Abstract:

Introduction: In recent years, deep learning models use multi layers of nonlinear processing information to extract and transform features, as well as analyze and classify patterns. Image classification is a fundamental topic in computer vision. It is described as the process of categorizing images into one of several specified groups. It serves as the foundation for other computer vision tasks such as localization, detection, and segmentation [i-1]. Since Krizhevsky's [i-2] winning entry in the 2012 ImageNet competition [i-3], their network "AlexNet" has been successfully applied to a wider range of computer vision tasks. Since 2014, as a result of the use of deeper and larger networks, the quality of network architectures has improved significantly [i-4]. Olaf Ronneberger, Philipp Fischer, and Thomas Brox developed U-Net in the paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" published in 2015. It's a step forward for FCN: Evan Shelhamer, Jonathan Long, and Trevor Darrell (2014). "Semantic segmentation using fully convolutional networks" **[m-1].** VGGN network was invented by Simonyan and Zisserman, Visual Geometry group at the University of Oxford in 2014. This family of architectures achieves 2nd place in the 2014 ImageNet Classification competition.This model is just a simple linear chain of layers, which is noteworthy **[s-1]**. A Microsoft research group introduced ResNet, a deep convolutional neural network, in 2015 and in the ILSVRC 2015 classification competition, they took first place. ResNet is similar to VGG net in terms of depth **[m-13],** although ResNet is around eight times deeper **[m-14].** The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017 awarded ‘Densely Connected Convolutional Networks' the Best Paper Award. DenseNet was created specifically to address the vanishing gradient's effect on high-level neural networks' accuracy. In simply, the information evaporates before it reaches its destination due to the long journey between the input and output layers. Inception v3 is a convolutional neural network for assisting in image analysis. It starts as a module for GoogleNet. It has three editions and it is the third edition. It is just as ImageNet as a database of a classified visual object. It helps to classified the objects [s-1]. In this paper we will focus general principles, optimization ideas, applications and limitations of CNN image classification architecture.

Literature Review:

