An Evaluation of Noise Tolerance on Various Convolutional Neural Networks’ Architectures in Natural Images Classification using Transfer Learning Techniques

Shila Mosammami , Eduart Uzeir

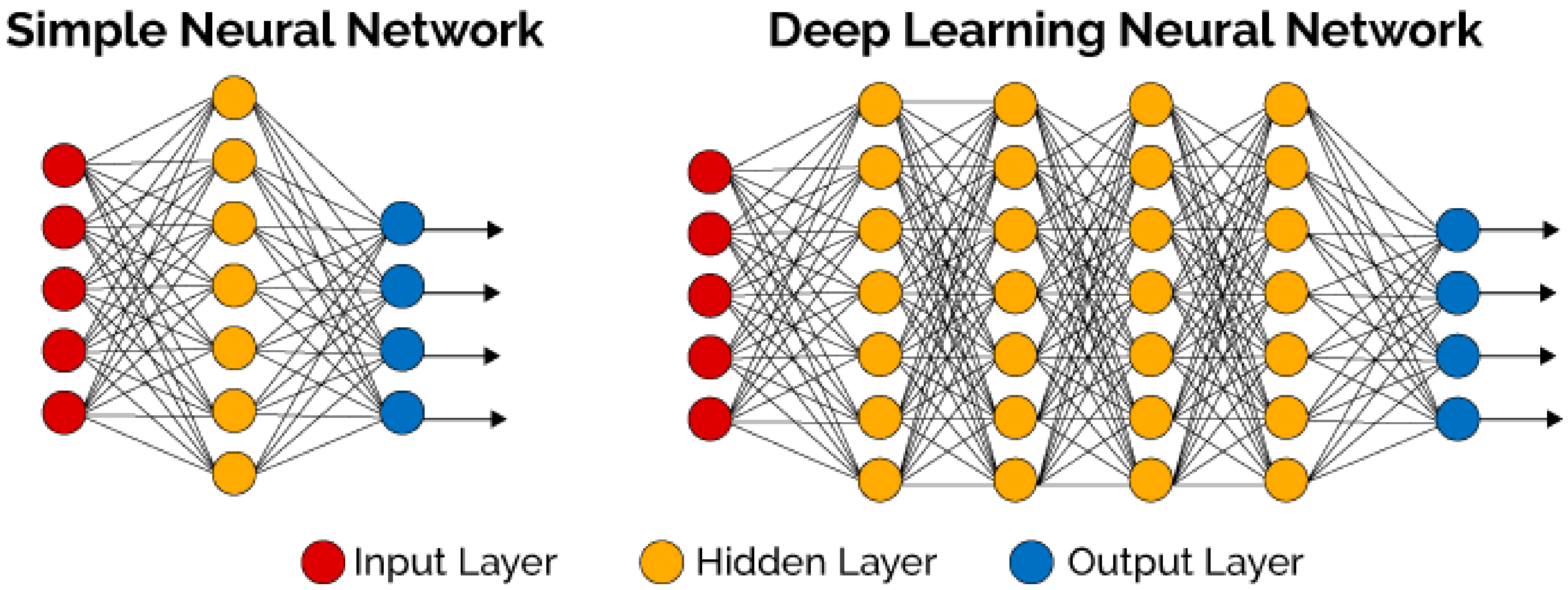
February 2019

[Corso di Laurea Magistrale in Informatica]

Introduction

Deep learning

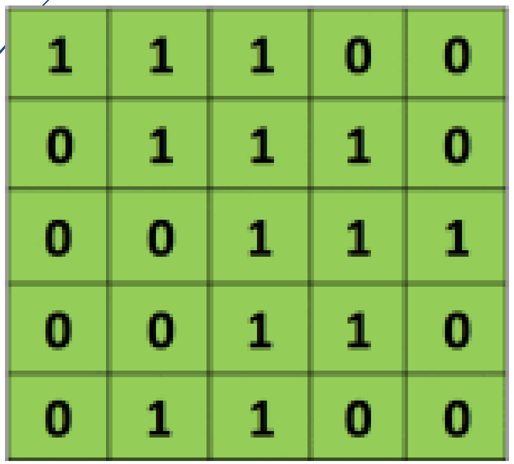
- A highly performed method for feature extraction and pattern recognition.
- Convolutional Neural Network (CNN) makes Image processing computationally manageable by applying filtering

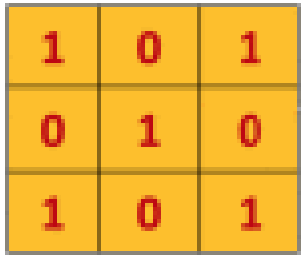


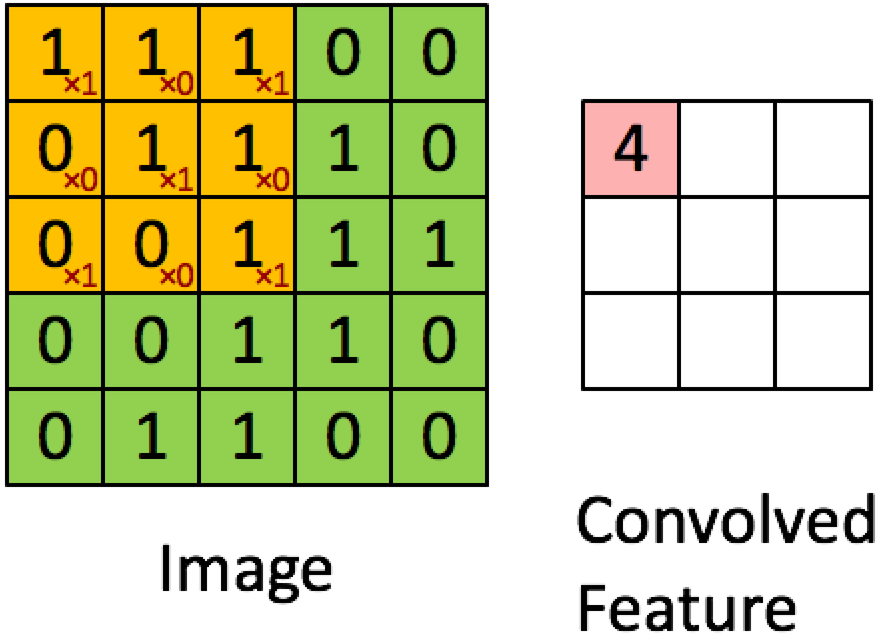
Introduction

CNN

- The filter (orange matrix) slides over the original image (green) by 1 pixel (also called ‘stride’) and for every position.
- Element wise multiplication (between the two matrices) are computed
- The summation forms a single element of the output matrix (pink).







The output matrix is called Convolved Feature or Feature Map

Introduction

CNN (Cont.)

