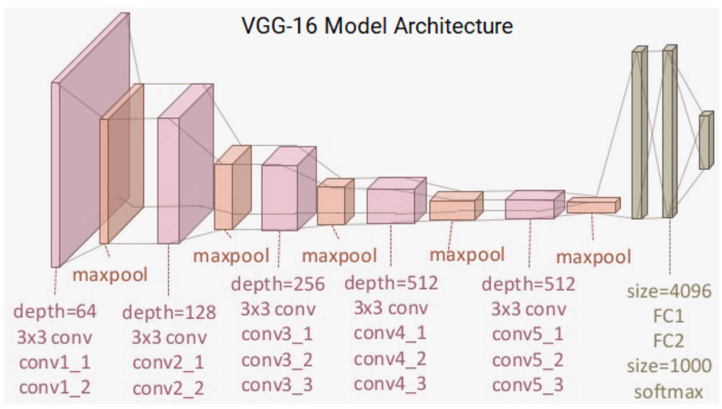

Figure 3: VGG-16 Model Architecture

Models

Five models were created using a Keras framework:
1. Simple CNN with regularization
2. CNN with regularization and image augmentation
3. TL using a pre-trained feature extractor with frozen layers
4. (3) with image augmentation
5. (4) but use fine-tuning instead of freezing layers

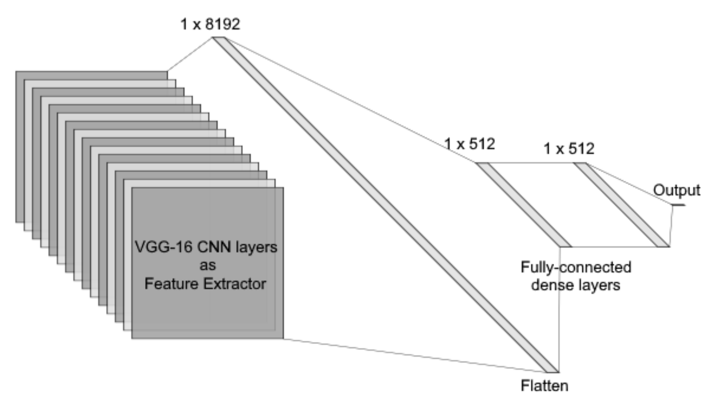

Figure 4: Model (3), (4), (5) architecture using VGG-16 as a feature extractor

High-level results

Model	(1)	(2)	(3)	(4)	(5)
Accuracy	76.3	85.4	89.2	89.6	94.9

Accuracy increases across models

Table 1: Accuracy (%) of each model against 1000 test images.

Results

Three major metrics were found via analysing at activation level

1. Neural Network Utilization (NNU)

- A non-insignificant number of **deactivated neurons, ϑ,** existed for all models.
- i.e. for all 1000 test images, specific neurons (or feature maps for CNN layers) **produced zero.**
- NNU can be deduced as the percentage of deactivated neurons within a network, indicating total network utilization:

