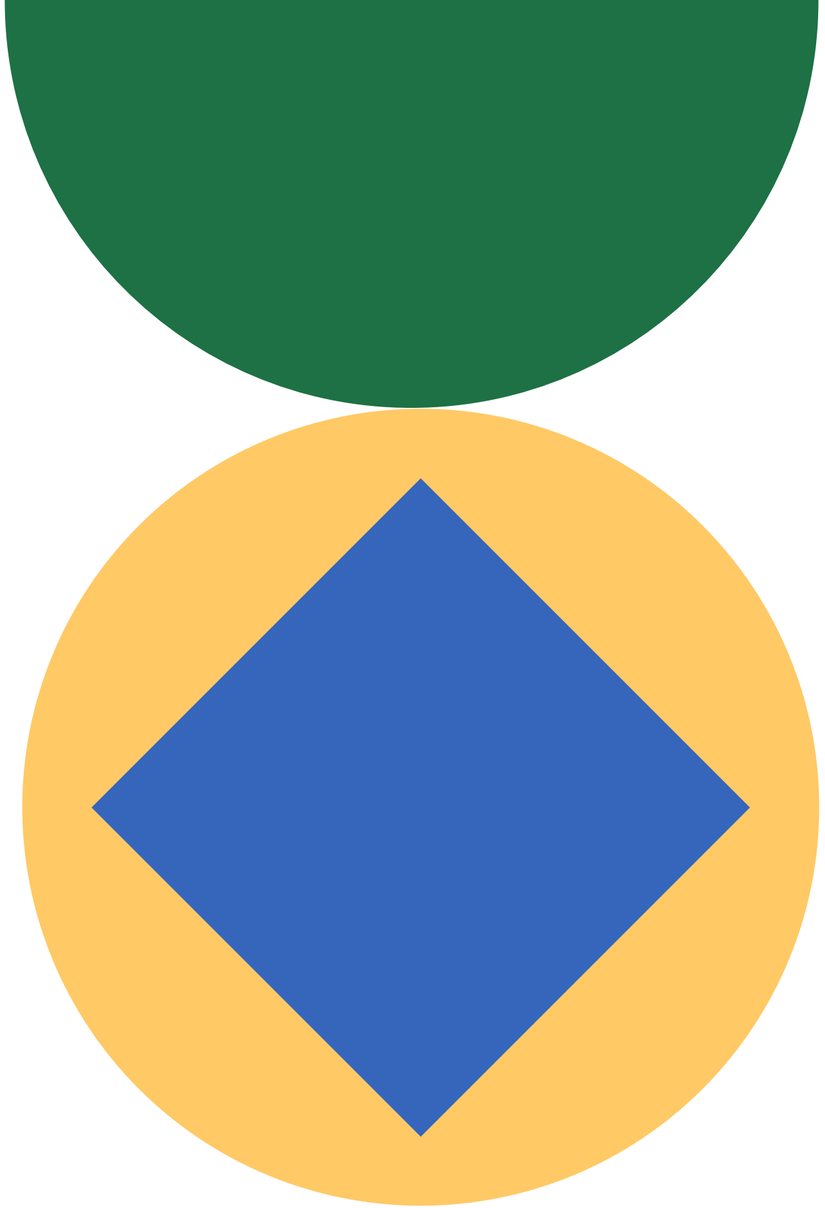
SQUEEZENET: ALEXNET-LEVEL ACCURACY WITH 50X FEWER PARAMETERS AND <0.5MB MODEL SIZE

By
Anup Joseph



Background

A large CNN model has fundamental limitations in deployment.
To solve this problem a variety of heavily engineered models with small sizes have emerged.

