4. Convolutional neural networks

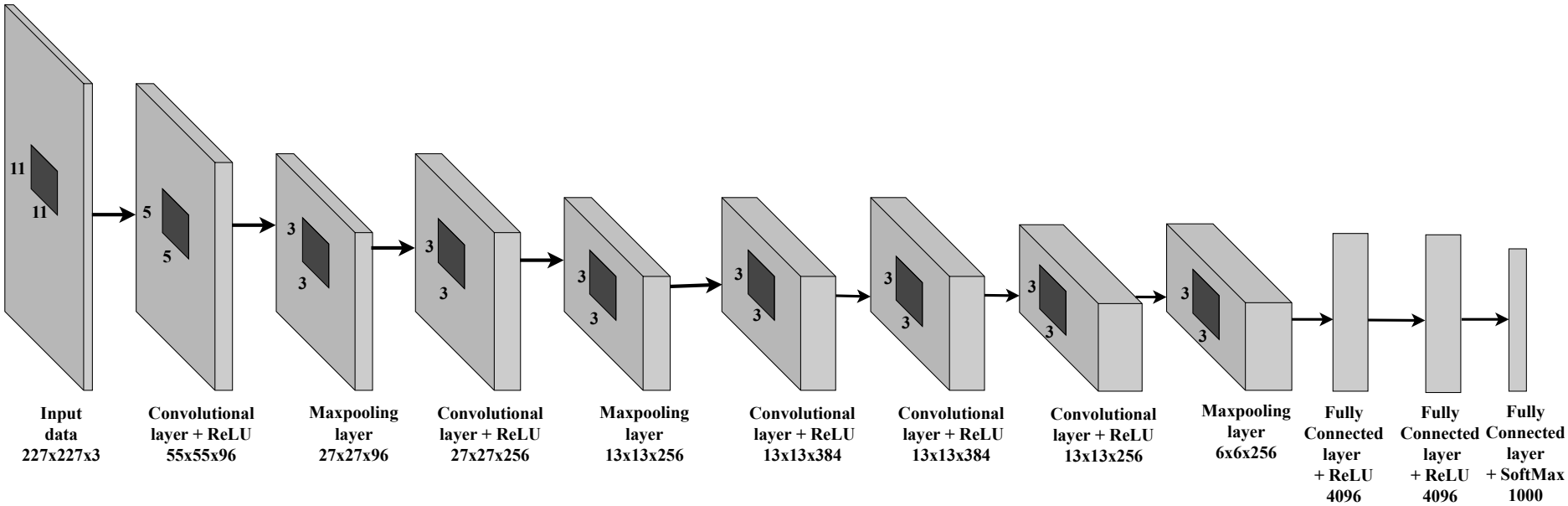
4.3. Advanced CNN structures

Manel Martínez-Ramón
Meenu Ajith
Aswathy Rajendra Kurup

Alexnet

- Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton (2012).
- Able to solve the Imagenet image classification problem with 1000 classes.
- Won the 2-12 ImageNet Large Scale Visual Recognition Challenge (LSVRC) conmpetition.
- Structure:
  - 8 layers and 60M learnable parameters.
  - 5 are convolutional layers along with max-pooling layers and 3 fully connected layers and ReLU non-linearity.
  - SoftMax with 1000 outputs.

Alexnet


11
11
Input data 227x227x3
5
5
Convolutional layer + ReLU 55x55x96
3
3
Maxpooling layer 27x27x96
3
3
Convolutional layer + ReLU 27x27x256
3
3
Maxpooling layer 13x13x256
3
3
Convolutional layer + ReLU 13x13x384
3
3
Convolutional layer + ReLU 13x13x384
3
3
Convolutional layer + ReLU 13x13x256
3
3
Maxpooling layer 6x6x256
Fully Connected layer + ReLU 4096
Fully Connected layer + ReLU 4096
Fully Connected layer + SoftMax 1000