1. **U-Net:** The model U-net is called U-net because of its U-shaped architecture. The left side of this model is called the contracting path, and the right side is called the expansive path **[m-2].** Moreover, four concatenations occurred between the expansive path and its corresponding contracting path in the network. The contracting path started with one channel input image, consisting of (572\*572) pixels. The network then goes with an unpadded convolution with a kernel size of (3\*3) for two repeated times. As this is an unpadded convolution, the pixels reduce to (570\*570) in the first step and then (568\*568) in the second step. For these two convolutions, the channel number has been set to 64. The next step started with down-sampling with (2\*2) max-pooling, which reduces pixel size to (284\*284), i.e., half, and again the unpadded convolution occurs two times as the previous step, but this time the channel number has been increased to 128 from 64. This continues three more times and finishes with pixel size (28\*28), consisting of 1024 channels, which is the end of the contracting path and the start of the extensive path. From this step, instead of max pooling, up-convolution takes place, which is the opposite of max pooling. The (2\*2) up-convolution increases the pixel size from (28\*28) to (56\*56), i.e., double. This time concatenation occurs with the channel of the contracting path with its corresponding expansive path. Therefore, the channel increased by 1536(1024+512). Through expansive path, unpadded convolution happens two more times with (3\*3) kernel size, reducing the pixel size to (52\*52). The channel number has been reduced to 512 in this step. Next, again up-convolution, as well as concatenation, happens and repeats the previous step three times. The network completed its expansive path and came up with the 64-channel image with a pixel size of 388\*388. Finally, (1\*1) convolution happens to reduce the channel from 64 to 2. So, the model lastly shows the two-channel output image of (388\*388) pixels.
2. **VGGNet-19:** The input to VGG based ConvNet is a (224\*224) RGB image. Preprocessing layer takes the RGB image with pixel values in the range of 0-225 and subtracts the main image value which is calculated over the entire ImageNet training set. The input images after prepossessing are passed through these weight layers. The training images are passed through these weight layers. The training images are passed through a stack of convolution layers. VGG-19 has 19 weight layers consisting of 16 convolutional layers with 3 fully connected layers and the same 5 pooling layers. VGG-19, there 2 fully connected layers with 4096 channels which are followed by another fully connected layer with 1000 channels to predict 1000 labels. The last fully connected layer uses the SoftMax layer for classification purposes **[s-1]**.
3. **ResNet:** The input layers of this network are made up of many residual blocks, and the operating principle is to optimize a residual function **[m-9].** This unique architecture allows for greater accuracy when layer depth is increased. The authors proposed residual mapping to accommodate the adding layers in their research. If we designate *H(x)* by the underlying mapping, then *F(x): = H(x) - x* determines the residual mapping. The residual block function is defined by: *y = F (x, Wi) + Wsx* when the input x and the output *y = H(x)* have the same dimension. All convolutional layers in ResNet models use the same convolutional window of size (3\*3), and the number of filters rises with network depth, from 64 to 512 (for ResNet-18 and ResNet-34), and from 64 to 2048 (for ResNet-50, ResNet-101, and ResNet-152). Only one max-pooling layer with pooling size (3\*3) is used in all models, and a stride of 2 is applied after the first layer. As a result, reducing the resolution of the input throughout the training phase is severely constrained. The average pooling layer is used to replace completely linked layers at the end of all models. This alternative has a few advantages. Firstly, there are no parameters to optimize in this layer, hence it aids in the reduction of model complexity. Secondly, this layer is more native in terms of enforcing feature map and category correspondences. The number of neurons in the output layer corresponds to the number of categories in the ImageNet dataset, which is 1000. In addition, in this layer, a SoftMax activation function is used to calculate the likelihood that the input belongs to each class.
4. **DenseNet:** The vanishing gradient problem describes how gradients aren't sufficiently back-propagated to the network's original layers as networks go deeper. As one moves backward into the network, the gradients become smaller, and the earliest layers lose their ability to learn basic low-level features. Regardless of how networks are built, they all attempt to build channels for information to travel between the initial and final levels. DenseNet establishes paths between the network's layers. Each layer in a dense block gets feature maps from all preceding levels and transfers their output to all following layers, given to the network's feed-forward structure. Concatenation is used to join feature maps from different levels (like in ResNet). Each dense layer is divided into two convolutional operations: (1 \* 1) CONV (standard conv operation for extracting features) and (3 \* 3) CONV (reducing feature depth/channel count). The growth rate (used K=32) is the number of channels output by a dense layer (1\*1 conv to 3\*3 conv). This means that a dense layer (l) will get 32 features from its preceding dense layer (l-1). Because 32 channel characteristics are concatenated and provided as input to the following layer after each layer, this is referred to as the growth rate. With the same number of parameters, the DenseNet model has a considerably smaller validation error than the ResNet model. These tests were carried out on both models with hyper-parameters that were more suitable for ResNet. After rigorous hyper-parameter searches, the authors claim that DenseNet will perform better [r-1].
5. **Inception-V3:** The inception v3 model was introduced to explore the inception architecture.Inception-v3 is a convolutional neural network architecture from the Inception family that makes several improvements including using Label Smoothing, factorized (7\*7) convolutions, and the use of an auxiliary classifier to propagate label information lower down the network. Convolutions with larger spatial filters (5\*5) or (7\*7) tend to be disproportionally expensive in terms of computation. For example, a (5\*5) convolution with n filters over a grid with m filters is 25/9 = 2.78 times more computationally expensive than a (3\*3) convolution with the same number of filters. Of course, a (5\*5) filter can capture dependencies between signals and between activations of units in earlier layers, so reducing the geometric size of the filters comes at a significant cost in expressiveness. However, we can ask whether a (5\*5) convolution could be replaced by a multi-layer network with fewer parameters with the same input size and output depth. If we zoom into the computation graph of the (5\*5) convolutions, we see that each output looks like a small fully-connected network sliding over (5\*5) tiles over its input. Since we are constructing a vision network, it seems natural to exploit translation invariance again and replace the fully connected component with a two-layer convolutional architecture: the first layer is a (3\*3) convolution, the second is a fully connected layer on top of the (3\*3) output grid of the first layer (see Figure 1). Sliding this small network over the input activation grid boils down to replacing the (5\*5) convolutions with two layers of (3\*3) convolutions [s-1].

Discussion:

Deep Convolutional Neural Networks have demonstrated excellent performance in object recognition, detection, and localization and many other computer vision tasks. Despite all the progresses shown by the various proposed architectures still there were few insights and logical reasoning about how they had achieved the stat-of-the-art records, which makes further improvements subject to the trial and error strategies, for example one of our benchmarks findings is that ResNet152 architecture [4] which has a depth of 152 layers yet scored higher speed record than VGG-19 architecture [5] which has depth of 19 layers only. DenseNet have shown to be the best architectures in terms of parameter space utilization, squeezing up to 4x less parameter space size with reference to AlexNet model, and 10x less with respect to VGG-19 \cite{i5].