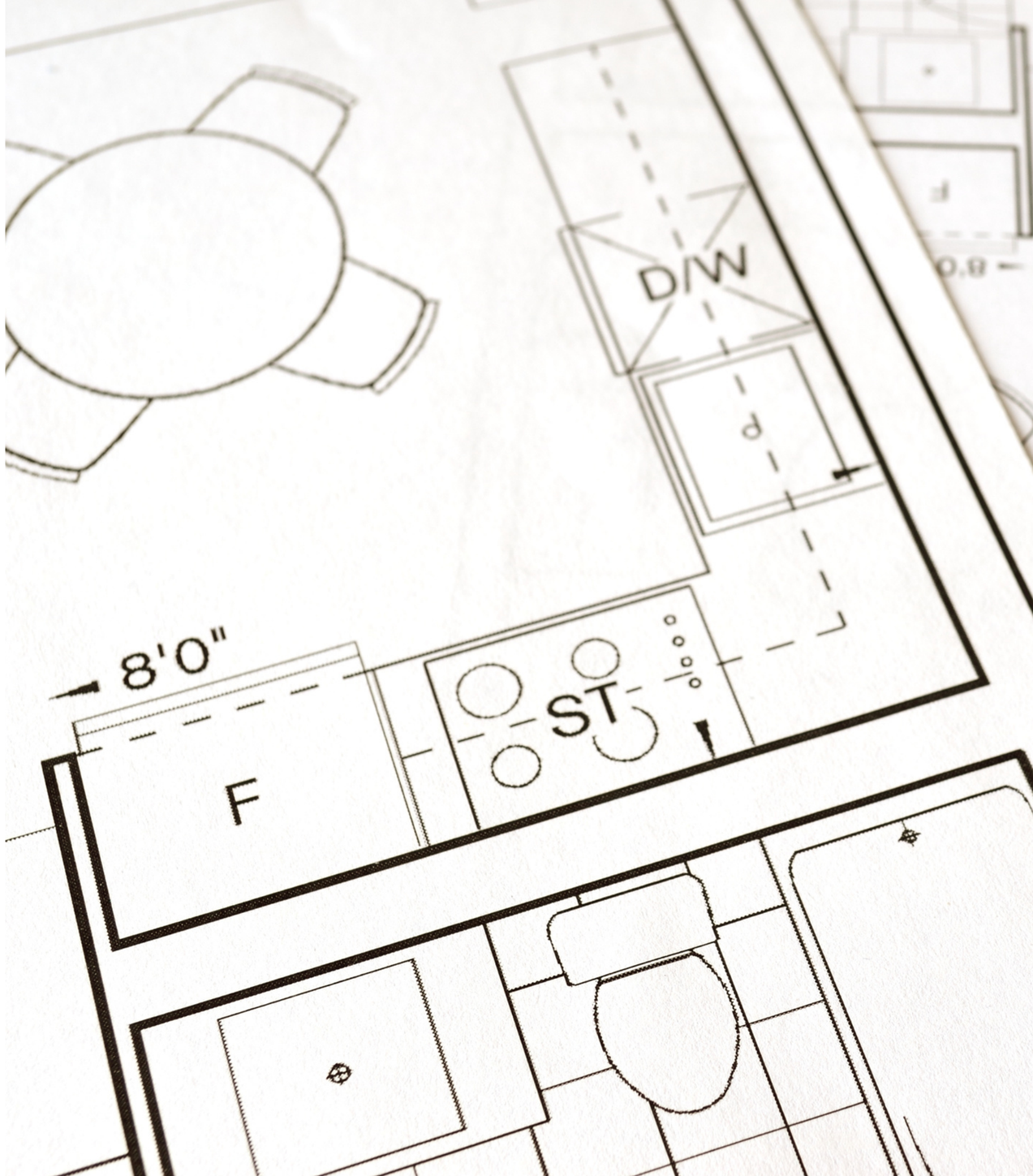
Advantages of smaller CNNs

- MORE EFFICIENT DISTRIBUTED TRAINING
- LESS OVERHEAD WHEN EXPORTING NEW MODELS TO CLIENTS
- FEASIBLE FPGA AND EMBEDDED DEPLOYMENT