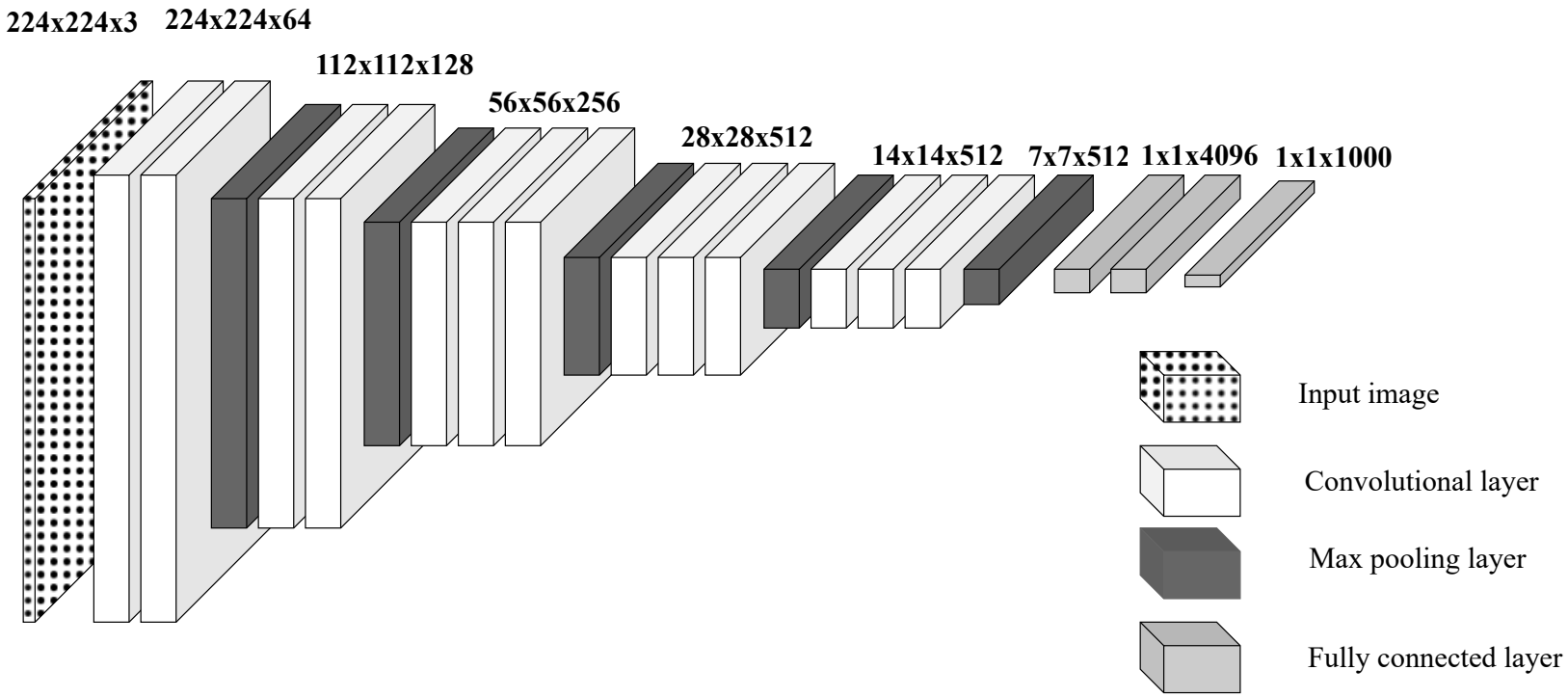
Alexnet

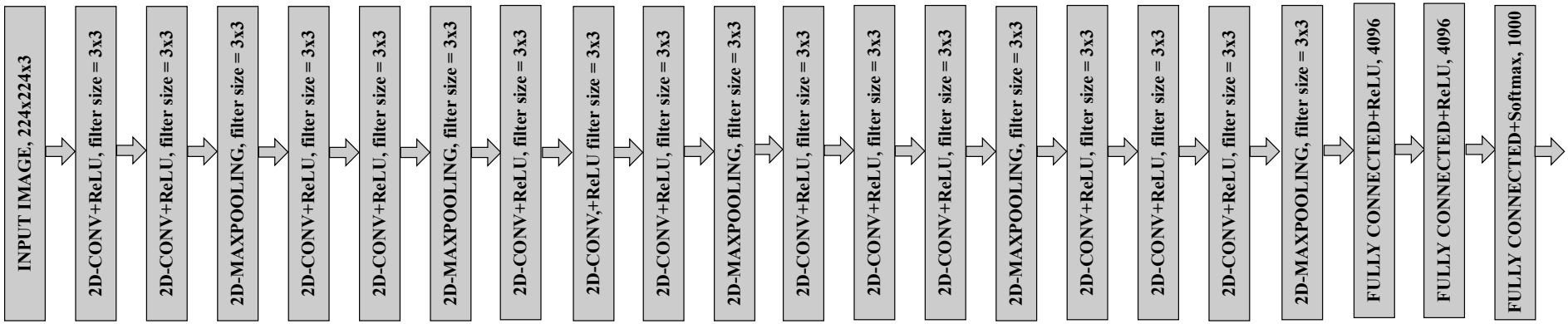
- RGB images, $224 \times 224 \times 3$
- The convolutions include a stride of 4 pixels.
- A padding of 3 pixels are possibly used.
- The fifth convolutional layer contains 256 kernels of size $3 \times 3 \times 192$
- Each of the hidden fully connected layers has 4096 neurons.
- Novelties:
  - Data augmentation by a factor of 2048 by extracting random patches from the images and altering the intensities of the RGB channels.
  - Dropout, ReLU, overlapping pooling, multi-GPU.

VGG

- VGG stands for Visual Geometric Group. This network was developed in the year 2014 by Karen Simonyan and Andrew Zisserman (Oxford University, 2014).
- Second place, 2014 ImageNet ILSVRC competition.
- Main idea: increase the depth of the convolutional network in large-scale image recognition settings.
- Smaller convolutional filters were used for increased depth.

VGG


224x224x3
224x224x64
112x112x128
56x56x256
28x28x512
14x14x512
7x7x512
1x1x4096
1x1x1000
Input image
Convolutional layer
Max pooling layer
Fully connected layer


INPUT IMAGE, 224x224x3
2D-CONV+ReLU, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-MAXPOOLING, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-MAXPOOLING, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-CONV,+ReLU filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-MAXPOOLING, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-MAXPOOLING, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-CONV+ReLU, filter size = 3x3
2D-MAXPOOLING, filter size = 3x3
FULLY CONNECTED+ReLU, 4096
FULLY CONNECTED+ReLU, 4096
FULLY CONNECTED+Softmax, 1000

VGG

- Depth is achieved through a stack of convolutional layers.
- Convolutional kernels of smaller size (3×3).
- five max-pooling layers with a window size of 2×2 and a stride 2.
- ReLU activations.
- Three fully connected layers with 4096, 4096 and 1000 nodes, with softmax activation.
- Incorporates 1×1 convolutions, that are just linear transformations.