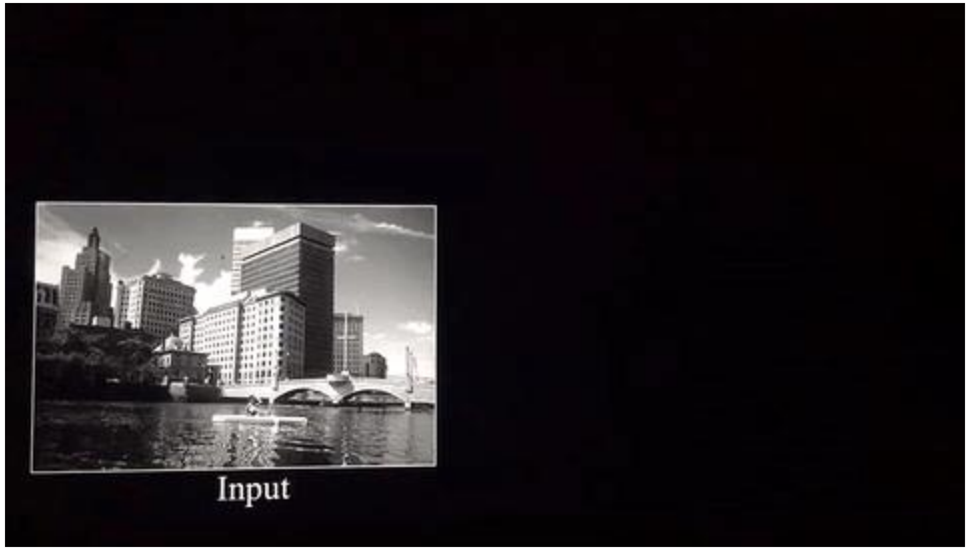
- CNN **learns the values of these filters** on its own during the training process
- We need to specify parameters such as ***number of filters***, ***filter size***, ***architecture of the network***
- The size of the Feature Map (Convolved Feature) :

$$\text{output width} = \frac{W - F_w + 2P}{S_w} + 1$$

$$\text{output height} = \frac{H - F_h + 2P}{S_h} + 1$$

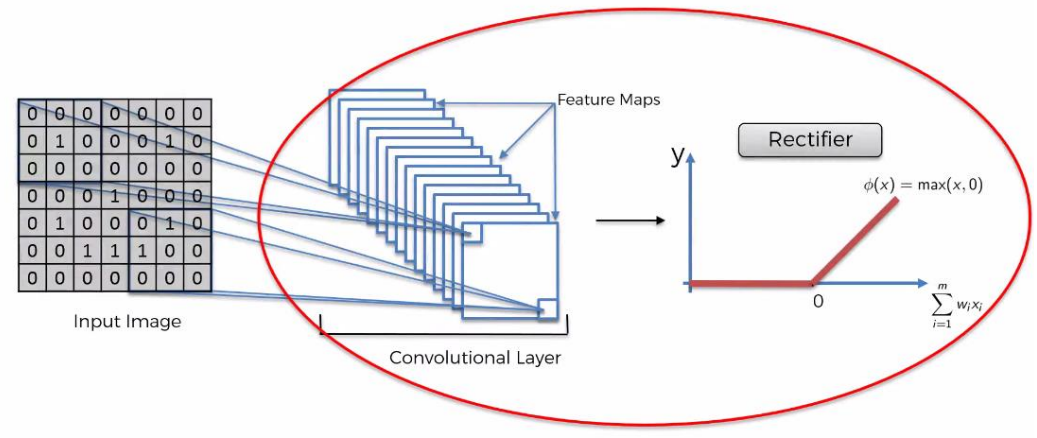
S = Stride
P= Zero-padding
W= Width
H= Height
F=Filter

And the output Depth would be the same as the depth of filters



Introduction

Concepts in CNN



Activation Function (RelU: Rectifier Linear Unit)

- Activation functions are also known as Transfer Function and basically decide whether a neuron should be activated or not.
- The Activation Functions can be basically divided into 2 types: 1) Linear Activation Function. 2) Non-linear Activation Functions
- They enable the networks to learn and perform more complex tasks.
- Relu is an element wise operation (applied per pixel) and replaces all **negative** pixel values by **zero** in the feature map .

Introduction

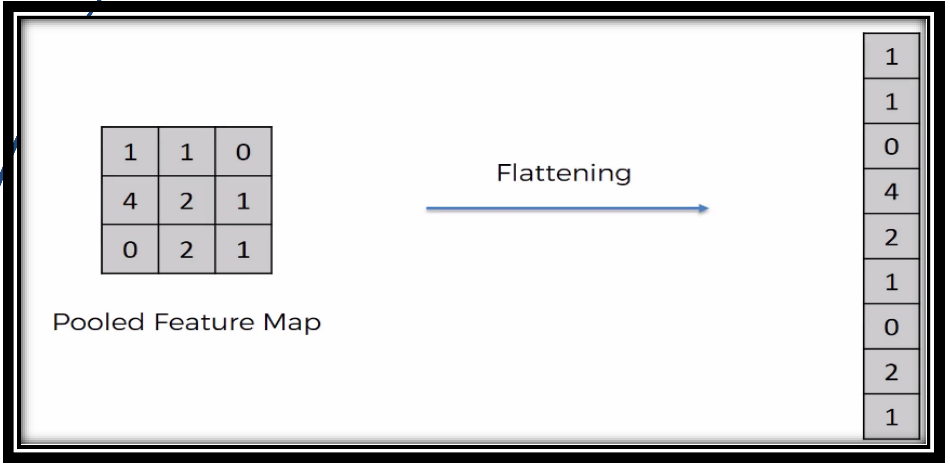
Concepts in CNN (Cont.)

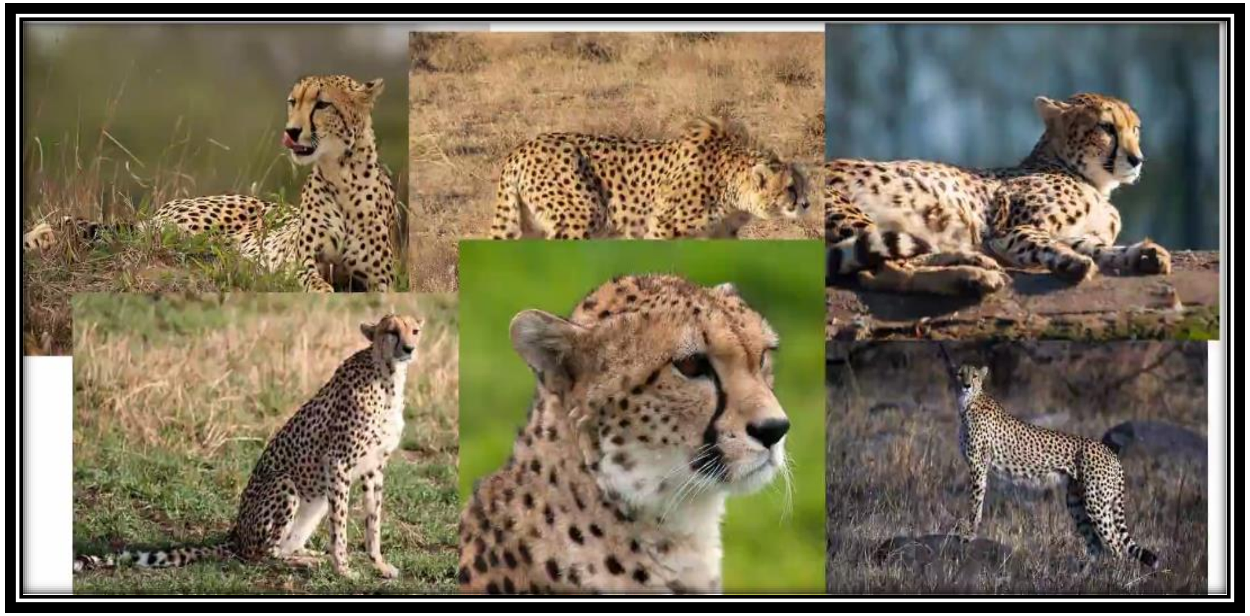
Pooling:

- By definition, Pooling (also called subsampling or down-sampling) reduces the dimensionality of each feature map but keeps the most important information. Pooling can be of different types: Max, Average, Sum etc.
- So that even if same object image, but in different position is analyzed, the recognition of the image still results correct output. check below multiple cheetah images in various positions. A small distortion in input will not change the output of Pooling – since we take the maximum / average value in a local neighborhood.
- Pooling reduces the number of parameters and computations in the network, therefore, controlling **overfitting**

Flattening:

It's about converting matrix into columnar form.