August 2021

Architectural Design Strategies

PAPER PRESENTATION



Replace 3x3 filters with 1x1 filters

1x1 filters have 9x fewer parameter than 3x3 filters

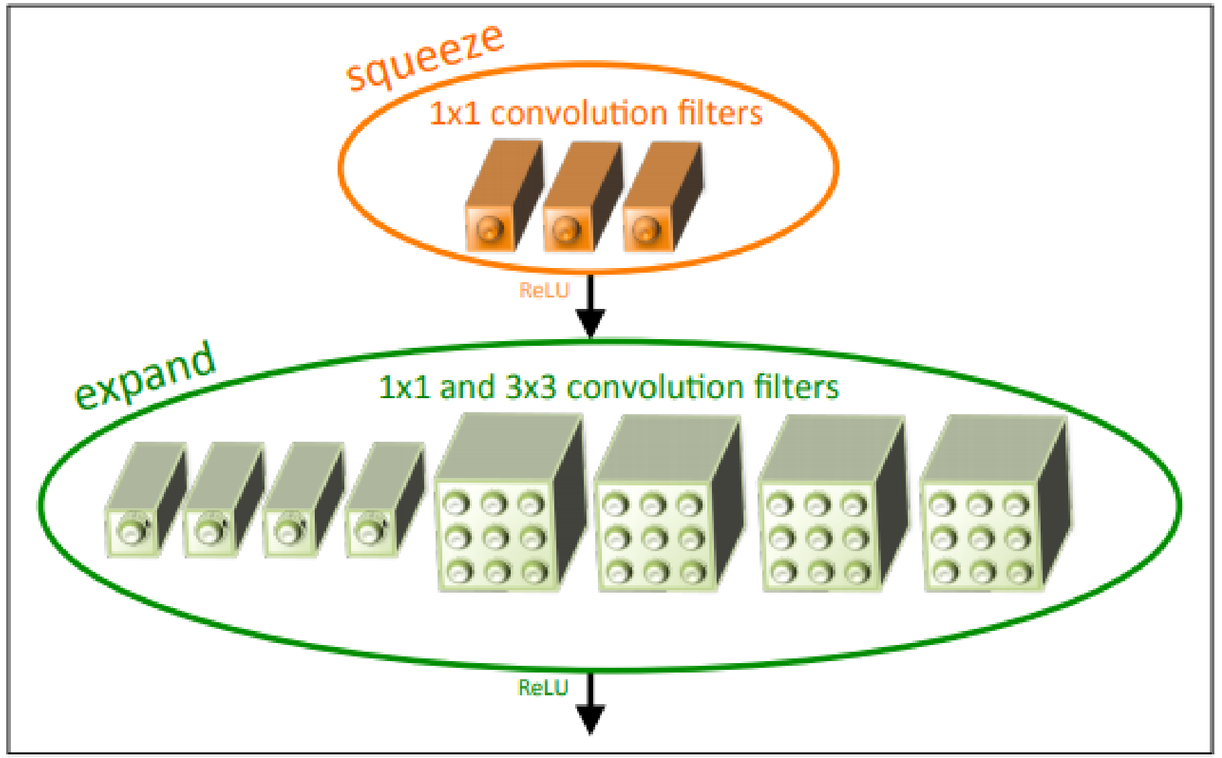
Decrease the number of input channels to 3x3 filters

To decrease the total number of parameters its important to decrease the number of input channels to the 3x3 filters

Downsample late in the network so that convolution layers have large activation maps.

August 2021

The Fire module



Fire module consists of two sections

- squeeze layers
- expand layers

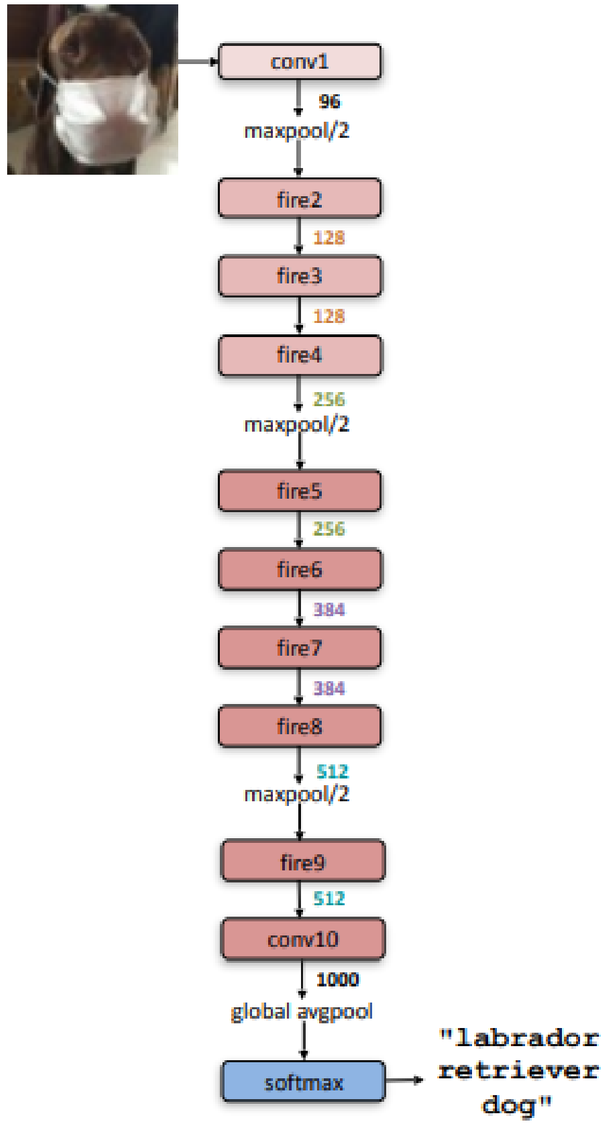
The Fire module has three tunable hperparameters