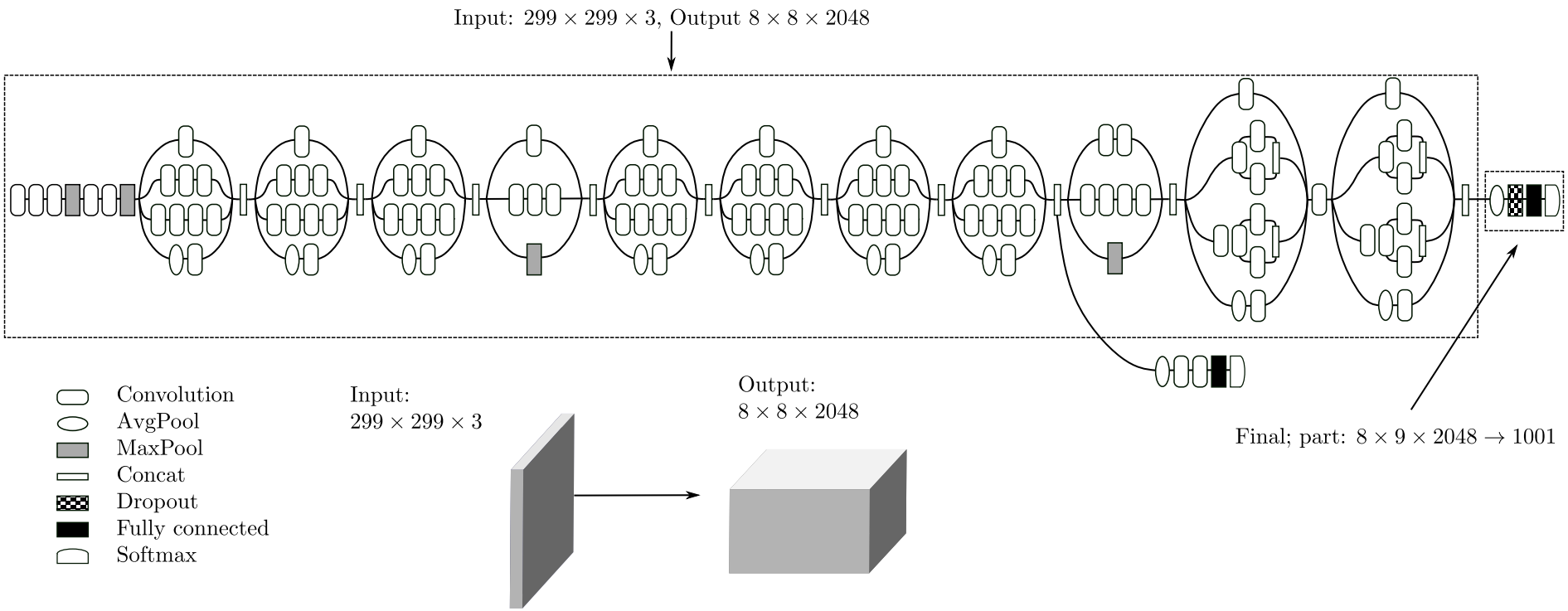
VGG

- First to use a p[reprocessing block that crops the image to size 224×224, extracts the mean RGB value.
- Shows that 3×3 kernels stacked several times is as effective as a 5×5 or 7×7 kernel.
- 6 versions with 11 layers, 16 layers (VGG16) and 19 layers (VGG19), containing different number of convolution layers, but all of them kept the same number of fully connected layers and nodes.

Inception

- Inception-v1/GoogleNet has 22 layers and 5 million parameters.
- Winner of the 2014 ImageNet LSVR Competition.
- Previous structures are focused on increasing the depth. Inception has recurring blocks of convolutional designs called Inception modules.
- Around 100 layers.
- ReLU activations.
- Auxiliary classifiers in order to prevent the vanishing gradient problem.

Inception


Input: 299 × 299 × 3, Output 8 × 8 × 2048
Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmax
Input:
299 × 299 × 3
Output:
8 × 8 × 2048
Final; part: 8 × 9 × 2048 → 1001

ResNet

- Winner of the 2015 ImageNet localization, ImageNet detection, in the LSCR competition and segmentation and detection challenges in the 2015 Common Objects in COntext (COCO) competition.