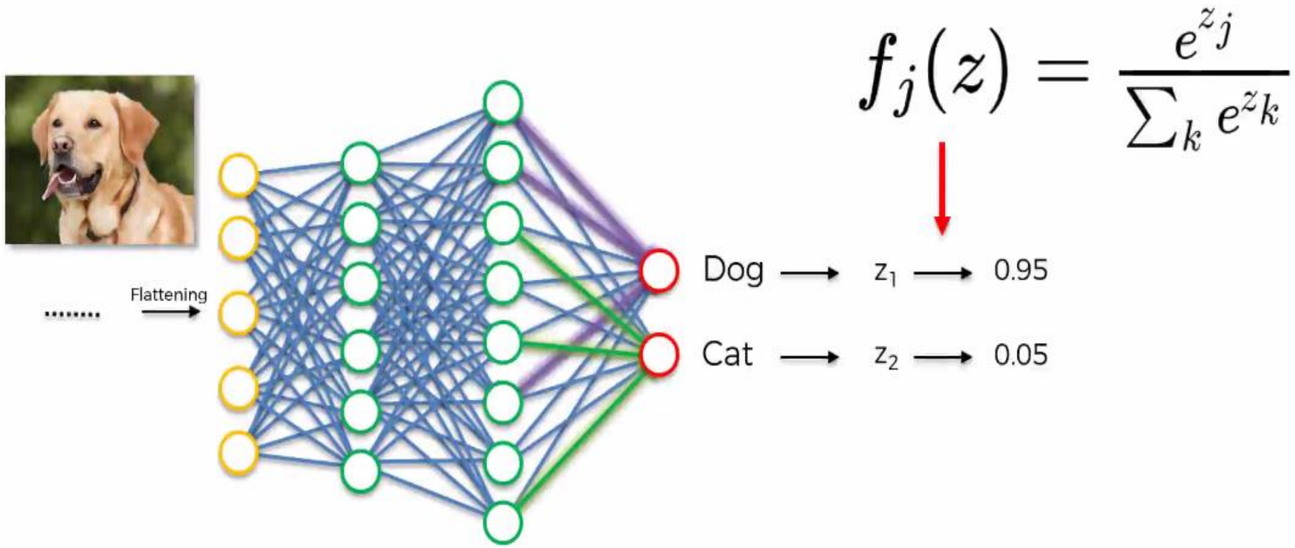


Introduction

Concepts in CNN (Cont.)

Fully Connection or Dense:

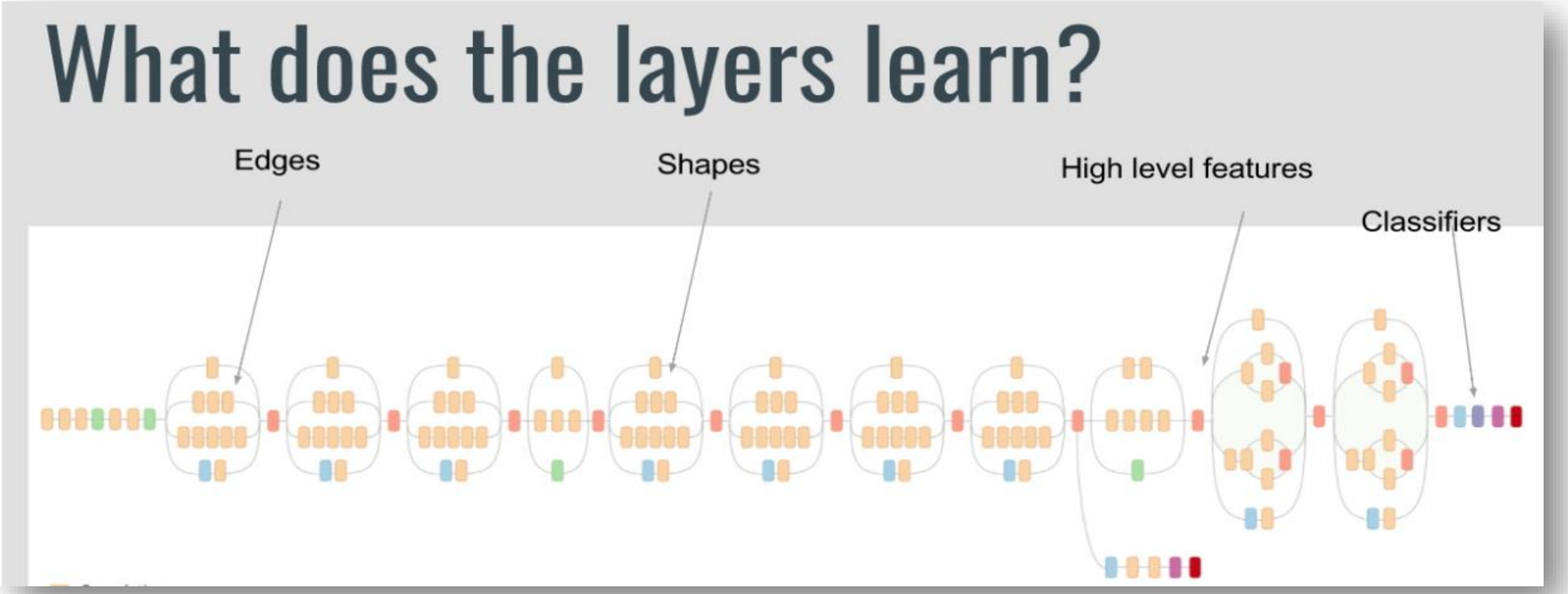
- The term "Fully Connected" implies that every neuron in the previous layer is connected to every neuron on the next layer
- The purpose of the Fully Connected layer is to use the features from previous layers for classifying
- Apart from classification, adding a fully-connected layer is also a (usually) **cheap way of learning non-linear combinations** of these features.
- The sum of output **probabilities** from the Fully Connected Layer is always **1**. This is ensured by using the **Softmax** as the activation function in the output layer of the Fully Connected Layer.



Introduction

Transfer learning

- Very Deep Networks are expensive to train. The most complex models take weeks to train using hundreds of machines equipped with expensive GPUs.
- Storing knowledge gained while solving one problem and applying it to a different but related problem
- Weights of pretrained related to some highly performed CNNs on the ImageNet challenge are available in Keras core library



VGG16 on ImageNet

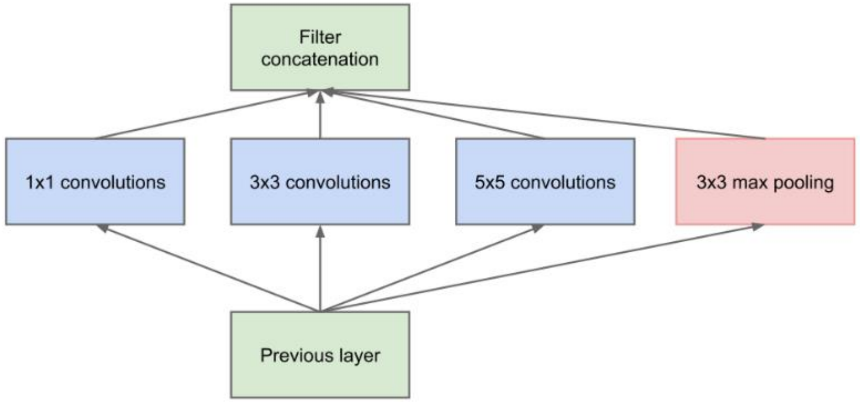
The network has been organized by 3*3 Convolutional networks (stride is considered 1; activation is ReLU) followed by 2*2 max pooling layer (stride 2) to decrease the size for each block. There are also fully connected layers which the first two have 4096 channels and the third which is a Softmax classifier, has 1000 channels equal to the number of required classes

The only preprocessing which was done is subtracting the mean RGB value from each pixel, computed on the training set.