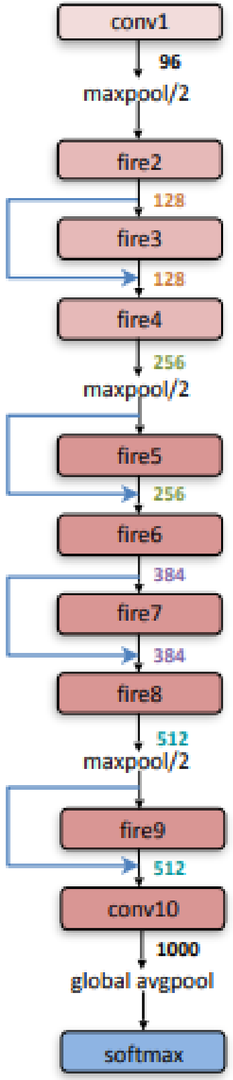
- no. of filters in the squeeze layer(all 1x1)
- no. of 1x1 filters in the expand layer
- no. of 3x3 filters in the expand layer

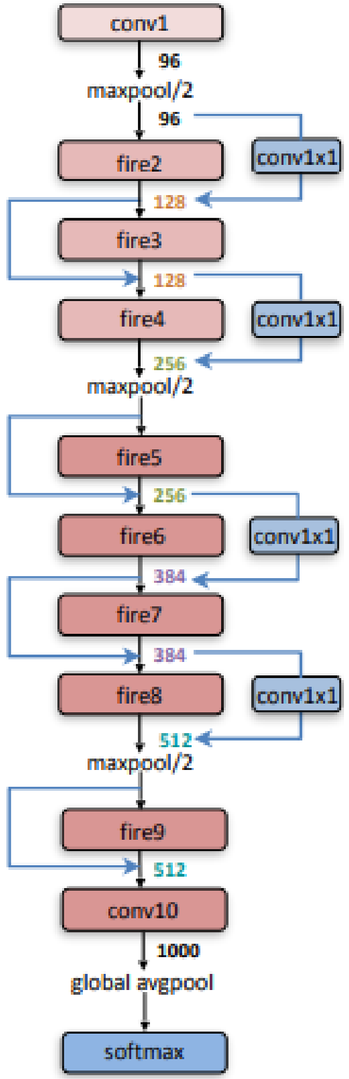
Configurations

Training Settings

- Learning Rate - 0.4 and then linearly decrease throughout the network.
- ReLU is applied to activations from squeeze and expand layers.
- The number of filters per fire module are gradually increased to the end of the network
- The arcitecture was implemeted using the Caffe2 framework but ports foor most popular frameworks exists







