Survey paper on Alex Net, VGG16, VGG19, CNN, ReLu

Jihad,MD.Miah (17-35375-3), Prosad Bilash (17-35349-3), Arnob, MD. Abrer Shariar (17-35408-3) Jumo, Kazi Shanzida (17-34318-1)

Department of Computer Sciences, American International University-Bangladesh

Abstract

One of the fastest growing components in the machine learning family is deep learning. Deep learning uses Convolutional Neural Networks (CNN) to classify images because they will provide the most accurate results when solving real problems. [1] CNN has many pre-training projects, such as Alex Net, Google Net, Dense Net, Squeeze Net, Res Net, VGG Net, etc. [2] In this article, we compare CNN, Alex net, and VGG by collecting information from different articles to compare architectures.

Keywords: Deep Learning, Computer Vision, Object detection, NN, CNN, VGG, Alex Net, Machine Learning

Introduction

The purpose of this article is to discuss the possible trade-offs between CNN, Alex Net, and VGG. In the past ten years, convolutional neural networks have made revolutionary advances in various areas of pattern recognition; from image processing to speech recognition.[3] The most beneficial aspect of CNN is to reduce the number of parameters in ANN, which prompts researchers and developers to adopt larger models to solve complex problems that classic ANN cannot solve. The main assumptions of the tasks performed by CNN should not have location-related properties. In other words, for example, in a face recognition application, we don't need to pay attention to the position of the face in the image.[4] The only problem is how to identify them, regardless of their position in a given image. VGG is a classic convolutional neural network architecture. It is based on an analysis of how to increase the depth of these networks. The network uses a small 3 x 3 filter. The rest of the network is simple: the only other components are the packet layer and the fully connected layer. Alex Net is a convolutional neural network. The architecture consists of eight layers, of which five are folded and three are fully connected.[5]

Literature review

Alex Net is the first large-scale convolutional neural specification very suitable for ImageNet classification. Alex Net participated in the competition and is ready to significantly surpass any previous model that did not rely on deep learning. Normalization, convoy norm, allows several layers of convolution, one grouping layer, and then many fully connected layers. It is really like LeNet. There are many overall levels. Before the last fully linked layer enters the output class, there are 5 such folded layers and 2 fully linked layers. Alex Net is trained on ImageNet and the input image size is 227 x 227 x 3. If we look at the first layer, it may be the convolutional layer of Alex Net, with 11 x 11 filters, 96 of which are applied in step 4. It has 55 x 55 x 96 in the output parameters, during which there is a 35K first layer. The second layer is the bundling layer, in this case, we applied 3 filters 3 x 3 in step 2. The output volume of the binding layer is 27 x 27 x 96, and zero is checked. The grouping layer knows nothing, because the parameters are the weights they want to find. The evolution layer has the weights we check, but what we do by grouping is just a rule: we look at the grouping area and take the maximum value, so there are no parameters to check, the first eleven filters, and then five times five, about three-to-three filters. We end up with about 4096 fully connected layers. The last layer is FC8 into SoftMax, which passes through thousands of ImageNet classes. This design is based on the fact that the nonlinearity of ReLu was originally used.[6]

Hyperparameter

This architecture is the first application of ReLu nonlinearity. Alex Net also uses a normalization layer. When adding data, flipping, dithering, cropping, color normalization, etc. in Alex Net, the initial learning rate is 1e2, and when the verification accuracy is unchanged, it decreases by 10. This network uses L2 regularization, and the weight loss is 5e4. It was trained on a GTX 580 GPU with 3 GB of memory. 16.4 In the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

In 2014, there were many architectures that were significantly different and achieved another leap in performance. The main difference between these networks is the deeper network. VGG 16 is a 16-layer architecture with a pair of convolutional layers, an array layer, and finally a fully connected layer. The VGG network is the idea of a deeper network with smaller filters. VGGNet adds eight levels to Alex Net. At the time I had a model with 16-19 layers of VGGNet options. The bottom line is that these models use very small 3 x 3 convection filters everywhere, which is basically the smallest convolution filter size and looks slightly different from the neighboring pixels. They just kept this very simple 3 x 3 structure and grouped them regularly throughout the network. Due to fewer parameters, VGG uses smaller filters and stacks larger filters. A deeper filter instead of a large filter. In the end, you will have the same effective perceptual field as the 7x7 convolutional layer. VGGNet has a folding layer and a grouping layer, several more folding layers, one grouping layer, several more folding layers, etc. The VGG architecture has 16 fully connected and fully connected folding layers. In this case, VGG 16 has 16 and then VGG 19 has 19, which is just a very similar architecture, but with some additional conversions. Floor. There so, this is a very expensive calculation, a total of 138MB, each picture has 96MB of memory, much more than ordinary pictures. The error rate of ILSVRC requests is only 7.3.