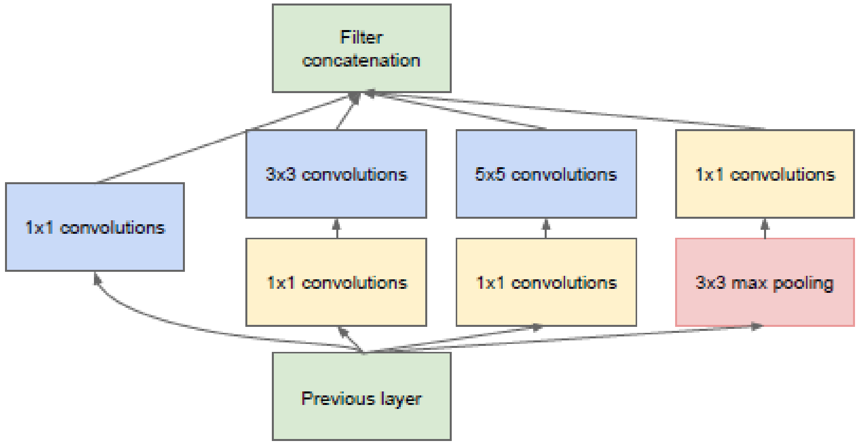
VGG16 Configuration : input (224*244 RGB image); 16 weight layers	
Block 1	Conv3-64 + ReLU
	Conv3-64 + ReLU
	Maxpool2
Block 2	Conv3-128 + ReLU
	Conv3-128 + ReLU
	Maxpool2
Block 3	Conv3-256 + ReLU
	Conv3-256 + ReLU
	Conv3-256 + ReLU
	Maxpool2
Block 4	Conv3-512 + ReLU
	Conv3-512+ ReLU
	Conv3-512+ ReLU
	Maxpool2
Block 5	Conv3-512 + ReLU
	Conv3-512 + ReLU
	Conv3-512 + ReLU
	Maxpool2
Classification block	FullyConnected-4096
	FullyConnected-4096
	FullyConnected-1000(Softmax)

Inception V3 on ImageNet

- The point is in this architecture 3*3 or 5*5 or 1*1 convolutions are done in parallel and concatenated resulting feature maps are sent for the next layer.
- This architecture allows to obtain both local features through the smaller convolutions and more abstracted features via larger convolutions.
- In Inception V3, each 5*5 convolution layer is replaced by two 3*3 convolution layers because of hypothesize that indicates strong correlation between adjacent units results in less information loss during dimension reduction.