Applications:

Using live images collected at various stages of the dragon fruit, the RESNET 152 deep CNN-based model was constructed to determine the mellowness of the dragon fruit [**m-10].** A mobile application is developed to detecting early banana diseases comparing ResNet-152 (accuracy 99.20%) and Inception V3 (accuracy 95.41%) **[m-11].** A Medical Monitoring System was presented to minimize information loss in the traditional pooling layer and layer-by-layer dimension reduction **[m-12].** U-Net has a wide range of applications in biomedical image segmentation, including brain and liver image segmentation. Variations of the U-Net have also been used to rebuild medical images **[m-3].** Sparse Annotation-Based Dense Volumetric Segmentation was achieved using a 3D U-Net **[m-4].** U-Net with VGG11 Encoder for Image Segmentation TernausNet **[m-5].** To estimate fluorescent stains, image-to-image translation is used **[m-6].** VGG-19 can be used to detect melanoma thickness in skin cancer patients **[s-2].**It can also be used to classify Wilson disease tissue using brain MRI **[s-3]**. An improved version of VGG-19 can be used to determine if someone is wearing a mask or not **[s-4].** From the idea of ResNet, dense connections have inspired optimizations in many other deep learning areas such as picture super-resolution [r-2], image segmentation [r-3], medical diagnosis [r-4] [r-5], and so on. This work exhibits the architecture's proficiency in image categorization [r-6]. Inception v3 enables health experts take some sample tests to determine systemic diseases from patients, but the research has studied the analysis of systemic diseases through digital image processing methods based on color analysis of nails. It enables to detection of diseases without painful sampling. Terry's nail is one of the nail disorders that can indicate systemic diseases such as liver cirrhosis [s-2].

limitations:

In U-Net architecture training takes a long time because there are so many layers. For larger photos, there is a relatively substantial GPU memory footprint. The diversity of features is lost due to the fixed receptive field of the convolution kernel **[m-7].** U-Net succeeds to segment the polyp with a low-level performance and a bad segmentation shape in polyp segmentation **[m-8].** VGG-19 has very high weight parameters **[s-1]**. The models are very heavy and take a lot of space, which increases inference time **[s-5]**. Takes a lot of training time **[s-3]**. Vanishing gradient is a big issue **[s-4]**. Deeper networks mean higher test errors which make VGG-19 vulnerable **[s-2].** The dense shortcut avoids the downside of demanding additional GPU resources while avoiding the problem of representational capacity decline in ResNet **[m-15].** Follow-up research has looked into the identity shortcut's flaws. The identity shortcut bypasses the residual blocks in order to retain features, which may limit the network's representation power **[m-16].** The disadvantage of the identity shortcut is that it produces the collapsing domain problem, which decreases the network's learning capacity **[m-17].** Excessive connections not only reduce the computing and parameter efficiency of networks but also make them more prone to overfitting in ResNet [r-7]. The inception v3 is generally designed by an iterative trial-and-error process, which requires a large amount of labeled data during the training phase, and also in this model, a huge number of neuron connections would bring in heavy computation expense. [s-3].

Conclusion:

A literature review of image classification architectures is presented in this paper. It classifies the growth and contribution to the deep learning renaissance during the last few years. It focuses on the progress in particular by debating and examining the designs, supervisory components, regularization processes, optimization strategies, and computation. This paper doesn't categorize in terms of popularity, performance in terms GPU/CPU and experimental results with comparisons. These are the limitations of this paper and could be solved in future update.

citations:

@article{[i-1],

title={Cs231n convolutional neural networks for visual recognition},

author={Karpathy, Andrej and others},

journal={Neural networks},

volume={1},

number={1},

year={2016}

}

@article{[i-2],

title={Imagenet large scale visual recognition challenge},

author={Russakovsky, Olga and Deng, Jia and Su, Hao and Krause, Jonathan and Satheesh, Sanjeev and Ma, Sean and Huang, Zhiheng and Karpathy, Andrej and Khosla, Aditya and Bernstein, Michael and others},

journal={International journal of computer vision},

volume={115},

number={3},

pages={211--252},

year={2015},

publisher={Springer}

}

@article{[i-3],

title={Imagenet classification with deep convolutional neural networks},

author={Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E},

journal={Advances in neural information processing systems},

volume={25},

pages={1097--1105},

year={2012}

}

@inproceedings{[i-4],

title={Rethinking the inception architecture for computer vision},

author={Szegedy, Christian and Vanhoucke, Vincent and Ioffe, Sergey and Shlens, Jon and Wojna, Zbigniew},

booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},

pages={2818--2826},

year={2016}

}