$$NNU\ (\%) = \frac{\vartheta}{\textit{total number of activations}}$$

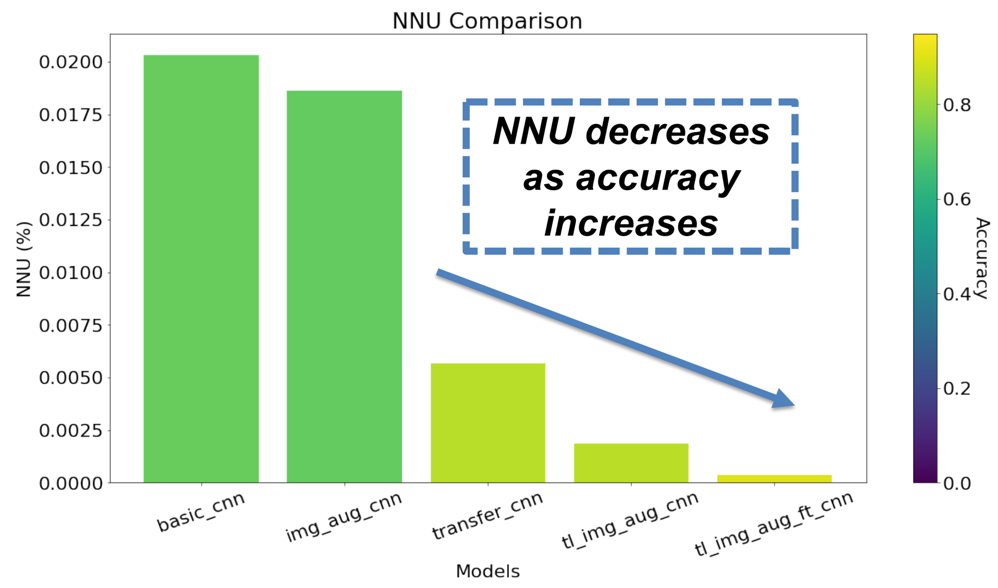

Figure 5: NNU comparison of models 1-5 showing the inversely proportional relationship of NNU to accuracy

2. Activation Spectrum (AS)

- Plotting activations as a spectrum across entire model provide qualitative insight. (see figure 6)
- Spectrums revealed **right-skewed** distributions with minimal deactivated neurons produced the superior results.

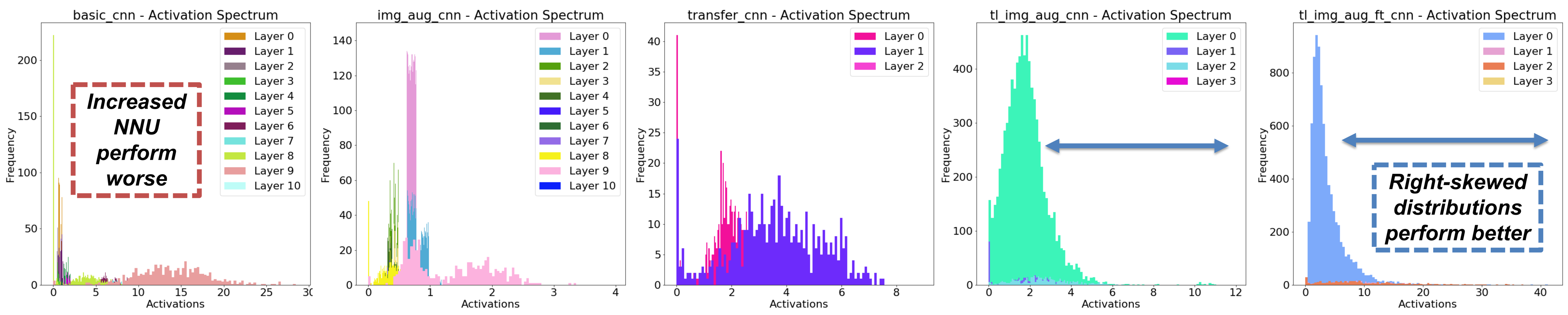

Figure 6: Models 1-5 (left to right) AS showing right-skewed distribution with minimal deactivated neurons indicates highest performance.

3. Activation Range (AR)

- Figure 7 shows max activation for each individual layer within a model.
- More consistent max-activation across layers seem to suggest better performance.