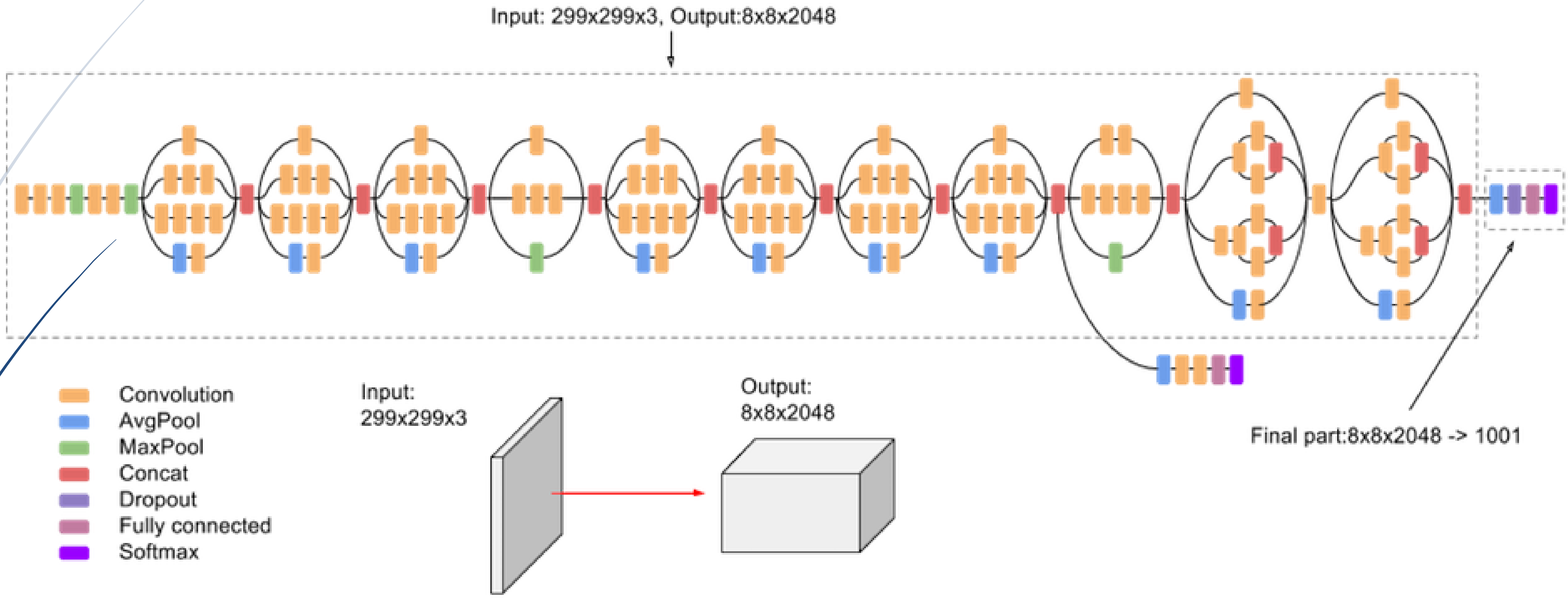
Inception module Naïve Version



Inception module with dimension reductions

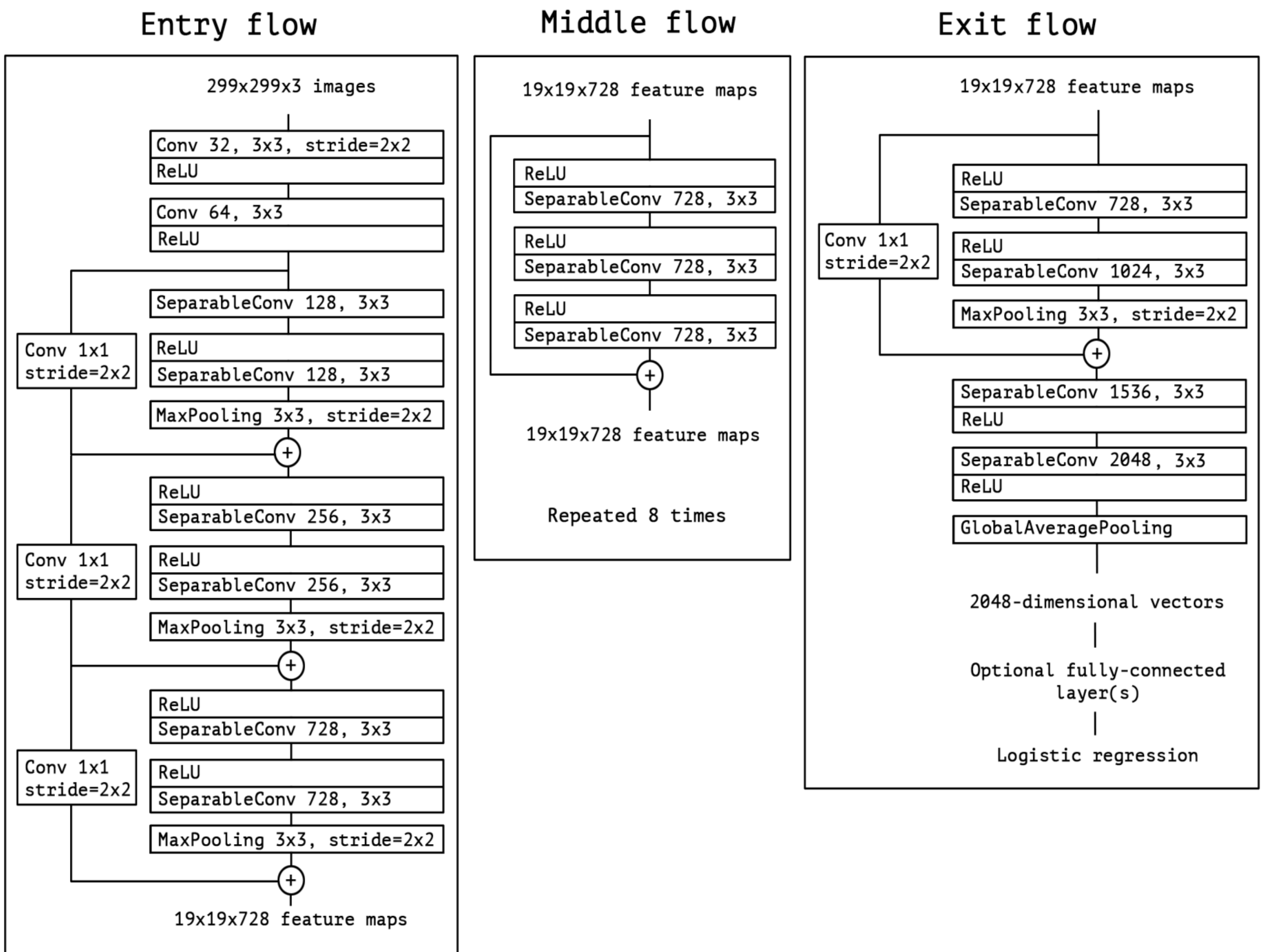
Inception V3 on ImageNet

- Look at the Architecture



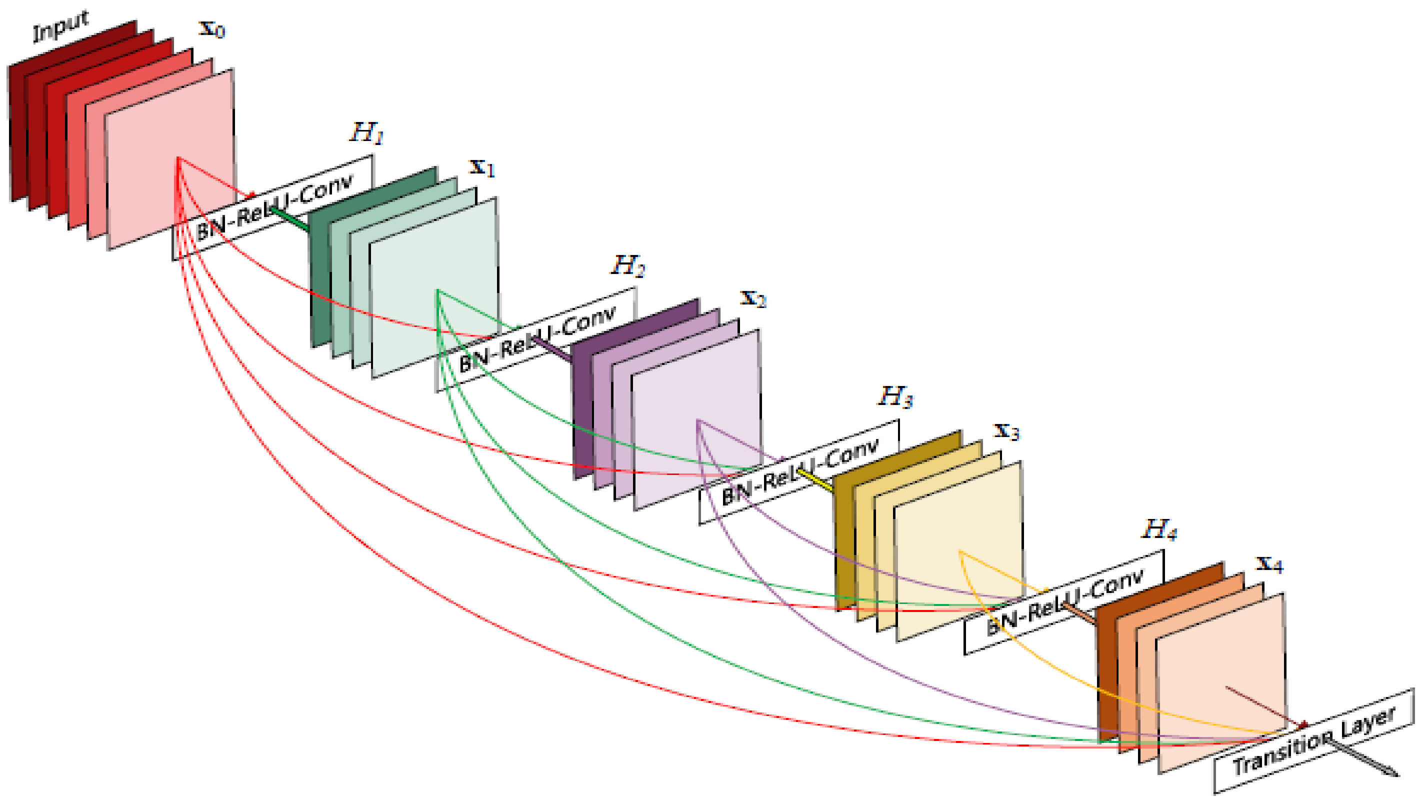
Xception on ImageNet

- Depth-wise separable convolution layers, plus residual connections.
- The hypothesis is that the cross channel (depth) correlation and spatial correlation in the feature maps of CNNs can be decoupled. Simply speaking, it is not needed to consider both the image region and the channels at the same time.
- The Xception architecture has structured in 14 modules all which apart from the first and last, have a linear residual layer.
- Note that after all Convolution and Separable Convolution layers there is **batch normalization**.



What are DensNets?

- DensNets are very deep CNN architecture characterized by direct connections between each layer allowing maximal information flow. The number of connections is L(L+1)/2, where L is the number of layers



How are they created?

- Dense connectivity: xl = Hl([x0, x1, ..., xl-1])

  l-th layer recieves the feature-map of all preceding layers x0, x1, ..., xl-1 and then use the concatenation of them.

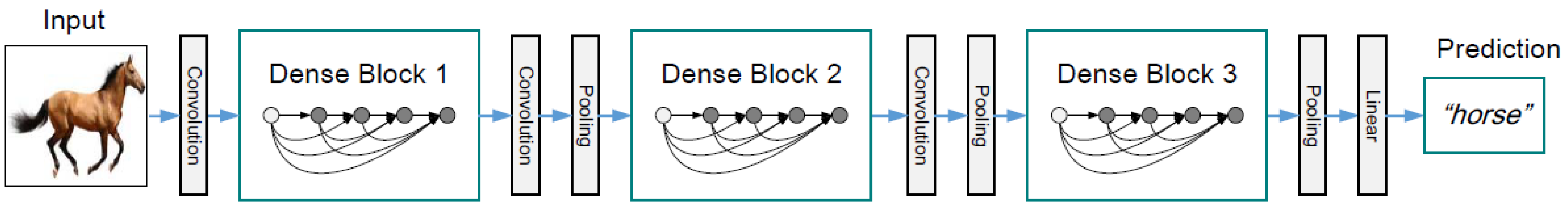
  Non-linear transformation Hl is a composite function of three consecutive operations: Batch Normalization, ReLU and 3x3 Conv.