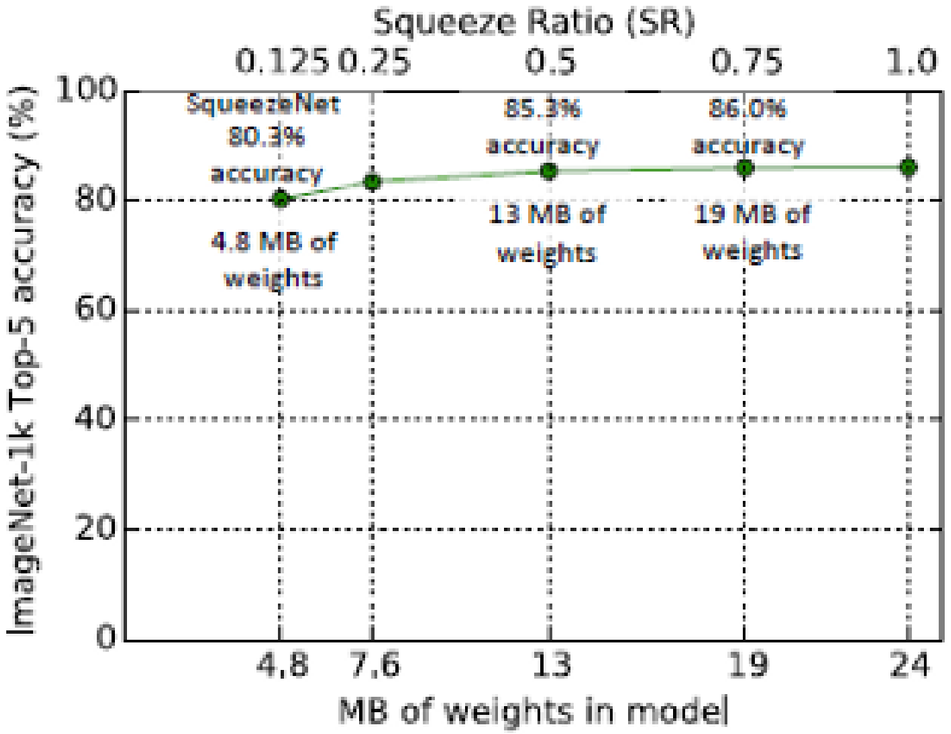
Model Architecture

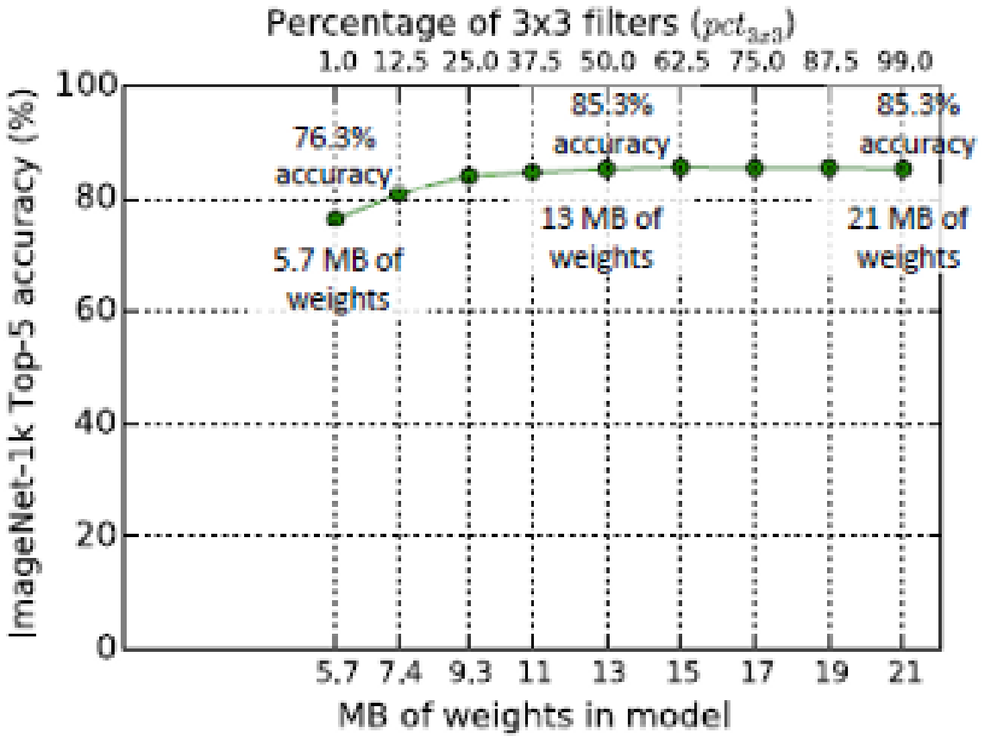
Results

CNN architecture	Compression Approach	Data Type	Original → Compressed Model Size	Reduction in Model Size vs. AlexNet	Top-1 ImageNet Accuracy	Top-5 ImageNet Accuracy
AlexNet	None (baseline)	32 bit	240MB	1x	57.2%	80.3%
AlexNet	SVD (Denton et al., 2014)	32 bit	240MB → 48MB	5x	56.0%	79.4%
AlexNet	Network Pruning (Han et al., 2015b)	32 bit	240MB → 27MB	9x	57.2%	80.3%
AlexNet	Deep Compression (Han et al., 2015a)	5-8 bit	240MB → 6.9MB	35x	57.2%	80.3%
SqueezeNet (ours)	None	32 bit	4.8MB	**50x**	57.5%	80.3%
SqueezeNet (ours)	Deep Compression	8 bit	4.8MB → 0.66MB	**363x**	57.5%	80.3%
SqueezeNet (ours)	Deep Compression	6 bit	4.8MB → 0.47MB	**510x**	57.5%	80.3%

Squeezenets outperform all model compressions approaches before it with the smallest model being 0.5MB registering a 510x decrease in model size compared to AlexNet

Hyperparameters





SR - the ratio between the no. of filters in squeeze to the no of filters in expand.

Squeezenet variants

Architecture	Top-1 Accuracy	Top-5 Accuracy	Model Size
Vanilla SqueezeNet	57.5%	80.3%	4.8MB
SqueezeNet + Simple Bypass	**60.4%**	**82.5%**	4.8MB
SqueezeNet + Complex Bypass	58.8%	82.0%	7.7MB

The bypass connections are same as skip connections in the Resnet architecture.
The simple bypass is just a wire i.e. it adds the output of past layer as an input further in the network.
The complex bypass on the other hand has a 1x1 convolution with the number of filters set equal to the number of output channels that are needed.


Thank
you