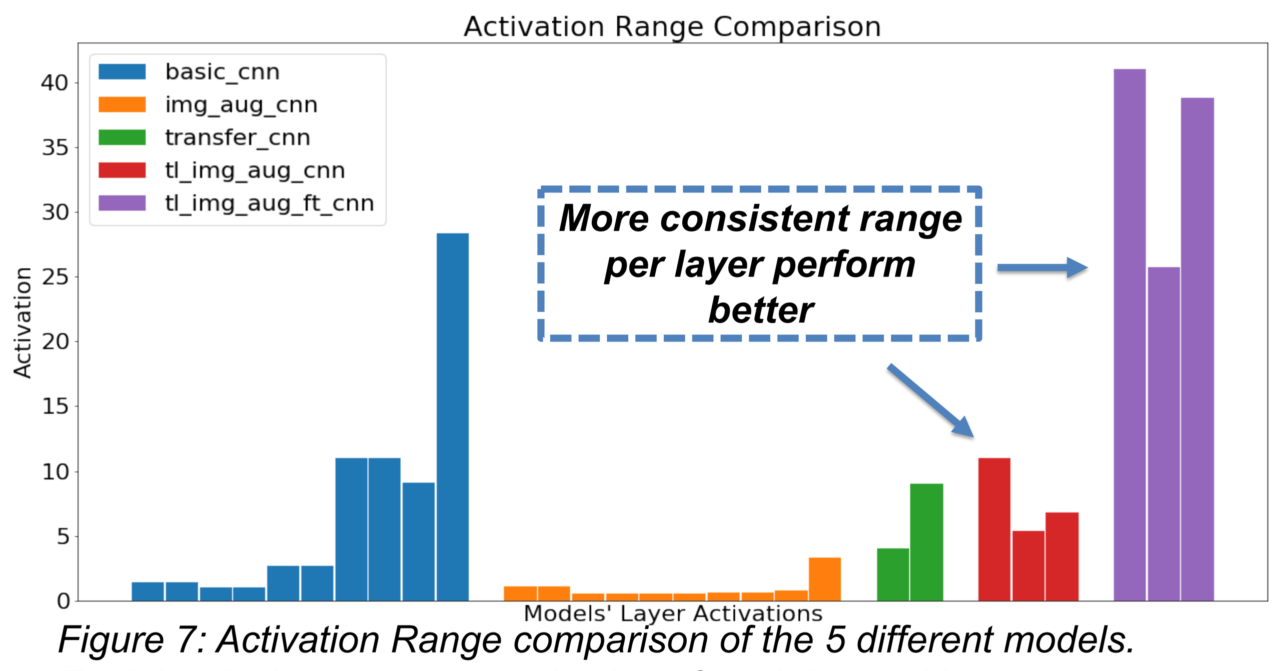

Figure 7: Activation Range comparison of the 5 different models. Each bar is the maximum-activation of each internal layer.

Conclusions

NNU, Activation Spectrum and Activation Range are three metrics that qualitatively assess TL applications

- From figure 5, $NNU \propto \frac{1}{Accuracy}$
- **Decreasing NNU** from DL application to the TL application is vital.
- Improve NNU: Reassess regularization techniques (i.e. dropout) and employ data augmentation
- From figure 6, **right-skewed distributions** while minimising NNU ⇒ successful TL
- Improve AS: minimise NNU, re-evaluate activation function and weight distribution.
- From figure 7, **consistency** from max activations per layer may be indicative of performance.
- Improve **AR:** similar to AS improvement

Future Research

- Proposed metrics (particularly NNU) could qualitatively assess the performance of TL in **any domain**, such as audio, NLP or computer vision.
- A starting point for **unpacking 'black-box'** nature of TL.
- A potential area to investigate other metrics is **exposing specific features** which trigger significant activation within networks (see figure 8).
- Investigation of **degrees of weight shift** within TL applications would also provide potential metrics

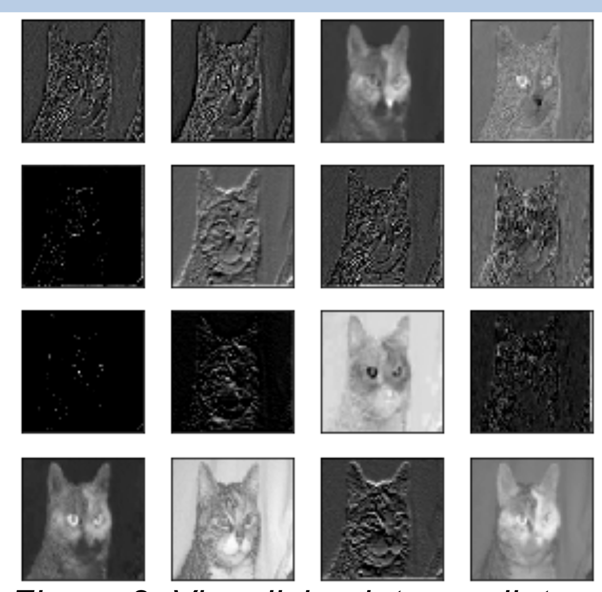

Figure 8: Visualizing intermediate activations within CNN