  Polling and convolution layers are implemented like transition layers between dense blocks.

  Low growth rate (k = 12) indicates the amount of new information that each layer contributes to the global state.

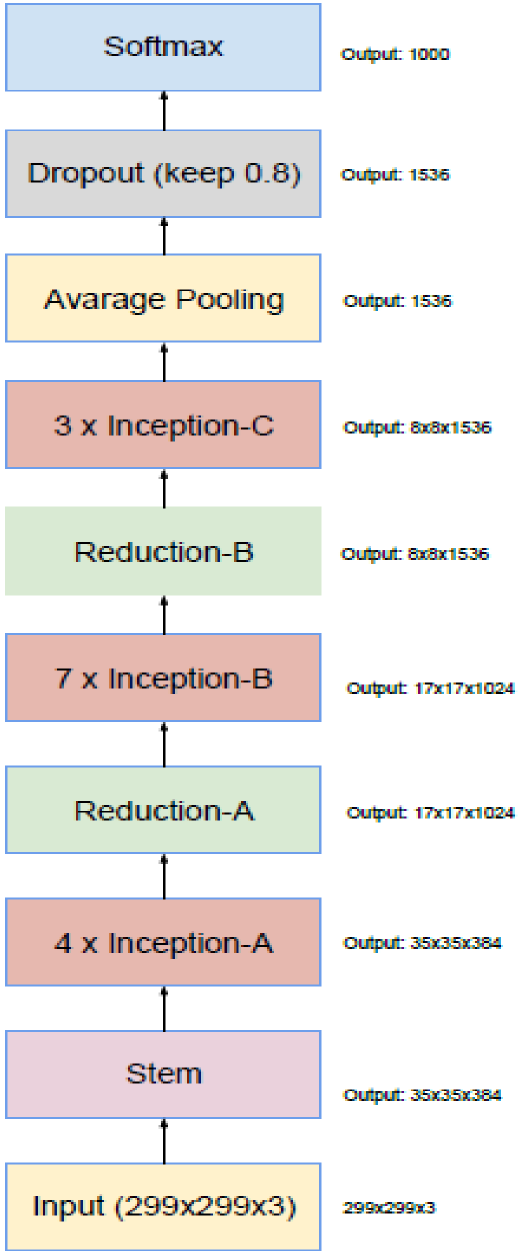
  Use of 1x1 Conv. as bottelneck layer to reduce input feature-maps leading to an improve computational efficiency.

  Use compression methods to further reduce the feature-map at transition layers.



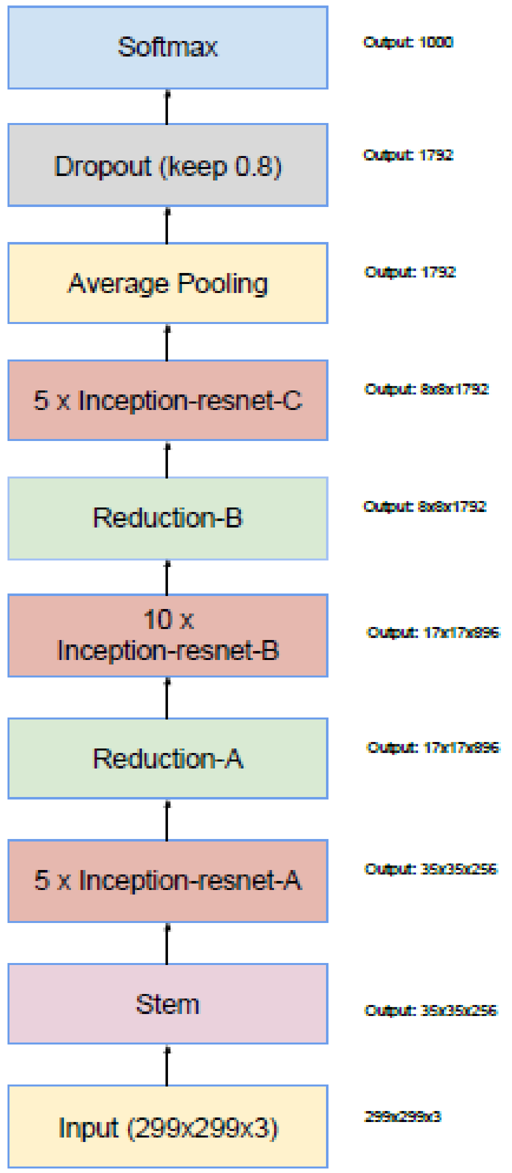
Inception-v4

- Is an evolution of InceptionV3, consisting of more Inception modules and also simplifies the overall architecture.
- Do not use residual connections.



Inception-ResNet

- Is the residual version of the Inception architecture.
- Two models have been developed Inception-Resnet-v1 that has computational cost as InceptionV3 and Inception-ResNet-v2 that is more costly and accurate in recognition performance.



Dataset & Implementation

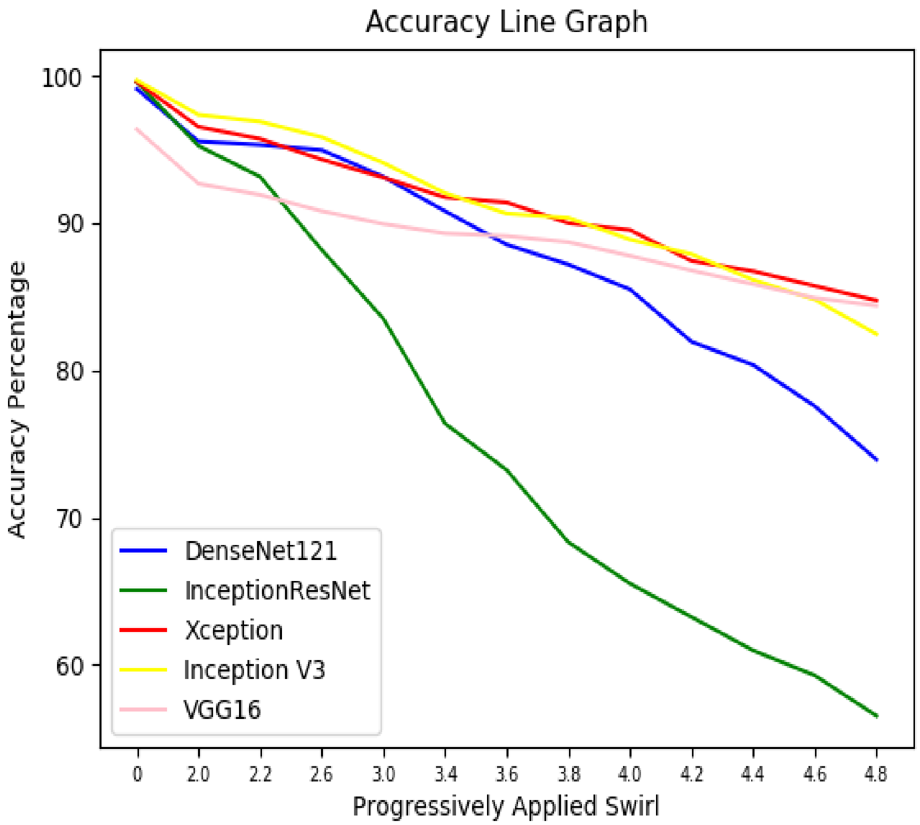
- Dataset: Kaggle Natural Images containing 8 classes