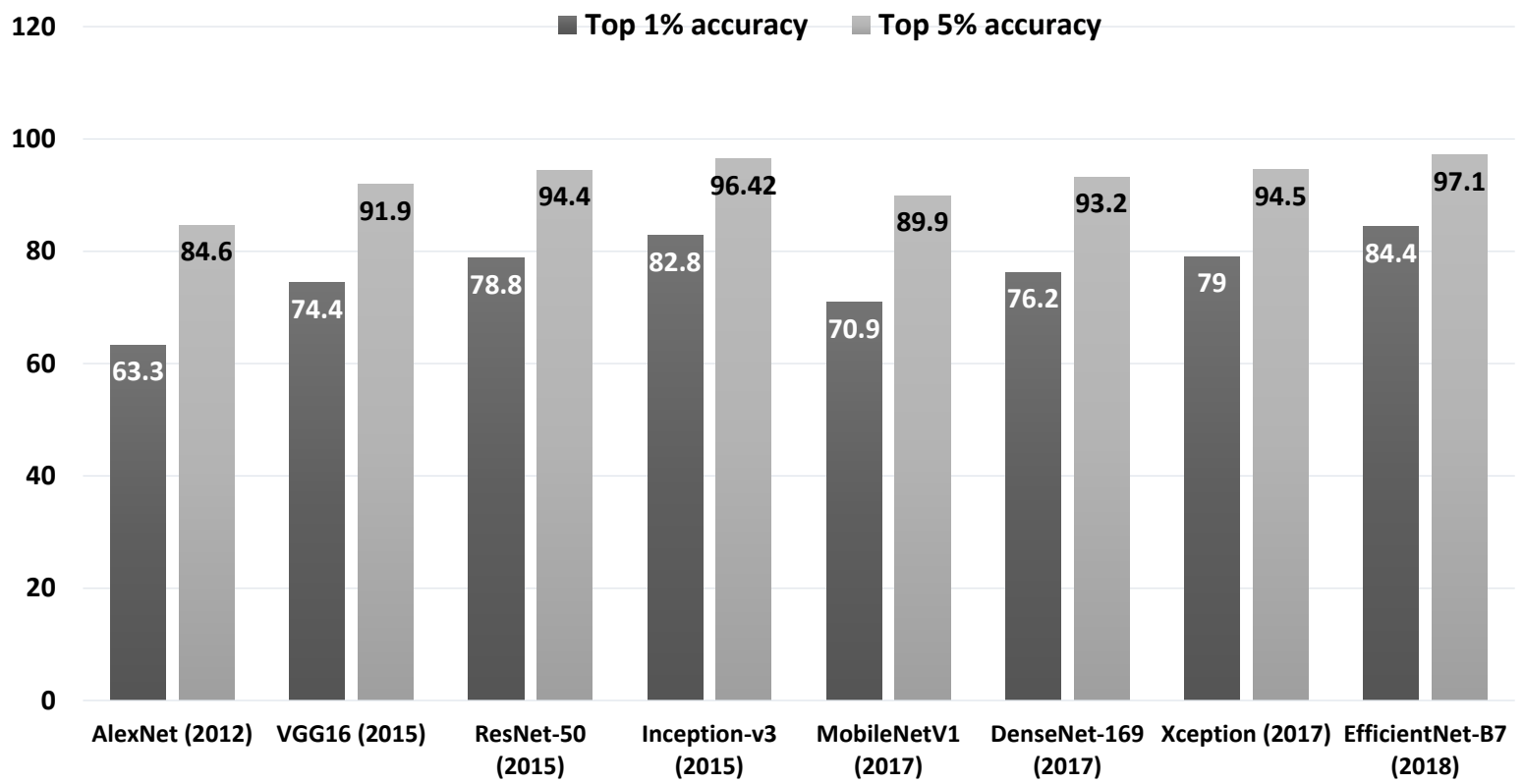
ResNet


residual connection
x
weight layer
relu
weight layer
$\mathcal{F}(x)$
$\mathcal{F}(x)+x$
relu
x
identity
Residual Network
Plain Network
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128
3x3 conv, 256
3x3 conv, 512
3x3 conv, 128, /2
3x3 conv, 256, /2
3x3 conv, 512, /2
avg pool
fc 1000