Discussion

The general basis for image recognition is ImageNet, which is a dataset of 15 million images and provides more than 22,000 categories. ImageNet is based on Web-based image capture and crowdsourcing of human tags, and even has its own competitor: the aforementioned ImageNet Large-scale Visual Recognition Challenge (ILSVRC). Researchers all over the world are faced with the task of introducing a method that can identify the first 1 and the last 5 errors (the first 5 errors). The speed is a percentage of the image, where the corresponding label is not one of the 5 possible labels in the model. The competition provided a training set of 1,000 categories and 1.2 million images, a validation set of 50,000 images, and a control set of 150,000 images; there is a lot of knowledge. Alex Net won the competition in 2012, and the model who supported her style won the competition in 2013.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Although the previous Alex Net derivatives focused on smaller window sizes and first-level convolution steps, VGG handles another very important aspect of CNN: depth.

VGG accepts 224 x 224-pixel RGB images. For the ImageNet competition, the author cropped a patch at the center of 224x224 on each image to keep the input image size the same. The folded layer in VGG uses a very small sensing area (3x3, still covering the smallest possible size of left/right and top/bottom). There are also 1x1 convolution filters, which act as the linear transformation of the input, and then the ReLU block. Authorization is retained after folding. VGG has three fully linked layers: the first two layers each have 4096 channels, the third has 1000 channels, one for each class. All hidden VGG layers use ReLU (a great innovation of Alex Net that can reduce learning time). VGG usually does not use local response normalization (LRN), because LRN will increase memory consumption and training time, but the accuracy is not much improved.

VGG is based on Alex Net, but there are some differences with other competing models. VGG does not use a large receptive field like Alex Net (11x11 in increments of 4), but uses a very small receptive field (3x3 in increments of 1). Since there are now three ReLUs instead of one, the decision function is more selective. There are even fewer parameters (27 times the channel instead of 49 times the Alex Net channel) VGG contains a 1x1 convolutional layer, making the decision function more non-linear without changing the perceptual field. A lot of weight material; of course, the more layers, the better the performance. However, this is not uncommon. Google Net, another model with deep CNN and small convolution filter, also appeared in the 2014 ImageNet competition.[7]

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	[256; 512]	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**

Conclusion

In terms of the verification scale, VGG's Top1 error rate was 25.5%, while the partner's Top5 error rate was 8.0%. On several test scales, VGG achieved 24.8% of Top1 errors and 7.5% of Top5 errors. The error rate of the top 5 in the ImageNet 2014 competition was 7.3%, which reduced the presentation to 6.8b. VGG is an innovative object recognition model that supports up to 19 levels. VGG is regarded as a deep CNN and performs well on various tasks and datasets outside of ImageNet, even in terms of baselines. VGG is still one of the most widely used image recognition architectures today. Therefore, from the summary of this evaluation article, it is often concluded that VGG recognizes images better than Alex Net.[8]

References

[1]https://web.stanford.edu/class/cs231a/prev_projects_2016/example_paper.pdf

[2]https://ieeexplore.ieee.org/abstract/document/8977648/references#references

[3]https://www.researchgate.net/publication/319253577_Understanding_of_a_Convolutional_Neural_Network

[4]https://www.researchgate.net/publication/5847739_Introduction_to_artificial_neural_networks

[5]https://www.robots.ox.ac.uk/~vgg/research/very_deep/

[6]https://towardsdatascience.com/architecture-comparison-of-alexnet-vggnet-resnet-inception-densenet-beb8b116866d

[7]https://towardsdatascience.com/vgg-neural-networks-the-next-step-after-alexnet-3f91fa9ffe2c

[8]https://arxiv.org/pdf/1409.1556.pdf