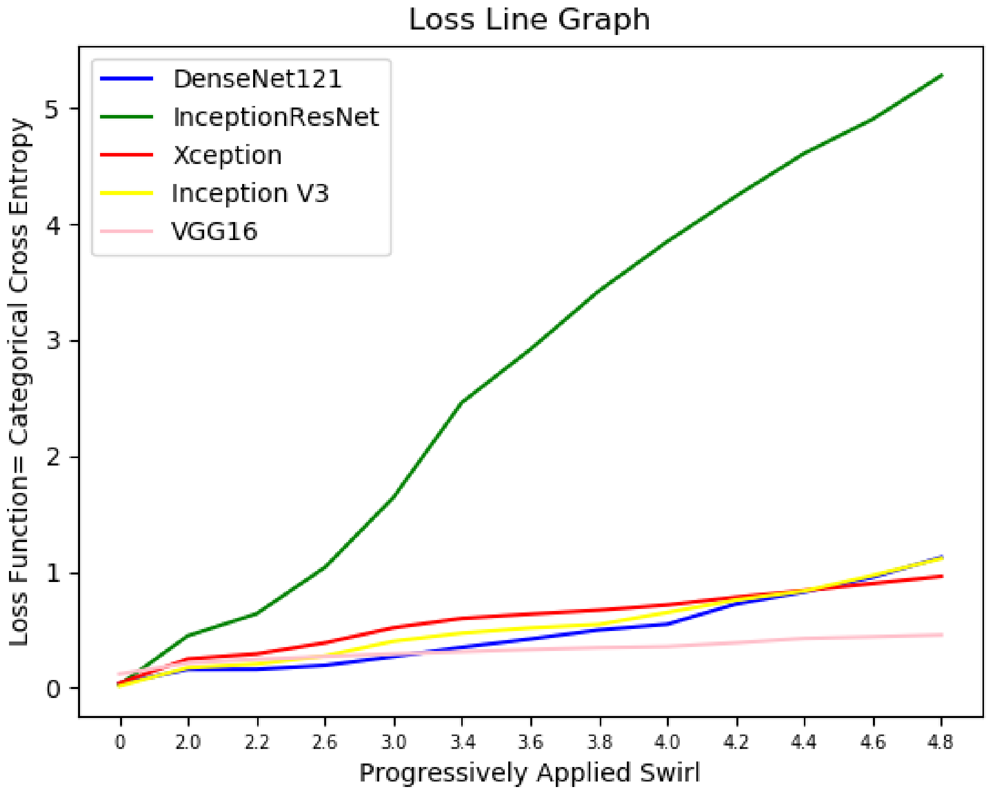
Accuracy and Loss Comparison on Test Data after Transfer Learning		
	Accuracy on Test Data	Loss on Test Data
VGG16	96.38	0.1202
Inception V3	99.71	0.0162
InceptionResNet	99.59	0.0250
Xception	99.59	0.0408
DensNet121	99.12	0.0370

Applied Noises and Related Results

Swirl

A swirl as a nonlinear deformation is essentially like a rotation and shift over a pixel point. In this report fixing the center of noise with the center of image and the radius with the height of it, we deformed images by changing the strength in various levels