Others

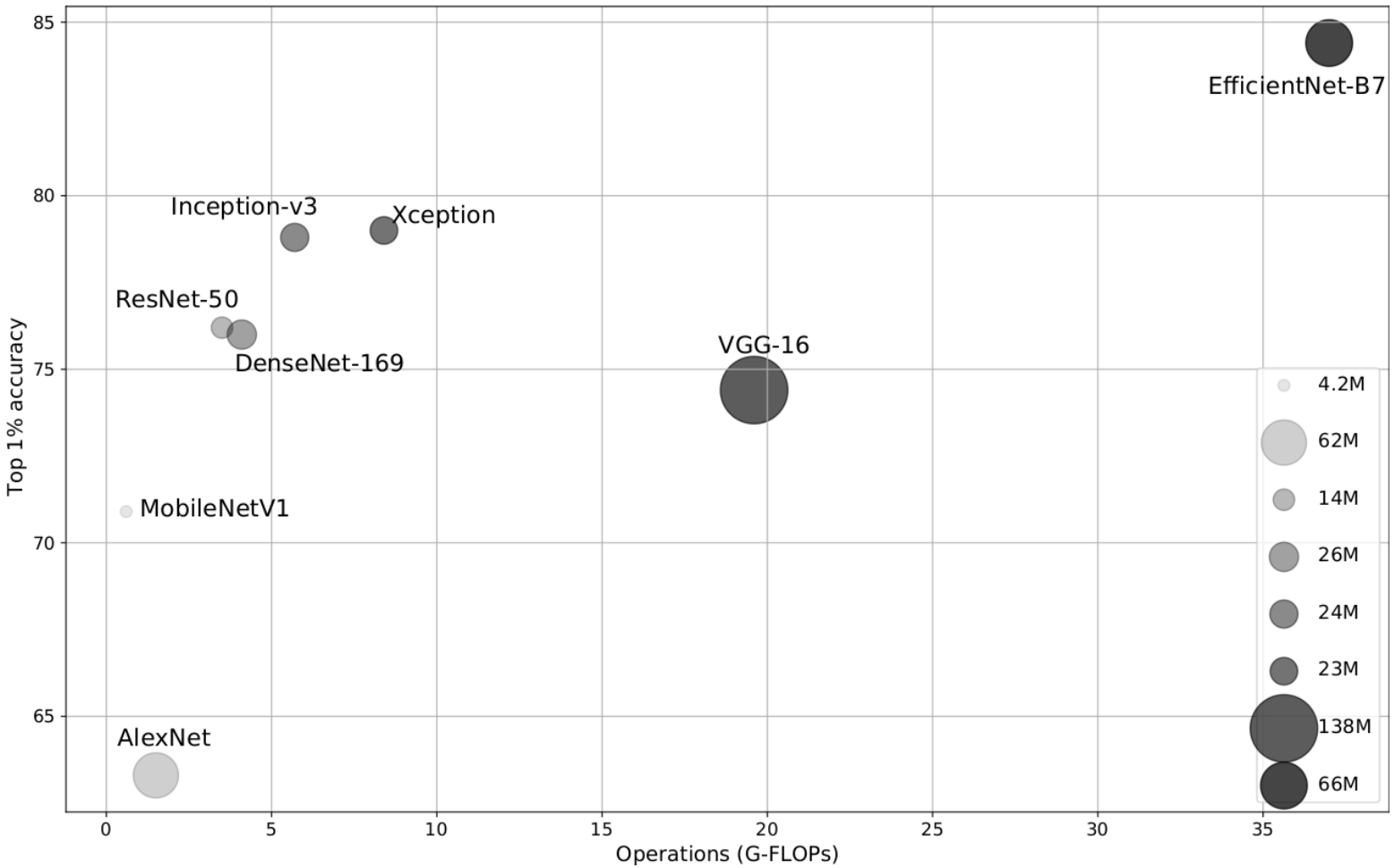
- Xception, François Chollet (2017).
- MobileNet, Google (2017).
- Densenet, Cornell University, Tsinghua University, Facebook, 2017.
- EfficientNet, Mingxing Tan and Quoc Le (2019).

Comparisons



Comparison of top 1% and top 5% accuracy of different CNNs.

Comparisons



Comparison of top 1% accuracy, Number of parameters and operations (G-FLOPS) of different CNN architectures for image classification.