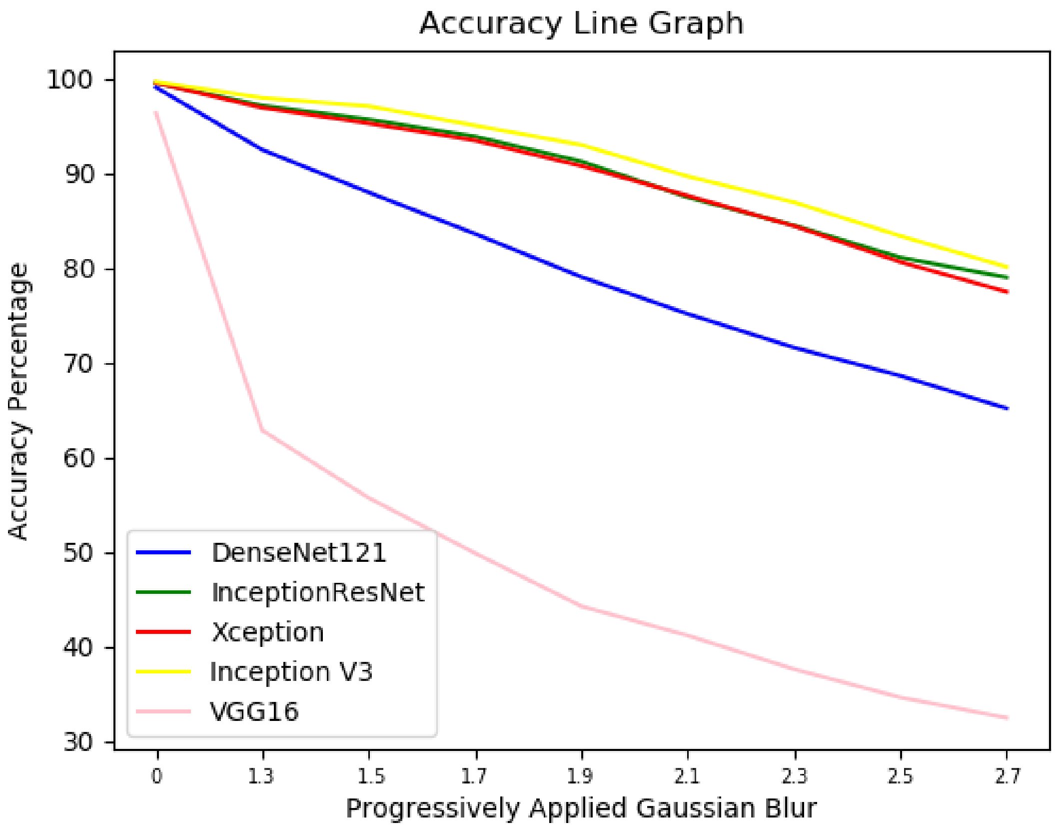


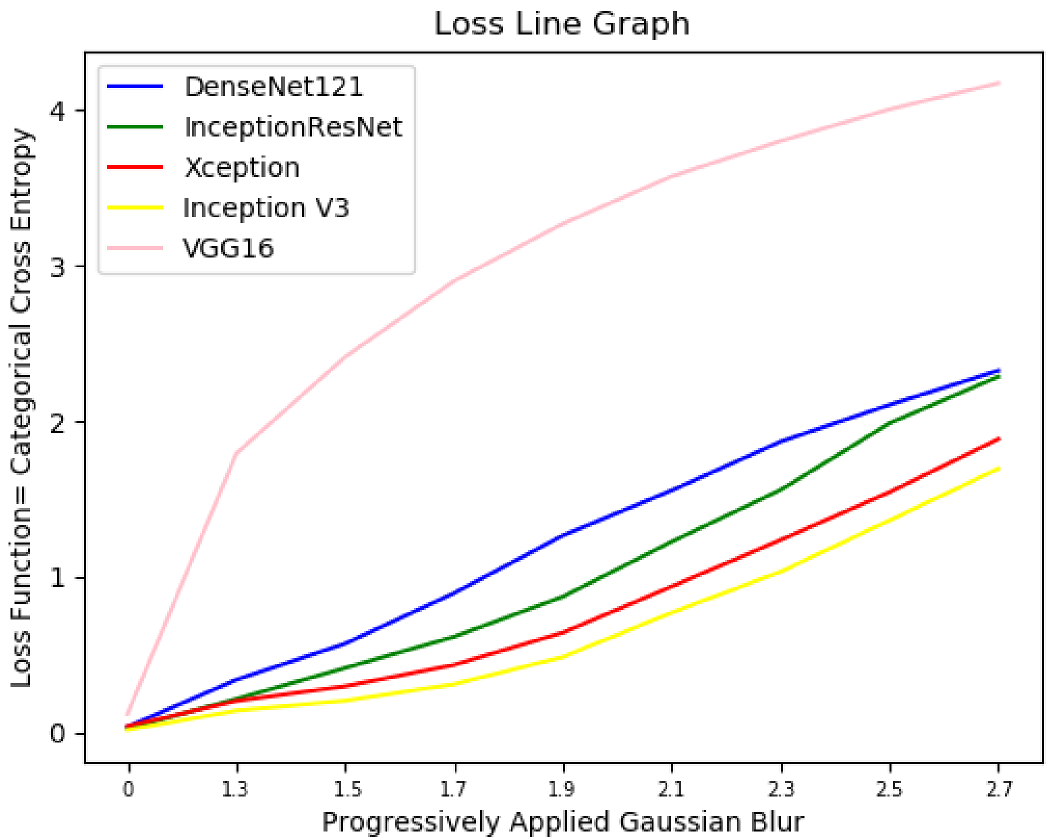


Applied Noises and Related Results

Gaussian Blur

It is a method applied to images to make them smoothed and hence a reduction in details. The severity of blurring in sigma had various progressive levels.

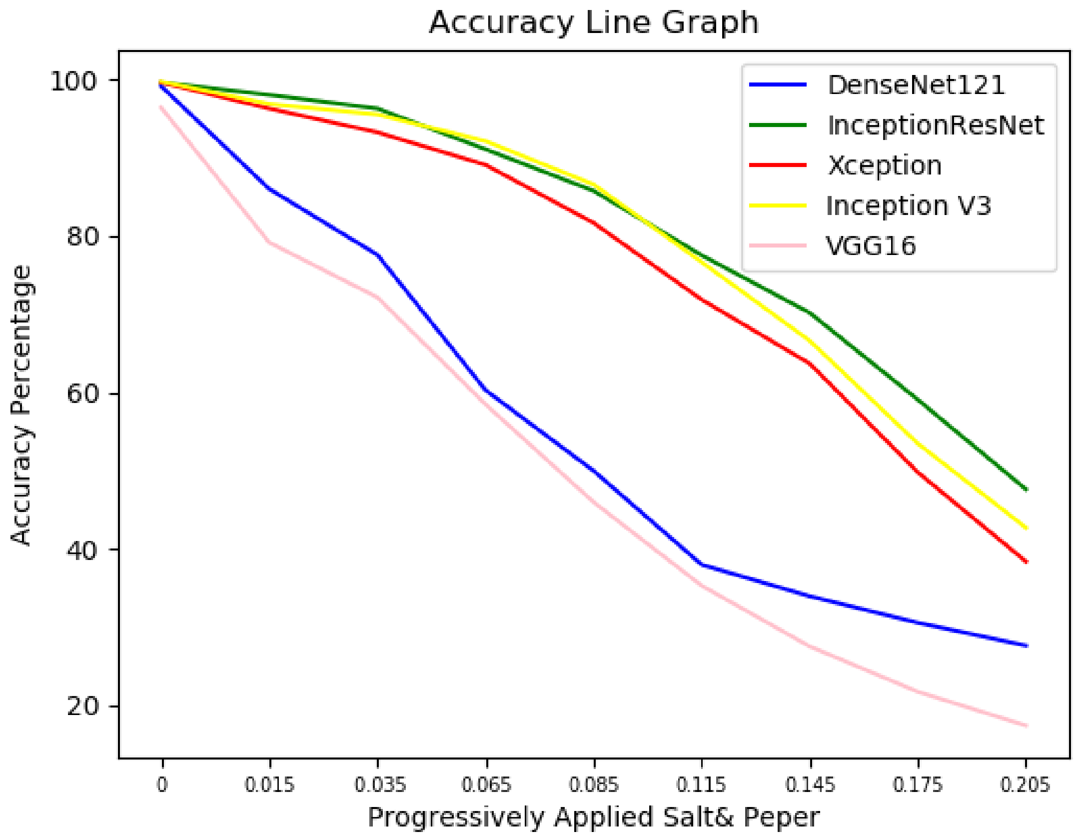


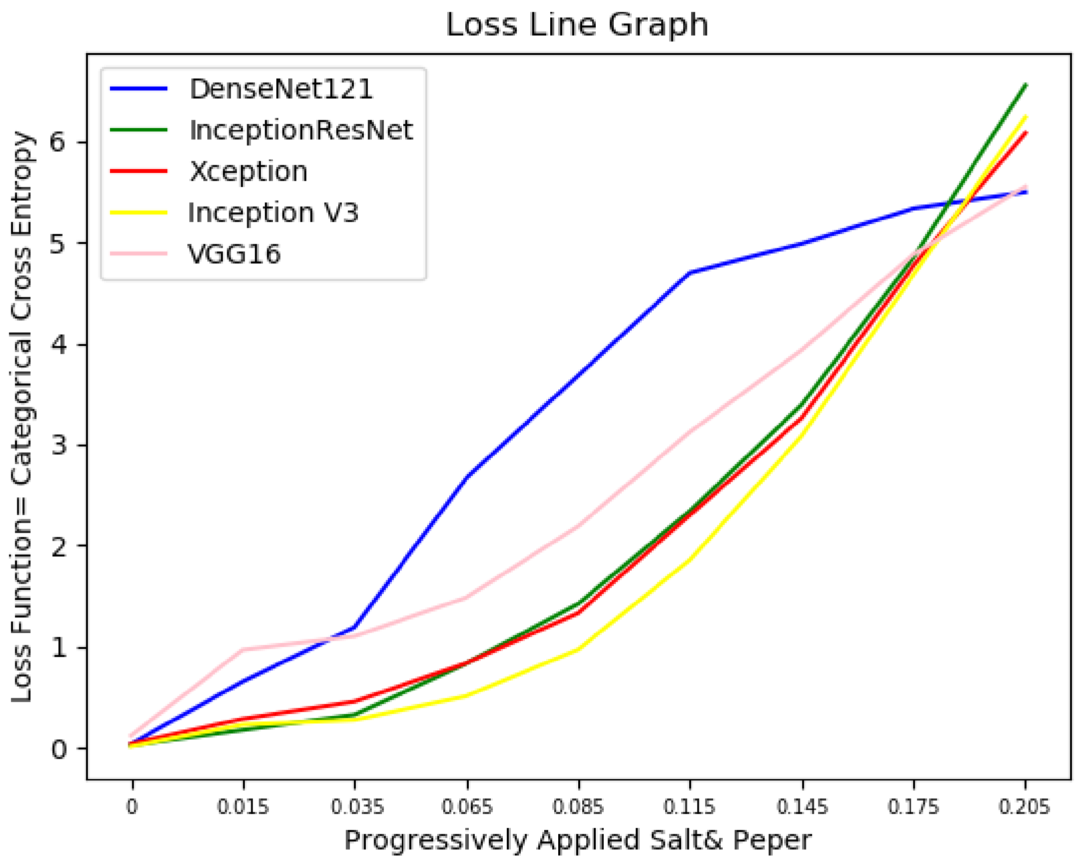


Applied Noises and Related Results

Salt&Pepper

It is the sparsely presence of black and white pixels in images. The probability of noise occurrence (how sparse) was considered progressively





References

- [1] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.
- [2] http://image-net.org/
- [3] https://www.kaggle.com/datasets
- [4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
- [4] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI (Vol. 4, p. 12).
- [5] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. arXiv preprint, 1610-02357.
- [6] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015, June). Going deeper with convolutions. Cvpr.
- [7] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).
- [8] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2818-2826).